

---

THESIM DOCTORALEM DE PHILOSOPHIA TITULO

---

# A procedure for loss-optimising the timing of loan recovery under uncertainty

---

By

ARNO BOTHA  
(15247199)

Supervised By

PROF. CONRAD BEYERS  
PROF. PIETER DE VILLIERS



UNIVERSITY OF PRETORIA  
Department of Actuarial Science

Thesis submitted to the University of Pretoria in accordance with the requirements of the degree DOCTOR OF PHILOSOPHY in *Actuarial Science* from the Faculty of Natural & Agricultural Sciences.

JULY 2021



## ABSTRACT

The point at which a loan is in default is posited to be a portfolio-specific, probabilistic, and risk-based "point of no return" beyond which loan collection becomes sub-optimal if pursued any further. A method is presented for finding a delinquency threshold at which the overall loss of a given portfolio is minimised, i.e., loans are forsaken neither too early nor too late. This method, called the *Loss-based Recovery Optimisation across Delinquency* (LROD) procedure, incorporates the time value of money, risk-adjusted costs, and the fundamental trade-off between accumulating arrears versus forsaking future interest.

The procedure is demonstrated across a range of portfolio compositions and credit risk scenarios using a simulation-based testbed. The computational results show that threshold optima can exist across all reasonable values of both the payment probability (default risk) and the loss rate (loan collateral). Furthermore, the procedure reacts positively to portfolios afflicted by either systematic defaults (due to economic downturns) or episodic delinquency (cycles of curing and re-defaulting).

For real-world loans, which are typically right-censored, a forecasting step is proposed during which the remaining cash flows of each censored account are first 'completed' before applying the LROD-procedure. This approach is illustrated using residential mortgage data from a large South African bank. The empirical results show that riskier scenario-based forecasts of credit risk yield smaller threshold optima. Furthermore, censored cash flows are iteratively forecast in an additional Monte Carlo-based step, thereby analysing the stability of threshold optima yielded by the procedure.

In conclusion, this work can enhance relevant business strategies, improve related modelling, and help revise the policy design of most banks, especially in tweaking the quantitative aspects of collection policies.

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## DEDICATION AND ACKNOWLEDGEMENTS

**N**ever have I struggled more than to abridge the tempered zeal and relentless inquiry that is the doctoral pursuit; or the anguish of its perpetual cycles of review, rethink, review, scrap, and rework, while self-doubt grows ever deeper and pride fractures into a mirage called progress ... but above all else, never have I struggled more than to convey the unbridled joy of *finally* conquering my doctoral journey ... as when I sat down to write this dedication. Through it all, I had to balance the academic rigours of this doctorate against the professional demands of a high-strung career in banking. Though utterly spent, I can now look back (with perhaps a *little* bit of pride) on what has truly been an indelible rite of passage, one that I shall never forget.

To both my mother, Maryna Botha, to whom I owe everything, and to my husband-to-be, Dirk Nel, I dedicate this thesis. I would not have survived the periods of intense self-doubt and countless urges to quit had it not been for both of you and your unwavering belief in my potential.

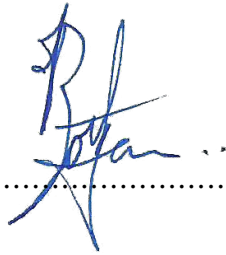
To all my cherished friends, thank you for the many moments spent together, fond memories, and sincere hugs when I needed it the most.

To professors Conrad Beyers and Pieter de Villiers, you have my utmost gratitude and profound respect for your tutelage, riveting debates, patience, and incisive feedback – without which I would not have emerged the scientist that I am today. One day, I hope to pass the torch with as much finesse and kindness as you have shown me these last few years.

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## AUTHOR'S DECLARATION

I hereby declare that this thesis, which I, ARNO BOTHA (15247199), submit for the degree PHILOSOPHIAE DOCTOR to the UNIVERSITY OF PRETORIA, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.



SIGNED: ..... DATE: 14TH JULY 2021

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*



## TABLE OF CONTENTS

	<b>Page</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Definitions</b>	<b>xiii</b>
<b>List of Symbols</b>	<b>xv</b>
<b>List of Abbreviations</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview of the thesis . . . . .	7
<b>2 Bank and borrower: a treatise of trust and its erosion</b>	<b>9</b>
2.1 Trust as the historic bedrock of banking . . . . .	10
2.2 The rise of consumer credit in modernity . . . . .	23
2.3 Financial intermediation and its <i>raison d'être</i> . . . . .	32
2.4 The quest for bank liquidity and overall system stability . . . . .	35
2.4.1 The mechanics of a modern bank and its funding . . . . .	36
2.4.2 Managing the fundamental risk of illiquidity . . . . .	39
2.4.3 A model for managing a bank's reserves . . . . .	43
2.4.4 Two interventions to reduce bank fragility . . . . .	45
2.5 Maintaining capital: the Basel Capital Accords . . . . .	48
2.6 The management of financial risk in banking . . . . .	61
2.6.1 A trifecta of risks: credit, market, and operational . . . . .	62
2.6.2 Common risk management strategies in banking . . . . .	66
<b>3 The banker's gauge of eroded trust</b>	<b>71</b>
3.1 Default definitions: a servant of many masters . . . . .	73
3.1.1 A regulatory overview of default definitions . . . . .	73
3.1.2 Delinquency: the leitmotif in risk models . . . . .	83
3.1.3 Roll rate analyses as decision-support tools . . . . .	87

3.2	Towards opportune loan recovery: analysing true ‘default’ . . . . .	94
3.3	Measures of loan delinquency . . . . .	101
3.3.1	Contractual Delinquency ( <i>CD</i> ): the $g_1$ -measure . . . . .	102
3.3.2	Macaulay Duration ( <i>MD</i> ): the $g_2$ -measure . . . . .	105
3.3.3	Degree of Delinquency ( <i>DoD</i> ): the $g_3$ -measure . . . . .	106
3.4	Optimising loan recovery times: the LROD-procedure . . . . .	108
3.5	Concluding remarks . . . . .	114
<b>4</b>	<b>Optimising loan recovery timing: a computational study</b>	<b>117</b>
4.1	Portfolio generation: a testbed for the LROD-procedure . . . . .	117
4.2	Computational results of recovery optimisation . . . . .	120
4.2.1	Random defaults . . . . .	120
4.2.2	Episodic defaults . . . . .	123
4.2.3	Markovian defaults . . . . .	124
4.2.4	Applying the LROD-procedure on real-world data . . . . .	127
4.3	Concluding remarks . . . . .	128
<b>5</b>	<b>Recovery optimisation using real-world data with forecasting</b>	<b>131</b>
5.1	Two techniques to forecast future loan receipts . . . . .	132
5.1.1	Random defaults with empirical truncation . . . . .	132
5.1.2	Markovian defaults . . . . .	133
5.2	Calibrating the forecasting techniques to mortgage data . . . . .	135
5.2.1	Calibrating the random defaults technique . . . . .	137
5.2.2	Calibrating the Markovian defaults technique . . . . .	138
5.2.3	Assessing the quality of forecasts . . . . .	139
5.3	Optimising the recovery decision: an empirical illustration . . . . .	140
5.3.1	Optimisation results using $S_1$ , $S_2$ , and $S_3$ respectively . . . . .	141
5.3.2	Monte Carlo simulations for analysing the variance of optima . . . . .	145
5.4	Concluding remarks . . . . .	148
<b>6</b>	<b>Conclusion</b>	<b>151</b>
	<b>Appendix A Ancillary material on various unrelated subtopics</b>	<b>157</b>
A.1	An example of loss reservation using Markov theory . . . . .	157
A.2	Illustrating three delinquency measures: a case study . . . . .	159
A.3	Failing to forecast before recovery time optimisation . . . . .	162
A.4	Fitting statistical distributions to the truncation parameter $k$ . . . . .	164
	<b>Bibliography</b>	<b>167</b>

## LIST OF TABLES

<b>TABLE</b>	<b>Page</b>
2.1 A simplified balance sheet of a bank . . . . .	37
2.2 A simplified income statement of a bank . . . . .	39
3.1 A comparison of external credit risk ratings . . . . .	82
4.1 Conceptual transition matrix for Markovian defaults . . . . .	119
5.1 Maximum likelihood estimates for the Markov chain estimated from $S_2$ . . . . .	138
5.2 Maximum likelihood estimates for the Markov chain estimated from $S_3$ . . . . .	138
5.3 Calibration and accuracy results of the forecasting techniques . . . . .	139
5.4 An experimental setup to approximate different risk compositions . . . . .	141
A.1 A two-loan case study of possible repayment histories . . . . .	159

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## LIST OF FIGURES

FIGURE	Page
1.1 The trade-off associated with two extreme delinquency thresholds . . . . .	4
1.2 The LROD-procedure in three steps . . . . .	5
1.3 Forecasting a portfolio to completion . . . . .	6
2.1 Total household debt (USA) . . . . .	24
2.2 Aggregate household debt-to-income ratios of various countries . . . . .	25
2.3 Finaly’s five-phase credit management model . . . . .	26
2.4 Two sample window trade-offs: risk immaturity vs. market irrelevance . . . . .	29
2.5 Stylised metamodel of a bank’s model-driven decisions . . . . .	31
2.6 The information brokerage function of a bank . . . . .	33
2.7 The asset-transformation function of a bank . . . . .	35
2.8 The typical structure of debt seniority . . . . .	37
2.9 Bank liquidity gap analysis . . . . .	41
2.10 Two types of liquidity gaps . . . . .	42
2.11 A Value-at-Risk approach for capital estimation . . . . .	55
2.12 The typical resolution of the default process . . . . .	59
2.13 Three stages of IFRS 9 loan impairment . . . . .	64
2.14 Three-step risk management process . . . . .	68
3.1 The process of curing from the default state . . . . .	81
3.2 Common types of risk models & exercises in retail banking . . . . .	86
3.3 The role of the outcome period within cross-sectional models . . . . .	88
3.4 Cohort analysis across candidate outcome periods . . . . .	89
3.5 Default rate development across candidate outcome periods . . . . .	90
3.6 An example of a roll rate analysis . . . . .	91
3.7 Stylised speeds of increasing and decreasing arrears . . . . .	98
3.8 Illustrating an approach for loss-optimising the recovery decision . . . . .	109
4.1 Losses across thresholds by measure $g$ (Random Defaults) . . . . .	120
4.2 Losses across thresholds for $g_1$ by truncation point $k$ (Random Defaults) . . . . .	121
4.3 Losses across thresholds for $g_1$ by repayment probability $b$ (Random Defaults) . . . . .	122

4.4	Losses across thresholds for $g_1$ by loss rate $r_A$ (Random Defaults) . . . . .	123
4.5	Losses across thresholds for $g_1$ by $k$ (Episodic Defaults) . . . . .	124
4.6	Losses across thresholds by measure $g$ (Markovian Defaults) . . . . .	125
4.7	Losses across thresholds for $g_1$ by $P_{PP}$ and $P_{DD}$ (Markovian Defaults) . . . . .	126
5.1	The empirical difference between observed loan tenure and contractual maturity . . .	135
5.2	A Venn diagram of three segments of mortgage accounts . . . . .	136
5.3	Histogram of the maximum delinquency observed per account . . . . .	137
5.4	Empirical losses across thresholds using sample $S_1$ in the procedure . . . . .	142
5.5	Empirical losses across thresholds using sample $S_2$ in the procedure . . . . .	143
5.6	Empirical losses across thresholds using sample $S_3$ in the procedure . . . . .	144
5.7	Monte Carlo-based estimates of the average loss curve for the $s_{11}$ scenario . . . . .	145
5.8	Monte Carlo-based estimates of the average loss curve for the $s_{22}$ and $s_{33}$ scenarios .	147
A.1	Two-loan case study: $g_1$ -measure . . . . .	160
A.2	Two-loan case study: $g_2$ -measure . . . . .	161
A.3	Two-loan case study: $g_3$ -measure . . . . .	162
A.4	Loss-optimal default thresholds using an untreated loan portfolio . . . . .	163
A.5	Statistical distributions fit to $\max g_1(t)$ using samples $S_2$ and $S_3$ . . . . .	165

## LIST OF DEFINITIONS

DEFINITION	Page
3.1 Loan delinquency . . . . .	84
3.2 Payments in arrears ( $g_0$ -measure) . . . . .	85
3.3 Contractual Delinquency ( $CD$ ): the $g_1$ -measure . . . . .	104
3.4 Macaulay Duration ( $MD$ ): the $g_2$ -measure . . . . .	106
3.5 Degree of Delinquency ( $DoD$ ): the $g_3$ -measure . . . . .	108
3.6 The notion of $(g, d)$ -defaulting accounts . . . . .	111
3.7 A simple portfolio loss model for a given $(g, d)$ -configuration . . . . .	112
3.8 Loss-optimising the recovery decision: the LROD-procedure . . . . .	113

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*



## LIST OF SYMBOLS

$A_t$	Accumulated amount in arrears at time $t$
$A(i, t)$	Present value of the arrears summed up to $t$ for loan $i$
$b$	Probability of payment ( <i>Random defaults</i> )
$\hat{b}$	Estimator for the probability of payment $b$
$\mathcal{D}_g$	Set of chosen thresholds specific to the measure $g$
$D : R_t = 0$	Delinquent state assumed by $X_t$ ( <i>Markovian defaults</i> )
$d \in \mathcal{D}_g$	Delinquency threshold for the measure $g$
$d_N \geq 0$	Specified maximum for constraining thresholds in $\mathcal{D}_g$ using $g_1$
$d^{(g)}$	Threshold associated with the minimum loss $m^{(g)}$ using $g$
$d_1(t) \in \{0, 1\}$ for $t = 1, \dots, T$	Function indicating if the repayment ratio $h_t$ is less than $z$
$d_2(t) \in \{0, 1\}$ for $t = 1, \dots, T$	Function indicating zero delinquency at time $t - 1$ using $g_1$
$d_3(t) \in \{0, 1\}$ for $t = 0, \dots, T$	Function indicating if $t$ is less than or equal to $t_c$
$d_4(t) \in \{0, 1\}$ for $t = 0, \dots, T$	Function indicating if actual duration exceeds expected duration
$f_{AD}(t)$	Actual duration, used in $g_2$ and $g_3$
$\tilde{f}_{AD}(t)$	Inflated actual duration, used in $g_3$
$f_{ED}(t)$	Expected duration, used in $g_2$ and $g_3$
$g \in \{g_0, g_1, g_2, g_3\}$	A particular measure of delinquency
$g_0(t) \geq 0$ for $t = 1, \dots, T$	Number of payments in arrears (unweighted <i>CD</i> -measure) at $t$
$g_1(t) \geq 0$ for $t = 1, \dots, T$	The $z$ -weighted <i>CD</i> -measure of loan delinquency at time $t$
$g_2(t) \geq 0$ for $t = 0, \dots, T - 1$	The <i>MD</i> -measure of loan delinquency at time $t$
$g_3(t) \geq 0$ for $t = 0, \dots, T - 1$	The <i>DoD</i> -measure of loan delinquency at time $t$
$g(i, t)$ for $t = 0, \dots, T_i$	Delinquency measurement using $g$ at time $t$ for loan $i$
$g^*$	The measure that yielded the lowest loss $m^{(g)}$ at threshold $d^{(g)}$
$h_t \quad \forall t = 1, \dots, T$	Repayment ratio between $R_t$ and $I_t$ , used in $g_1$
$\mathbf{I}$	Instalment vector, populated by $I_t$ up to $t_c$
$I'_{(T)}$	The $T^{\text{th}}$ element in a recursively updated variant of $\mathbf{I}$
$I_t \quad \forall t = 0, \dots, T$	Instalment expected at time $t$
$I > 0$	Level expected instalment
$I_c > 0$	Calculated instalment that amortises a given balance at $t_c$
$I_t^i$	Instalment expected at time $t$ for the $i^{\text{th}}$ account

$\mathcal{I}_t^i$	Function indicating payment for loan $i$ at time $t$ , i.e., $R_t^i \geq I_t^i$
$i = 1, \dots, N$	Loan index in an $N$ -sized portfolio
$j = 1, \dots, N$	Delinquent loan index given $p_D$ ( <i>Episodic defaults</i> )
$k$	Parameter that controls the extent of $(k, g)$ -truncation
$\hat{k}_i \geq 0$	Randomly sampled truncation parameter for loan $i$
$L_P$	Loan amount (or principal)
$L_M$	Maximum loan size offered by a lender
$L_g(d)$	Discounted total portfolio loss given a $(g, d)$ -policy
$l(i, t)$	Discounted risk-adjusted loss at time $t$ for loan $i$
$l_j \in [1, k]$	Number of consecutive non-payments sampled for loan $j$
$m(t) \geq -1$ for $t = 1, \dots, T$	Magnitude by which delinquency should be reduced
$m^{(g)}$	Minimum loss attained at threshold $d^{(g)}$ for $g$ using $L_g(d)$
$N$	Number of loan accounts in a portfolio
$O(i, t)$	Present value of all remaining instalments at $t$ for loan $i$
$o_j \in [1, t_c - l_j]$	Delinquency episode's starting time, sampled for loan $j$
$P_{ij}$ with $(i, j) \in \{P, D, W\}$	One-period transition probability between states $i$ and $j$
$P: R_t = I$	Paid state assumed by $X_t$ ( <i>Markovian defaults</i> )
$p$	Annual compounding period, usually monthly, i.e., $p = 12$
$p_D \in [0, 1]$	Proportion of accounts designed to become delinquent
$\mathbf{R}$	Receipt vector, populated by $R_t$ up to $t_c$
$\mathbf{R}'$	A $(k, g)$ -truncated variant of the receipt vector $\mathbf{R}$
$R_t \quad \forall t = 0, \dots, T$	Receipt (cash inflow) at time $t$
$R_t^i$	Receipt (cash inflow) at time $t$ for the $i^{\text{th}}$ account
$R(i, t)$	Present value of receipts summed up to $t$ for loan $i$
$r \in [0, 1]$	A generic nominal monthly client interest rate used in $v_j$
$r_A \in [0, 1]$	Loss rate on the arrears amount $A(i, t)$
$r_E \in [0, 1]$	Loss rate on the expected balance $O(i, t)$
$\mathcal{S}_D$	Subset of loans in a portfolio considered as $(g, d)$ -defaulting
$\mathcal{S}_P$	Subset of loans in a portfolio considered as $(g, d)$ -performing
$s \in [0, 1]$	Delinquency sensitivity, used in $g_3$
$s_{ij}$ with $(i, j) \in \{1, 2, 3\}$	A setup using training sample $i$ for optimising sample $j$
$T$	Observed current age/tenor of a loan
$\mathcal{T}$	Maximum between $T$ and $t_c$
$t_c$	Contractual term of an amortising loan
$t_{c_i}$	Contractual term of loan $i$
$t \in \mathbb{Z}$	Time index (in months), observed up to $T \geq t$
$t_i^{(g, d)} \quad \forall i \in \mathcal{S}_D$	Earliest time at which loan $i$ has $(g, d)$ -defaulted
$t_0 \leq t_c$	Observable loan tenure prior right-censoring

$t_0(i) \leq t_c$	Loan tenure observed from data for loan $i$
$t_1 \leq t_c$	Starting point of right-censored remaining cash flows
$t' \leq t_c$	Starting point of $(k, g)$ -truncation, provided that $g(t') \geq k$
$t_k \in [t_0, t_c]$	Right-censored starting point of $(k, g)$ -truncation with $g_1(t_k) \geq k$
$t_w \in [t_1, t_c]$	Truncation starting point that may exist if $X_{t_w} = x_7$
$u_t \in [0, 1]$ for $t = 1, \dots, t_c$	Randomly generated number for time $t$ ( <i>Random defaults</i> )
$v_j$	Function for discounting back $j$ periods using a generic rate $r$
$v_t^{(a)}$	Function for discounting back $t$ periods using a risk-free rate
$v_t^{(b)}$	Function for discounting back $t$ periods using the client rate
$W : R_{t \geq t'} = 0$	Write-off state assumed by $X_t$ ( <i>Markovian defaults</i> )
$X_t \in \{P, D, W\}$	Three-state random variable at time $t$ (computational study)
$X_t \in \{x_0, \dots, x_7\}$	Multi-state random variable at time $t \in [0, t_0]$ (empirical study)
$x_0, \dots, x_6$	$g_1$ -based delinquency states assumed by the multi-state $X_t$
$x_7$	Write-off state assumed by the multi-state $X_t$
$z \in [0, 1]$	Specified boundary for the repayment ratio $h_t$ , used in $g_1$
$\Delta_t \quad \forall t = 0, \dots, T$	Difference between $I_t$ and $R_t$
$\delta$	Continuously compounded annual rate
$\delta^{(p)} = \delta \div p$	Nominal variant of $\delta$ per annual compounding period $p$
$\delta_t \in \mathbb{Z}$	One-period delinquency movement using the $g_1$ -measure
$\lambda(L_M, L_P, s)$	Multiplier function that inflates $f_{AD}(t)$ , used in $g_3$

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## LIST OF ABBREVIATIONS

<i>ALM</i>	Asset & Liability Management
<i>APR</i>	Absolute Priority Rule
<i>ARM</i>	Adjustable-Rate Mortgage
<i>ASRF</i>	Asymptotic Single Risk Factor
<i>AUC</i>	Area Under the Curve
<i>BCBS</i>	Basel Committee on Banking Supervision
<i>BIS</i>	Bank for International Settlements
<i>CAR</i>	Capital Adequacy Ratio
<i>CCF</i>	Credit Conversion Factor
<i>CRR</i>	Capital Requirements Regulation
<i>CD</i>	Contractual Delinquency (measure of delinquency)
<i>DFE</i>	Delinquency Forecast Error
<i>DI</i>	Deposit Insurance
<i>DoD</i>	Degree of Delinquency (measure of delinquency)
<i>DPD</i>	Days Past Due
<i>EAD</i>	Exposure At Default
<i>EBA</i>	European Banking Authority
<i>EC</i>	Economic Capital
<i>ECL</i>	Expected Credit Loss (IFRS 9)
<i>EL</i>	Expected Loss (Basel)
<i>GFC</i>	Global Financial Crisis (2008)
<i>HHI</i>	Herfindahl-Hirschmann Index
<i>IAS</i>	International Accounting Standards
<i>ICAAP</i>	Internal Capital Adequacy Assessment Process
<i>IFRS</i>	International Framework of Reporting Standards
<i>IRB</i>	Internal Rating-Based approach
<i>LCR</i>	Liquidity Coverage Ratio
<i>LGD</i>	Loss Given Default
<i>LLR</i>	Lender of Last Resort
<i>LROD</i>	Loss-based Recovery Optimisation across Delinquency

<i>MAE</i>	Mean Absolute Error
<i>MD</i>	Macaulay Duration (measure of delinquency)
<i>MDP</i>	Markov Decision Process
<i>MLE</i>	Maximum Likelihood Estimate
<i>NII</i>	Net Interest Income
<i>NIR</i>	Non-Interest Revenue
<i>NPL</i>	Non-Performing Loan
<i>NSFR</i>	Net Stable Funding Ratio
<i>OTD</i>	Originate-To-Distribute
<i>P2P</i>	Peer-to-peer
<i>PAR</i>	Portfolio Arrears Rate
<i>PD</i>	Probability of Default
<i>RWA</i>	Risk-Weighted Assets
<i>SME</i>	Small-to-Medium-sized Enterprise
<i>SARB</i>	South African Reserve Bank
<i>SCRA</i>	Specific Credit Risk Adjustment
<i>SICR</i>	Significant Increase in Credit Risk
<i>UK</i>	United Kingdom
<i>UL</i>	Unexpected Loss
<i>USA</i>	United States of America
<i>VaR</i>	Value-at-Risk

## INTRODUCTION

The practice of borrowing and lending has existed for over 4,000 years, having started with a farmer borrowing seed-grain and promising to repay from his future harvest with interest. This simple transaction, written on a Sumarian clay tablet, represents the first-ever codification of trust between two parties. The lender assumes that the farmer will honour his repayment obligation at some future date, effectively exchanging a modicum of ‘trust’ for a sum of money. In fact, the very word *credit* is derived from the Latin word *creditum*, meaning “to have trusted”. Unsurprisingly then, the biggest risk for the lender is that of the borrower *not* honouring his obligation, which is formally known as *credit risk* and explained in Van Gestel and Baesens (2009, pp. 24–29) and Thomas (2009a, pp. 1). Another aspect of any credit agreement, as discussed in Finlay (2010, pp. 31–32), is that of *time* and the uncertainty that it brings to repayment in general. Beyond repayment, the topic of trust (and its erosion) is deeply embedded across many banking functions, starting with exchanging deposited commodities for some kind of receipt that is trusted to be equal in value, e.g., the modern-day ‘promissory’ bank note. As trade flourished during peace-time (before inevitably crashing again during wars), so too did the need increase for these custodian banks and their services, as I shall examine in section 2.1 and section 2.3. Increased trade almost surely brings with it a multitude of currencies to be exchanged, safeguarded, and reissued as loans; all of which reinforce the role of banks throughout history as trusted intermediaries, financiers, and custodians of wealth.

The non-payment of basic amortising loans typically occurs gradually over time, which suggests that ‘trust’ also erodes gradually between bank and borrower. Indeed, quantifying credit risk relies fundamentally on first measuring the extent of eroded trust (or loan delinquency)

using past records. A few candidate measures exist that are variations of simple accountancy ratios, of which a few are discussed in Rosenberg and Christen (1999) and in Sah (2015). As a prominent example, the so-called *Contractual Delinquency* (CD) measure<sup>1</sup> is typically constructed from the number of days past due (DPD), yielding the number of payments in arrears. In turn, this CD-measure is extensively used when building statistical models (called credit scorecards) that predict the risk of a borrower reaching a certain delinquency level. These models are often implemented in modern-day computer systems that decide automatically whether or not to grant credit to a new applicant. In fact, the astronomical growth in consumer credit over the last few decades could not have been possible without a degree of automation, as discussed in section 2.2.

Loan delinquency and measures thereof have become increasingly embedded over time into a growing array of models beyond the scope of simple scorecards, as I will argue in subsection 3.1.2. Today, delinquency measurement is often the broad backbone on which banks enact credit and pricing decisions, devise debt collection strategies, and perform overall risk management, in addition to its use within risk modelling. However, most applications of the CD-measure require setting a delinquency threshold beyond which a loan is deemed as in ‘default’. Banks have commonly specified three payments in arrears (or 90 DPD) as a pragmatic point of ‘default’, long before the introduction of relevant regulations. That said, this threshold generally ranges between 30–180 days, supported by managerial discretion and some types of analyses, as explored in subsection 3.1.3. However, the direct financial implications of any chosen threshold are generally not considered during typical analyses, especially when developing credit scoring models. Therefore, and as originally argued in Hand (2001), pursuing modelling excellence becomes questionable when the constructed outcome variable, itself determined by the default definition, is inherently quite arbitrary.

Other than simply breaching the aforementioned threshold, default definitions often contain more qualitative criteria. The definitions may further differ based on the portfolio type and the context of credit risk modelling: either *unexpected* or *expected* credit losses. However, the international standards that govern either context are enforced to varying degrees by individual regulators, with some examples thereof explored in subsection 3.1.1. Specifically, unexpected loss modelling is largely regulated by the Basel II Capital Accords<sup>2</sup>, while expected loss modelling is mostly managed by the IFRS 9 accounting standard<sup>3</sup>, which focuses less on extreme risk events than Basel II. Accordingly, ‘default’ (and/or the regulatory threshold itself) may differ across

---

<sup>1</sup>This measure is known by a few names, e.g., payments (or months/time) in arrears, arrears category, and missed payments. However, it is called the ‘CD’-measure throughout this study, or referred to by its mathematical form as the  $g_0$ -measure (defined later).

<sup>2</sup>Amongst other things, these accords prescribe the way of setting capital aside that is intended to absorb unexpected losses during liquidity crises, as discussed in sections 2.4–2.5.

<sup>3</sup>IFRS 9 articulates a set of principles for modelling the expected credit loss. Fundamentally, future loan write-offs should be offset by keeping an adequate level of loss provisions in advance, barring catastrophic *unexpected* losses that should rather be covered by Basel’s capital buffer.



---

competing jurisdictions, which certainly complicates any related modelling for multinational banks. Yet even if the threshold is decreed to be the same value across all nations, there is little objective evidence to support such a value beyond simple discretion and crude analysis. Furthermore, the acquisition, merging, and sale of loan portfolios (or portions thereof) present another challenge. So-called 'legacy' definitions from the previous owners can certainly conflict with that of the new owner, which implies that multiple definitions may run concurrently in the same portfolio. Consequentially, the very idea of 'default' has arguably become a vague and incoherent concept in trying to serve so many 'masters' at once.

The original premise of a default definition is to reach a so-called "point of no return", beyond which loan repayment becomes extremely doubtful. Every unpaid instalment (or portion thereof) erodes the trust between bank and borrower, which is only tolerable up to a point, as argued in section 3.2. This ambiguous point may itself differ across portfolios and even banks, presumably due to differing risk appetites and market conditions. Having reached this point, the bank effectively assumes that the obligor's delinquency will perpetuate if the loan is kept. Therefore, the lender now pursues the immediate and maximal recovery of debt (including seizing any collateral), instead of retaining the credit relationship any further. However, a loan may 'cure' from default whenever a borrower repays the arrears (regardless of reason), which further casts doubt on a chosen default threshold as the supposed "point of no return". The challenge hereof is to find the ideal switching point, i.e., the *best* time at which the lender should abandon all hope of repayment. Finding this point using a delinquency measure is convenient since past loan performance can be projected into scale-invariant delinquency progressions across loans, without losing any behavioural information.

Owing to the difficulties of defining 'default' precisely, I explore a more fundamental meaning of 'default' in section 3.2 as the portfolio-dependent, probabilistic, and risk-based "point of no return" beyond which loan collection becomes sub-optimal if pursued. The 'default' state is simply based on breaching a certain delinquency threshold using a given delinquency measure, so that the "net cost" of each candidate threshold can be assessed. A loss basis (instead of profit) is sensible since forsaking lent capital will generally incur a loss of some kind, rarely a profit. Regardless, too strict a threshold will surely marginalise accounts that would have resumed repayment (or cured from 'default'), had the bank not foreclosed (or charged-off) that soon. A loan may also experience multiple episodes of 'redefaulting' and curing, which is further exacerbated by a threshold that is too strict. Conversely, too lenient a threshold will naively tolerate increasing arrears at the cost of greater liquidity risk and bigger capital buffers, possibly becoming capital-inefficient. The goal now becomes to devise an expert system in which these two extremes can be appropriately offset against each other. Doing so can form a proverbial 'Goldilocks-zone' that contains the ideal delinquency threshold for a portfolio, which translates into the 'best' time for loan recovery. This concept is illustrated in Fig. 1.1 using the arrears amount (proportional) as a

high-level threshold, including two extreme choices thereof.

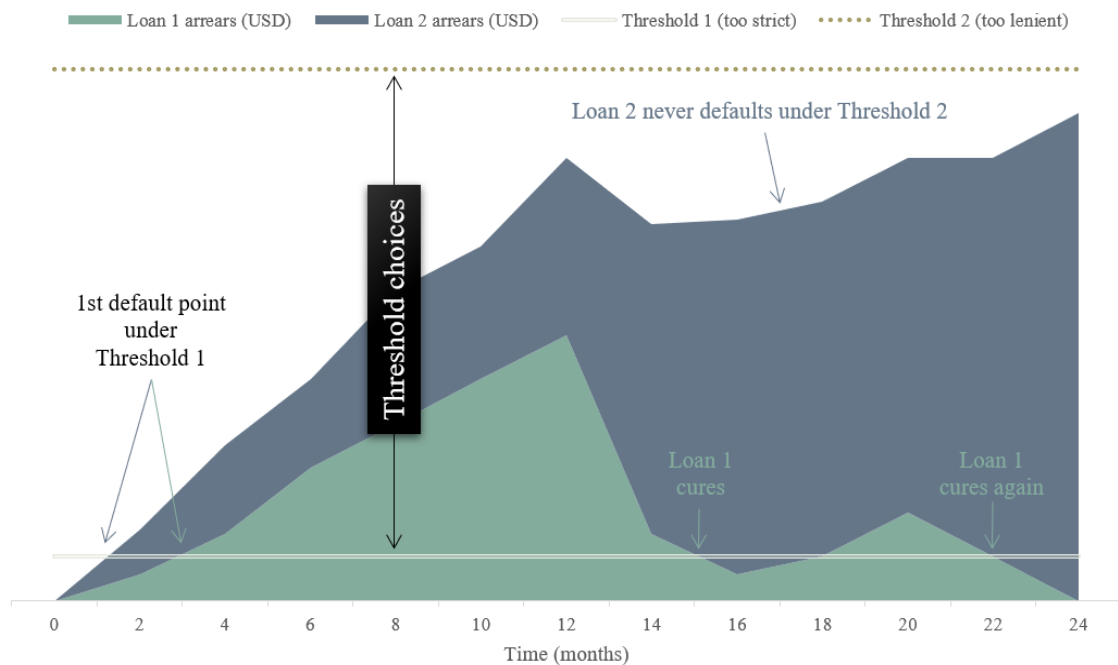


FIG. 1.1: Illustrating the trade-off associated with two extreme arrears-based thresholds for two fictional loans of the same size. Threshold 1 is overly strict for loan 1 given that it cures later; but suitable for loan 2 since it never cures. Conversely, threshold 2 is overly naive for loan 2, though suitable for loan 1.

The *CD*-measure is but one way of quantifying loan delinquency and surely alternative measures exist or can be formulated that better suit a particular portfolio. To this point, a few flaws of the *CD*-measure are discussed in section 3.3, followed by giving an improved variant thereof. Two alternative measures are presented as well, one of my own design, in trying to quantify delinquency more precisely. In particular, both partial payments and ‘prepayments’ (or underpayments and overpayments respectively) can make measured delinquency less precise due to rounding in the *CD*-measure. Lastly, different measures will likely have different measurement domains, which suggests that a measurement can mean different things. The choice of measure therefore presents another dimension when designing an optimisation procedure for this study.

The recovery (or foreclosure/charge-off) decision is therefore conjectured to be a portfolio-wide optimisation problem of competing risks (and costs). The decision variable is the choice of a threshold on the domain of a given measure. Within this context, the following set of research questions are explored:

- 
- Question 1** Can an optimal default point sensibly exist? Is 90+ DPD such an optimal point, thereby explaining its widespread use?
- Question 2** How can the exact timing of the loan recovery decision be optimised, if at all?
- Question 3** Are there alternative measures of delinquency (other than the *CD*-measure) that can better suit recovery optimisation?
- Question 4** How can different delinquency measures be feasibly compared to one another, especially if their domains (and output) differ fundamentally?
- Question 5** Given some optimisation procedure, what are some of the factors that affect recovery optimisation in general?
- Question 6** Is it feasible to optimise the recovery decision for a real-world loan portfolio? What underlying challenges (if any) exist that may impede optimisation?
- Question 7** What factors affect recovery optimisation using real-world data? Given uncertainty, how can the stability of the threshold optima be assessed?

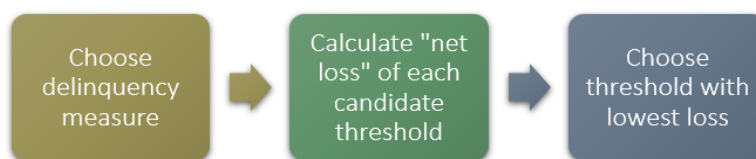


FIG. 1.2: High-level steps of the contributed LROD-procedure.

A method is developed to explore these research questions, called the *Loss-based Recovery Optimisation across Delinquency* (LROD) procedure, presented in section 3.4 and summarised in Fig. 1.2. Essentially, an ideal portfolio-wide threshold is sought such that loans are forsaken neither too early nor too late, if at all. To examine recovery optimisation from "first principles", a simple simulation-based setup (or testbed) is devised in section 4.1. Basic amortising loan portfolios are randomly generated given a specifiable risk profile. The testbed itself facilitates drawing quick managerial insight on a portfolio's optimisation potential, before conducting any deep data work. This is achieved simply by tweaking the testbed's simulation parameters. Finally, the LROD-procedure is demonstrated in section 4.2, having conducted a broad computational study on the testbed across different parametrisations and delinquency measures. Threshold optima are successfully found across most levels of default risk and loss risk, as measured by the probability of payment and loss rate respectively. Furthermore, the procedure reacts positively to portfolios that suffer from systematic pattern-like defaults (due to economic downturns), as well as portfolios with episodic delinquency (cycles of curing and re-defaulting).

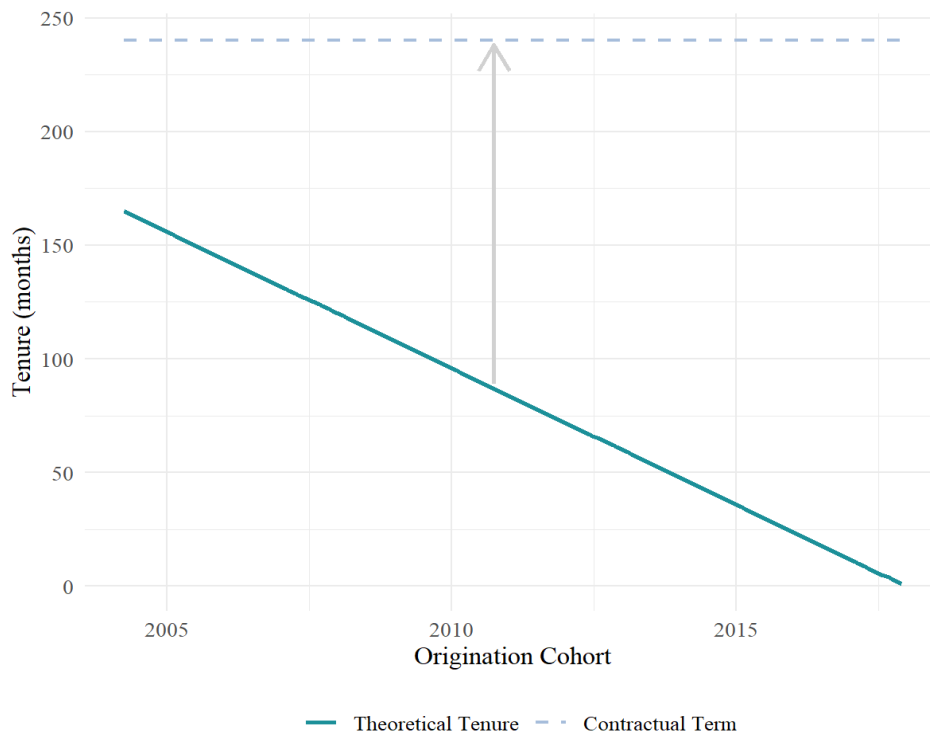


FIG. 1.3: Illustrating the increasing right-censoring effect of newer cohorts, which requires more forecasting than older cohorts in completing a hypothetical 240-month term portfolio.

Real-world loan portfolios are typically right-censored in that some accounts may not have reached contractual maturity yet. Older loan cohorts will therefore have more observable data than newer cohorts, which is unsurprising since most portfolios are actively being grown every month. However, the LROD-procedure assumes that the underlying portfolio is uncensored. As a solution, the residual cash flows of each censored account can first be ‘completed’ using an appropriate forecasting method, as illustrated in Fig. 1.3. To this end, a few forecasting techniques are outlined in section 5.1. Each technique is calibrated in section 5.2 using a South African mortgage portfolio, followed by extensive testing and retraining. Recovery optimisation is then conducted and discussed in section 5.3. Lastly, censored cash flows are iteratively forecast in an additional Monte Carlo-based step in the procedure. This allows one to analyse the stability of the optimisation results, which can inspire greater confidence in any found optima when the variance is low.

## 1.1 Overview of the thesis

The thesis is structured as follows. Chapter 2 reviews the history of banking from the perspective of broken trust to provide broad context for the present study. This review includes the system-wide implications of bank failures, the subsequent regulatory quest for system stability, and the increasing role of mathematical modelling in managing financial risks in banking. Within the ambit of this modelling, the concept of ‘default’ is dissected in chapter 3 towards developing a more comprehensive and dynamic theory of loan default. The discussion culminates in the aforementioned LROD-procedure, which formulates the recovery decision’s timing as a delinquency-based optimisation problem under uncertainty. This problem is then illustrated in chapter 4 across various portfolio types and dynamics using a comprehensive simulation study. The results from chapter 4 along with parts of chapter 3 are associated with a research article, accepted for publication in the journal *Expert Systems with Applications* with a preprint available in Botha et al. (2021).

While chapter 4 demonstrates the theoretical viability of the LROD-procedure, chapter 5 examines the real-world application thereof using mortgage data. In particular, a forecasting step is proposed that remedies some of the challenges related to real-world data, thereby enabling the practical use of the LROD-procedure. The results from chapter 5 along with parts of chapter 3 are associated with a second research article, accepted for publication in the *Journal of Credit Risk* with a preprint available in Botha et al. (2020). In summary, the quantitative aspects of any retail bank’s collection policy can be improved, perhaps substantially so, by using the LROD-procedure. Finally, chapter 6 concludes the study and outlines both the wider implications thereof for credit risk modelling, as well as possible areas of future research.



*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## BANK AND BORROWER: A TREATISE OF TRUST AND ITS EROSION

Trust lies latent in the triad of depositor, banker, and borrower. This becomes self-evident when exploring the rich history of banking across millennia, as surveyed in section 2.1. Regardless of the exact role of the banker, the prerequisite of financial intermediation remains that of ‘trust’; whether it be intermediating as a wealth custodian, transactor, exchanger, or lender. Moreover, history suggests that trust – and indeed trade activity – can only exist within a stable environment, propped up by the rule of law and its enforcement. Beyond simple banking functions, I discuss the more recent meteoric rise of consumer credit over the last hundred years in section 2.2. This astonishing growth is largely attributed to the advances in technology, as well as improved risk assessments afforded by statistical techniques. In fact, the use of mathematical and statistical models have become increasingly prevalent in driving a bank’s decision-making. Some authors have even hailed the current era as a so-called "third revolution" in modelling a bank’s decisions, especially as various models progressively overlap one another and become ever more sophisticated.

Beyond decision-making, modelling advances have demonstrated that the fundamental reason for banking’s continued existence is due to *asymmetrical information* between bank and borrower, which I shall review in section 2.3. In particular, a bank develops specialised expertise and generates risk information when fulfilling its many roles. This means that subsequent lending is more risk-sensitive and strategically superior when compared to the case of no intermediation, i.e., the individual that lends directly. Without the ability to produce risk information *en masse*, it is doubtful that banking success would have reached its pinnacle in modernity as it clearly had. That said, financial intermediation requires funding and trust to be successful, and I discuss

the mechanics and underlying risks thereof in section 2.4. As part of its operation, a bank faces the elemental risk of not being able to fund withdrawal requests from depositors during a panic. This illiquidity implies a balancing act between placating depositors on their funds' safety and acting credible to procure debt-funding from other lenders. Failure of either objective can cause a liquidity crisis or exacerbate an unfolding crisis, which can in turn propagate across the banking system and cause universal turmoil.

To help safeguard the financial system and correct possible market failures (driven by asymmetrical information), many governments have since created specific interventions. A particular pertinent intervention is that of imposing minimum capital requirements to curb excessive risk-taking, as discussed in section 2.5. This includes the internationally well-known Basel Capital Accords, which are underpinned by a considerable mathematical literature and a broad range of models. While capital ought to absorb unexpected losses that may otherwise induce a liquidity crisis (or bank failure), it is not sufficient to cover more 'expected' levels of credit risk. Providing for these more frequent losses, as governed by another well-known standard (IFRS 9), is equally crucial to the base survival of a bank. As such, loss provisioning is briefly reviewed together with other broad risk types in section 2.6, followed by discussing a few common risk management strategies.

## **2.1 Trust as the historic bedrock of banking**

Most banks throughout history can be characterised by two interdependent functions: the brokerage role and the asset transformation function. A bank attracts deposits and investments from which it funds lending activities to those borrowers deemed creditworthy. Doing so effectively 'transforms' idle deposits into more useful debt by acting as the financial intermediary between depositors and borrowers. Moreover, as custodians of deposits, banks are uniquely positioned to facilitate payments and exchange currencies, effectively brokering the flow of money amongst parties. These functions largely came about from the invention of money as a system of account in which the majority of trading activities took place, as opposed to bartering directly. Bartering is challenged by first having to determine the differences in the disparate value of various goods and services, not to mention baskets thereof, before transacting. This transactional friction is solved by using a highly divisible currency in which to denominate value, as discussed in Van Gestel and Baesens (2009, pp. 1–3, 9–12).

In most civilisations throughout history, currencies were commonly denominated in rare and durable metals such as gold and silver. In fact, the earliest known coinage dates back to the seventh century BC during which the Lydian kings (present-day Turkey) stamped emblems on small standardised ingots forged from electrum (an alloy of gold and silver), which signified their supposed value. As explained in MacDonald and Gastmann (2001, pp. 24), these coins were



backed by the authority of the king as having value by fiat, without needing to verify this claim using scales (to determine the metal's purity) when transacting. Naturally, an element of trust was already imbued within these authoritative seals themselves. That said, money itself can take (and has taken) any widely agreed-upon form, including cattle, amber, grain, ivory, salt, rice, and various metals. A particularly interesting form of money was that of the cowrie; the shell of a mollusc from the Pacific and Indian oceans. Of all the historical objects used as currency, cowries were used for far longer and more universally than any other currency, including coinage. This, according to Davies (2002, pp. 36–37), is attributed to the cowrie's durability, ease of cleaning, relative uniformity (or fungibility), and difficulty to counterfeit; also some of the attributes for deeming something as 'money'.

While the introduction of money meant wider proliferation of banking, the latter predates the invention of coinage by about two thousand years. From Davies (2002, pp. 34–55), simple banking operations have surfaced multiple times throughout antiquity. In fact, the first banks were temples in Mesopotamia circa 3,000–2,000 BC that served as sanctuaries for a depositor's wealth. These priests accepted common deposits such as grain, cattle, fruits, agricultural implements – and later, gold and silver money. However, gold and silver money should not be confused with 'coinage' as invented by the seventh century BC Lydians. According to Davies (2002, pp. 61–64), these gold and silver *pre-coins* date from as early as 2,250 BC and underwent various stages of invention, ranging from large silver blobs, to bars, to rods, and to elongated nails. Fearing common thievery and desiring convenience, the depositor (or merchant) believed the temple to be a safe place due to its bustling and devout crowds, as well as believably hosting the righteous divine. Moreover, these temples were central locations in the city-states of Babylonia and therefore the perfect locale for conducting business activities, as discussed in MacDonald and Gastmann (2001, pp. 22–23) and Davies (2002, pp. 48–50).

The art of writing originated from Mesopotamia first as a method of bookkeeping. As such, the local temples often issued depositors with clay-based receipts of the deposited wealth. Having multiple depositors, it was not long until the same temples supervised transacting merchants, simply by transferring holdings from one merchant to that of another. This practice saw the eventual establishment of state-owned "grain banks" into which harvests were pooled from several farmers as general deposits, centralised across Babylonia and Egypt. Having risen to prominence during the reign of the Egyptian Ptolemaic dynasty in the fourth century BC, as recounted in Davies (2002, pp. 52–54), these grain banks facilitated debt repayments amongst various parties. Such a payment<sup>1</sup> was enacted simply as an accounting entry that transcended locale without any money physically changing hands. The value thereof was offset against the deposited grain of the payer at one location and credited to that of the recipient at another site.

---

<sup>1</sup>This is also known as a *giro* transfer, or giro credit, and refers to a direct transfer of money between two account holders at the same bank.

Channelling debt repayments via these grain banks (especially larger payments) became quite popular amongst the ancient Egyptians. In fact, the resulting transactional records were widely considered as official and even used as evidence whenever disputes were litigated. Using payment receipts in this way clearly highlights the inherent trust that were placed in these grain banks, having fulfilled their entangled roles as custodian, transactor, and later, as lender.

While they first brokered simple payments amongst their depositors, these temples and granaries later started to issue loans from the wealth stored in their vaults. For grain banks especially, this meant lending seed-grain to farmers, having agreed to repay the borrowed grain with interest from their future harvests. Another example of institutionalised lending is the emergence of a private mercantile bank called the House of Egibi that operated for over a few centuries during the first millennium BC, as narrated in Davies (2002, pp. 51) and Hudson (2010). Although the House of Egibi secured a wide range of deposits, they only ever used their own wealth when funding loans, albeit with no arbitrage between deposit and lending rates (both usually 20%). Furthermore, both Mesopotamia and Egypt lacked timber and certain stones (e.g., marble) that were required to develop their societies further, despite being blessed with fertile lands and abundant water sources. In addition, the upper royal classes desired more luxury goods beyond those that were locally available, according to MacDonald and Gastmann (2001, pp. 20–21). This presented opportunities for ambitious traders who then sourced these goods from international markets. However, increased trade brought with it the challenge of foreign and multiple currencies, even though coinage was not yet that widespread at the time. As a result, the merchants started to use these temples and banks as "clearing houses", who both facilitated the auction of goods as well as brokered the credit-based transactions thereafter.

The invention of coinage during the seventh century BC quickly spread throughout the Persian empire, the Aegean islands, Greece, and northward to Thrace, Macedonia, and the Black Sea. Moreover, currency exchanges became necessary largely due the prolific Greek traders who transacted in multitudes of metal-based coins, as discussed in MacDonald and Gastmann (2001, pp. 25–26), Davies (2002, pp. 66, 71–74), and Rigas and Riga (2003). In fact, multi-currency trade saw the rise of Greek bankers (or *trapezitai*<sup>2</sup>), who epitomised money-changing as the most common form of banking during the Graeco-Roman period. These services proliferated amidst a deluge of different coinages that varied in both quality and type, as a by-product of trade prosperity. For their service, Greek bankers charged a commission of 5%-6% of the currency value. Unfortunately, this prosperity also led to a rise in fraud and counterfeiting, which only made the Greek bankers even more instrumental in 'sanctifying' these currency exchanges.

---

<sup>2</sup>The Greek word for banker *trapezitai* is likely derived from the trapezium-shaped tables (a *trapeza*) upon which Greek bankers exchanged various currencies. These tables had a series of lines and squares, believed to have aided calculations.

However, it was the lending side of Greek banking that became the most lucrative form of intermediation at the time. According to MacDonald and Gastmann (2001, pp. 25–26), Davies (2002, pp. 66, 71–74), and Rigas and Riga (2003), ship masters and merchants alike wanted more funding with which they could then undertake more rewarding voyages, which certainly posed a significant risk to the coffers of the Greek bankers. The resulting loans were largely funded using the current accounts of the bank's merchant depositors and carried interest rates between 6 to 30 percent, based on the assessed risk. In stark contrast to the earlier Egyptian House of Egibi (who only used their own capital for lending), the Greeks secured their lending activities on the short-term deposits of their borrowers, including copper, silver, gold, and sometimes even slaves. This early example of asset transformation exemplifies the high level of trust that these Greek bankers enjoyed. That said, Demosthenes (an early banking lawyer at the time) once remarked that any banker who self-funded his lending was destined for bankruptcy, which again suggests that banking success is predicated on first attaining trust.

Although the Greeks were instrumental in furthering banking innovations (particularly the Athenians), their position was soon rivalled by the small offshore Aegean island of Delos during the period of 200–100 BC. According to Davies (2002, pp. 78–79) and Rigas and Riga (2003), this island had but two assets: its great harbour and its famous temple of Apollo. However, both of these assets helped the island of Delos rise as one of the principal clearing houses of Macedonian trade, which included tar, pitch, timber, silver, oriental wares from Arabia and India, as well as slaves. The bankers of Delos retained the inherent trust of the antique world for well over 400 years. While the previous Greek bankers brokered transactions purely on a cash basis, the newly established Bank of Delos opted for credit instead as its transactional basis. Accordingly, trade volumes increased further and the Aegean coffers of these bankers grew ever larger. Moreover, they safeguarded wealth within the temple of Apollo, itself surrounded by the ocean, thereby further deterring robberies. This explains the strategic allure of trusting these particular bankers, which subsequently attracted to their vaults substantial levels of state and private wealth. In fact, this wealth helped to fund the rise of the Roman empire, with the Bank of Delos itself later serving as the model for Roman banks. From Delos, the refined practices of conducting credit-based transactions and giro transfers soon spread to Rome. Perhaps paramount to Rome's success was one particular practice borrowed from Delos: the strategic centralisation of deposit contracts across vast banking networks, akin to the grain banks of contemporary Egypt and earlier Mesopotamia.

While the Romans assimilated the advances in money and banking from the Greeks and the Egyptians, the Roman achievements were more militaristic and administrative rather than economic. In particular, the Romans contributed its legal discipline by which contracts and property rights were strictly enforced, as well as the equitable settlement of disputes, as discussed

in MacDonald and Gastmann (2001, pp. 26–29). While the Code of Hammurabi<sup>3</sup> was indeed an earlier legal system circa 1700 BC in Mesopotamia, the Roman rule of law was more effective given its military might and bureaucratic strength. According to Davies (2002, pp. 51), this legal discipline instilled predictability and widespread trust that induced a rudimentary though functional credit system across the empire. Furthermore, the Romans considered the minting of coinage (and its centralisation) to be more important than advancing the Ptolemaic ideas of central banking. In fact, the empire's most famous and universal silver coin – the *Denarius* – was exclusively minted in Rome itself, with provincial towns only allowed to mint bronze coinage. From Davies (2002, pp. 89–93), the Roman emperors controlled mints directly and relied heavily on taxation for revenue instead of issuing national debt. Furthermore, the Romans often debased their own currency to fund their expanding armies by using more impure metals when minting coins. Despite the tenfold increase in silver coins that were in circulation during the great expansionary period of 150–50 BC, the inflow of raw bullion could simply not compete with the considerable demand for Roman coins struck from it.

The Roman empire eventually began its slow economic decline until its eventual collapse, as discussed in Howgego (1992) and Davies (2002, pp. 94–112). This decline was characterised by the ceaseless debasement of the Roman currency, rampant inflation, inadequate taxation to support a growing welfare state, maintaining the increasingly unaffordable military, and the exhaustion of Roman mines. In fact, the emperor Gallienus debased the Roman coinage during his reign (260–268 AD) to such an extent that the *Denarius* contained but 4% silver. Similarly, Aurelian introduced two whole new coins into circulation during his reign (270–275 AD) as Roman emperor. Both coins were valued by dictate at 2.5 times the previous nominal value of similar coins, i.e., the basis of inflationary finance. Unsurprisingly, these actions generated rapid inflation and brought about the temporary state seizure of Roman banks who refused to accept these inferior coins. While it is certainly true that the Roman empire ultimately fell to barbaric invasions, one can argue that the underlying cause of a weakened military is the chronic economic chaos endured up to the fifth century AD. In particular, continued economic turmoil meant the breakdown of trust in the Roman credit system, already battered by a weak and mistrusted currency. After the empire's collapse, money-based trade itself broke down across Europe, which persisted through most of the Dark Ages. As such, it is argued in Van Gestel and Baesens (2009, pp. 2) that banking itself became mostly irrelevant during this time when there was little trade to intermediate, wealth to store, currencies to exchange, or loans to fund from empty vaults. The sporadic kingdoms, having sprouted from the remnants of imperial Roman power, first had to relearn minting coins during the next few centuries, let alone install the rule of law again, before banking could hail its previous glory.

Various coinages surfaced repeatedly across Europe and Arabia over the next few centuries.

---

<sup>3</sup>This Code contained almost 300 laws, some of which pertain to banking operations and its ethical practice.

However, it was the adoption of the tally stick by the medieval English treasury (or the Exchequer) in the twelfth century AD that saw the widespread return of a credit system. This tally stick, according to Davies (2002, pp. 148–152), is a piece of wood (or ‘slip of wood’ from its Latin origin *talea*) that was originally used as evidence (or a receipt) of a payment, commonly cut from 20 cm lengths of hazel wood. A specific notch, signifying the exact amount owed in taxes (or other debt), was cut across the breadth of the stick. Thereafter, the stick was partially split across its length in twain up to the handle, with one piece broken off. The larger piece with the handle, called the ‘stock’, was kept by the creditor, while the smaller piece, called the ‘foil’, was given to the debtor along with the loan. When put together, both pieces would match the original shape of the stick, therefore ‘tallying’ the debt and serving as legal evidence of the loan transaction. In fact, verifying (or ‘checking’) tax payments were carried out on the Exchequer tables, which were ten by five feet in size and were adorned with a chequered cloth. This cloth not only inspired the name of the Exchequer but also later gave its name to the common bill of exchange, the ‘cheque’.

Taxes owed to the king, as represented by a tally stock kept in the Exchequer, could subsequently be used by the king to pay someone else, simply by transferring the tally stock itself. In a time when charging interest (or ‘usury’ as it was more commonly known) was forbidden on religious grounds, these tallies became wooden ‘cheques’ and were used to raise state funding based on Exchequer-held debt. As a result, the English monarchs soon began to issue tallies in anticipation of collecting taxes in future, which indicates an early form of modern-day government bonds. The overall money supply was significantly increased by bartering Exchequer debt for funds, based on the inherent trust that the monarch will repay his debt using future tax revenue. This catered for the growing demand amongst traders for coins, without resorting to the gross debasement of coinage as enacted in earlier Roman times. These tallies were particularly opportune since the limited European supply of gold and silver at the time hindered any large-scale minting of coinage.

While medieval finance in Western Europe slowly re-emerged, earlier Islamic conquests saw the establishment of a *Pax Islamica* in the eleventh century AD. This brought about a virtual free-trade zone across Western Asia, North Africa, and the Mediterranean, which was supported by a more sophisticated credit system than that in Western Europe at the time. In servicing the Islamic empire’s vast trade routes, credit flourished in the form of early bills of exchange as a method of payment. These bills were then drawn from cooperating merchant bankers that were flung far across the empire, as discussed in MacDonald and Gastmann (2001, pp. 34–36, 41–44). Another Islamic credit instrument was the *mudaraba* arrangement under which the investor entrusted cash or goods to the trader. In turn, this capital was eventually repaid along with an agreed upon share of profit, as a common way of circumventing the prevailing sin of usury. However, it was Jewish bankers who reinvigorated lending and the widespread use of credit across medieval Europe, especially since their religion permitted usury when dealing with

non-Jews. Moreover, Jewish merchant families were networked across rival kingdoms in both Europe and the Islamic world, which facilitated the flow of credit-based trade across oceans and port-cities. Specifically, the Jewish bankers in Baghdad issued an early form of a letter of credit known as a *suftadja*, therein committing payment to other parties on behalf of their Jewish merchant clients. These familial bankers that cooperated across large distances were effectively similar to the erstwhile Roman banks and its centralised credit system.

International trade soon flourished amongst Jews, Muslims, and Europeans alike, partially due to the trade fairs held across France and Italy from the twelfth century AD onward. Although these fairs originally celebrated a local religious saint, rulers at the time – particularly the Counts of Champagne in France – quickly realised the tax potential of these fairs. The noblemen subsequently helped to establish the rule of law during these fairs, which promoted greater business confidence; increasing both trade and the associated tax revenue in turn. Based on MacDonald and Gastmann (2001, pp. 59–62), these fairs subsequently attracted a diversity of enterprising merchants and with them, a barrage of bi-metal bullion<sup>4</sup> as money. Apart from exchanging these currencies, the hassle of transporting money *en masse* – along with the associated costs of hiring guards and caravans – made credit a useful and alternative payment system, especially for conducting larger transactions. In fact, international merchants frequenting these fairs soon innovated a credit instrument called *lettres de faire* (or fair letters) based on credit. These documents recorded the sale of goods, whilst promising payment at a future fair. This delay allows sufficient time for tallying the total debits and credits amongst participating merchants plying their trades to one another during a particular fair.

In truth, the fair letters were but one short step away from becoming so-called *bills of exchange*, which historians commonly consider as the greatest financial innovation of the late Middle Ages. An example is given in MacDonald and Gastmann (2001, pp. 61) of such an early bill that was issued (or drawn) by a fourteenth century Tuscan merchant (the drawer) who instructed his bank (the drawee) to pay the bearer of the bill (the payee) the amount inscribed on the bill itself. This is similar to the preceding Jewish *suftadja* and, indeed, a precursor to the modern-day cheque (or more formally, a "negotiable instrument"). Naturally, these bills soon became currency in themselves given their securability against the drawer's assets, which are implicitly trusted to cover payment sufficiently. Bills of exchange were far safer to handle and made transactions quicker relative to keeping coins and bullion at hand. Though the trade fairs themselves eventually stopped in the late thirteenth century AD, these bills – and the credit system in which they operated – soon spread across the medieval world, as discussed in MacDonald and Gastmann (2001, pp. 65). Ironically, the trade fairs that introduced these

---

<sup>4</sup>Medieval Europeans only started using the Italian *florin* as a central currency in 1252 AD. Preceding this, gold and silver bullion itself were used as unedified money, although many other coins circulated at these fairs that were, however, struck by multiple sovereigns.

bills were made redundant by the very same, since networked merchants transacted across ever increasing distances using these bills instead, with little need for visiting a fair once established.

The close relationship between stable power and credit is perhaps best exemplified by the Knights Templar and the Knights Hospitallers during the Crusades, particularly during the twelfth and thirteenth centuries AD. Initially, these two religious orders of warrior monks protected holy pilgrimages and provided medical care to war casualties. In time, these knights acquired their own ships, castles, armies, storehouses, as well as held strategic points of presence that spanned England to Egypt and Spain to Syria, according to MacDonald and Gastmann (2001, pp. 66–68) and Davies (2002, pp. 153–158). It is therefore unsurprising that the Templars soon expanded into storing wealth and intermediating payments amongst merchants and princes of the realm throughout Europe and the Holy Land. Ennobled with the perceived trust of the Christian faith, they were excellent candidates for safely transferring goods and money across their vast network to facilitate the war efforts. This was not unlike the paved trade routes that were protected by the earlier Roman military, which enabled their credit system. Moreover, the heavily fortified bases of the Templars soon attracted deposits from monarch and merchant alike. Naturally, the Templars soon began to exchange copious amounts of currencies and to lend the amassed wealth out again at large, thereby establishing a new credit system. Specifically, the knights expedited payments using bills of exchange, acted as tax collectors for both kings<sup>5</sup> and popes, transferred ransom payments, granted loans to crusaders and pilgrims, and even legally coined their own currencies. However, their banking successes abruptly ended once their credibility was grossly undermined by the French King Philip IV and his undignified greed. Eventually, the relentless accusations and political machinations of the king led to Pope Clement V abolishing the Templar order in 1312.

The Templars' demise proved to be fortunate for their long-time rivals, the Italian bankers, who readily met the credit demands across Europe and the Middle East in the absence of the knights. The two Italian city-states of Venice and Genoa became financial superpowers; originally spurred by the trade fairs, competing Jewish bankers, increasing international trade, and the Crusades. As a result, these Italian bankers intermediated progressively more transactions and established a competing credit system from as early as the twelfth century AD. From De Roover (1963, pp. 1–2) and MacDonald and Gastmann (2001, pp. 73–76), the Genoese *bancherii* (or bankers) innovated interbank transfers between customers who held accounts at competing Italian banks. This was largely facilitated by their invention of double-entry bookkeeping and its creative use in avoiding the Christian sin of usury. In fact, the Genoese *bancherii* invented foreign exchange contracts, which were similar to bills of exchange though issued across two

---

<sup>5</sup>Perhaps the greatest supporter of the Crusades was the English King Henry II (1154–89) who enacted so-called "crusading taxes". These revenue streams were mainly funnelled into his accounts that were held at both the Templars and Hospitallers, instead of the London Exchequer, as discussed in Davies (2002, pp. 157).

currencies at a preset exchange rate, as explained in Van Gestel and Baesens (2009, pp. 3–4). These bills were later exonerated by the Roman Church, having realised that the element of risk justified compensation. The Church argued that the potential profits were uncertain given that they arose from trading these foreign bills at fluctuating prices once issued. Moreover, many merchant banks of the time went bankrupt by investing too great a portion of their assets in risky commercial ventures that subsequently failed. The potential of bankruptcy further justified the charging of interest rates when lending, even if somewhat obscured in the ledger.

The dangers of default risk continued to be exemplified by even some of the most successful European banks during the fourteenth and fifteenth centuries AD, as discussed in De Roover (1963, pp. 2–5) and MacDonald and Gastmann (2001, pp. 79–82). The Hundred Years' War between England and France led to the failure of both the Bardi and Peruzzi familial banks of Florence. Specifically, the English King Edward III defaulted on his various unsecured loans from both Bardi and Peruzzi. Most monarchs at the time demanded loans from bankers in order to fund and sustain their military campaigns, which these monarchs fielded to retain or expand their sovereignty. Yet despite having perfectly legal loan agreements, the Italian bankers soon realised that their borrowers' ability and willingness to repay were tenuous and uncertain at times, even if they were powerful monarchs. To this point, the word 'bankrupt' has its roots in the Italian phrase *banca rotta* (or "broken bench"); see Van Gestel and Baesens (2009, pp. 4). The idiomatic suggestion is that the wooden bench on which Italian bankers conducted their business was physically ruptured upon default, likely in frustration.

On the other hand, the common man was largely unsympathetic to these bank failures as induced by monarchical defaults. In fact, according to De Roover (1963, pp. 2–5) and MacDonald and Gastmann (2001, pp. 79–82), Italian bankers were already despised at the time. An especially hated group were the *lombardii* in Northern Italy; though small-scale, they were opportunistic pawnbrokers and enterprising moneylenders who openly charged 'usurious' interest. Another more prominent example is the Medici family who tried to collect upon a large loan in 1477 AD, held by the deceased Duke of Burgundy, Charles the Bold. That said, the Medici family came to be the most powerful European bankers following the fall of Bardi and Peruzzi, eventually governing Florence itself. Having masterfully combined politics and finance, the Medici Bank strategically called in large debts from their enemies, Naples and Venice. With their coffers suddenly depleted, these city states could no longer fund their mercenary armies, who deserted as a result, much to the Medicis' delight.

As with the larger Bardi and Peruzzi banks, the Medici Bank eventually declined as well in the late fifteenth century AD, again affected by royal defaults. The prevailing political theory was that a monarch ruled by divine right and were subject to no temporal authority. Naturally, this assertion severely conflicted the bankers when calling in the earthly loans of a divine prince



of the realm. From De Roover (1963, pp. 5–6, 331–334) and MacDonald and Gastmann (2001, pp. 85–89), both the English King Edward IV and the French Duke Charles the Bold defaulted on their various debts during the Wars of the Roses. Along with deteriorating management, it was nigh impossible for the London and Bruges bank branches to remain afloat thereafter. A few other branches soon failed as well (notably those in Venice, Milan, and Lyons), which is ascribed in De Roover (1963, pp. 358–367) to gross maladministration, lack of coordination, economic recession, and both the lavish lifestyles and political entanglements of the Medicis. The severe deterioration became self-evident when the bank governor, Lorenzo de' Medici, even misappropriated public funds; presumably to stave off impending insolvency. The bank finally conceded defeat in 1494 upon the Italian invasion of the French King Charles VIII.

The now-familiar pattern reasserts itself with the subsequent rise (and eventual fall) of two other banking families in the sixteenth century AD. Specifically, the Fuggers and Welsers rose to prominence as German banks and even eclipsed the Medicis, having absorbed most of their remaining assets. From MacDonald and Gastmann (2001, pp. 101–107), the Fuggers primarily served the intermediation needs of the European royals, most notably the Hapsburgs. Both German banks financed the growing international spice trade from Asia and the newly discovered Americas, which diversified their revenue streams considerably. However, as with the erstwhile Italian bankers, the German lenders suffered the same sovereign defaults from fickle princes and their wars. In particular, the bankruptcy of the Spanish King Philip II in 1577 AD shook the foundations of German banking and preempted the fall of the Fuggers. For additional context, more than twenty large banks declared bankruptcy in Spain and Italy during the strikingly short two-year period 1587–1589. Widespread private bank failures induced by sovereign defaults led to the establishment of public banks. These institutions offered credit to both private clients and the state, whilst being better secured by shares in public debt than by overly leveraged private capital.

Up to the sixteenth century AD, merchant banks historically serviced a speculative niche, though a new entity would soon enter the stage in the form of an early stock exchange in Antwerp, called the *bourse*. As a common clearing house, the swap rates for commercial bills of exchange (amongst England, France, Italy, and Germany) as well as government bonds could now be regularly published, negotiated, and traded at the same central location. Moreover, this Antwerp bourse availed to sovereigns a larger source of speculative credit, which could help fund conquests and royal expeditions to the New World, as discussed in MacDonald and Gastmann (2001, pp. 98–100, 107–114). This is perhaps best exemplified by Spanish public debt being sold to and traded amongst third parties, ultimately funding the Spanish Conquistadors in the Americas. However, religious conflicts between Catholics and Protestants made sieges and rebellions a reality in Antwerp (commonly known as the Spanish Fury and the Dutch Revolt), which caused the local merchants and bankers to flee to more stable environments. Moreover, the Dutch closed off the

Scheldt river that previously fed trade into the embattled city of Antwerp. With less trade activity, the Antwerp bourse naturally declined.

During the rebellion, the Dutch strategically diverted the flow of merchant ships to Amsterdam, which was far more politically stable than Antwerp. Moreover, the seventeenth century AD Dutch Republic was governed by an oligarchy of merchants (known as the Dutch regents), instead of an authoritarian monarch. This, as argued in MacDonald and Gastmann (2001, pp. 110), led to the national interest shifting from typical territorial conquest, to the more capitalist pursuit of prosperity. While the influx of multinational merchants from Antwerp undoubtedly increased trade and profits, the various coinages in circulation became problematic. From Quinn and Roberds (2005), coins may devalue over time to less than their stated nominal value simply due to abrasion. Another factor at the time was the various minting ordinances that decreed nominal values for coinages against the florin in Amsterdam. However, the Dutch inadvertently encouraged the deliberate debasement of coins given the arbitrage<sup>6</sup> in value between finer/heavier and rougher/lighter varieties of coins, supposedly equal in value by ordinance. In turn, worn or debased coins can artificially devalue a nation's own currency when exchanging the debased foreign coins for freshly minted domestic currency. In foreshadowing the modern central bank, the Dutch founded the *Wisselbank* (or the Exchange Bank of Amsterdam) at which one could safely store wealth and exchange multiple currencies with less devaluation. More importantly, the *Wisselbank* transformed stored coins and bullion into special florin-denominated credit known as *bank money* that became more valuable than any physical currency, thereby discouraging the circulation of debased coinage.

Apart from the *Wisselbank*, the Dutch regents established the Amsterdam bourse (or *Effecten-Beurs*) in the same period. This early stock exchange attracted excess capital by speculating on (or investing in) commodities, government bonds, and shares in company profits. Common commodities at the time included grain, whale oil, and spices; while equity shares were predominantly in the Dutch East Indian Company. The national credit system was in fact augmented by the Amsterdam bourse since the stock investments served as collateral in secondary cycles of borrowing and lending. Thus, the hallmarks of a modern financial world came to be. Financiers not only had to trust their borrowers to repay, but also consider the dynamic value of the underlying collateral, assuming it to be sufficiently tradable for offsetting potential credit losses. Moreover, modern stock market manipulations, e.g., short-selling and bear raids, were invented by the same enterprising Dutch at the time, according to MacDonald and Gastmann (2001, pp. 114–116). As a result, the overarching credit system became progressively more sensitive to market movements as the value of the underlying collateral increasingly depended on the same movements. The

---

<sup>6</sup>Two currencies are compared in Quinn and Roberds (2005): the cross rixdollar (minted in the Spanish Netherlands) against another variant of the rixdollar (issued in the Republican Netherlands), both valued at 2.5 florins, though the former was of a higher quality metal, thereby creating arbitrage.

danger hereof is perhaps epitomised by the market crash in 1637 as a direct result of the Dutch Tulipmania.

The Dutch impact upon the financial world was considerable during the seventeenth century AD. From Quinn and Roberds (2005), Amsterdam became the *de facto* hub for drawing, paying, and trading bills of exchange across all of Europe. The *Wisselbank* aside, this feat was partially due to another Dutch innovation at the time. In particular, the Dutch have improved upon the earlier bill of exchange by rendering it transferable beyond the original payee. Each successive bearer of the Dutch bill enjoyed the same payment claim on the original debtor (or drawer) than the last. Upon exchange, the new bearer endorses payment responsibility of the previous bearer. According to MacDonald and Gastmann (2001, pp. 114), this establishes a chain of drawers that are sequentially accountable for the remittance, should the previous drawer default on his payment obligation. Effectively, these bills became a credit-backed form of payment, which is a precursor to the modern-day bank note.

Dutch finance soon visited English shores, whose banking activities in the seventeenth century AD were predominantly split amongst scriveners, tax-farmers, and pawnbrokers. Scriveners were respected lawyers entrusted with large deposits; while tax-farmers advanced loans to the Exchequer and collected repayment from tax-payers directly, as licensed by the monarch. However, it was the goldsmiths in London who eventually overtook even Amsterdam as the new financiers of not only Europe but also other parts of the world. As discussed in Davies (2002, pp. 249–252), the goldsmiths' affinity with precious metals made them natural currency exchangers and their armoured gold stores soon attracted demand deposits during a time when war, plague, and fire were common. As was the case with the *Wisselbank* receipts and the Dutch bills of exchange, the deposit receipts issued by these goldsmiths became quasi-money in their own right. Unlike the Dutch, however, the goldsmiths started issuing loans that were directly denominated in these receipts instead of physical coins, thereby introducing the first bank note as it were. These 'promissory notes' were widely accepted as payment since creditors trusted that the coins kept in the vaults of the goldsmiths would be readily available when presenting these notes. In turn, the goldsmiths upheld this credit system by keeping a fractional reserve of coins for such withdrawals, which is a practice that largely continues to this day (although using bank money instead of coins).

Aside from the goldsmiths at the time, England saw another significant shift in the balance of power between the British Parliament and the English Crown. From Nichols (1971), monarchs used to control the public purse exclusively (and whimsically) throughout history, often debt-funding ambitious military conquests, sometimes to the ruin of banks. However, the British Parliament won equal rights to governmental finance in 1688, helped by the sanctioned invasion by the Dutch Prince William III, who later ascended the English throne. Thereafter, both monarch

and parliament were jointly held responsible for repaying government bonds, which stabilised the inherent credit risk of these bonds. National debt became a ‘perpetual’ source of loans to the government whose repayment is ‘bonded’ to the nation’s future tax revenue. This balance of power between monarch and parliament promoted using national debt increasingly as a major funding source, which further fuelled the broader credit system, as discussed in MacDonald and Gastmann (2001, pp. 130–134) and Davies (2002, pp. 255–263). Thus the modern age of banking was begun when the Bank of England was founded in 1694, inspired by the earlier Dutch *Wisselbank* and necessitated by the British government’s lack of funding.

In addition to facilitating government bonds, the Bank of England attracted private deposits with favourable rates, most notably that of the Dutch in search of greater yield. Naturally, the Bank’s extensive capital base escalated its subsequent lending capabilities. To this point, the Bank purposefully charged lower interest rates than the goldsmith bankers in trying to dismantle the latter’s monopoly on lending. Soon, even the goldsmiths’ receipts fell into disuse since the Bank’s promissory notes carried with them the credibility of the British government itself. The passage of the Bank Charter Act (1844) entrusted to the Bank of England the exclusive power to print money, a right it still enjoys to this day in England and in Wales. Overall, the Bank played an instrumental role in financing (and greatly profiting from) the worldwide Industrial Revolution, as discussed in MacDonald and Gastmann (2001, pp. 135–138). Moreover, the excess capital of the Bank helped fund many smaller private banks over time. The capital investment in these ‘subsidiary’ banks gave the Bank of England an emerging supervisory role as a *central bank* and ultimately serving as the lender of last resort. The Bank played a stabilising role in the credit system and facilitated the ease at which money was borrowed amongst private banks. In turn, this liquidity and mutual trust ensured the full usage of all financial resources of the British economy across all of its colonies, with little wastage posed by idle savings.

The history of banking continued largely unimpeded throughout the nineteenth and twentieth centuries AD across the new world, with many a bank rising and falling (see section 2.5). Throughout millennia, it is clear that the bedrock of banking has remained largely unchanged, being grounded in mutual trust to this day. Even the modern-day promissory bank notes are ennobled by the public’s trust that their wealth is safeguarded within banks and readily available from banks. Banking itself will likely experience significant stress once the rule of law becomes destabilised, as perhaps best shown by the fall of the Roman Empire or the many royal defaults and wars that bankrupted the Italian and German banks during the Renaissance. Furthermore, history proves that banking operations follows naturally from flourishing trade activity and the wealth it generates. Ancient temples, the Egyptian Ptolemaic grain banks, the Greek *trapezitai*, the Bank of Delos, the Knights Templar and Hospitallers, and the goldsmiths of London all first started as deposit-takers, safeguarding the merchants’ profits. Naturally, this custodianship led to another role: offsetting payments between accounts in giro transfers, followed later by conducting

interbank transfers amongst the Italian banks of the Renaissance. Trade prosperity brought multiple currencies with it and again banks were the trusted intermediaries to sanctify these exchanges. As trading activities grew and the demand for coins rose, banks stepped in yet again to transform idle deposits into more useful loans. Banks advanced credit to selected borrowers from these deposits, relying on a fractional reserve and the public's trust in the credit system. The proliferation of banking means an expansion of the same credit system, which both causes and is caused by booming commerce. This symbiotic relationship is perhaps best exemplified by the explosive growth of British lending during the Industrial Revolution, in turn largely financed by the Bank of England. The ensuing cycle of trust and the flow of credit amongst the triad of depositor, banker, and borrower continues *in perpetuum*; at least until this trust is sundered, usually with disastrous effect on the other agents, or even the nation's prosperity itself.

## 2.2 The rise of consumer credit in modernity

In more recent history, consumer credit has grown at a truly exponential rate over the last few decades, shown in Fig. 2.1 for the United States of America (USA). This credit growth has its origins in the 1920s when Henry Ford and A.P. Sloan started financing vehicle sales for their customers in an effort to boost sales. The later introduction of the credit card in the late 1950s saw the use of credit becoming widespread amongst consumers. Its current estimate of approximately \$11 trillion consists of mortgages, credit cards, personal loans, vehicle financing, overdrafts and other revolving loans for the individual, at least according to the Board of Governors of the Federal Reserve System (2020). For perspective, consumer debt levels in 2007 was 40% greater than total industry debt (\$9.2 trillion) and more than double than corporate debt (\$5.8 trillion) at the time, as discussed in Thomas (2009a, pp. 1–3). Although greatest in the USA, consumer debt in other countries are not far behind, e.g., the United Kingdom (UK) had debt levels in 2007 at £1.4 trillion – a staggering £400 billion growth within the span of a mere three years. From an affordability perspective, US household debt constituted 130% of total annual income at the same time. In fact, this trend of debt levels exceeding household income is true for quite a few countries over the last twenty or so years, of which a few examples are shown in Fig. 2.2.

As an explanation, consider the life cycle theory of consumption of Modigliani (1986), which states that consumers generally have a greater appetite for risk early in their lives; then revert to saving in their middle years; and then draw from these savings during retirement. As such, the greater risk appetite of younger consumers directly translates to a greater level of borrowing relative to older age groups. The growth of the 20–34 years age group rose sharply during the period 1960–1990, which, as noted in Thomas et al. (2002, pp. 24), likely explains the rapid rise in overall household debt levels during the same period, in accordance with this theory. Conversely, the downward pressure seen in Fig. 2.1 from 2008 is directly attributable to the sub-prime mortgage crisis that heralded a global recession, following the spread of foreclosures

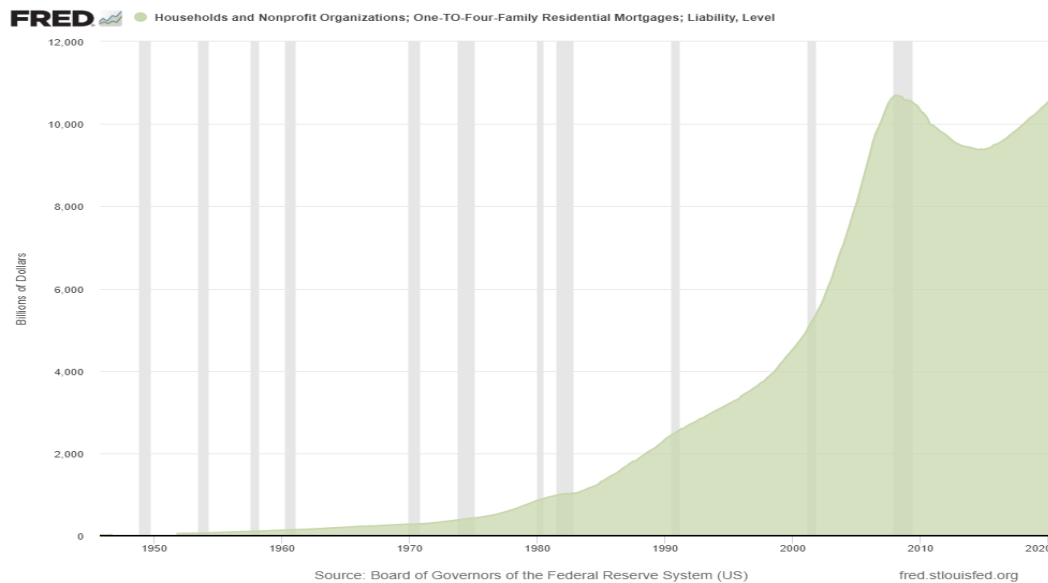


FIG. 2.1: Total household debt in the USA across time, with shaded areas indicating recessions. From the Board of Governors of the Federal Reserve System (2020).

and the number of subsequent bank failures. In retrospect, it seems that this historic financial crisis has only caused a slight blip in aggregate debt levels rather than signalling any structural change, especially since current debt levels have returned to their previous 2008-peak, albeit in the USA.

According to Thomas (2010), the considerable credit expansion could not have been possible without a degree of automation during the credit approval process. Indeed, the development of such automated decision-making models, called credit scorecards, greatly facilitated this rapid growth in consumer credit by rendering consistent approve/decline decisions on high volumes of credit applications. Prior to these formalisations in retail banking, bank managers either approved or declined credit applications by conducting applicant interviews and subjectively assessing the underlying credit risk. They did so using guiding principles known as the five ‘Cs’ of granting credit. As discussed in Finlay (2010, pp. 83) and Van Gestel and Baesens (2009, pp. 93–94), this includes the Capacity to repay (affordability), the applicant’s Character to repay (intent), current macroeconomic Conditions, and Capital or Collateral as possible security. However, this judgemental approach was largely inconsistent over time, typically varying by the daily mood of the bank manager, as discussed in Hand (2001) and Thomas et al. (2002, pp. 9–10). At the same time, one cannot deny that at least some of these credit decisions were fraught with the irrational personal prejudices of these managers. Making a credit decision is often described in literature as an art rather than a science, which is perhaps why it was understandably difficult to teach the craft at the time. Overall, the judgemental approach cannot easily scale with increasing

## 2.2. THE RISE OF CONSUMER CREDIT IN MODERNITY

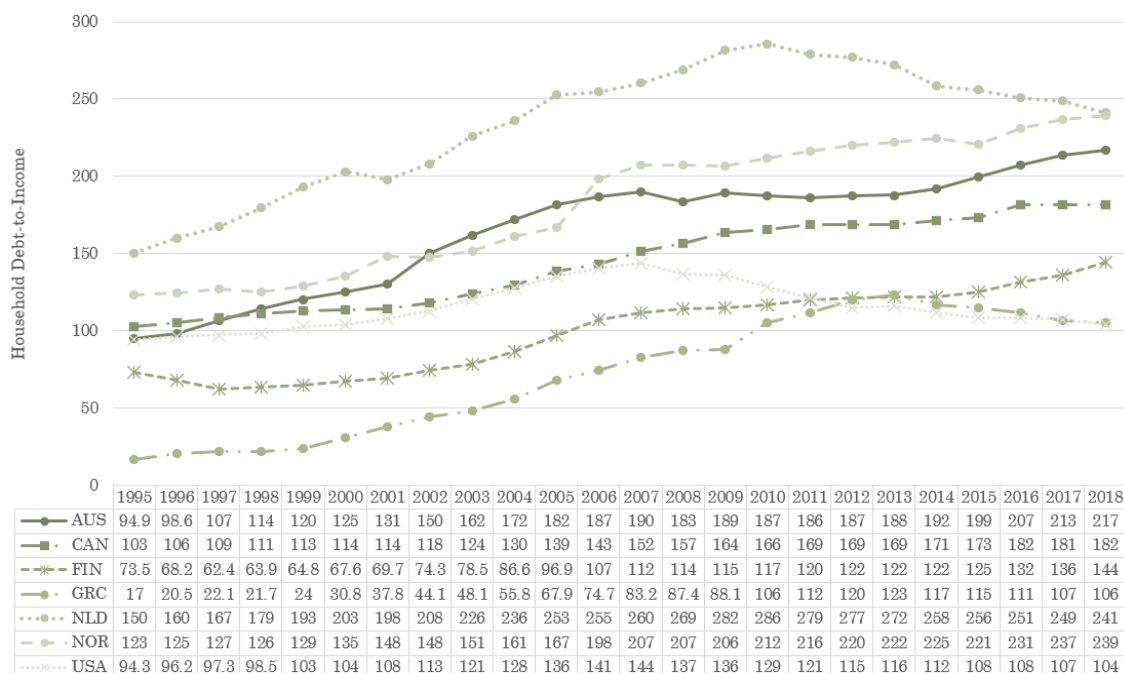


FIG. 2.2: Consumer household debt-to-income over annual periods by country, including Australia (AUS), Canada (CAN), Finland (FIN), Greece (GRC), Netherlands (NLD), Norway (NOR), and the United States of America (USA). Reproduced from OECD (2020).

application volumes; it relies heavily on human decision-makers, which makes the approach expensive; and it yields inconsistent and biased decisions for the same reason.

This credit decision occurred during a very specific step that still exists to this day within the typical five-phase credit management process. As explained in Finlay (2010, pp. 11–13) and reproduced in Fig. 2.3, a lender first devises campaigns to solicit new customers into applying for a pre-designed credit product, during phase one (Marketing). Naturally, the goal is to maximise the pool of potential customers in an effort to maximise the eventual and so-called take-up (or conversion) rate, i.e., the proportion of loan applicants who became borrowers. During phase two (Customer Acquisition), the lender assesses the creditworthiness of applicants in an exercise called "credit scoring". For those deemed creditworthy, the lender prices the loan, i.e., compiling a loan offer that contains a specific interest rate, loan amount, contractual term, and credit limit, as applicable to the specific type of credit product. The goal of this second phase is to minimise bad debt that has yet to develop as a result of granting credit today, i.e., selectively granting credit to those deemed as sufficiently trustworthy in repaying their debts. Of course, this goal may naturally conflict<sup>7</sup> with that of the first phase. Thereafter, phase three (Customer Management)

<sup>7</sup>These two conflicting goals are often balanced against each other by either adopting specific growth strategies at

generally involves monthly housekeeping of the account, e.g., preparing and delivering regular statements, as well as possibly offering further advances and/or cross-selling other credit products, provided the account is in good standing.

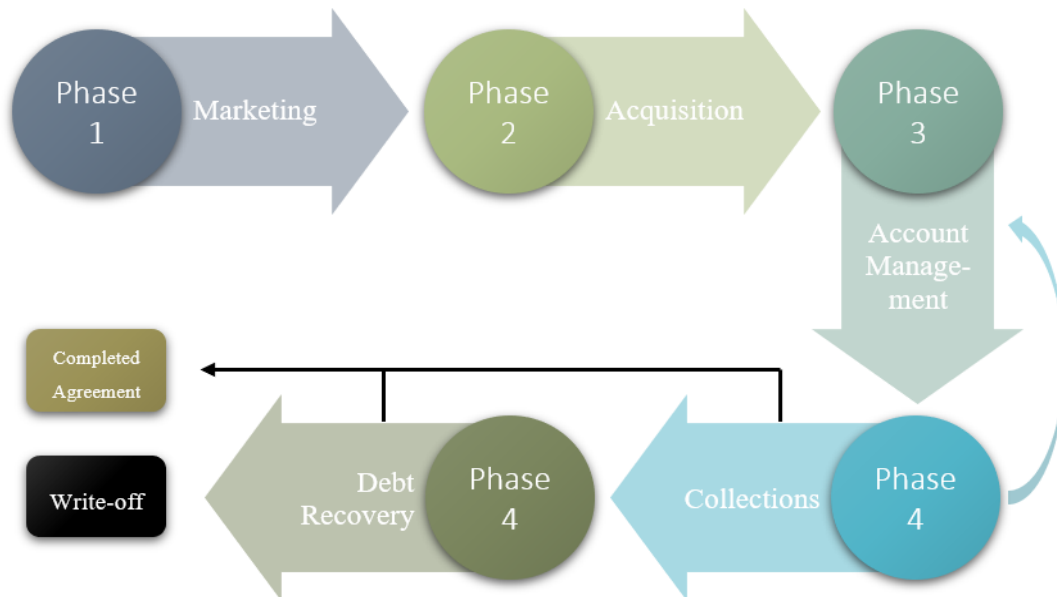


FIG. 2.3: The five-phase credit management model through which all credit agreements progress during their lifetimes. Loan applicants are solicited in 1), credit scored in 2) and serviced in 3) until the ‘natural’ contractual end. However, delinquent accounts are nursed in 4) while doubtful debts are recovered in 5) prior write-off. Reproduced from Finlay (2010, pp. 11).

However, if the account accrues any arrears, it enters phase four (Collections) during which various initiatives are launched in an attempt to nurse the strained relationship between borrower and bank back to health. This may include temporarily lowering instalments, zeroing interest rates, or perhaps extending a payment ‘holiday’ during which the lender suspends instalments as a token of good faith. If the borrower, despite these brokered arrangements, continues to renege on his repayment obligation despite collection efforts, the account enters phase five (Debt Recovery). No longer is it the goal to salvage the broken trust between borrower and bank, but rather to recover as much as possible of the outstanding debt during what is called the *workout period*. This includes seizing any underlying assets that served as loan collateral to the original agreement. In summary then, a credit agreement typically ends in one of two ways: either naturally after the successful repayment of all outstanding debts, or with the remaining debt written (or charged) off after all debt recoveries are taken into account.

It is during this second phase (Customer Acquisition) that the practice of automated credit <sup>the cost of an increased credit risk appetite, or scaling back market share amidst economic turmoil or satiated risk levels.</sup>



scoring became strategic and even critical. Credit scoring both catered for the rapidly increasing volumes of credit applications; as well as fuelled the fiery demand for further credit thereafter as a result of expanded decision-making capacity. From Hand and Henley (1997), Hand (2001), Crook et al. (2007), Thomas (2009a, pp. 5), and Louzada et al. (2016), statistical credit scoring is fundamentally a classification task in which loans (new or existing) are predicted to become either good or bad risks, regarding their future repayment. This is achieved by sampling the repayment performances of past borrowers and extrapolating from it the future performances for new borrowers, based on the similarity in the common characteristics between new and old borrowers. So-called 'good' and 'bad' classes are created as performance polarisations by which old loans are first assessed in retrospect. The purpose thereof is to find a statistically formulated relationship between borrower characteristics and these good/bad classes. The degree to which a new applicant belongs to either class is then given by this relationship as a "credit score", which quantifies the applicant's credit risk on a typical scale of 0 to 999 (higher scores = lower risk). The resulting credit scoring model (or scorecard) is then implemented within a lender's computerised application system, capable of rendering a far greater number of credit decisions based on more factors than what would have been humanly possible. Moreover, given the statistical nature of the scorecard, these credit decisions are markedly more objective than those preceding the era of automated credit scoring.

The work of Durand (1941) first used a statistical method – the Fisher linear discriminant function – to classify loans as either good or bad in what became an early scorecard. From Thomas et al. (2002, pp. 2–4, 41–42), statistical scoring was first explored in response to the rise of metropolitan clothing mail-order companies, which sent goods to customers on credit during the 1930s. Thereafter, the formation of *Fair, Isaac, and Company* (now known as *FICO*<sup>8</sup>) in 1956 saw the practice of credit scoring gain momentum. One particular milestone is the advent of the credit card, including BankAmericard in the USA (known today as VISA) and Barclaycard in the UK in 1966. This feat was largely made possible due to the previous successes of credit scoring, which was enabled by a simultaneous growth in computing power, as discussed in Thomas et al. (2002, pp. 3–4) and Thomas (2009a, pp. 4–5). However, it was arguably the promulgation of the Equal Credit Opportunity Acts of 1975 in the USA (later amended in 1976) that saw credit scoring being wholly accepted as a decision-support tool throughout the banking industry. These Acts prohibited discrimination in the credit decision, unless it was "*empirically derived and statistically valid*", thereby embedding the use of statistical modelling within the bedrock of modern retail banking.

It is perhaps unsurprising that the practice of credit scoring, given its success in credit

---

<sup>8</sup>FICO was founded in San Francisco by engineer Bill Fair and mathematician Earl Isaac. It sold its credit scoring systems widely to American lenders, based on the belief that data – when used intelligently – can enhance business decisions. See <https://www.fico.com/en/about-us#history>.

cards, soon spread to other lending products during the 1980s, including term loans, mortgages, and revolving credit. To this day, the philosophy underlying credit scoring remains rooted in pragmatism and empiricism; it merely seeks to predict the risk of nonpayment and not explain the risk structurally, as discussed in Thomas et al. (2002, pp. 4–6) and Thomas (2009a, pp. 5–6). This pragmatism implies that any characteristic of a borrower – or that of the borrower’s environment or life stage – that strengthens prediction accuracy, ought to be considered within the model itself. These variables include those obviously associated with creditworthiness, e.g., the number of times a borrower has defaulted in the past on other credit products. Other variables pertain to the overall stability of an applicant, e.g., the time spent at the current employer or the tenure based at the current address. Some variables explain the financial sophistication and resourcefulness of an applicant, e.g., possessing credit cards, the tenure at the current bank, being married or not, the number of financial dependants, or the number of other credit agreements held by the applicant. However, some variables are prohibited by law in some jurisdictions – despite any discovered statistical relevance to predicting credit risk – as some legislators commonly believe their use will lead to unfair discrimination in the credit decision, according to Van Gestel and Baesens (2009, pp. 98).

Other than the choice of variables, the spirit of pragmatism also surfaced in various other areas of modelling default risk. An excellent example hereof is that lenders historically estimated a very specific risk: that of the applicant becoming precisely 90 days past due within the next twelve months, if approved. The modelling setup is typically that of cross-sectional models in that two ‘snapshots’ of information are taken at different time points and merged: applicant information and the subsequent loan performance thereafter. From Thomas (2009a, pp. 6–7) and Thomas (2010), varying some aspects hereof – specifically, the period between the two snapshots, or even the default definition – were never of real interest to lenders. The accuracy of the predicted default risk was not nearly as important as the model’s ability to order applicants by relative estimates of default risk, i.e., its risk-ranking ability. However, the recent introduction of IFRS 9 (see subsection 2.6.1) certainly changed this perspective according to Skoglund (2017), having stressed the importance of accuracy over risk-ranking ability.

Given a set of risk-ordered applicants, lenders then tried to find a suitable cut-off score above which credit is granted and beneath which an application is rejected. This cut-off score was again set quite subjectively (and changed quite infrequently) using strategic business factors, e.g., growing market share or tightening the credit supply, thereby presenting another example of pragmatism. For a replacement credit scorecard, the cut-off was often chosen such that the new model theoretically yielded the same number of accept-decisions as that of the previous model. According to Thomas et al. (2002, pp. 145–146), this strategy seeks to instil confidence in the new model, regardless of the new model’s supposedly superior discrimination ability between the good/bad risk outcomes. Once confident, the lender would typically want to realise the benefit of

better discrimination by steadily lowering the cut-off score, thereby accepting a greater proportion of applicants whilst maintaining the same risk appetite. However, the fairly recent work of Jung et al. (2013) provided a more rigorous and dynamic approach to informing this cut-off score more frequently, based on the good:bad odds ratio and how this ratio itself can vary over time.

Another aspect of the inherent pragmatism in credit scoring is that of the recency of application data that is used in model development. While the so-called *sample window* across which application data is extracted typically varies from one to five years, longer periods are generally preferred to encompass as much of the prevailing economic cycle as possible, according to Siddiqi (2005, pp. 31–33). Moreover, the overall recency of this sample (controlled by its exact starting and ending points) is vital since it incorporates not only information on economic conditions at the time, but also the portfolio composition and the effects of a lender’s policies, as discussed in Thomas et al. (2002, pp. 121–122) and Kennedy et al. (2013). A model estimated from training data that is observed during a particularly favourable economic cycle may quickly degrade in its performance as the cycle worsens (or, at least, changes). Similarly, if the development sample overwhelmingly contains data from the distant past, then the model’s predictions may no longer agree with the present reality (market conditions or lender strategies), given the inherent sampling bias towards a long-gone era. On the other hand, too recent a period may prohibit sufficient data maturity to enable a reasonable forecast. As a trivial example hereof, last month’s approved applicants, whilst being representative of current market conditions, clearly will not have enough repayment history from which to model default risk. This trade-off is illustrated in Fig. 2.4.

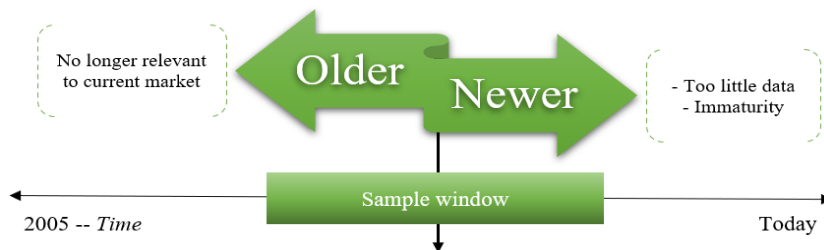


FIG. 2.4: The trade-off when choosing the time period for sampling application data. Recent data may be too immature (or unavailable) while older data may no longer be relevant to current market conditions or representative of a lender’s portfolio or policies.

Given the success of scoring the creditworthiness of new applicants, also known as an *application scoring*, the late 1970s gave rise to a variant thereof called *behavioural scoring* that focuses on existing borrowers. The objective remains that of predicting future nonpayment as in application scoring. However, this prediction task is complemented with additional data observed after credit approval, which theoretically enhances prediction accuracy, as explained in Van

Gestel and Baesens (2009, pp. 101–102) and Finlay (2010, pp. 115, 123–130). This richer data can include subsequent repayment behaviour, income information, and general spending patterns for those customers with a transactional/cheque account. Behavioural scoring is generally used as a decision-support tool in the third phase (Customer Management) within Finlay’s model. There are two broad goals within this phase: 1) to provide customer service (called *operational management*) and; 2) to maximise the return on the borrower relationship over its lifetime (called *relationship management*). Examples of the decisions serviced by behavioural scoring include advancing more credit to the borrower, cross-selling other products, encouraging greater product-use (such as an overdraft add-on to a cheque account), and adjusting the pricing for existing customers in managing attrition risk. This last example is particularly important since there is little benefit in acquiring customers (usually at great cost), only to lose them thereafter to competitor banks.

In principle, changes in the customer’s profile over time often require strategic responses, which may benefit from an updated view on creditworthiness as estimated by a behavioural score. Consider an existing borrower (in good standing) that received a salary increase at some point during the loan life. Naturally, there is expanded scope at this point for raising the credit limit of this now-wealthier customer, as part of maximising the lifetime return on this relationship. Conversely, a borrower who becomes unemployed at some point has a reduced capacity to honour existing debt obligations, even if only temporary. The impact of these events on overall creditworthiness is certainly dynamic, which in turn advocates the use of behavioural scoring. That said, it remains unclear how exactly risk-ordering these borrowers by default risk *directly* impacts profit-optimality within the wider macroeconomic reality and lender policies at play, as argued in Thomas (2009a, pp. 7) and Thomas (2010). Perhaps the practice of behavioural scoring simply fits within the theme of pragmatism, instead of purporting to profit-optimize the aforementioned strategic decisions mathematically.

More recent developments in literature outline a mind-shift to modelling the likelihood of a borrower generating profits instead in an exercise called *profit scoring*, as discussed in Van Gestel and Baesens (2009, pp. 105), Thomas (2009a, pp. 216–220), with an illustration given in Stewart (2011). By replacing the default risk application scorecard, an applicant is either approved or rejected based on a required profit margin, as predicted by the profit scorecard. While this approach is certainly appealing given its closer alignment with business objectives, profit scoring is also plagued by many questions, which leaves the practitioner little choice but to be pragmatic for the time being. Perhaps most notable of these challenges is the base definition of ‘profit’ on the account-level. The notions of direct and indirect costs become challenging to attribute to the individual account since these costs typically depend on the situational context of the lender. As examples hereof, consider fluctuating portfolio sizes, system infrastructure (including the cost of downtime), and staffing costs. Another challenge is the choice of the time horizon over which profit is measured before modelling it. Since profitability is generally perturbed by macroeconomic

conditions, choosing an appropriate time horizon becomes fraught with balancing biases towards periods of economic booms against maintaining a sample that is still representative of current market conditions. Despite these challenges, the inherent appeal of profit scoring as the next logical evolution in credit scoring may spur research initiatives in the near-future, which may very well solve some of these problems.



FIG. 2.5: Stylised metamodel of overlapping factors in model-driven decision-making of a modern bank.

From what is already described as the "third revolution" in credit scoring more than ten years ago in Thomas (2009a, pp. 7), it is clear that the modern-day lender faces increasing competition on two intertwined fronts: the growing demand for consumer credit, as well as the proliferation of consumer preference. The latter is especially important when one considers the ease at which the financially unencumbered consumer can switch to a competing bank and/or product offering. As a result, the lender becomes more amenable to the idea of profit-optimising its many strategic decisions using mathematical rigour, instead of relying just on pragmatism alone. At least from the perspective of making model-driven decisions, the loan amount (or credit limit); the price (interest rates and/or fees); the price sensitivity of a borrower (competitor offerings); the market-appealing mix of product features and overall design; customer selection; the macroeconomic backdrop and timing of the credit offer; managing attrition risk *ex post* acquisition; and risk management regarding capital reserves (section 2.5) and loss provisions (section 2.6) – all of these factors may eventually be modelled together as one dynamical system,

illustrated in Fig. 2.5. A worthwhile avenue of future research may very well explore the intricate relationships amongst these components in a bank's decision-making. It is not hard to imagine the benefits of a rigorous, all-encompassing, and sophisticated profit-optimisation 'supermodel' (or metamodel). The pursuit hereof may soon become tractable, especially when considering the advancements made in machine learning and artificial intelligence.

### **2.3 Financial intermediation and its *raison d'être***

One of the fundamental reasons for the continued existence of banking is due to the so-called *asymmetrical information* that exists between bank and borrower. As argued in Leland and Pyle (1977) and Bhattacharya and Thakor (1993), a bank can generate cheaper and better quality information on the riskiness of borrowers than individual lenders. This advantage is a historical by-product of merchant banks having specialised as trade brokers and intermediaries amongst many agents. These trade brokers generally facilitate trade by matching the buying and selling sides of two parties, usually in exchange for a service fee, as explained in Van Gestel and Baesens (2009, pp. 10). In its role as a trade broker, a bank is able to accumulate and exploit subtle 'signals' (behavioural or market insight) across customers and over time. Moreover, the monitoring of loan repayments on a large scale provides additional information, which further refines the intermediary's subsequent risk analyses. To this point, the work of Diamond (1984) first explored and modelled the role of banks as delegated repayment monitors, which compared favourably to the individual that lends directly.

Consider that a single wealthy investor (or saver) does not have access to the risk information ordinarily held (and refined) by an intermediary. As a result, the individual would therefore need to procure it at great cost to analyse the underlying credit risk of the lending proposition. For every subsequent loan, the individual lender would need to repeat this laborious assessment, which multiplies the previous costs. Moreover, the individual lender's assessments are not corroborated by or supplemented with risk information from institutional lenders, who may already have surveyed more complete information on the prospective borrower. In a multi-agent economy, there is clearly a significant degree of duplication when various individuals try to conduct these risk assessments, which may quickly escalate and introduce gross inefficiencies. This is perhaps exacerbated by the fact that these individual assessments can be subjectively biased, incomplete, or even based on false 'information'. In fact, Santomero (1984) argues that the lack of adequate and trustworthy risk information within a credit market likely necessitates a few firms that are dedicated to sourcing and evaluating said risk information. In this regard, banks are ultimately well-positioned for such a role given the overlap thereof with their classical roles as wealth custodians, transactors, lenders, and exchangers, as illustrated in Fig. 2.6.

Instead of conducting rigorous risk reviews on each prospective borrower, the individual

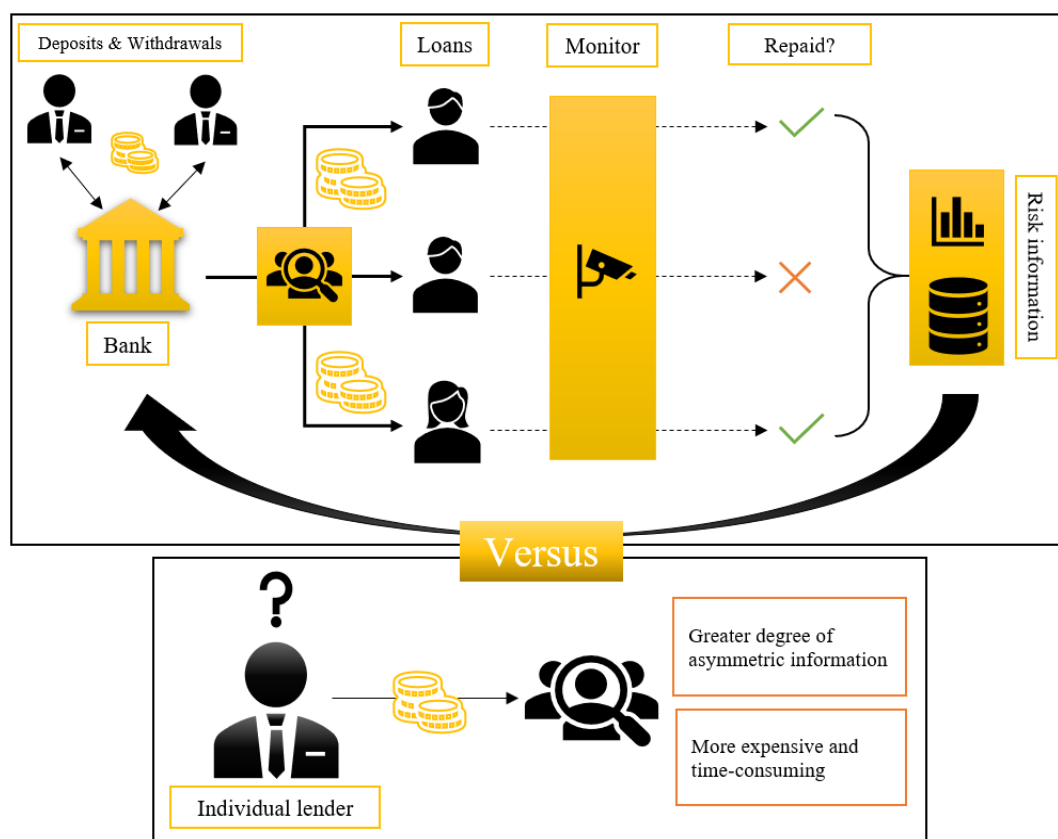


FIG. 2.6: Illustrating the information brokerage function of a bank and its benefit over direct lending and borrowing. Banks can manufacture superior risk information over time as delegated monitors and by eliminating duplicate and/or incomplete risk assessments. Risk information is continuously refined by repeating this process when granting new loans, all of which reduce lending inefficiencies compared to direct lending.

lender can simply charge a single risk-insensitive interest rate as an easier alternative. However, and as argued in Bhattacharya and Thakor (1993) and Van Gestel and Baensens (2009, pp. 12), a single price will likely discourage low-risk borrowers as they will seek better rates elsewhere. Having exited the applicant pool, the lender is left with only higher-risk borrowers that are typically more desperate for funding relative to low-risk borrowers – an acquisition bias called *adverse selection* in microeconomic theory. This bias arises naturally in credit markets due to imperfect information and often leads to a phenomenon called *credit rationing*. In particular, banks offer credit at a certain price level such that loan demand exceeds supply. However, a bank's expected return can actually *reduce* if the loan interest rate (or collateral requirement) increases beyond a certain point when trying to cater for the higher demand, as originally modelled in Stiglitz and Weiss (1981). Credit rationing therefore implies that some applicants will simply never be credit-approved and that risk-insensitive rates will likely bankrupt the individual

lender.

Intermediation theory remains valid in modernity even when considering the recent advent of electronic marketplaces ‘replacing’ traditional banks, i.e., so-called P2P (peer-to-peer) lending. To this point, Berger and Gleisner (2009) studied the role of intermediation on the digital micro-finance platform called *Prosper.com*. These P2P-platforms serve as electronic markets that mediate amongst individual borrowers and small-scale individual lenders, including the subsequent sale of these loans to other interested parties on the same platform. However, the embedded credit-screening process produces risk information that is functionally similar to that of a traditional intermediary (or bank). Furthermore, the *Prosper.com* platform hosted informal and decentralised social networks of borrowers and lenders, who can vouch for one another. The resulting friend endorsements are then useful in assessing creditworthiness for future loans via the platform. In fact, Freedman and Jin (2008) showed that the monitoring of loan repayments (as conducted by these small online communities) encouraged loan repayment overall and contributed to lower credit risk, which is again similar to traditional banks as delegated monitors. Both of these studies align with the propositions of Sarkar et al. (1998) in that the proliferation of electronic markets will simply lead to new forms of intermediation (e.g., ‘cybermediaries’) instead of displacing intermediation theory entirely, as initially espoused.

As trusted financial intermediaries amongst transacting agents, banks are mainly in the business of pooling shorter term liquid cash deposits and transforming these into longer term illiquid loans. Put differently, the surplus funds of savers are made productive by lending it to borrowers/consumers. This feat is generally achieved when banks fulfil their so-called *asset-transformation* function, as illustrated in Fig. 2.7 and reviewed in Santomero (1984). A bank offers to pay interest on these deposit contracts, thereby attracting buyers (or ‘depositors’). The cost thereof is offset by then charging interest and fees when lending some of these deposits to borrowers, after which the bank retains the difference as revenue. When compared to purchasing fixed-income securities with similar yields, deposit contracts may have cheaper transaction costs, provides liquidity, and offers convenience. This last point is important since depositors would alternatively need to assume the role of an investment analyst, as argued in Merton (1977). By implication, a depositor would need to hunt for a risk-equivalent security from each competing firm; then scrutinise the balance sheet and management (amongst other factors) of each security, before selecting one to purchase. Clearly, this is an arduous and inconvenient process compared to purchasing a simple deposit contract with a similar yield.

Another important facet of these deposit contracts is their maturity profiles, which can range from shorter term current accounts (or *demand deposits*) to longer term fixed deposits (or *term deposits*). Lending is only truly viable as long as a bank can fund the sporadic withdrawals (and maturing term deposits) of depositors. A bank attempts to manage this cash flow ‘traffic’ by



holding a sufficiently stable fractional reserve of cash, as discussed in Van Gestel and Baensens (2009, pp. 9–13, 20–21). Moreover, a bank uses its risk expertise to ensure that its subsequent lending activities are sufficiently diversified. This diversification includes balancing the appetites and needs of borrowers (and the bank’s profit potential) against the risk of capital loss associated with lending to the very same borrowers. Risk-based lending therefore demonstrates how the previous information brokerage function complements the transformation of deposits into loans.

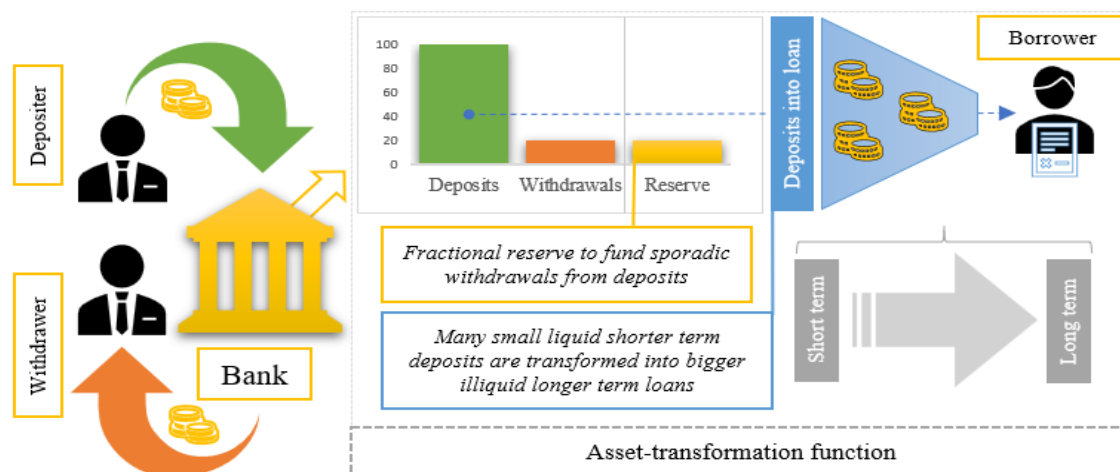


FIG. 2.7: Illustrating the asset-transformation function of a bank wherein numerous liquid deposits (usually shorter term) are collected and used to fund more illiquid (usually longer term) loans to borrowers. A fractional reserve is maintained to fund sporadic withdrawals. Arrows indicate the flow of funds.

## 2.4 The quest for bank liquidity and overall system stability

To better understand the role of trust (and its collapse) in modern banking, the basic mechanics of a typical bank and its funding are discussed in subsection 2.4.1. Sources of funding are broadly categorised across a depositor franchise, debt-based instruments, retained earnings, and equity originally invested into the bank. These funds are primarily re-issued as risk-bearing loans, though banks are increasingly diversifying their revenue streams across fee-producing services. Of course, subsequent deposit withdrawals and maturing debt obligations pose a fundamental risk to a bank in the form of a liquidity shortage. The strategic management of this particular risk, reviewed in subsection 2.4.2, is made non-trivial largely due to unexpected market movements, behavioural dynamics of borrowers, and the prevailing business strategy. In light of this complexity, some authors have devised mathematical models to help manage bank liquidity, as discussed in subsection 2.4.3. However, managing a single bank’s reserves independent of other banks may still fail in fending off a widespread liquidity crisis. As such, many governments have since devised stopgap solutions such as deposit insurance and lender

of last resort strategies. Though these interventions can indeed stabilise a system somewhat, they also carry inherent trade-offs (see subsection 2.4.4) and other costs. Ultimately, the task of maintaining adequate liquidity will likely remain an endearingly complex and never-ending quest for any bank, which arguably justifies their unique role in modern society even further.

### 2.4.1 The mechanics of a modern bank and its funding

Historically, a bank was largely funded by the deposits collected from individuals (called *retail deposits*), corporates, small-to-medium-sized enterprises (SMEs), governments and parastatals. Other external liabilities include interbank funding (since banks lend to one another) as well as debt securities issued by the bank to raise capital from willing buyers, as explained in Dermine (2007, pp. 495–497) and Van Gestel and Baesens (2009, pp. 27). The funding mix between the deposit franchise and the issued debt securities (or institutional funding) can vary by bank, time, and even the markets in which they operate. As an example, FirstRand Bank in South Africa maintained a 64%-36% split between deposit and debt instruments respectively; the competing Standard Bank group had a similar funding mix of 63%-37% – see the financial results of FirstRand Bank Limited (2019) and the Standard Bank Group (2018). These two South African banks purposefully aim to fund their operations primarily from an extensive depositor franchise, given the particular structure of the South African credit market. Specifically, the high degree of contractual savings held in pension/provident funds and asset managers, poses as an attractive funding source for South African banks. Moreover, ZAR-denominated transactions are entirely cleared and settled within the South African banking system, which further supports the idea of using domestic deposits instead of relying on foreign funding.

Debt-based funding programmes issued by a bank are typically hierarchical, with some debt types enjoying greater priority regarding their repayment than others when facing bankruptcy. This hierarchy can range from senior tranches down to junior debt<sup>9</sup>, as illustrated in Fig. 2.8. From the investor's perspective, the higher debt tranches are safer bets in general than their lower-tranche counterparts exactly due to the former's elevated repayment priority. This was evidenced in Schuermann (2004b) when comparing bond repayments during the 1990–91 and 2001 recessions, as well as in Schuermann (2004a) when studying empirical losses suffered on defaulted corporate bonds. However, this presumption of low risk is not always the case, with violations in the so-called *absolute priority rule* (or APR) more common than one would think, as studied empirically in Longhofer, Carlstrom et al. (1995). In essence, this principle states that a distressed debtor shall receive no value from his assets until all creditors have been repaid, with priority given to senior creditors. The authors modelled circumstances where the strategic violation of APR would actually be optimal for all bankruptcy participants, including

---

<sup>9</sup>Subordinated debt may also qualify for up to 50% of tier 2 regulatory capital held against unexpected losses, as discussed in Van Gestel and Baesens (2009, pp. 351) and later in section 2.5.

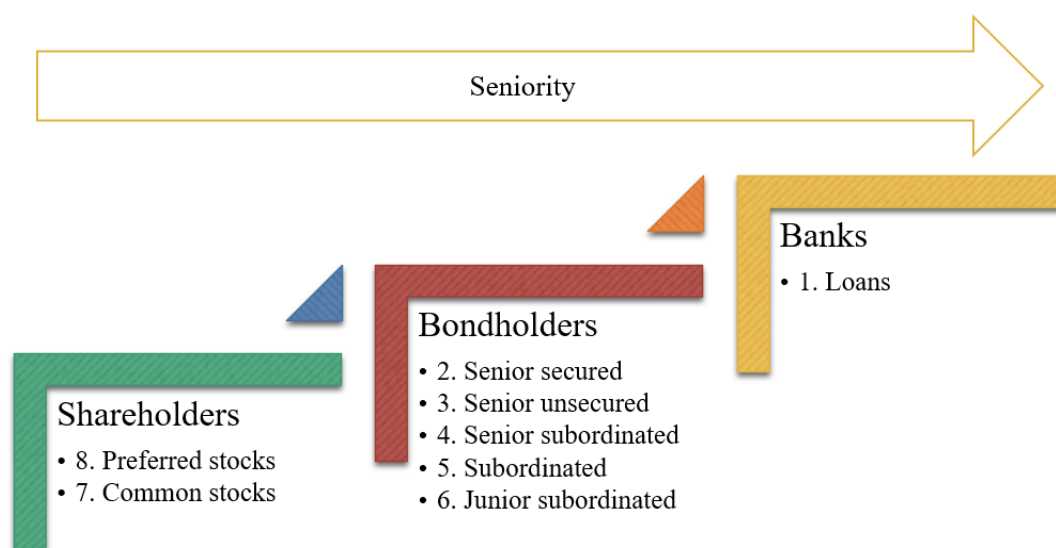


FIG. 2.8: Illustrating the seniority of different types of debt holders, given default. After bankruptcy proceedings, bank loans are honoured first from the proceeds, followed by the bondholders, then the shareholders. Adapted from Schuermann (2004a) and Van Gestel and Baesens (2009, pp. 27, 67).

the bondholders themselves. Primarily, the prevalence of APR violations can depend on the cost of bankruptcy versus the benefit of reorganising a distressed firm (instead of pursuing its liquidation).

<b>Assets</b>	<b>Liabilities and equity</b>
Loan assets (advances)	Demand deposits
Cash and government bonds	Term deposits
Interbank loans	Interbank deposits
Investment securities	Debt securities
Property and equipment	Equity (capital & reserves)
	Equity (shareholder's equity)

TABLE 2.1: A simplified balance sheet of a bank.

Other than debt and deposits, a bank has liabilities to its owners in the form of reserves, as well as retained earnings and capital originally invested by its shareholders. Capital reserves may act as financial buffers against economic headwinds in the future and absorb unexpected losses, as discussed in section 2.5. For expositional purposes, typical assets and liabilities are shown in a simplified version of a bank's balance sheet in Table 2.1. A bank may further hold certain off-balance sheet items, which can create a cash flow contingent on some event in the future, as explored in Dermine (2007, pp. 492). Examples hereof include loan commitments, guarantees, and financial derivatives (such as forwards, options or swaps) of which the payoffs

are related to movements in interest rates, exchange rates, commodity or equity prices.

Apart from holding adequate capital reserves, most of the bank's procured funding is transformed into credit assets, which will then generate interest income and fees for the bank over time. The type of borrowers can range from retail customers, SMEs, larger corporates, governments, as well as other banks. From Van Gestel and Baesens (2009, pp. 19), Finlay (2010, pp. 2–8), and Phillips (2013), a bank may offer various types of loans, which can be characterised using the following common factors. A loan may be secured by underlying collateral or be completely unsecured (e.g., mortgages and auto loans vs. term loans); it may have a fixed repayment term or be open-ended (e.g., amortising loans vs. credit cards); it may have different fee schedules that apply conditionally (e.g., restructuring fee vs. monthly account fees); and it may have different repayment schedules (e.g., instalment finance vs. bullet loans) that provide the bank with revenue streams across different time horizons. Furthermore, a bank may securitise<sup>10</sup> various credit assets into more liquid and marketable securities to be sold to other agents. The proceeds thereof can be used as an additional funding source with which to finance new loans that may yet again be securitised. Other than investing in credit assets, a bank may use the funds to buy equity investments in other companies as well as hold derivative instruments. However, these investment decisions are not strictly unique to core banking (or lending) activities and therefore outside of the scope of this study.

In addition to the interest income and endowments earned by a bank, another main source of income are fees that are levied for services rendered to the bank's customers, sensibly called *non-interest revenue* (NIR). Income derived from trading activities and payouts received from insurance contracts are typically included in a bank's NIR, as discussed in Dermine (2007, pp. 495–498) and Van Gestel and Baesens (2009, pp. 17–23). Furthermore, Allen and Santomero (2001) showed that modern banks have deliberately diversified their traditional asset-transformation role (from which they derive interest income) to include more fee-producing activities. These services can include the management of trusts, mutual funds, mortgage banking, transaction services, insurance brokerage, underwriting annuities, and trading. As an example, the financial results of FirstRand Bank Limited (2019) demonstrate the active pursuit of revenue diversification, with an NIR reported at 42.5% of total income.

All of these funding sources attract an expense of sorts, e.g., interest on debt or dividends for equityholders. In particular, a bank will likely have to pay interest to depositors in exchange for using deposits as a loan funding source. Another form of payable interest is to holders of any debt securities that the bank may have issued in the past. More importantly, a bank has a special expense directly related to realised credit risk called an impairment charge (or loss). This

---

<sup>10</sup>See Bhattacharya and Thakor (1993), Van Gestel and Baesens (2009, pp. 76–81), Vento and La Ganga (2009), and subsection 2.6.2.

Item	Symbol
+ Interest income and endowments	$a$
- Interest expense	$b$
- Impairment charge	$c$
<i>Net Interest Income (NII)</i>	$x_1$
+ Fees and commission income	$d$
+ Trading and investment income	$e$
+ Insurance income	$f$
<i>Non-Interest Revenue (NIR)</i>	$x_2$
<i>Income from operations</i>	$x_3$
- Operating expenses	$g$
<i>Earnings before tax</i>	$x_4$
- Tax	$h$
<i>Earnings after tax</i>	$x_5$

TABLE 2.2: A simplified income statement of a bank. Plus-signs depict income and minus-signs represent expenses, all of which are respectively denoted by  $a, \dots, h$ . Italicised line items are subtotals, expressed as  $x_1 = a - b - c$ ,  $x_2 = d + e + f$ ,  $x_3 = x_1 + x_2$ , and  $x_4 = x_3 - g$ . Net profit is expressed as  $x_5 = (a - b - c) + (d + e + f) - g - h$ .

charge is usually offset against the difference between interest earned and interest paid, which is in turn called *net interest income* (NII). Lastly, a bank incurs various operating expenses when conducting its risk-based business, e.g., staff salaries, marketing, audit fees, computer expenses, repairs and maintenance, insurances, lease charges, and donations. These various income and expense items are shown in a simplified income statement in Table 2.2.

#### 2.4.2 Managing the fundamental risk of illiquidity

Transforming deposits into illiquid loans certainly poses the fundamental risk of not being able to fund withdrawals beyond a particular level. This is particularly pertinent when depositors rush *en masse* to reclaim their funds, thereby triggering a possible liquidity crisis. However, the definition of ‘liquidity’ can be ambiguous at times and certainly contextual, as explored in Vento and La Ganga (2009) and Sekoni (2015). So-called *market* liquidity generally refers to the speed at which an asset can be converted into cash without significantly affecting its price. In contrast, *funding* liquidity (or bank liquidity) is implicitly described by Basel Committee on Banking Supervision (2006b) as a type of ‘reservoir’ from which a bank can draw to support financial intermediation. Vento and La Ganga (2009) describes ‘liquidity’ more broadly as the ability of a bank to coordinate an equilibrium between financial inflows and outflows over various time periods; a definition adopted in this study. These various cash flows are inextricably connected to the asset-liability mix of a bank. Accordingly, the inability of a bank to meet maturing short-term

obligations from available funding is then called *liquidity risk*, where ‘obligations’ and ‘funding’ can both relate to either assets or liabilities. In particular, paying maturing liabilities and meeting scheduled draw-downs on previously-granted credit assets are both obligations.

The main difficulty in managing a bank’s liquidity risk is due to the differences in the maturity dates between extended loans and a bank’s funding liabilities, both of which are typically interest-sensitive. A so-called ‘gap-analysis’ (or maturity ladder) attempts to find significant differences between total cash outflows versus inflows at each successive future period – a trivial exercise for fixed cash flows across fixed timelines, with an example thereof given in Fig. 2.9. However, uncertain cash flows complicate this exercise due to behavioural elements (e.g., defaulting), as discussed in Dermine (2007, pp. 495, 516–525), Vento and La Ganga (2009), Van Gestel and Baesens (2009, pp. 33–37), and Sekoni (2015). To incorporate the inherent uncertainty underlying such a gap-analysis, one may try to forecast loan ‘production’ (or sales) into the future; alternatively, try to forecast future draw-down levels on credit facilities. However, it remains a non-trivial exercise that depends on product design and price elasticity. Another complicating factor is that of the chosen business strategy, which generally has either a value-driven or a growth-driven focus. According to Van Gestel and Baesens (2009, pp. 43–44), the former tries to maximise long-term profitability by preferring good credit quality, whilst the latter may sacrifice credit quality in exchange for short-term growth in market share.

Fundamentally, obligations should not exceed available funding from a cash flow perspective. While an equilibrium would be ideal, a so-called *negative* liquidity gap is preferred over its converse, as illustrated in Fig. 2.9 for periods 2 and 3. To this point, a *positive* gap implies that total outflow exceeds total inflow, which can signal a liquidity crisis at the relevant periods. In such an event, a bank may have little choice but to procure expensive emergency funds by issuing debt or by selling some of its assets, in an effort to remain liquid and a going concern. However, a negative liquidity gap carries an opportunity cost in that excess liquid funds are unproductive or even unprofitable. From Allen and Gale (2017), this opportunity cost manifests when compared to the better returns of more productive long-term loan assets as an alternative to holding liquidity. In addition, excess funding may carry interest-rate risk related to the banking ledger, i.e., adverse movements in funding rates that negatively affect the financial statements. Note that this discussion excludes interest-rate risk affecting market positions held in the trading ledger, which are typically covered when managing *market risk* (see section 2.6). According to Dermine (2007, pp. 516–519), Van Gestel and Baesens (2009, pp. 35–37), and Finlay (2010, pp. 164–166), the most notable of these interest-rate risks is that of *repricing risk*. Having matched the maturing cash flows of assets against maturing obligations at each period, subsequent rate movements can cause net losses (or shortfalls) after the fact. In addition, repricing risk can arise when there are differences in rate type (floating vs. fixed) or in maturity profiles (short-term vs. long-term), as measured between assets and liabilities in both cases.

## 2.4. THE QUEST FOR BANK LIQUIDITY AND OVERALL SYSTEM STABILITY

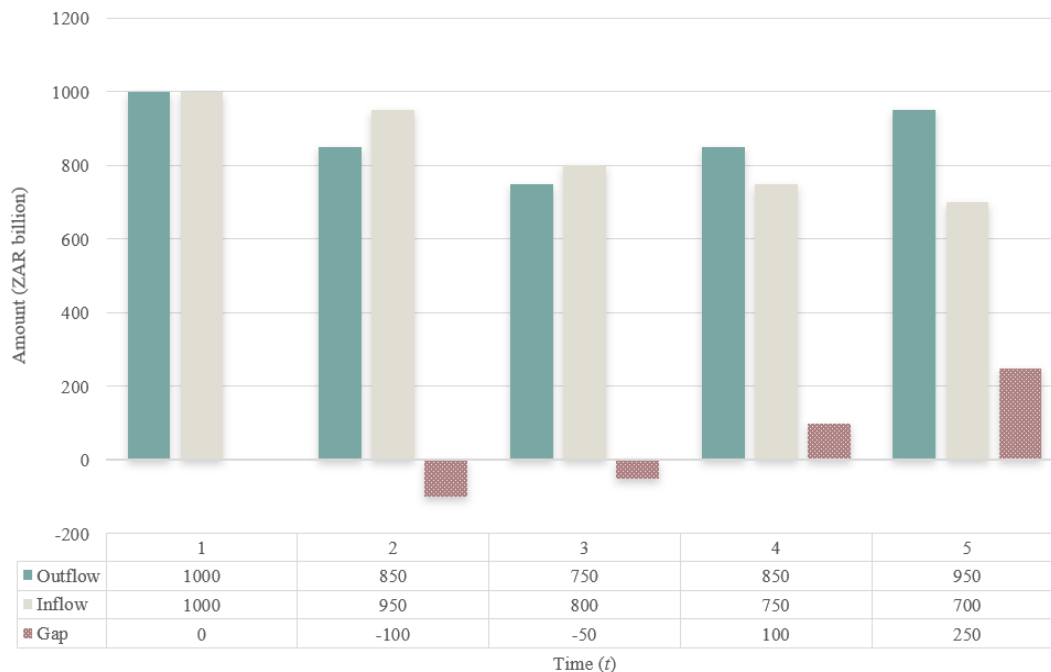


FIG. 2.9: Bank liquidity gap analysis, showing a widening shortfall in available funding to support intermediation from  $t \geq 4$ . For example, the cash outflows from total loan assets at  $t = 4$  are forecast to be ZAR 850 billion. At the same time, cash inflows from scheduled new liabilities or revenue from credit assets are forecast only to be ZAR 750 billion. Additional funding will be needed to cover the shortfall of ZAR 100 billion. Recreated from Van Gestel and Baesens (2009, pp. 34).

Banks generally prefer to borrow short-term and lend long-term, since short-term funding is typically cheaper than long-term funding regarding rates. However, short-term funding can heighten liquidity risk as these liabilities become due more frequently than the longer-term variety, even if the latter is more expensive. Consider financing a single long-term fixed-rate mortgage using many short-term smaller deposits. Apart from securitisation (see subsection 2.6.1), the bank cannot truly liquidate the loan asset to fund maturing deposits, which implies sourcing funds from elsewhere lest a liquidity crisis is triggered. If the maturation is uncertain, then the bank may decide to maintain a light positive liquidity gap at times to realise higher returns. Given the business strategy and market conditions, banks may very well switch between these two types of liquidity gaps in realising either benefit, with the trade-offs thereof summarised in Fig. 2.10. Ultimately, funding sources are often blended across differing maturity profiles to balance cost efficiency against overall liquidity risk.

Another challenge in managing liquidity risk is that of optionality, which is the risk of sudden and unexpected changes in either the maturities or the balances of loan assets. Optionality is largely unavoidable due to the discretionary nature of borrower behaviour, commonly presenting

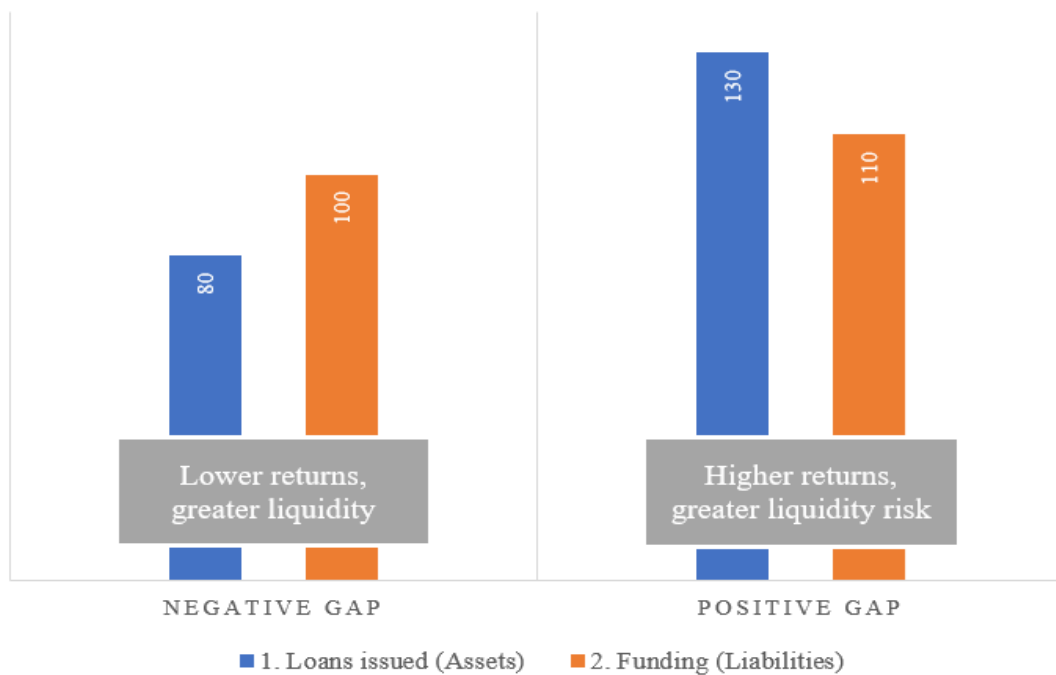


FIG. 2.10: Two types of hypothetical bank liquidity gaps, with advantages and disadvantages. Denominated in ZAR billions.

as the early settlement (or ‘pre-payment’) of a loan. In this case, early settlement debilitates the force of interest and thereby affects overall bank profitability, as discussed in Van Gestel and Baesens (2009, pp. 36–37). Borrower behaviour itself is more difficult to model than even relatively cyclical or seasonal cash flows (when aggregated). As such, banks often use macroeconomic scenario analyses that model borrower behaviour as a function of consumption, prime lending rates, inflation, economic growth, etc. These scenarios help inform a range of different liquidity needs, for which a bank will hold a general liquidity buffer. The exact level, according to Dermine (2007, pp. 522–525) and Van Gestel and Baesens (2009, pp. 47–51), generally depends on the overall business strategy of a bank, seasonality of future cash flows, ease of borrowing and its related costs, and expected macroeconomic turbulence. This rather short-term buffer should not only service the average level of immediate depositor withdrawals, but also be large enough to fund new loans that are expected to be granted over the short-term. Moreover, the liquidity buffer acts as financial grease to the friction of payment delays, and is first consumed by maturing contingent liabilities before designated assets are liquidated. Both liquidity and interest-rate risks are sufficiently significant that they are attributed to its own organisational function, known as *asset & liability management* (ALM).



### 2.4.3 A model for managing a bank's reserves

Via banks, depositors and borrowers are able to invest and consume at will and at random times, thereby sharing both liquidity and liquidity risk amongst themselves. This was first mathematically modelled in Diamond and Dybvig (1983) across multiple equilibria (or outcomes) for simple demand deposits. The authors found that 1) when calm prevails, liquidity and risk-sharing can be efficient; 2) when agents panic and withdraw their funds, a *bank run* occurs that disrupts both liquidity and risk-sharing, which impacts overall production in the economy. Banks that are funded purely with demand deposits will likely be even more concerned with maintaining confidence amongst agents. To this point, sources of angst can include any commonly observed event in the economy, e.g., a bad earnings report, a widely publicised run at another bank, a pessimistic government forecast, or even sunspots (given the possible interference with electronic systems). Historically, banks temporarily suspended withdrawals as a defence mechanism against bank runs, e.g., the US banks imposed a week-long banking 'holiday' during the 1930s. This strategy is more formally known as *suspension of convertibility* in that deposits are temporarily banned from being converted back into cash. However, suspending convertibility proved to be sub-optimal for risk-sharing since banks do not know exactly how many withdrawals will occur at future periods; the threat of such a suspension may nonetheless erode confidence further and impact liquidity thereafter.

In optimising a bank's balance sheet, there are two main literature branches outlined in Baltensperger (1980) and Santomero (1984), namely, models for reserve (or liquidity) management and for portfolio composition. The first branch, discussed in this section, deals with optimising the quantity of primary/secondary reserves held by a bank to offset stochastic reserve *losses*, caused by inadequate deposit levels. Reserve modelling has its origins in Edgeworth (1888) who studied periodic withdrawals from the Bank of England and conjectured that the "*calculus of probability*" may govern the limit of this reserve kept by a bank. Today, liquidity management is essentially treated as a problem of inventory optimisation given stochastic demand in the field of operations research.

As a basic model of managing this reserve, consider two asset types: non interest-bearing liquid reserves  $R$  and interest-earning assets  $A$  yielding  $r_A$  net of all costs. Assume that the bank is a so-called *price-taker* in the credit market and cannot adjust  $r_A$  based on the volume of loans it extends. Alternatively, a modification is provided in Baltensperger (1980) wherein a bank relates  $r_A$  negatively to the amount of credit extended. However, this modification does not change the reserve model fundamentally and is therefore discarded in this study. Assuming price-taking then, let  $W$  denote the stochastic withdrawals, net of all deposits, credit line usage, and loan repayments, with its associated probability density function given as  $f(W)$ . If the reserves are insufficient to offset net withdrawals, i.e.,  $W > R$ , additional emergency funds must be procured

at a cost  $c$ . For simplicity, assume this cost is proportional to the reserve deficiency  $W - R$  itself, i.e., the cost per additional unit of currency needed. The optimisation problem is then to balance the amount of reserves kept for withdrawals against the amount with which to fund new loans, each with its associated costs. To maximise expected profit from deposits  $D$ , a one-period<sup>11</sup> model is accordingly specified as

$$\tilde{R} = r_A R - \int_R^\infty c(W - R)f(W)dW. \quad (2.1)$$

The function  $\tilde{R}$  in Eq. 2.1 has two main terms. First, there is the opportunity cost  $r_A R$  on holding bigger reserves  $R$ , which could have been used for funding loans yielding  $r_A$ . Second, there is the expected cost of having a deficient reserve, i.e., the liquidity cost. As such, for each extra unit of currency held in the reserve, the bank not only incurs a marginal opportunity cost of  $\frac{d}{dR}(r_A R) > 0$ , but also enjoys a marginal reduction in liquidity cost  $-c \int_R^\infty f(W)dW < 0$ . Minimising the sum of these cost items means equating them, i.e., the first-order condition of this optimisation problem will give a reserve amount that satisfies

$$r_A = c \int_R^\infty f(X)dW. \quad (2.2)$$

In other words, reserves are set such that the probability of a deficient reserve  $\int_R^\infty f(X)dW$  equals the ratio  $\frac{r_A}{c}$ , as a fundamental condition to be met in most reserve models. However, one should be careful when interpreting these parameters at face-value. The cost of a deficient reserve  $c$  may not be linear in reality and can be affected by uncertain access to funding markets at the time. Furthermore, the net yield  $r_A$ , although net of all costs and inclusive of all fees charged by the bank, implies that optimal reserves only depend (relatively) on interest rates, though not (absolutely) on the interest amounts themselves, which seems unrealistic.

Should it be necessary, regulatory reserve requirements can be incorporated into the aforementioned model without drastically altering the economics thereof. The major effect of regulatory requirements would simply amount to altering the critical value against which withdrawals are compared. That is, the new reserve deficiency  $W > R^*$  is used instead of the previous  $W > R$  where  $R^*$  may incorporate various legal requirements, as originally modelled in Poole (1968) and reviewed in Baltensperger (1980). Furthermore, better information on a bank's customers may reduce the variance underlying the withdrawals  $W$ , which may decrease overall reserve costs. This was previously incorporated exogenously in Santomero (1984) as a third 'information-cost' term, which was simply subtracted from Eq. 2.1. However, its quantification may be challenging in reality.

The reserve model developed so far clearly depends a great deal on the withdrawal distribution  $f(W)$  as a proxy for deposit fluctuations. It is argued in Baltensperger (1980) that a Gaussian

---

<sup>11</sup>Most of these models were designed for short-term horizons, e.g., managing the bank's reserve position daily and supplementing shortfalls from the interbank market or the central bank, as explained in Poole (1968).

distribution can approximate  $f(W)$  since  $W$  is the sum of a large number of (presumably) independent changes across the balances of individual deposit accounts. This Gaussian assumption can only ever be an approximation since the upper limit of  $W$  will be equal to all deposits initially held by a bank, which is a *finite* amount, thereby contrasting the Gaussian distribution's *infinite* domain. The author further argues that if one assumes that the net withdrawal distribution is symmetrically anchored around 0 (for simplicity), i.e.,  $E(W) = 0$ , then optimal reserves  $R$  can be defined using a multiple  $b$  of the standard deviation of  $W$  (denoted as  $\sigma_W$ ), where  $b$  may be related to the previous ratio  $\frac{r_A}{c}$ . This  $R$  is then expressed as

$$R = b\sigma_W. \tag{2.3}$$

A few intuitive remarks on  $f(W)$  are in order. It is quite reasonable that  $f(W)$  ought to depend on the characteristics (e.g., volume, maturity structure, and costs) of the deposits held by a bank, as formalised in Miller (1975). Should initial deposits  $D$  increase overall, it is likely that  $\sigma_W$  will also increase and, by extension, so too shall the optimal reserve  $R$  – though not proportionally. This relationship is sensible as long as the increased  $D$  occurs alongside an increase in the number of (independent) sporadic withdrawals. Finally, if  $D$  is redistributed towards more volatile deposit types (e.g., more demand deposits than term deposits), then both  $\sigma_W$  and  $R$  will reasonably increase as well, and *vice versa*.

#### 2.4.4 Two interventions to reduce bank fragility

Previous banking failures and the erosion of trust through the ages (as reviewed in section 2.1) can motivate the design of more focused interventions to safeguard both bank and borrower. Apart from a bank suspending its deposit convertibility, there are two broader government-run instruments to help guard against liquidity shocks reverberating across the financial system itself. These instruments include *deposit insurance* (DI) and a *lender of last resort* (LLR) strategy. The work of Merton (1977) first explored DI schemes and the costing thereof using option pricing theory. Whilst successful, DI schemes can be expensive with the costs thereof often borne directly by taxpayers following a liquidity crisis, leading to a deadweight loss. Apart from costs, DI schemes pose an inherent disincentive for depositors to demand interest rates that are commensurate with the bank's risk appetite. Moreover, if the DI premiums charged to banks are risk-insensitive, as is commonly the case according to Bhattacharya and Thakor (1993), then DI schemes may inadvertently encourage excessive risk-taking amongst some banks in the system. Some authors have, however, found little empirical evidence between DI schemes and greater risk-taking as a result thereof, while others found a worrying relationship between DI schemes and bank failures, as reviewed in Santos (2006). In fact, the work of Anginer et al. (2014) studied the empirical impact of DI in averting contagious bank runs across 96 countries during (and preceding) the so-called *2008 Global Financial Crisis* (GFC). While systemic stability was greater

during the crisis, DIs generally had a detrimental effect on the risk of bank failure during calm periods. Ultimately, the latter outweighed the benefit of stability during the crisis across the full sample period.

Amongst the many proposals to improve DI schemes, most notable is that of pricing their premiums fairly given a bank's underlying risk profile. According to Bhattacharya and Thakor (1993) and Santos (2006), this is a non-trivial exercise primarily because of asymmetric information between bank and insurer. Moreover, risk-sensitive DI schemes will be particularly difficult to implement across the banking system *without* causing 1) safer banks cross-subsidising riskier banks; 2) increased regulatory inspection to characterise banks' portfolios. The first outcome distorts the efficient market allocation of deposits amongst banks, as demonstrated in Taggart and Greenbaum (1978) and Chan et al. (1992). The second outcome mandates intrusive monitoring, which is not without its costs as well. That said, linking capital requirements intrinsically to DI premiums can satisfy the aforementioned conditions. In particular, the work of Chan et al. (1992) showed that an equilibrium exists when riskier banks select lower capital levels at the cost of higher premia per dollar of insured deposits; while safer banks choose lower premia afforded by higher capital requirements.

As an alternative to DI schemes, the prevailing central bank can instead provide emergency funding to distressed banks under an LLR arrangement. Given the historical size of most central banks, it was only natural for them to step in during liquidity crises and lend to distressed banks. According to Santos (2006) and Allen and Gale (2017), LLR setups generally predate DI schemes with the Belgian National Bank being the first to fulfil an active LLR role in the 1850s, followed by the Bank of England (amongst others) in the 1870s. A common rule at the time was that rescued banks must still be solvent even if illiquid, which theoretically reduces the LLR's risk exposure somewhat. However, many authors have since questioned the mythical separation between illiquidity and insolvency in practice. Furthermore, the whole purpose of an LLR strategy is to insure against liquidity shocks, which is counteracted when the central bank withholds credit from some distressed banks but not others. Similar to DI schemes, liquidity assurances from a central bank can also result in a perverse incentive for banks to increase credit risk deliberately and/or maintain greater positive liquidity gaps, as argued in Diamond and Dybvig (1983). Ultimately, both DI and LLR schemes are not perfect and are still actively researched and refined by governments to this day.

Negating the perverse incentives (or moral hazards) that are exerted by both DI and LLR setups is difficult, especially at the system level. Perhaps a better course of action is their joint incorporation into broader bank regulation and integrated monetary policy; at least according to various authors, as reviewed in Santos (2006) and Allen and Gale (2017). In fact, a broader problem with the design of most DIs, LLRs, and even capital regulation, is that of leaving systemic

risk as an exogenous factor when modelling the risk of an individual bank failure. The research on incorporating any single bank's contribution to the overall risk of system failure is still fairly limited, as is the case with liquidity regulation itself. That said, high-level liquidity requirements, such as those in the recently introduced Basel II Capital Accord (see section 2.5), were prudently promulgated in the 2010s, largely in response to the 2008 GFC. These requirements are centred mainly on two measures: the *Liquidity Coverage Ratio* (LCR) and the *Net Stable Funding Ratio* (NSFR). The LCR measures a bank's ability to weather a deep liquidity crisis for at least 30 days, while the NSFR quantifies the degree of maturity mismatches between assets and liabilities. Along with capital regulation, many policymakers believe that minimum liquidity requirements can help stabilise the banking system.

Although holding excess liquidity carries an undeniable opportunity cost, banks may offset this cost by strategically lending excess liquidity to the interbank market during crises. When liquidity becomes scarce, asset prices necessarily become volatile as the market tries to reach an equilibrium, according to Allen and Gale (2017). That said, interbank loans carry a form of credit risk (called *counterparty risk*) since the borrowing bank in distress can fail to repay even the emergency funding in due time. In turn, adverse selection can occur during a liquidity crisis wherein only the riskier banks are actively seeking interbank loans, to which lenders respond by charging higher interest rates. However, it was demonstrated in Heider et al. (2015) that lenders may fearfully continue to hoard liquidity in some extreme cases despite the allure of higher interest rates, primarily due to the asymmetric information amongst banks. Alternatively, the interest rates of interbank loans may become too high even for distressed borrowers, which results in a similar market breakdown. In these cases, a central bank may yet again have to provide emergency<sup>12</sup> liquidity, even though doing so causes moral hazard.

In conclusion, a bank's quest of maintaining sufficient liquidity within a fragile and dynamic system is never-ending and non-trivial. At the one end, depositors must be assured lest the collapse of their trust triggers a bank run and a liquidity crisis. At the other end, creditworthiness must be maintained if banks are to access the interbank market for debt-based funding. Add to this mix the nature of banks as profit-seeking firms within dynamic markets, then the inherent complexity of banking quickly manifests. Moreover, interventionist schemes such as DIIs and LLRs are well-intentioned in safeguarding banking fragility, though can introduce secondary challenges in addition to being costly. In some cases, these setups may even exert the opposite effect than intended, as demonstrated in Anginer et al. (2014). On the other hand, regulatory requirements such as minimum capital levels can rebuffer unexpected losses and protect bank liquidity during crises, though again at a cost. The fairly recent idea of regulating liquidity itself

---

<sup>12</sup>This was indeed the case during the 2008 crisis; both the US Federal Reserve and the European Central Bank injected liquidity into their respective markets, albeit in different forms - see Vento and La Ganga (2009) and Heider et al. (2015).

is quite rational, though even less well-understood when compared to capital regulation. As phrased in Allen and Gale (2017), it is unclear even what exactly to argue about when it comes to liquidity regulation.

## **2.5 Maintaining capital: the Basel Capital Accords**

Even though banks are commercial organisations, they function differently from other firms and uniquely affect the economy at large. From Santos (2006), Dermine (2007, pp. 530–532), Van Gestel and Baesens (2009, pp. 53–55), and Allen and Gale (2017), the classical argument for governmental intervention is largely premised on the risk of a banking system failure. Public fear or reputational damage may trigger bank runs, possibly prompting a liquidity crisis that may in turn propagate across the banking system. Moreover, general market dysfunction can manifest due to asymmetric information between bank and depositor, especially regarding a bank's risk-taking levels. Depositors may very well request higher interest rates if they knew the risks underlying the use of their funds. That said, a bank's risk-taking may be naturally counteracted by the possibility of reputational damage, as quantified in Bhattacharya and Thakor (1993). The untarnished reputation of a bank partly assures the safety of the deposited funds, which is self-evident when reviewing banking history in section 2.1. However, reputation can only deter excessive risk-taking when the associated publicity of choosing lower-risk projects becomes rewarding in its own right, which may be unsustainable in a competitive market. Ultimately, there is clearly a case to be made for at least some type of public oversight to maintain confidence in the banking system.

Many countries have drafted various types of governmental interventions to protect retail deposits and bank liquidity, including DI and LLR schemes (as reviewed in Santos (2006) and section 2.4). Of these interventions, perhaps the most prominent type is that of capital regulations imposed on all national banks by some government agency. While bank capital can serve as a source of funds (in the form of equity) to enable lending, it is the risk-bearing function thereof that is more important for regulators. According to Taggart and Greenbaum (1978), bank capital can absorb any deterioration in loan asset quality, thereby stabilising the bank's overall asset values that may otherwise be in flux. By doing so, the probability of insolvency is directly affected, thereby protecting depositors and shoring up the public's confidence in the bank, which in turn limits the scope for a bank run. Moreover, the shareholders of a bank enjoy risk reduction since capital will ultimately reduce potential losses in the event of insolvency. Lastly, a common belief amongst bankers is that holding capital can deter excessive risk-taking when lending.

However, instating national mechanisms to protect banks from liquidity shocks will unavoidably interfere with the free market and disturb its equilibrium, possibly inhibiting economic growth objectives. In particular, regulated reserve requirements will impose a 'tax' of sorts when

trying to raise funding from depositors, thereby inhibiting financial intermediation, which was demonstrated in Taggart and Greenbaum (1978). Moreover, many studies have since shown that capital requirements are not as effective in controlling risky lending or curbing liquidity crises as one would otherwise believe – see Bhattacharya and Thakor (1993) and Santos (2006). In fact, the general equilibrium models of Besanko and Thakor (1992) showed that greater capital requirements can reduce a bank’s reliance on deposit funding. In turn, this reduction then decreases the interest rates offered by banks for deposits due to the weaker demand. Simultaneously, the cheaper cost of funding means equilibrium loan rates decrease as well, which leads to declining revenue at first, though is counteracted by greater loan sizes. This implies that higher capital requirements benefit borrowers but hurt depositors and shareholders, even if such requirements can induce safer lending.

Apart from imposing capital and liquidity requirements, a central bank (or a relevant public agency) commonly regulates three other aspects of banking: market participation (by issuing/revoking a banking licence), the money supply (by setting the prime lending rate), and information availability (by requiring the public disclosure thereof in annual reports). However, the need for cooperation amongst central banks became increasingly clear as cross-border credit flows surged and international lending grew, as discussed in Van Gestel and Baesens (2009, pp. 55–57) and Baesens et al. (2016, pp. 5–6). This need was further underscored by subsequent liquidity crises and large international banking<sup>13</sup> failures, particularly those in 1974 of Bankhaus Herstatt in Germany and Franklin National Bank in the USA. Soon thereafter, the *Basel Committee on Banking Supervision* (BCBS) was established in 1975 by a board of central bank governors of the G10 countries. The BCBS is head-quartered at the *Bank for International Settlements* (BIS) in Basel, Switzerland, which was itself previously established in 1930 as *the* bank for all central banks. The BIS remains a natural host for the BCBS to this day, given its goal of sustaining monetary and financial stability and cooperation across the globe.

Initially, the BCBS only facilitated cooperation amongst central banks, though soon expanded its scope to providing minimum supervisory standards to be enforced by central banks. Chief amongst these standards is the issue of banks’ *capital adequacy*, which must enable them to weather *unexpected* losses (UL) in excess of *expected* losses (EL). Capital adequacy is primarily measured as the ratio between capital held and the risk-weighted loan asset balances. In fact, the 1980s saw widespread decreases in the capital ratios of many large multinational banks, largely due to increased lending in riskier emerging markets, as discussed in Van Gestel and Baesens (2009, pp. 55–57, 344–345), Thomas (2009a, pp. 289), and in Baesens et al. (2016, pp. 6). Subsequently, many central banks agreed that, as a basic principle of lending, the amount of capital ought to be reserved based on the risk profile of a bank’s loan assets. This consensus culminated in a regulatory framework that was published in 1988, better known as the Basel

---

<sup>13</sup>See Van Gestel and Baesens (2009, pp. 84–92) for a list of notable crises during the twentieth century.

Capital Accord (or simply Basel I). This framework prescribed the minimum regulatory capital to be at least 8% of *risk-weighted assets* (RWA), which is believed to have been the average capital ratio at most banks at the time. Basel I also introduced fixed risk weights<sup>14</sup> based on the asset class itself: 0% for cash exposures, 50% for mortgages, and 100% for other commercial exposures. Incorporating the appropriate risk weight for a single exposure (or loan)  $i$ , the minimum capital is then simply expressed as

$$\begin{aligned}\text{Capital}_i &= 8\% \times (\text{Risk-weighted asset}) \\ &= 8\% \times (\text{Risk weight} \times \text{Exposure}).\end{aligned}\tag{2.4}$$

Eq. 2.4 implies that riskier positions need greater capital to offset catastrophic default risk, according to Van Gestel and Baesens (2009, pp. 344–345). On the portfolio-level, the total regulatory capital is then simply the sum of these loan-level capital calculations. Clearly, total capital will not necessarily be equally distributed amongst all exposures, simply by virtue of differing risk weights. This discrepancy is especially relevant for lenders that have different classes of loan exposures (e.g., mortgages and corporate loans), in contrast to smaller mono-line banks offering but one loan product. Finally, Basel I requires that the total *Capital Adequacy Ratio* (CAR) be equal or greater than 8%, expressed as

$$\frac{\sum \text{Capital}_i}{\sum (\text{RWA})_i} \geq 8\% .\tag{2.5}$$

Basel I differentiated only broadly by asset class without considering the underlying credit risk of each borrower, which typically varies across the portfolio. Although the first Accord was later amended to cater for bilateral netting of derivative products (in 1995) and to cover market risk (in 1996), the BCBS eventually decided to revise the framework entirely in 1999. Published in 2006 following extensive industry consultations, the reworked Basel II Capital Accord<sup>15</sup> mostly refined the measurement of credit risk as well as made the modelling thereof more rigorous. In this regard, credit risk is now quantified using a statistical approach based on the individual loan's *expected loss*. This stochastic quantity depends on three risk parameters: the borrower's default risk, the loss rate given a default event, and the associated exposure size, according to Van Gestel and Baesens (2009, pp. 25–29, 274–277) and Baesens et al. (2016, pp. 10–11). At the loan-level, let  $D$  denote a Bernoulli-distributed random variable such that  $D = 1$  indicates a default event and  $D = 0$  signifies the complement thereof, both expressed across some outcome period (typically twelve months). The expectation thereof,  $\mathbb{E}[D]$ , equals the so-called *Probability of Default* (PD) since  $\mathbb{E}[D] = \mathbb{P}[D = 1] \cdot 1 + \mathbb{P}[D = 0] \cdot 0 = \mathbb{P}[D = 1]$ . Assuming default, the associated

---

<sup>14</sup>A more detailed table of specific risk weights is given in Van Gestel and Baesens (2009, pp. 346).

<sup>15</sup>The revised framework is formally called the "*International Convergence of Capital Measurement and Capital Standards, A Revised Framework, Comprehensive Version*", though it is more generally known as "Basel II" – see Basel Committee on Banking Supervision (2006a).



stochastic loss is expressed as the product of the estimable loss rate  $l$  called the *Loss Given Default* (LGD); and the related at-risk exposure  $e$  called the *Exposure At Default* (EAD). More technically, the EAD represents the average proportion of the loan balance or credit limit that is at risk of loss at the time of default. By further assuming independence, the various loan-level quantities  $\{D_i, l_i, \epsilon_i\}$  are assembled into the stochastic loss  $L_i$  of loan  $i$  as

$$L_i = D_i \cdot l_i \cdot \epsilon_i \implies L_i = D_i \cdot \text{LGD}_i \cdot \text{EAD}_i. \quad (2.6)$$

Overall, Basel II recognised the mitigatory effects of certain risk management practises, e.g., credit derivatives, collateral, and insurance guarantees, which should rightfully reduce regulatory capital. Since these effects typically reflect in the internal credit data of a bank, Basel II promotes greater risk sensitivity (compared to Basel I) by better leveraging this data when estimating the underlying risk parameters (PD, LGD, EAD). In turn, greater sensitivity affords superior differentiation, which can reduce the capital charge and thereby improve bank profitability, as discussed in Van Gestel and Baesens (2009, pp. 347, 392) and Baesens et al. (2016, pp. 9–10). In this regard, Basel II allows two broad levels of sophistication and flexibility when modelling credit risk: the simplest though least flexible Standardised approach and the more flexible *Internal Ratings-Based* approach (IRB). The Standardised approach expands upon the different risk weights first used in Basel I and relies more heavily on estimates from external credit rating agencies (e.g., Moody's, Fitch, and Standard & Poor's). In contrast, the IRB<sup>16</sup> approach allows a bank to use its own models and credit risk estimates, which is generally more accurate. In fact, the work of Jankowitsch et al. (2007) demonstrated that more accurate credit rating systems can reduce regulatory capital as well as yield significant economic value, simply by better pricing loans based on their actual credit risk.

Eq. 2.6 can be estimated across all eventualities and aggregated into the so-called Expected Loss (EL), thereby obtaining the mean value of the individual loan loss probability distribution, expressed as

$$\mathbb{E}[L_i] = \mathbb{E}[D_i] \cdot \mathbb{E}[l_i] \cdot \mathbb{E}[\epsilon_i] \implies EL_i = \text{PD}_i \cdot \overline{\text{LGD}}_i \cdot \overline{\text{EAD}}_i. \quad (2.7)$$

Apart from Basel II capital modelling and loss provisioning, estimating the EL is especially useful in loan pricing contexts. In these cases, a loan's risk premium must theoretically cover  $\mathbb{E}[D] \times \mathbb{E}[l]$ , proportional to the loan amount, as explained in Van Gestel and Baesens (2009, pp. 274–276) and Thomas (2009a, pp. 278–288). While loan-level estimates are certainly useful, the portfolio-level variants are often more practical to use in managing the overall portfolio. These activities may include assessing and tweaking the portfolio's profitability, facilitating its securitisation, and reserving Basel II-compliant capital for the portfolio. Moreover, any inherent diversification

<sup>16</sup>The IRB approach itself has two sub-levels: foundation (IRB-F) and advanced (IRB-A), both differentiated again by the level of complexity. Note that banks are required to use either the Standard or the IRB-A approach for retail exposures, as discussed in Van Gestel and Baesens (2009, pp. 392) and Baesens et al. (2016, pp. 10–11).

benefit wherein one loan's loss is offset by another loan's profit becomes more tractable (or even visible) at the portfolio-level than at the loan-level.

As such, the portfolio loss distribution can simply be obtained across  $N$  loans by summing together the individual loss quantities from Eq. 2.6, i.e.,

$$L_P = \sum_{i=1}^N L_i = \sum_{i=1}^N (D_i \cdot l_i \cdot \epsilon_i) \implies L_P = \sum_{i=1}^N (D_i \cdot \text{LGD}_i \cdot \text{EAD}_i). \quad (2.8)$$

Similar to Eq. 2.7, the portfolio's expected loss estimate is expressed as the sum of the various expected losses at the individual loan-level, given by

$$\mathbb{E}[L_P] = \sum_{i=1}^N \mathbb{E}[L_i] = \sum_{i=1}^N (\mathbb{E}[D_i] \cdot \mathbb{E}[l_i] \cdot \mathbb{E}[\epsilon_i]) \implies \mathbb{E}L_P = \sum_{i=1}^N (\text{PD}_i \cdot \overline{\text{LGD}}_i \cdot \overline{\text{EAD}}_i). \quad (2.9)$$

While the portfolio loss distribution  $L_P$  contains all credit risk information, its shape can be very different to that of the individual loss distribution  $L_i$ . This is largely as a result of convolution, i.e., the distribution of the sum of independent random variables ( $L_1 + L_2 + \dots$ ) corresponds to the distributions of the summands ( $L_1, L_2, \dots$ ), demonstrated in Van Gestel and Baesens (2009, pp. 276–277). Consider lending ZAR 1,000 in total across  $N$  accounts with a PD of 5% while assuming that both LGD and EAD equal 100% for expositional simplicity, thereby yielding  $\mathbb{E}[L_P] = \text{ZAR } 50$ . When lending to  $N = 1,000$  accounts, then the actual portfolio loss will likely be close to ZAR 50 mainly due to the central limit theorem. That said, if  $N = 1$ , then the actual loss will either be ZAR 0 or ZAR 1,000, despite the fact that the mean loss will remain ZAR 50 for both  $N = 1,000$  and  $N = 1$ . While this example from Thomas (2009a, pp. 278) clearly demonstrates the importance of larger more granular portfolios, it also alludes to the potential disconnect between portfolio-level and loan-level distributions due to distributional convolution.

In addition to distributional convolution, the shape of the portfolio loss distribution  $L_P$  can be affected by the extent of default correlation and loan concentration found within a portfolio. From Van Gestel and Baesens (2009, pp. 285–287), the benefit of diversified lending erodes away as correlated default events become more widespread, which would roughly translate into holding more capital depending on the portfolio size  $N$ . In this regard, smaller portfolios have a more linear relationship between UL and correlation strength while larger portfolios typically exhibit a slower ramp-up effect in UL as default correlation increases. Similarly, capital increases exponentially as lending becomes more concentrated, which is typically measured by the so-called *Herfindahl-Hirschmann Index* (HHI) as demonstrated in Van Gestel and Baesens (2009, pp. 287–291). Moreover, the risk parameters may themselves be stochastically correlated, thereby expressing a joint behaviour, e.g., both PD and LGD increase linearly during economic downturn periods. As such, Basel II introduced a fourth risk parameter  $R$  that represents the degree of correlated defaults. According to Van Gestel and Baesens (2009, pp. 317–319, 377) and

Finlay (2010, pp. 184), Basel II suggests the following  $R$ -values<sup>17</sup> for retail exposures: 4% for qualifying credit lines, 15% for mortgages, and a PD-based exponentially weighted interpolation for other retail exposures that ranges between 3% and 16%.

The prevalence of distributional convolution and default correlation means that modelling the portfolio loss from the lower loan-level quickly becomes complex. Furthermore, the various interdependencies amongst risk parameters imply that the eventual level of capital will itself depend on the portfolio's composition. While this relationship seems intuitive, any inherent volatility in the risk estimates will undoubtedly diffuse throughout the capital base in that capital estimates vary wildly over time, as argued in Thomas (2009a, pp. 293–295) and Van Gestel and Baesens (2009, pp. 312). In turn, volatile capital levels suggest that the credit decision will become overly volatile at the loan-level since the same loan may be granted one day but declined the following day. This volatility is impractical from an operational perspective and certainly not conducive to growing (or even maintaining) market share due to the inevitable backlash from customers. As a solution, the principle of *portfolio invariance* was devised in Gordy (2003) wherein the capital per loan must only depend on the particular loan's own risk profile. This (necessarily) restrictive condition requires two assumptions. Firstly, the portfolio must be as finely-grained as possible, composed of a large number of loans such that no single loan dominates the portfolio in terms of loan size. Secondly, there must be a single systemic risk factor that affects all loans in the portfolio. Both assumptions taken together imply that any idiosyncratic risk factors amongst loans tend to negate one another, leaving only the systemic risk factor as a source of uncertainty at the portfolio level.

To achieve portfolio invariance, Basel II therefore requires that the portfolio be subdivided into various risk grades or segments, at least for capital modelling purposes. Each risk grade should contain loans that are largely homogeneous in their overall risk profile. According to paragraphs 404–409 in Basel II and Van Gestel and Baesens (2009, pp. 160, 258–259, 399), a minimum of seven risk grades are required for wholesale PD-models, whilst no minimum is explicitly specified for retail PD-models or any LGD/EAD-models – though segmentation should certainly still be pursued where possible. The principle is to devise a meaningful segmentation scheme  $S$  such that the resulting groups  $s \in S$  are adequately variegated across the portfolio whilst constraining risk variance within each risk class, i.e., between-class differentiation and in-class uniformity. In turn, the  $PD_i$ -estimates of the individual loans within risk grade  $s$  can be replaced with a single value  $D_s$  that represents the segment-level default risk  $PD_s$ , i.e.,  $\mathbb{E}[D_i] := D_s$  for all  $i \in s$ . Having conditioned the PD-estimate per segment, it is not unreasonable to assume that all loss risks  $LGD_i$  are now independent of one another. As such, the portfolio

---

<sup>17</sup>Note that these correlation values were reverse-engineered from historical loss data of G10 supervisory databases, which relate to corporate lending instead of retail lending. After all, individuals are not stock-listed nor can their personal assets and liabilities be as easily assessed as in the case of corporates. For more detail, refer to Van Gestel and Baesens (2009, pp. 318–319).

$EL_P$  from Eq. 2.9 simplifies considerably into

$$\begin{aligned} \mathbb{E}[L_P] &= \sum_{i=1}^N (\mathbb{E}[D_i] \cdot \mathbb{E}[L_i] \cdot \mathbb{E}[\epsilon_i]) \iff \sum_{s \in S} \left( \sum_{i \in S} \mathbb{E}[D_s] \cdot \mathbb{E}[L_i] \cdot \mathbb{E}[\epsilon_i] \right) \\ &= \sum_{s \in S} D_s \left( \sum_{i \in S} \mathbb{E}[L_i] \cdot \mathbb{E}[\epsilon_i] \right) \implies EL_P = \sum_{s \in S} PD_s \left( \sum_{i \in S} \overline{LGD}_i \cdot \overline{EAD}_i \right). \end{aligned} \quad (2.10)$$

The portfolio EL-measure, defined in either Eq. 2.9 or Eq. 2.10, summarises the portfolio loss distribution into a single quantity. Though it cannot provide information on the distributional shape, the EL-measure exhibits all four properties of a so-called *coherent* risk measure  $\rho$ , as originally formulated in Artzner et al. (1999) and discussed in Van Gestel and Baesens (2009, pp. 278–279). Consider two bounded random variables  $X$  and  $Y$  that represent losses on two individual loans. Firstly, **subadditivity** holds when the risk of the sum may be less than the sum of the risks, i.e.,  $\rho(X + Y) \leq \rho(X) + \rho(Y)$ , which implies a diversification benefit when adding together individual risks. Secondly, **monotonicity** requires that riskier positions be measured as such, i.e., if  $X \leq Y$ , then  $\rho(X) \leq \rho(Y)$ . Thirdly, **positive homogeneity** applies when both a risk measurement and its input scale linearly, i.e.,  $\rho(\lambda X) = \lambda \rho(X)$ . Lastly, **translation invariance** means that a risk measurement decreases when adding a risk-free investment  $\alpha$  to a portfolio, i.e.,  $\rho(X + \alpha \cdot r_f) = \rho(X) - \alpha$  with  $\alpha \in \mathbb{R}$  and  $r_f$  denoting a risk-free discount factor. The converse is true as well, i.e.,  $\rho(X - \alpha \cdot r_f) = \rho(X) + \alpha$ , according to Artzner et al. (1999).

Another (perhaps more important) risk measure is the so-called Value-At-Risk function  $\text{VaR}_\alpha$ . This measure yields the lowest extreme percentile  $L$  at the far-right tail of the  $L_P$  distribution such that the probability of attaining even greater losses than  $L$ , i.e.,  $\mathbb{P}(L_P > L)$ , is at most  $1 - \alpha$  over a given time horizon. More formally, the  $\text{VaR}_\alpha$ -estimate of  $L_P$  is defined as

$$\text{VaR}_\alpha(L_P) = \min \left( L \in \mathbb{R} : \mathbb{P}(L_P > L) \leq 1 - \alpha \right). \quad (2.11)$$

The chosen confidence level  $\alpha \in [0, 1]$  is usually very high, e.g.,  $\alpha = 99.9\%$  that corresponds to a 1-in-thousand year failure event. At  $\alpha$ , one can be sure not to lose more than  $\text{VaR}_\alpha(L_P)$  in  $(1 - \alpha)\%$  of times. A thorough discourse hereof is given in Artzner et al. (1999), Van Gestel and Baesens (2009, pp. 282–285), Thomas (2009a, pp. 288–293), Finlay (2010, pp. 185–187), and Baesens et al. (2016, §9). The portfolio VaR was adopted in Basel II to help set regulatory capital levels and is typically expressed over one year for covering unexpected credit risk. In fact, the related idea of *economic capital* leverages the portfolio VaR to calculate the necessary capital for supporting all portfolio risks in excess of loss provisions, as illustrated in Fig. 2.11. Denoted as  $\text{EC}_\alpha$ , the economic capital for covering the portfolio's loss distribution  $L_P$  is simply defined as

$$\text{EC}_\alpha(L_P) = \text{VaR}_\alpha(L_P) - EL_P. \quad (2.12)$$

While the VaR-measure is quite popular in industry, it is not a coherent risk measure since it does not possess the subadditivity-property, as demonstrated in Artzner et al. (1999). This implies

that the combination of portfolios will yield a greater VaR than that of each individual portfolio simply added together, therefore discarding diversification. A variant of the VaR-measure that is fully coherent is the so-called *Expected Shortfall* function (or the conditional VaR), as explained in Van Gestel and Baesens (2009, pp. 285).

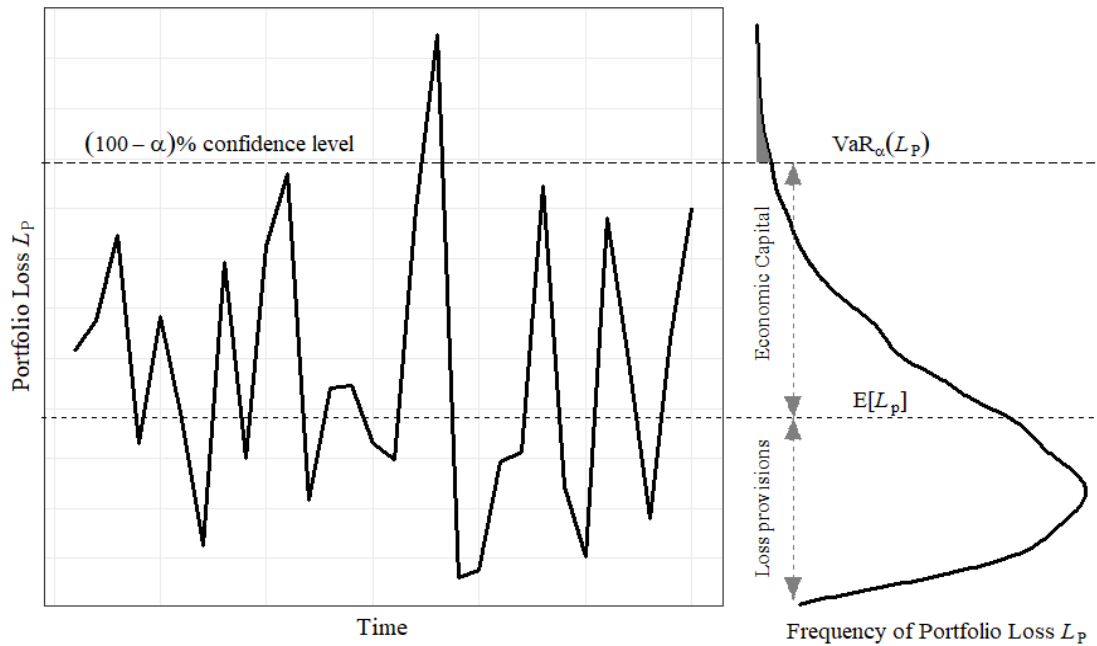


FIG. 2.11: A Value-at-Risk approach for estimating the portfolio's unexpected loss, denoted as  $\text{VaR}_\alpha(L_P)$ , at a  $(1-\alpha)\%$  confidence level. Hypothetical portfolio loss rates are shown over time, from which the portfolio loss distribution  $L_P$  is calculated. While loss provisions ought to cover  $E[L_P]$ , the difference  $\text{VaR}_\alpha(L_P) - E[L_P]$  constitutes economic capital.

Regulatory capital is meant to cover extreme losses and, as such, one cannot simply use averages of the PD-parameter, even when segmented. The seminal work of Vasicek (2002) devised a mechanism by which  $\text{PD}_s$  may be stressed to an appropriate level for capital reservation. From Thomas (2009a, pp. 296–298), Van Gestel and Baesens (2009, pp. 294–296), Witzany et al. (2013), and Baesens et al. (2016, pp. 240–242), the starting point of Vasicek's structural default model, also known as the *asymptotic single risk factor* (ASRF) model, is to restate the loan-level default indicator  $D_i$  following Merton's original model. Specifically, assume that default occurs whenever a borrower's underlying assets  $Y_i$  (or overall credit quality) fall below a certain value  $c_i$ , i.e.,

$$D_i = \begin{cases} 1 & \text{if } Y_i \leq c_i \\ 0 & \text{otherwise} \end{cases} . \quad (2.13)$$

Furthermore, Vasicek assumes that  $Y_i$  is standard normally distributed, which is to say that the

unconditional probability of default for loan  $i$  becomes  $\mathbb{P}(D_i = 1) = \mathbb{P}(Y_i \leq c_i) = \Phi(c_i)$  where  $\Phi$  is the cumulative standard normal distribution function. If equated to an average default rate  $\bar{d}$  as  $\Phi(c_i) = \bar{d}$ , then  $c_i = F^{-1}(\bar{d}) = \Phi^{-1}(\bar{d})$  where  $\Phi^{-1}$  represents the inverse cumulative standard normal distribution function (or quantile function).

Vasicek further incorporates in-segment default correlation by structurally decomposing each  $Y_i$  into two independent components: 1) a portfolio-level *systemic* risk factor  $V$  such as a macroeconomic index; and 2) an *idiosyncratic* loan-specific risk factor  $Z_i$ . Assume that the random variables  $Y_i$  are jointly standard normally distributed with equal pairwise correlations, i.e.,  $\text{corr}(Y_i, Y_j) = \rho$  for  $i \neq j$ . Accordingly,  $V$  and  $Z_i$  are also mutually independent standard normally distributed variables and related to  $Y_i$  by a Gaussian copula as

$$Y_i = \sqrt{\rho} \cdot V + \sqrt{1-\rho} \cdot Z_i. \quad (2.14)$$

The decomposition in Eq. 2.14 supposes that default correlation between the market factor  $V$  and the credit quality of each loan  $i$  is constant and equal to  $\sqrt{\rho}$ . By substituting Eq. 2.14 into the default assumption of  $Y_i \leq c_i$ , a necessary condition is obtained that must hold for loan  $i$  (as measured via  $Z_i$ ) in order for default to occur, expressed as

$$Z_i \leq \frac{c_i - \sqrt{\rho} \cdot V}{\sqrt{1-\rho}}. \quad (2.15)$$

By assumption, the random variables  $Z_i$  are independent from the market factor  $V$ , even though  $Y_i$  are actually correlated amongst themselves. For simplicity, Vasicek conditions Eq. 2.15 to a certain value of the market shock  $V$ , thereby relieving the degree of correlation amongst  $Y_i$ . For a realisation  $v$  from  $V$  and estimating  $c_i$  with  $\Phi^{-1}(\bar{d})$  under portfolio invariance, the conditional probability of default  $\theta(v)$  becomes

$$\mathbb{P}\left(Z_i \leq \frac{\Phi^{-1}(\bar{d}) - \sqrt{\rho} \cdot V}{\sqrt{1-\rho}} \middle| V = v\right) = \Phi\left(\frac{\Phi^{-1}(\bar{d}) - \sqrt{\rho} \cdot v}{\sqrt{1-\rho}}\right) := \theta(v). \quad (2.16)$$

Given that  $Z_i \sim \mathcal{N}(0, 1)$ , another quantity that becomes useful later is the inverse of Eq. 2.16, i.e.,

$$\frac{\Phi^{-1}(\bar{d}) - \sqrt{\rho} \cdot V}{\sqrt{1-\rho}} = \Phi^{-1}(\theta(v)). \quad (2.17)$$

Lastly, the market shock  $v$  can be substituted with an extreme percentile from the standard normal distribution since  $V \sim \mathcal{N}(0, 1)$ . This is to say that Eq. 2.17 is evaluated at an extreme probability level  $1 - \alpha$ , e.g.,  $\alpha = 0.999$ , which is simply given by  $v = \Phi^{-1}(1 - \alpha)$ , as derived in Baesens et al. (2016, pp. 240–241). By symmetry, this quantity is the same as  $v = -\Phi^{-1}(\alpha)$ , which is substituted into Eq. 2.16, thereby yielding the classical Vasicek ASRF model, expressed as

$$\Phi\left(\frac{\Phi^{-1}(\bar{d}) - \sqrt{\rho} \cdot (-\Phi^{-1}(\alpha))}{\sqrt{1-\rho}}\right) = \Phi\left(\frac{\Phi^{-1}(\bar{d}) + \sqrt{\rho} \cdot \Phi^{-1}(\alpha)}{\sqrt{1-\rho}}\right). \quad (2.18)$$

For Basel II capital reservation under an IRB-A approach, a given  $PD_s$ -value can be stressed by incorporating Eq. 2.18 into a function  $K$ . The default correlation  $\rho$  is substituted with Basel's  $R$ -values at  $\alpha = 0.999$  and  $K$  is then defined as

$$K(PD_s) = \Phi \left[ \frac{\Phi^{-1}(PD_s) + \sqrt{R} \cdot \Phi^{-1}(0.999)}{\sqrt{1-R}} \right]. \quad (2.19)$$

For each loan  $i$ , the Basel II capital level  $C_i \in [0, 1]$  is then simply defined as the difference between the stressed expected loss  $K(PD_s) \cdot \overline{LGD}_i$  and the presumably provision-covered expected loss  $PD_s \cdot \overline{LGD}_i$ , expressed as

$$\begin{aligned} C_i &= K(PD_s) \cdot \overline{LGD}_i - (PD_s \cdot \overline{LGD}_i) \\ &= \Phi \left[ \frac{\Phi^{-1}(PD_s) + \sqrt{R} \cdot \Phi^{-1}(0.999)}{\sqrt{1-R}} \right] \cdot \overline{LGD}_i - (PD_s \cdot \overline{LGD}_i) \end{aligned} \quad (2.20)$$

For convenience, the minimum CAR of 8% from Eq. 2.5 is algebraically reshuffled as  $12.5 \cdot 8\% = 1$ . Accordingly, the loan-level risk-weighted assets  $RWA_i$ <sup>18</sup> is computed using  $C_i$  as

$$RWA_i = 12.5 \cdot \overline{EAD}_i \cdot C_i. \quad (2.21)$$

In modelling each component, the first risk parameter  $D$  – default risk (PD) – may depend on a number of factors, including borrower-centric, portfolio-based, and macroeconomic-related input variables, as explained in Van Gestel and Baensens (2009, pp. 24–25), Thomas (2009a, pp. 282–283), and Baensens et al. (2016, §5–6). Its estimation is principally the same exercise as behavioural credit scoring, which was previously discussed in section 2.2. However, there is greater scope to focus on prediction accuracy when developing PD-models than application/behavioural scorecards, simply due to the former's relatively more direct impact on loss provisioning (see section 2.6) and capital reservation. In particular, greater accuracy can more easily translate into cost efficiency than risk-ranking ability alone. The latter has typically been the focal point of most application/behavioural scoring models instead of pursuing accuracy above all else. Furthermore, PD-models are generally developed at the product-level in retail banking, loosely motivated by the product's securability (e.g., mortgages vs. personal loans) and/or its broader design. To this point, the loan account (held by a borrower) usually forms the base granularity for risk models in retail banking. However, in wholesale banking, the borrower (e.g., corporate or other large counterpart) becomes the base-level for risk models instead of the account.

Modelling default risk at the product-level (or asset class) will often reveal significant differences in PD-estimates across different products, especially when the same borrower (or counterpart) holds multiple products at the same bank. Many practitioners can attest to the lower default rates of more secure products when compared to their less secure counterparts,

<sup>18</sup>Basel II applies an additional scaling factor of 1.06 on the RWA in the interest of prudence.

which is curious at first glance. In principle, a counterpart that defaults on one loan may likely default on other concurrent loans, thereby causing a so-called ‘contagion’ effect in product-level default rates. However, when the chain of defaults seizes at certain products – specifically more secure portfolios – then it suggests some other dynamic in play. Furthermore, a product’s securability certainly affects its overall LGD (and therefore the portfolio EL), but one wouldn’t expect securability itself to influence the PD. It turns out that retail borrowers in financial distress may choose to default rather selectively by product type, as discussed in Van Gestel and Baesens (2009, pp. 25). More ‘critical’ debts are paid first, usually those debts directly related to financial and job security, e.g., mortgages and auto loans. At the very least, this phenomenon suggests an ‘order’ in which default propagates across products amongst highly credit-leveraged borrowers.

The second risk parameter  $l$  – loss risk (LGD) – is commonly defined as the proportion of a defaulted exposure to be written-off. This quantity varies widely based on the type of default resolution (e.g., write-off vs. cure) and underlying loan collateral (if any). From Van Gestel and Baesens (2009, pp. 26–28, 217–222) and Baesens et al. (2016, §10), more secure products typically have very low LGD-values since seized collateral can offset credit losses. Furthermore, the LGD can even become negative in secured lending since the collateral may have been auctioned at a higher value than that of the outstanding debt, or due to recouped penalty fees and interest. Conversely, loss rates can exceed 100% of the debt due to extra litigation and administrative costs but failed loan recovery. To this point, Basel II (paragraph 460) requires that all material direct and indirect costs be considered when estimating loss, including the time value of money.

Furthermore, calculating the LGD of a defaulted loan inherently depends on the account’s default resolution during the ‘workout’ period; an uncertain and often lengthy process itself. In fact, optionality greatly affects LGD in that a distressed borrower may recover financially and subsequently repay all arrears, thereby ‘curing’ the default event. Aside from liquidation/recovery and curing, a third type of default resolution is that of restructuring. A bank can strategically avoid costly (and uncertain) liquidation proceedings and instead maintain the credit relationship with the borrower, though at the cost of reduced income and a medium loss. This resolution is achieved by reorganising a debt in such a way that instalments become affordable for the distressed borrower, thereby ‘curing’. The remaining unresolved (or right-censored in that write-off/cured outcomes are still pending) defaults are typically excluded when calculating the realised/actual LGD, even though these cases are often the recent majority. These ideas are illustrated in Fig. 2.12.

The third risk parameter  $\epsilon$  – exposure risk (EAD) – is often closely related to the LGD definition and its calculation, such that  $\text{LGD} \times \text{EAD}$  denotes the economic loss upon defaulting. There are two broad approaches to defining EAD, as discussed in Van Gestel and Baesens (2009, pp. 28–29, 226–229) and Baesens et al. (2016, §11). Firstly, the EAD may be defined as time-



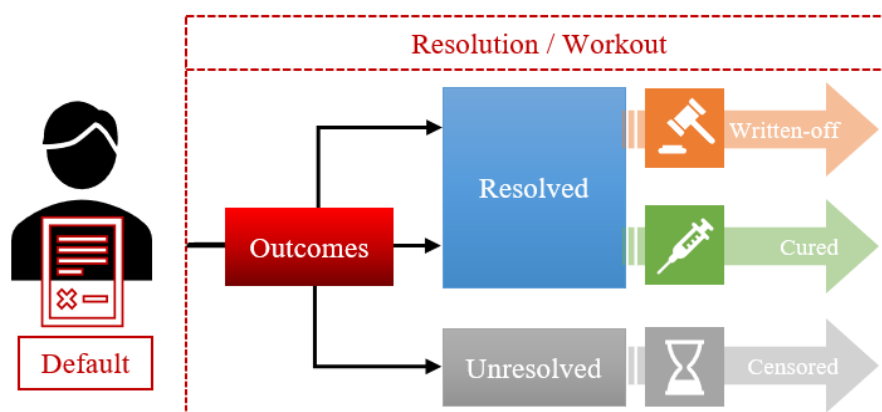


FIG. 2.12: Illustrating the typical process of resolving defaulted exposures during the workout period. Defaults are eventually resolved into either write-offs or cures, while the rest are still pending an outcome.

invariant in that it literally represents the debt balance at the default point  $\tau$ . In this case, the EAD is calculated as the starting balance  $b_0$  (or overall credit limit) relative to the default balance  $b_\tau$ , thereby yielding a simple ratio  $s = b_0/b_\tau$ . These loan-level factors are then averaged across the defaulted population and used quite generically in the EAD-estimate as  $\epsilon = s \cdot b_0$ . However, this rather deterministic approach may be too conservative for more flexible credit products wherein borrowers can draw and repay dynamically, e.g., credit cards and revolving loans. As such, another approach (favoured by Basel II as well) for measuring the EAD  $\epsilon$  is to model the so-called *Credit Conversion Factor* (CCF)  $\eta$  instead of  $\epsilon$  itself. The CCF is characterised by two quantities, the credit limit  $b_0$  and the drawn amount  $b_t$  that is measured at time  $t$ , assuming  $b_0 > b_t$ . In essence, the CCF tries to quantify the proportion of the undrawn limit that will yet convert into credit at the default time  $\tau$ . The EAD is then related to the CCF as  $\epsilon_\tau = b_{\tau-k} + \eta \cdot (b_0 - b_{\tau-k})$  across a chosen outcome period  $k$  with  $\tau - k > 0$ .

Additional Basel II requirements apply when modelling each risk parameter, as summarised in paragraphs 461–479 of Basel II, Van Gestel and Baesens (2009, pp. 259-262), and Baesens et al. (2016, pp. 279). In particular, the sampling period should ideally span a complete business cycle (or at least seven years) for wholesale LGD-modelling, five years for wholesale PD-modelling, and at least five years for all retail modelling. Paragraphs 473 and 479 in Basel II also affords some discretion to weigh more recent observations more heavily than older observations in retail models. Furthermore, the epoch in time from which data is sampled should ideally include a recession or economic downturn period. This requirement simply injects some conservatism into the eventual risk estimates to cover possible modelling deficiencies, as argued in de Jongh et al. (2017). Lastly, extensive modelling guidance for estimating any of these risk parameters is given in Van Gestel and Baesens (2009, pp. 174–201) and Baesens et al. (2016). The various modelling

techniques can broadly be categorised into:

1. **Financial models** that are more theoretical, e.g., Merton's structural model, the cash flow-based gambler's ruin model, and reduced-form models based on Cox processes – see Van Gestel and Baesens (2009, pp. 176–181);
2. **Statistical models** that are largely data-driven, e.g., generalised linear models, and more advanced machine learning techniques. These techniques may include Support Vector Machines (SVMs), Multilayer Perceptron (MLP) neural networks, and Bayesian Belief Networks (BBNs) – see Hastie et al. (2009);
3. **Expert models** that are committee-based or in the form of heuristic rule systems. Although similar to statistical models in structure, an expert model relies on *a priori* opinions and expert judgement in parametrising the model's components instead of data, which may be scarce – see Van Gestel and Baesens (2009, pp. 191–194).

Both Basel I and II require banks to reserve capital into at least 50% of highly liquid and quality loss-absorbing "core" Tier 1 capital, with the rest held as relatively less liquid Tier 2 supplementary capital, as discussed in Van Gestel and Baesens (2009, pp. 350–353) and Baesens et al. (2016, pp. 8–9). Tier 1 capital ought to be highly reliable and liquid (especially during adverse economic conditions), e.g., common stock, preferred stock, and retained earnings. On the other hand, the less liquid Tier 2 capital can include undisclosed reserves, revaluation reserves, general loan loss provisions, and subordinated term debt; with some allowed deductions from the total, e.g., goodwill and insurance investments. Following the 2008 GFC, the required Tier 1 weight was adjusted to at least 75% for Tier 1 capital (or 6% of RWA considering Eq. 2.5), thereby enhancing the rapid loss-absorption power of banks. This change formed part of the new Basel III Capital Accord (effective from 1 January 2013), which largely introduced additional liquidity requirements without changing the credit risk modelling process itself that underpin Basel II. For the most part, Basel III established an additional capital conservation buffer that consists of the 30-day LCR and the 1-year NSFR (see subsection 2.4.4), which must equal at least 2.5% of RWA. Moreover, Basel II seeks to stabilise capital levels across the macroeconomic cycle by imposing an additional counter-cyclical buffer. This add-on is controlled by the local regulator depending on the level of macroeconomic stress at any time and can range from 0% to 2.5% of RWA. Lastly, Basel III emphasised the need for banks to stress-test their internal risk models to a greater extent.

The Basel Capital Accords are broadly based on three overlapping sets of principles, called Pillars 1–3 with a thorough discourse thereof given in Thomas (2009a, pp. 290), Van Gestel and Baesens (2009, pp. 348–349, 418–427), Finlay (2010, pp. 176–177), and in Baesens et al. (2016,

pp. 7–8). Pillar 1, which encompassed much of the discussion so far, generally prescribes the calculation of minimum capital across various risk types. In this regard, Basel I only focused on credit risk, while Basel II expanded the risk scope to market and operational risks, to be discussed in section 2.6. Pillar 2 outlines the principles of both internal and external supervisory oversight in evaluating and monitoring a bank’s quantitative risk models or processes. Its purpose is to promote continuous improvements in both modelling methods and the overall design of risk processes, thereby ensuring adequate capitalisation. Pillar 2 is underpinned by four principles, the first of which is meant for banks and the remainder for regulators:

1. Banks should have an internal process<sup>19</sup> for assessing the adequacy of overall capitalisation against their risk profile;
2. Regulators should periodically review the aforementioned adequacy-process of banks;
3. Regulators should expect banks to hold capital in excess of the minimum level;
4. Regulators should intervene timeously when capital levels of banks fall beneath the minimum.

Lastly, Pillar 3 advocates the public disclosure of a bank’s modelled risks. This disclosure can include certain details of the capital calculation and certain risk management processes, all of which are aimed at alleviating asymmetrical information between bank and investor. Doing so will likely promote greater confidence in the solvency and risk management practices of a bank. In turn, greater confidence has a reciprocal effect in that it assists a bank in procuring funding at lower costs, which can have a bearing on profitability. Moreover, subjecting all banks in a system to these Capital Accords (I–III) will undoubtedly safeguard overall liquidity and stave off failure, albeit at a cost to depositors and shareholders. Simultaneously, ongoing research will likely expand the level of sophistication in modelling credit risk in both the EL and UL, as promoted by Pillar 2 of the Accords – an exciting prospect.

## 2.6 The management of financial risk in banking

A bank generally faces a myriad of risk types as an unavoidable result of its operation. The ever-present danger of illiquidity necessitates reserving sufficient capital against unexpected losses, as discussed in sections 2.4–2.5. However, some risk types are more probable than others and I shall accordingly examine three broad classes of risk in subsection 2.6.1 that, together with liquidity risk, constitute so-called *financial risk*. Moreover, the greatest factor hereof – credit risk – warrants not only capital to cover unexpected losses, but also necessitates a provisions

---

<sup>19</sup>This particular process is more commonly known as the *internal capital adequacy assessment process* (ICAAP).

account from which expected losses can be offset on a more frequent basis. As such, the discussion cannot be complete without presenting a brief introduction to the IFRS 9 accounting standard that governs loss provisions. Finally, common strategies for managing these identified risks are outlined in subsection 2.6.2.

### 2.6.1 A trifecta of risks: credit, market, and operational

It is widely accepted that credit risk presents the single largest source of bank risk, to such an extent that even a relatively small number of defaulting borrowers have the potential to bankrupt a bank. Credit risk itself consists of a few subcategories of specific risks, as outlined in Dermine (2007, pp. 498–499) and Van Gestel and Baesens (2009, pp. 23–25). All of these credit risk subclasses relate to the breakdown of trust and/or renegeing on the repayment of loans or other commitments. Between two banks, **counterparty risk** is the probability that the borrowing bank fails to pay as obligated by the lending bank, which may include bonds, derivatives, or insurance contracts. Another form of counterparty risk is that of **settlement risk**, which arises during the exchange of foreign currencies (or securities), wherein one party has already delivered while the other party has not. A defining example of settlement risk is the failure of Bankhaus Herstatt on 26 June 1974. According to Van Gestel and Baesens (2009, pp. 85), this German bank had its banking licence withdrawn whilst transacting with US banks, who had already and irrevocably paid their dues to Bankhaus Herstatt preceding its imminent failure. **Country risk** (or sovereign risk) materialises when a government defaults on its financial commitments or freezes foreign currency payments. However, the most common type of counterparty credit risk is so-called **retail and/or wholesale credit risk**, which is the potential loss due to a borrower not honouring a debt obligation to the bank within agreed-upon timelines.

A particularly important facet of credit risk is the low but persistent loss of lent capital due to the materialisation of ‘expected’ credit risk over time. All loan assets carry the inherent risk of becoming impaired and while capital covers catastrophic portfolio impairment, it is not meant to offset low-level loan impairments. To this point, there are at least two fundamental reasons for keeping a so-called *provision* account for offsetting expected losses, as explained in Dermine (2007, pp. 514–515) and Finlay (2010, pp. 167–169). Firstly, and as a central tenet of risk management, a bank effectively manages long-term solvency risk and smooths earnings volatility when providing for future expected credit losses earlier. In fact, a bank may better absorb these expected losses if they are spread out over time, instead of conducting a sudden and systemic write-down of impaired loan assets that may subsequently trigger a liquidity crisis. Secondly, loss expectations pose as an effective counterweight to a bank that may otherwise engage in riskier lending. Early loan performances may prematurely bias overall profitability in the absence of loss expectations, which is particularly incendiary when staff bonuses are linked to these early loan performances. In turn, this premature exuberance may very well lead to increasing the risk

appetite of a lender unwittingly.

Accordingly, the IFRS 9 (2014) accounting standard requires that the value of financial assets be comprehensively adjusted based on a bank's evolving estimates of expected credit risk. The underlying principle is that the bank willingly forfeits a portion of its income at each period into a central loss provision account that ideally offsets any amounts written-off in future on average. The provision value should be updated frequently based on the forecast of a loan's so-called *expected credit loss* (ECL), provided by an underlying statistical model. Similar to Basel's EL-measure, the ECL is generally expressed as the probability-weighted sum of all future cash shortfalls that the bank expects to loss over a certain horizon, according to IFRS 9 (2014, §5.5.8, §5.5.17, §B5.5.25–35). As credit risk evolves, the loss provision is adjusted either by raising more from earnings or releasing a portion thereof back into the income statement, respectively as an impairment loss or gain. Accordingly, loss provisions directly reduce bank profitability and are similar to depreciation since the gross carrying value of an asset is effectively decreased. Ultimately, providing for future bad debt rightfully recognises that loan assets recorded on the balance sheet are actually less than the value at which they are stated.

The ECL may be calculated using a staged approach based on the extent of deterioration in a loan's credit risk, according to IFRS 9 (2014, §5.5.3, §5.5.5). Each subsequent stage requires a more severe ECL estimate, as illustrated in Fig. 2.13. In particular, Stage 1 includes most loan assets, provided that these assets either have low credit risk or have not experienced a so-called *significant increase in credit risk* (SICR) event since origination. As a middle ground, Stage 2 contains those assets that have in fact deteriorated quite significantly in their credit quality, but do not yet qualify as fully credit-impaired. Finally, Stage 3 includes assets with objective evidence of credit impairment that may adversely affect their future cash flows, e.g., defaulted accounts. These impairment stages aim to reflect a broader pattern of deterioration (or improvement) in credit quality over time, thereby allowing for recognising credit losses more timeously when necessary, as discussed in IFRS 9 (2014, §B5.5.2) and Cohen, Edwards Jr et al. (2017). Another consideration is that of the time horizon underlying the ECL-estimate: 12 months for Stage 1 and lifetime for Stages 2–3. The latter horizon considers all possible default events over the remainder of the asset's expected life. The 12-month horizon then refers to the portion of the lifetime ECL that may occur over the next 12 months, specifically due to a single default event. Lastly, the manner in which a bank should calculate interest revenue differs by impairment stage as well: the gross carrying amount (or balance) is used when in Stages 1–2, but switches to the amortised cost (balance less the loss allowance) when in Stage 3.

Stage migrations are largely based on a loan accruing arrears or entering debt restructuring under distress, as summarised in Fig. 2.13. However, the SICR-component is arguably more important as a stage impairment classifier, presumably based on a "*multi-factor and holistic*

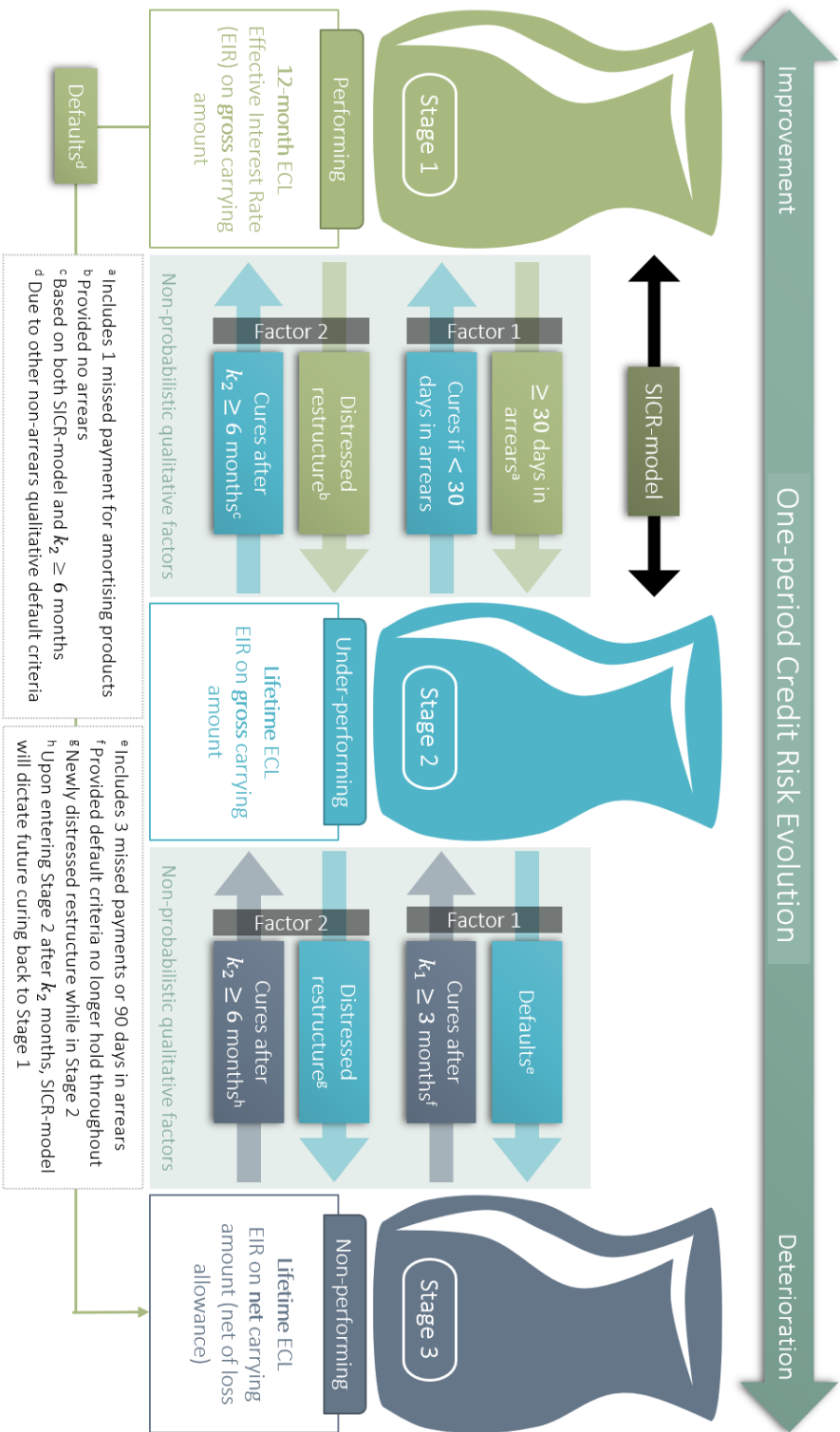


FIG. 2.13: Illustrating the one-period evolution of credit risk according to the IFRS 9 staged impairment framework. Each subsequent impairment stage implies a greater ECL estimate to reflect the deeper level of credit deterioration. Arrows indicate possible stage migrations, subject to meeting certain qualitative criteria after a certain probation period has lapsed. The exception is the probabilistic SICR-component that may include other more external factors.

*analysis*" using "*reasonable and supportable information*", as set out in IFRS 9 (2014, §5.5.9, §B5.5.16). At a minimum, flagging an account as an increased risk requires estimating and comparing the account's lifetime default risk at two points in time. Consider the probabilistic model  $p_D(x, t)$  that estimates the default probability  $\mathbb{P}(D|X, T)$ , where  $D$  is a default indicator random variable that is measured at a given time  $T = t$ , whilst observing all available information  $X = x$  at the same  $t$ . Per requirement, the account's default risk is then estimated twice: once at origination  $t = 0$  and again at the desired reporting time  $t > 0$  (or subsequent monthly period). It is self-evident that risk has increased whenever  $p_D(x, t') > p_D(x, 0)$  for some  $t' > 0$ . Having calculated the magnitude of this increased risk, i.e.,  $p_D(x, t') - p_D(x, 0)$ , the bank has to decide on a threshold<sup>20</sup> of sorts beyond which an account should transit into Stage 2 and below which an account remains in Stage 1. Lastly, IFRS 9 (2014, §5.5.11, §B5.5.17) provides a broad range of factors to assist with the SICR-decision, including so-called "*forward-looking information*", i.e., macroeconomic information (historic and forecast). This concludes the brief introduction and discussion of IFRS 9.

The second risk in the trifecta is that of **market risk**, which spans various subclasses of risk (similar to credit risk) that are introduced by bank participation in the market. This risk type is more relevant for investment banks as opposed to classical retail or commercial banks, as discussed in Van Gestel and Baesens (2009, pp. 29–30). Historically, the distinctions amongst retail, commercial and investment banks were highly dependent on the set of specialities of a bank. In this regard, the Second Banking directive (89/646/EEC) of the European Commission introduced the idea of *universal banking*, which removed the previous boundaries amongst bank types. Instead, the directive describes a complete list of acceptable banking activities, given in Dermine (2007, pp. 494). Regardless, some types of market risk include **equity risk** (downward price movements of equity holdings, e.g., common stock in other companies), **currency risk** (adverse rate movements in investments held in foreign currencies), **commodity risk** (devaluations in physical products like grain, gold, and gas; which impact derivatives), and **interest-rate risk** (downward rate movements in floating rate debt instruments held by a bank for trading purposes).

Market risks are typically measured using a much shorter time window (often measured in days) when compared to modelling credit risk, mostly since market prices are available much more frequently. The VaR-measure is typically used in expressing the maximum loss over these shorter time horizons at a certain probability, as explained in Van Gestel and Baesens (2009, pp. 30–31). Basel II then simply requires holding sufficient capital to cover the quantified level of market risk. Furthermore, some investments may be subject to both market and credit risk, especially when trying to hedge one's risk by investing in debts (e.g., corporate bonds). However, there is a difference between buying instruments intended for imminent trading versus holding

<sup>20</sup>This exercise is currently non-trivial, open-ended, and largely based on the subjective discretion of a bank.

instruments until redemption (or maturity). The former usually attracts market risk whilst the latter carries relatively more credit risk.

The third and final risk in the trifecta is called **operational risk**, which are potential losses as a result of failures in internal processes, IT systems, people, and external events. From Dermine (2007, pp. 500) and Van Gestel and Baesens (2009, pp. 31–33), operational risk generally refers to any other risk type beyond credit risk, market risk, and liquidity risk. Operational risk has many subcategories, which are too numerous and pervasive to list in this text. That said, a particularly significant subclass is that of **legal risk**, which include any regulatory fines or penalties, as well as losses from potential law suits. Another noteworthy subclass is that of **fraud**, which can materialise in ever-changing forms, both internally and externally. More technically, fraud generally transpires whenever illicit financial gain is sought or bank-owned property is misappropriated, including financial assets. Although related to fraud, **cyber security risk** is another subclass that has emerged during the last few years in its own right, according to Alghazo et al. (2017). Internet banking and the self-service offerings of so-called "digital banks" avail greater convenience to customers. In turn, banks have accelerated the adoption of technology and digitisation efforts to cater for this demand, though at the risk of security breaches. In this regard, digital attacks may be launched on the computer and network systems of a bank, usually to steal valuable and sensitive data and/or cause reputational damage.

Other more unintentional forms of operational losses include general negligence when servicing customers, e.g., fiduciary failures that may lead to fines from an ombudsman or legal suits. Operational losses further include avoidable costs due to failures in trade relations (or contractual agreements) as well as losses in the form of refunds to remedy process-based or human-related errors, or service disruptions. Less distinct forms include damage to property as a result of natural disasters or riots, the payment of personal injury claims due to health and safety violations, and payouts to settle discriminatory events. Most of these operational losses can be mitigated to a certain degree by conducting adequate risk management, which may include buying insurance, developing fraud detection systems, or reserving Basel II-compliant capital for covering operational risk. In addition, embedding an agreeable risk culture and devising adequate internal controls may further reduce the proclivity of human and process errors.

### **2.6.2 Common risk management strategies in banking**

A proper risk management strategy generally aims to smooth away significant volatility in earnings and to avoid large concurrent losses. In modern banking, the risk management function will typically partner with other more operationally-inclined departments (e.g., sales, finance, IT) when fulfilling its role, according to Finlay (2010, pp. 16–23). Risk management itself is



typically conducted in a three-step<sup>21</sup> continuous process, as illustrated in Fig. 2.14. This process generally starts off with risk identification, followed by measuring a risk, and ending with its mitigation by devising appropriate strategies. From Van Gestel and Baesens (2009, pp. 38–40), sources of potential risk or threats to the business model are continuously identified through careful analysis, discovery, and critical thinking. Once identified, the practitioner may conduct statistical analyses of these past events, which enables risk quantification, followed by the risk-ranking thereof based on the severity and/or underlying probability. Developing credit risk models to forecast credit losses is an example of risk measurement. Expert judgement or more theoretical/structural models can also be used in measuring risk, especially in data-poor contexts. The final step is to devise the strategic treatment of the measured risk. In this regard, Van Gestel and Baesens (2009, pp. 40–42) categorises various treatments into four broad groups: avoidance, reduction, acceptance, and transferal. While these treatments are tailored to each specific risk in practice, they will be explained here within the context of credit risk.

**Avoidance** strategies are centred on curtailing the investment decision itself. This is usually achieved either by selectively investing in (or lending to) certain counterparts, or by drastically limiting the exposure amount based on the level of perceived risk. As a secondary benefit, this strategy may reduce concentration risk by progressively restricting credit to ever riskier counterparts, thereby diversifying the loan portfolio. Application credit scoring, as previously discussed in section 2.2, is essentially an example of a credit risk avoidance strategy.

**Reduction** strategies generally try to assume but a part of the underlying risk whilst sharing the remainder thereof with other parties. An example hereof is petitioning another lender to help fund a loan (especially in large corporate lending), thereby sharing the inherent credit risk at the cost of reduced interest income. A simpler example is when a bank requires collateral from high risk borrowers that may be seized in the event of default, thereby reducing the overall credit risk exposure.

**Acceptance** strategies are based on assuming the underlying risk entirely, which is generally reserved for low risk cases or portfolios that are already well-diversified. An example of an acceptance strategy is loss provisioning since any potential loss is 'accepted' simply by offsetting it against the provision account. However, risk acceptance directly affects profitability, which may be unpalatable to investors. Regardless of the chosen risk management strategy, the strength and feasibility thereof must be regularly evaluated in line with new loss events (e.g., recent loan write-offs) and the overall business context. This includes 'back-testing' the predictions from risk models onto the recent past, as well as continuously refining the risk measurement process itself.

---

<sup>21</sup>Some will argue that ranking a measured risk as well as monitoring it thereafter poses two additional and interleaved steps. However, risk-ranking is an inherent part of risk quantification and is therefore included in the second step. Moreover, monitoring only 'treated' risks is not as sensible as monitoring *all* steps continuously instead, which is the more pervasive practice.



FIG. 2.14: The three-step continuous risk management process, starting with identification, then measurement, and then treatment of each risk. This process is supplemented by continuous monitoring, evaluation and further refinement.

The latter may take the form of redeveloping credit risk models that may have deteriorated over time, as well as conducting novel academic research in the field of credit risk modelling – not unlike this study.

**Transfer** strategies focus on buying insurance from guarantors for covering any future loss events, thereby transferring the underlying risk to a third party. An example of this is a derivative called a *credit default swap* wherein the insurer reimburses a lender the underlying credit loss in the event of default. Another more notorious example is that of securitisation, whereby originated debts are packaged into more liquid and tradable securities. From Bhattacharya and Thakor (1993), Van Gestel and Baesens (2009, pp. 76–81), and Vento and La Ganga (2009), widespread securitisation became a lucrative business model in the 2000s since securitised debts were no longer held on a bank’s balance sheet, which meant less risk and therefore reduced capital. Furthermore, originating banks used the proceeds from selling off these securities as a funding source itself, thereby promoting further lending activities and entrenching this so-called *originate-*

*to-distribute* (OTD) business model. In contrast, the traditional business strategy is to retain an originated loan on a bank's balance sheet up to maturity or write-off. However, as a risk transfer strategy, securitisation failed abysmally during the 2008 GFC largely due to the inability of those agents who bought securitised debts to perform the classical roles of a bank. Specifically, these entities were neither nearly as well prepared as banks to analyse credit risk, manage credit deterioration, or act as delegated monitors; nor did they keep adequate capital under regulatory supervision.

In conclusion, the main elements of financial risk in banking consist of liquidity risk, credit risk, market risk, and operational risk. The inherent dangers of each risk type is primarily mitigated by implementing an adequate management strategy, which include effective bank policies and holding sufficient capital. However, banks need to balance financial risk against business/performance risk, which generally refers to the failure of a business strategy and/or the erosion of profitability, according to Van Gestel and Baesens (2009, pp. 51–52). Mismanagement of either business or performance risks will likely impact the bank's overall return on equity for shareholders. High-level examples hereof include declining loan production, inadequate loan pricing (e.g., the inability of interest rate margins to cover expected losses or bank expenses), price-elasticity amidst competing banks, improper diversification or simply lavish spending. Ultimately, the principle is to fuse attaining sufficient profitability with adequately managing the risks of loan assets, both within the ambit of a sustainable business strategy. It is clear with no excessive stretch of the imagination that this principle alone can be rather daunting to implement – not to mention even the potential optimisation thereof.



*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## THE BANKER'S GAUGE OF ERODED TRUST

Estimating the frequency of any event in a given sample fundamentally depends on the definition of the event. While loan 'default' is intrinsic to credit risk (and its subsequent estimation), the default event itself has many definitions, used both historically and in modern times. These differences become self-evident when comparing the legal requirements of various prudential regulators, most notably that of the South African regulator<sup>1</sup> compared to the UK and EU counterparts. More importantly, these regulators (and many others) all subscribe to the Basel Capital Accords, which is a set of reasonable principles that were published by the Basel Committee on Banking Supervision (2006a) and upon which credit risk models are commonly built in practice. The recent introduction of the IFRS 9 accounting framework has some interaction with Basel II, especially regarding the former's impairment stage classification that ultimately leads to the notion of loan 'default'. These standards aside, some regulators are more prescriptive than others in enforcing elements thereof, especially regarding default definitions and related topics. However, prescription should never deter broader scientific inquiry in and of itself, especially since regulations are often amenable to (and dependent on) academic advances.

Estimating default risk is becoming increasingly interwoven in the many emerging statistical and mathematical models that drive decision-making in modern banking. Given this ubiquity, the stakes are certainly raised when the exact 'default' point is varied by altering the underlying definition, not to mention its impact cascading across other models. Besides specific regulatory prescriptions, there is limited scientific reason to neglect re-examining the suitability of the

---

<sup>1</sup>Known respectively as the *South African Reserve Bank* (SARB), the Bank of England's *Prudential Regulatory Authority* (PRA), and the *European Banking Authority* (EBA).

default definition and others, even less so when pursuing risk innovation. Admittedly, this is currently more viable in certain areas of banking than in others, e.g., application scorecards and collections modelling. However, the idea of 'default' remains quite firmly rooted in reaching a probabilistic "point of no return", beyond which repayment becomes extremely improbable. Operationally then, a 'default' is fundamental impetus for the lender to act, e.g., by initiating debt recovery and/or abandoning the credit relationship. Surely there must be varying consequences associated with the lender's action in this regard and, more importantly, its exact *timing*. If this is true, then modelling any aspect of the default event using potentially stale definitions thereof will likely be sub-optimal. As paraphrased from Hand (2001), it is reasonable to question the pursuit of modelling excellence when the constructed outcome variable itself, i.e., the *default definition*, is inherently quite arbitrary.

I begin this chapter by reviewing the relevant pieces of regulation and standards that relate to default definitions in section 3.1. This review includes two additional topics. Firstly, the functional areas in retail banking are explored wherein loan delinquency is typically used; secondly, a decision-support tool called a *roll rate analysis* is examined, which is commonly used to test default definitions outside of Basel II and IFRS 9 contexts. Two interrelated problems are then discussed in section 3.2, emerging from the gap in literature on default definitions versus loan collection. First of all, classical roll rate-based approaches are divorced from direct loss considerations and competing costs when varying the 'default' point. Moreover, these roll rates are highly sensitive to a few design parameters (e.g., outcome period length, sampling window), which can obscure idiosyncratic characteristics of the portfolio and enable confirmation biases when selecting a default definition. Secondly, the mere possibility of a loan recovering from a supposed "point of no return" chafes away at any confidence held in the underlying definition thereof, especially as curing becomes more frequent.

Finding a "point of no return" presupposes that loan delinquency is already quantified *a priori*. The measurement of delinquency itself becomes a crucial but often overlooked prerequisite, which, at the very least, warrants a discussion of the underlying limitations of current measurement practices. As such, three delinquency measures are mathematically reworked and discussed in section 3.3. These measures form an intrinsic but modular part in studying the ideal timing of the bank's recovery decision and, by extension, the "point of no return". It is against these thematic backdrops that the philosophy underlying loan recovery is primed as a variable point on the banker's imagined continuum of eroded trust. To this end, I contribute a novel optimisation method in section 3.4 that manifests this idea, called the *Loss-based Recovery Optimisation across Delinquency* (LROD) procedure, followed by the chapter's conclusion in section 3.5. At its core, this procedure attempts to find the ideal point for a given portfolio such that loan recovery occurs neither too early nor too late in aggregate loan life.

### 3.1 Default definitions: a servant of many masters

To collect on a distressed loan is to have breached a certain "point of no return" in the relationship between bank and borrower, at which the lender abandons all hope in the eventual repayment of a loan. This notion is arguably similar to that which underlie most default definitions found in practice. In turn, default definitions themselves become a valuable starting point for a discussion that is ultimately centred on *when* to abandon a troubled loan. Banks have used different default definitions throughout history, even differing amongst various portfolios held by the same bank. Similarly, the prescriptions of regulators relating to default definitions differ by jurisdiction; especially so in the degree of flexibility when interpreting international standards, such as Basel II and IFRS 9 (see sections 2.5–2.6). To this end, various regulations and standards relating to default are examined in subsection 3.1.1, followed by surveying the default definitions of well-known external credit rating agencies.

However, default definitions serve a few other masters beyond domestic regulations and international standards, most notably the contexts of credit scoring, loan pricing, and collection modelling. It is therefore worthwhile to review how both 'delinquency' and the subsequent notion of 'default' is used elsewhere in banking, most notably in basic credit and pricing decisions. Having reviewed these areas in subsection 3.1.2, the specifics of a decision-support tool called a *roll rate analysis*, which is commonly used in selecting default definitions when building application scorecard models, are discussed in subsection 3.1.3.

#### 3.1.1 A regulatory overview of default definitions

Credit risk is inherent to any credit agreement, though the manner in which the default event is defined can vary by product, customer type, and bank. Historically, these definitions include reaching a certain number of days in arrears (or overdrawn), filing for bankruptcy, claims that are not fulfilled up to a certain nominal value, negative net present values, as well as being three payments in arrears, as discussed in Van Gestel and Baesens (2009, pp. 203–207) and Baesens et al. (2016, pp. 137–138). The introduction of Basel II brought with it a greater degree of uniformity in the formulations of default definitions across banks world-wide, while still leaving some room for subjective discretion (subject to the local regulator's approval). Specifically, paragraph 452 of the Basel Committee on Banking Supervision (2006a) defines 'default' when either one or both of the following events has occurred:

1. The obligor has reached 90<sup>2</sup> days past due (or three payments in arrears) on a material loan balance, or has been in excess of an advised credit limit for at least 90 days;

---

<sup>2</sup>Paragraph 452 of the Basel Committee on Banking Supervision (2006a) concedes that some regulators allow up to 180 days past due as a default criterion for retail and public sector exposures.

2. The bank considers, *in its opinion*, that the obligor is unlikely to repay its obligations in full, without the necessary intervention of the bank, e.g., liquidating any collateral.

Basel II also lists a few reasonable indicators of default in paragraph 453 for those banks following the IRB approach, which are in many cases enforced verbatim by the regulator of a country. For example, Regulation 67 of the amended Banks Act of South Africa (2012, pp. 1201–1202, 1211) defines ‘default’ exactly the same way as in Basel II and lists the same indicators. In addition, Regulation 67 indirectly describes ‘default’ when it defines *non-performing debt* as those debts having reached the point when it is “no longer prudent to credit interest receivable to the income statement” of a bank. This definition reinforces the original notion of default as having reached a certain “point of no return”, i.e., unlikely to repay in Basel’s parlance. Accordingly, banks are afforded a modicum of discretion in formulating their own default definitions, which lends additional credibility for the present study. Finally, Basel II’s default indicators include the following, at a minimum:

- Indicator 1** The bank assigns a non-accrued status to the relevant credit obligation, thereby no longer charging interest;
- Indicator 2** The bank writes down a portion of the credit obligation, or raises a specific provision, as a result of the belief that the credit quality has significantly deteriorated since the inception of the credit obligation;
- Indicator 3** The bank resolves to sell the credit obligation at a material economic loss related to credit risk;
- Indicator 4** The bank files for the obligor’s bankruptcy;
- Indicator 5** The bank agrees to the restructuring of the credit obligation, which likely results in a materially reduced financial obligation;
- Indicator 6** The obligor files for bankruptcy (or is placed therein), which will likely either delay or avoid repaying the credit obligation.

Some of these indicators (of “*unlikeliness to repay*”), most notably, indicators 1–4, are retrospective in that they denote ‘default’ as a result of certain actions taken by a bank *ex post*. However, these actions are only reasonably pursued after a bank has already resolved that continuing the credit agreement is of decreasing (or little) financial benefit. This is to say that the trust between bank and borrower in honouring the credit agreement in full has already been eroded beyond a certain point, likely as a result of persistent non-payment. If one considers that reaching this particular precipice already reflects ‘default’ in essence, then these specific



default indicators do not preemptively signal ‘default’ as much as they merely reaffirm what a bank already considers to be obvious. This suggests the fallacy of circular reference, or *petitio principii*, on the premise of using these indicators in defining default when they themselves are deducible by presumably the same underlying default criteria, e.g., having accrued a certain level of arrears.

The remaining indicators 5–6 are uncertain and prospective in nature since they may not necessarily coincide with either previous or future non-payment, at least not with absolute certainty. Consider a financially-distressed obligor who is not yet in arrears, contacting the bank to restructure his obligation timeously. At this point of restructure, there can be no erosion in trust (or default) since there is no amount in arrears. This remains the case even though the restructured terms may eventually lead to a slight economic loss relative to the original agreement. In this scenario, statistical analysis may very well show a time-lagged relationship between the restructure event and subsequent non-payment, which supports the original premise of a restructure event denoting default. However, though default is reasonably likely, it is *not* an absolute certainty at the point of restructure itself. Consider the converse of a restructure or bankruptcy event indeed following a series of unpaid instalments. These indicators 5–6 merely reaffirm the observed erosion of trust incurred by non-payment instead of supposedly signalling it, similar to indicators 1–4. Therefore, indicators 5–6 ought to be considered more as possible predictors of default, rather than indicating definite default at a certain point in time.

The IFRS 9 (2014) accounting standard does not rigidly prescribe a fixed or singular default definition for all banks. Doing so would be challenging (and perhaps unwise) since banks (and their internal business units) differ from one another in their product offerings, risk appetites, operations and financing, markets in which they operate, and levels of sophistication. Instead, IFRS 9 more reasonably requires in paragraph B.5.5.37 that a chosen default definition simply be consistent with the definition used in other internal credit risk models and management. Furthermore, the same paragraph provides a rebuttable presumption of 90 *days past due* (or DPD) as a default definition. The rebuttal hereof can logically lead to any other default definition, provided that “*reasonable and supportable information*” demonstrates its appropriateness.

The recent guidelines (D403 of 2017), published by the Basel Committee on Banking Supervision (2017), intends to harmonise default definitions across banks by proposing the use of 90 DPD quantum as a universal definition of a *non-performing loan* (NPL). This criterion is supplemented by additional default indicators of unlikeliness to pay and applies to any exposure type (retail or otherwise). Collateralisation does not play a direct role, except perhaps as an indirect default indicator, should a bank think it necessary. Critically, this universal NPL-definition is not intended to replace any default definition currently used for loss impairment calculation or IRB capital estimation, as noted in the guidelines. Instead, a central definition will sensibly promote

the comparability of credit risk information amongst banks, as well as help regulators assess asset quality more accurately.

The regulators of some jurisdictions may grant concessions or impose additional requirements on member banks when interpreting Basel II's default definition. As an example, the South African regulator issued circular C2/2014 in SARB (2014) wherein it stated that Basel's 90 DPD criterion has broad equivalence to that of using a three-month missed instalments-based definition, which is widely used by South African banks. More importantly, the SARB acknowledged that the costs of implementing Basel's day-count definition will be too great for South African banks. Many of their computer systems are designed for instalment-based products, which integrates better an instalment-based default definition. Accordingly, the communique listed necessary steps for banks to follow when intending to use an alternative default definition that differs from Regulation (i.e., Basel II). Most notable of these steps is that of demonstrating the suitability of the proposed definition over time for a particular portfolio. This particular concession arguably provides academic scope for pursuing alternative definitions – particularly when varying the time-based element thereof – provided these definitions are demonstrably superior at the end of the day. Relatedly, Regulation 1 of the amended Banks Act of South Africa (2012) states that compliance with the Regulations should not prove costlier than the risk benefits accrued by being compliant in the first place. At the very least, this affords a mandate for experimentation, even if it may lead to an alternative default definition that yields fewer losses, but also contrasts the prescriptions of the SARB – or even Basel II.

Another regulatory example is Article 178 in Regulation (EU) No 575/2013, the so-called *Capital Requirements Regulation* (or CRR) as promulgated by the European Parliament (2013), which gives the default definition of an obligor across all EU-member banks. The European Banking Authority (or EBA) later amended Article 178 in EBA (2016) by providing guidelines on interpreting and applying the default definition within the EU jurisdiction. Firstly, the same six minimum default indicators as in Basel II's paragraph 453 are enforced when defining default. In particular, the EBA (2016, §2.3.1) clarified Article 178(3)(b) of the CRR, which relates to a perceived decline in credit quality (Indicator 2) serving as a default indicator. In line with IFRS 9, an obligor should be classified as having defaulted whenever a bank recognises a Specific Credit Risk Adjustment (or SCRA) due to the obligor's credit risk having deteriorated. These adjustments include

- (a) losses representing credit risk impairments that are recognised in the profit/loss account for all instruments measured at fair value;
- (b) losses due to current/past events affecting either a single though material exposure or a collection of less material exposures that are significant in the collective.

Although Stage 2 exposures under IFRS 9 already contain SICR-accounts, i.e., those accounts whose credit risk has potentially increased, the EBA (2016, §2.3.1) stresses that reaching Stage 2 does *not* constitute an indicator of default by itself. The guidelines do, however, seek to align reaching Stage 3 credit impairment under IFRS 9 as a broader sign of default, in an effort to harmonise some aspects between Basel-compliant capital modelling and IFRS 9-compliant expected loss modelling. However, under no circumstances should the treatment of SCRA under IFRS 9 overrule the CRR and its requirements or discretions. To this point, reaching Stage 3 under IFRS 9 will generally constitute a default event<sup>3</sup>, except for

- (a) cases where the discretionary use of up to 180 DPD is applicable as a default criterion, as per Article 178(1)(b) of the CRR and paragraph 452 of Basel II;
- (b) cases of *technical* (or false) defaults, i.e., having entered default purely due to system/data errors or evidenced failures in the payment system, as clarified by the EBA (2016, §2.2.2);
- (c) cases of *immaterial* (or too small) defaulted exposures, i.e., where the materiality threshold remains unbreached, as per Article 178(2)(d) of the CRR, and discussed hereafter;
- (d) exposures to central governments, local authorities, and public sector entities, as defined by the EBA (2016, §2.2.3).

Regarding the aforementioned materiality threshold, point 1 of Basel II's paragraph 452 requires at least 90 DPD to lapse on a *material* credit obligation before the exposure is considered as in default, which is enforced by Article 178(1)(b) of the CRR within the EU jurisdiction. To help clarify the meaning of a 'material' past due obligation, the EBA (2018) specified that all exposures should be subjected to a materiality test that consists of two components when deciding default. Firstly, the arrears balance  $A$  must exceed a specified limit on an absolute basis to be considered material. Secondly,  $A$  as a proportion of the outstanding balance  $B$  must exceed a given %-valued threshold to be considered material. Both of these thresholds may be differentiated between retail exposures and other non-retail exposures. Owing to differences in economic conditions, individual EU-member regulators are allowed some flexibility when setting the thresholds of the materiality test. The EBA suggests  $A > €100$  (or the equivalence thereof in domestic currency) and  $A/B > 1\%$  for an overdue retail exposure. For non-retail exposures that are overdue, the EBA retains the threshold of the relative component while suggesting a threshold of  $A > €500$ . More recently, the UK regulator rendered these thresholds moot for retail exposures when setting  $A > 0$  and  $A/B > 0\%$  in PRA (2019), though still adopting the EBA's suggested thresholds for non-retail exposures. Conducting these materiality tests when defining default is expected to enter into force from 31 December 2020 at the latest, at least for capital estimation.

---

<sup>3</sup>However, if Stage 3 credit impairment status is the only default indication, then such exposures may still be considered as performing, given that CRR supersedes IFRS 9.

Perhaps more interesting is that Article 178(1)(b) of the CRR allows IRB-banks to relax the standard 90 DPD criterion up to 180 DPD, where appropriate<sup>4</sup>. This discretion recognises that the quantum of this particular default criterion may differ amongst banks, presumably in line with the varying levels of risk tolerances of each bank as they compete with one another. However, the EBA recently announced its intention of withdrawing this discretion, primarily since only a small number of UK banks (and one French institution) are currently using it as their default criterion. The 2017/17 opinion piece of the EBA (2017b), with an annex given in EBA (2017a) containing an empirical analysis, argued that this withdrawal will harmonise reporting and remove 'unwarranted' variability in risk-weighted assets (or RWA) across EU banks. Its analysis was primarily based on measuring the change in RWA implied by adopting 90 DPD instead of 180 DPD as the default criterion. Using highly aggregated data from affected institutions, the EBA found that capital will likely increase for two thirds of these institutions by an average of 1.61%, or with a minimum and a maximum relative change of -20.3% and 23.57% respectively. However, the analysis assumes all other factors remain equal and ignores the fundamental opportunity costs at play and potential risk benefits to be had when varying the 'default' point as a function of arrears levels, which is explored later in this study. As such, it is unfortunate that both the EBA and the UK regulator, having recently enforced this opinion in PRA (2019) starting 31 December 2020, seek such a high degree of bureaucratic uniformity amongst banks in the formulation of risk itself that it may stifle risk innovation in this regard.

The Basel II default indicator, as specified in Article 178(3)(c) of the CRR relating to selling the credit obligation at a loss (Indicator 3), was further refined by the EBA (2016, §2.3.2). In particular, a bank may decide to sell an obligation at a loss due to liquidity concerns or changes in business strategy, which logically does not constitute a default event for the obligor. If, however, the bank sells an obligation  $E$  at an agreed price  $P < E$ , which results in a *credit-related* economic loss, then this event should be considered a default indicator. This should only apply provided that the %-difference is sufficiently material, e.g.,  $(E-P)/E > 5\%$  as suggested by the EBA (2016, pp. 27–28) in paragraphs 41–48. Moreover, the EBA (2016, §2.3.5) contends that both Basel II's paragraph 453 and Article 178(3)(a–f) of the CRR do not provide an exhaustive list of default indicators. Banks are expected to consider and pursue additional indicators of default (or measures thereof) if they think it necessary, based on their own experience. Lastly, the EBA (2016, §2.6) requires that a bank should strive to apply default definitions consistently within its internal risk management, as far as possible. However, the guidelines concede that some differences in definitions may be unavoidable, e.g., different legal entities within the banking group, or different jurisdictions across various geographical locations in which the banking group operates.

---

<sup>4</sup>This criterion applies specifically only to residential mortgages, commercial properties within the retail exposure class, or public sector entities. Other than these, this treatment should only be applied in exceptional circumstances for exposures to central governments, local authorities, and public sector entities, as noted by the EBA (2016, §2.2.3).

A particular default definition may be differently applied either at the facility-level (or product-level) or at the obligor-level. The latter case equates to triggering default on all facilities held if the obligor defaults on any one of his facilities. However, if a bank assesses some of the obligor's exposures at the facility-level instead, then triggering default on any of these specific exposures will not automatically propagate across other exposures held by the same obligor. Naturally, this is at odds with the spirit of applying the default definition at the obligor-level. The paradoxical reason is that paragraph 455 of Basel II and Article 178(1) of the CRR allows for applying the default definition at the facility-level for all *retail* exposures. The EBA (2016, §2.7) advises that banks applying their default definitions at different levels across different types of retail exposures, should demonstrate that the cases are negligible where the same borrowers are subjected to various default definitions applied at varying levels. More broadly, there is no presumption of default contagion amidst retail exposures, i.e., a defaulted exposure of one facility held by a distressed retail borrower will not automatically trigger default on another facility held by the same borrower. However, banks are encouraged to analyse the extent of default contagion (especially in retail banking), perhaps to formulate an additional default indicator, if appropriate. That said, such an exercise can quickly escalate into statistical modelling<sup>5</sup>, which detracts from framing the default definition first and foremost as a measurement problem.

Another Basel II default indicator, Article 178(3)(d) of the CRR regarding distressed restructures (Indicator 5), is rather unclear on what exactly constitutes a *materially* diminished financial obligation. In fact, guidelines posted by the EBA (2016, §2.3.3) sought to clarify this particular indicator by proposing a so-called 'impairment test', in line with IFRS 9. As a result, all cases of restructuring that will likely result in an overall diminished financial obligation due to credit forgiveness, forbearance measures, or postponement/suspension of principal, interest, or fees, should be considered a *distressed* restructuring. The extent of financial diminishing should be subjected to a materiality test, which is based on the difference in the net present value of expected cash flows before and after changing the contract's terms and conditions. Both net present values, denoted  $P_1$  (before restructuring) and  $P_2$  (after restructuring), should be compared at the same point of restructuring and discounted using the original effective interest rate (in line with IFRS 9). If  $(P_2 - P_1)/P_2$  exceeds a specified %-threshold, then the distressed restructure should be recognised as a sufficiently material default event, e.g.,  $(P_2 - P_1)/P_2 > 1\%$  as suggested by the EBA (2016, pp. 29–30) in paragraphs 49–55.

The South African regulator defines distressed restructures similarly as the EU regulator when defining forbearance, following that the Basel Committee on Banking Supervision (2017, pp. 12–14) in paragraphs 36–44. Directive D7/2015 of the SARB (2015) states that restructured exposures that are in arrears (excluding *technical* arrears) either at the time of restructure or

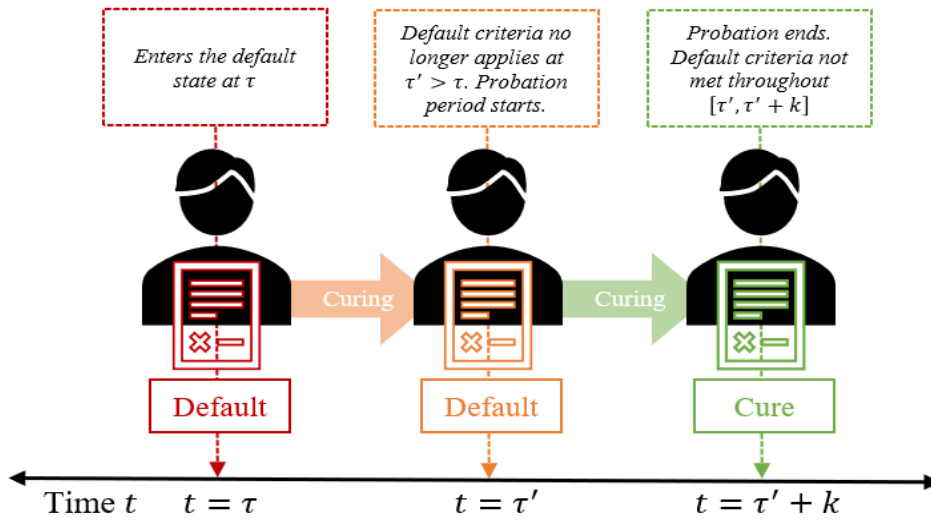
---

<sup>5</sup>A more prudent approach would be to include information on possible default contagion as input variables when modelling elements of credit risk.

during the past six months prior, should be considered as *distressed* restructures. This includes loans with no arrears that were preemptively restructured to prevent the obligor from going into arrears in the first place. However, this requirement seems overly punitive, especially when subsequent payments are dutifully paid by the obligor, even if these instalments are slightly reduced. Nonetheless, the South African regulator considers distressed restructures as objective evidence of credit impairment. Therefore, banks need to conduct regular impairment tests and, if necessary, raise a specific impairment against these restructured exposures. As long as a specific impairment exists, the exposure must be considered as in default. The South African regulator also condones (though does not prescribe) a more stringent policy wherein all distressed restructures are automatically classified as in default, regardless of recognising an impairment loss.

In general, the default state is not an indefinite state into which the obligor is forever trapped. Both paragraph 457 of Basel II and Article 178(5) of the CRR acknowledges this and requires banks to rate the obligor/facility as they would for a performing exposure whenever default criteria cease to apply. However, reclassifying a previously-defaulted exposure as performing, i.e., curing, was limited and made subject to additional requirements by the EBA (2016, pp. 34–36) in paragraphs 71–74. To ensure the curing assessment is sufficiently prudent and that the credit quality has indeed ‘permanently’ improved, a minimum probation period of three months applies. During this period (itself perhaps informed by the standard 90 DPD default criterion), the obligor’s behaviour and financial situation are carefully monitored. This period starts from the moment that the obligor no longer meets any default criteria, and must lapse in full and without pause before exiting the default state. Distressed restructures, however, warrant special attention since such an exposure will never cease being restructured until it is fully repaid or de-recognised (e.g., written-off). Therefore, the EU’s minimum probation period that applies to distressed restructures is 1 year from the latest of: i) the moment of restructure, plus any grace-period extended to the obligor or ii) the moment of default, provided that the exposure is no longer in default at the end of the probation period. The South African regulator is less stringent in this regard, having simply specified a minimum probation period of six months for distressed restructures in directive D7/2015 issued by the SARB (2015), with no minimum probation period prescribed for ‘regular’ defaulted exposures. These ideas and legal requirements (where applicable) are summarised in Fig. 3.1, along with a proposed minimum probation period of three months for normal South African defaulted exposures.

Aside from the various regulators, external rating agencies have their own (though not too dissimilar) default definitions, given in Table 3.1. These agencies regularly publish credit risk ratings on a wide range of counterparts, in guiding investment decisions. As discussed in Van Gestel and Baesens (2009, pp. 115–117, 149–151), the three most prominent rating agencies to date include Moody’s, Standard & Poor’s (S&P), and Fitch. While the original intent was



*Minimum probation periods*

Default type	UK / EU	South Africa
Normal	$k \geq 3$	$k \geq 3^*$
Distressed restructure	$k \geq 12$	$k \geq 6$

*\*Proposed only -- no legal minimum specified*

FIG. 3.1: Illustrating the curing process wherein a defaulted exposure rehabilitates and exits the default state, provided that default criteria are no longer met over a minimum period. Legal requirements are tabulated for both South Africa and the UK/EU, detailing the minimum length of the applicable probation period  $k$ .

to differentiate between investment grade and non-investment grade debt securities (mainly government bonds), modern ratings cover a spectrum of issuers and associated credit risk, which are frequently re-assessed. These ratings are generally assigned by a committee of domain experts using both quantitative and qualitative methods on public and private data. Examples of data include the financial statements of the entity being assessed, their debt structure, management quality, market position and growth prospects. These entities commonly include large companies, banks, state-owned enterprises, municipalities, and sovereigns themselves. Lastly, the eventual credit ratings assigned to these entities are no longer only used for investment decisions, but also in asset pricing and general portfolio management.

All three agencies measure default fundamentally based on the non-payment of interest and/or capital repayments, including the delay thereof within a grace period of certain length (one day for Moody's, 10–30 days for S&P and Fitch). However, it is clear from Van Gestel and Baesens (2009, pp. 208–209) that there are some nuanced differences to this principle. Moody's will consider a contract to be in default on the very first day of payment becoming delayed, while S&P and Fitch applies a grace period of 10–30 days. Furthermore, Moody's does not consider

Credit quality	Fitch	Moody's	S&P
Extremely strong	AAA	Aaa	AAA
	AA+	Aa1	AA+
Very strong	AA	Aa2	AA
	AA-	Aa3	AA-
	A+	A1	A+
Strong	A	A2	A
	A-	A3	A-
	BBB+	Baa1	BBB+
Adequate	BBB	Baa2	BBB
	BBB-	Baa3	BBB-
	BB+	Ba1	BB+
Speculative	BB	Ba2	BB
	BB-	Ba3	BB-
	B+	B1	B+
Highly speculative	B	B2	B
	B-	B3	B-
	CCC+	Caa1	CCC+
Vulnerable	CCC	Caa2	CCC
	CCC-	Caa3	CCC-
Highly vulnerable	CC	Ca	CC
Extremely vulnerable	C	C	C
Selective, restrictive default	RD	RD	SD
Default	D	D	D

TABLE 3.1: The long-term credit risk ratings published respectively by Fitch, Moody's, and Standard & Poor's. Investment grades refer to overall good creditworthiness, e.g., Aaa–Baa3 for Moody's, with the remainder of ratings located lower in the spectrum denoting speculative (or higher credit risk) grades, e.g., Ba1–C for Moody's. The exceptions to these otherwise *a priori* predictions of default risk include the observed/actual default states, e.g., RD–D for Moody's. Recreated from Van Gestel and Baensens (2009, pp. 116).

technical defaults (e.g., covenant violations), while S&P ignores the dividends due from preferred stock as financial obligations, which implies that unpaid dividend payments are ignored. On the other hand, there is some agreement amongst their definitions as well, such as considering bankruptcy proceedings as indicative of default, similar to Basel II's fourth and sixth indicators. Secondly, exchanging a debt security under distress, or repackaging of an existing obligation such that the overall financial position of creditors is reduced, are both considered as signs of default by all three agencies. This is similar to one of Basel II's default indicators relating to the distressed restructuring of a loan (Indicator 5).

The use of external data (such as the previous credit ratings) when defining default internally



is supported by both paragraph 456 in Basel II and by the EBA (2016, §2.4) in Article 178(4) of the CRR. However, banks following the IRB approach should rigorously assess the differences in default definitions and the impact thereof when using external datasets. Adjustments are required where necessary, otherwise it will suffice to show that the differences are negligible in their impact on the eventual risk parameters. If broad equivalence between internal and external definitions cannot be demonstrated, a larger margin of conservatism should be applied when estimating the subsequent risk parameters. The exact calculation of this margin, however, is not prescribed and should presumably be challenged and assessed internally by the relevant technical committee of a bank.

The discussion hitherto has made clear the differences amongst the default definitions of various institutions, as well as the relevant prescriptions of some regulators. Neither Basel II nor IFRS 9 define 'default' *absolutely* and both standards afford some discretion to lenders, perhaps wisely so. Regulators may enforce these international standards to different levels, depending on how flexible they wish to be in their interpretation thereof. That said, at least some of the default regulations seem to coalesce around the central idea of low repayment probability (or "*unlikeliness to repay*") in trying to define where this rather probabilistic point ought to be. Admittedly, the recent regulatory drive for standardising these definitions is quite understandable from a compliance and comparability perspective. However, in doing so, regulators dampen the probabilistic element of "*unlikeliness to repay*" by decreeing certain criteria (e.g., 90 DPD) as risk 'absolutes' beyond reproach. Instead of finding this threshold statistically using decision theory uniquely for each portfolio, a standardised default threshold devolves into little more than a static hurdle – perhaps useful for reporting and accounting purposes, but not much more than that. If reaching 'default' is indeed impetus for the lender to abandon the credit relationship in having reached a "point of no return" (as it has been historically), then surely there must be different consequences to varying the timing of this recovery decision. In turn, retaining possibly stale default definitions purely for the sake of regulatory compliance seems like a wasted opportunity when pursuing risk modelling innovation in the grander scheme of things. This is especially regrettable when there appears to be little objective evidence that supports fixing the default threshold to such a static and risk-insensitive value as 90 DPD.

#### **3.1.2 Delinquency: the leitmotif in risk models**

The advent of loan application credit scoring (as discussed in section 2.2) made necessary a more methodical manner of measuring credit risk. Bankers required an automated proxy of sorts for capturing what is essentially the development of mistrust between bank and borrower, as instalments go unpaid over time. In this sense, any 'accrued' mistrust (or delinquency) can abate over time once the borrower posts a series of overpayments, thereby reducing the arrears amount. In turn, this gradually restores some confidence that the original credit agreement will again

be honoured. The potential for either worsening or abating mistrust suggests that 'delinquency' itself is not a fixed state, but instead a flexible level of incurred mistrust over time, as defined in Def. 3.1 within this study.

### Loan delinquency

**Definition 3.1.** Loan delinquency is a time-dependent measurable quantity that represents the extent of eroded trust between bank and borrower in honouring the original credit agreement. Let  $g$  denote such a delinquency measure wherein an increased value of  $g$  signifies increased mistrust, and *vice versa* for a decreased value of  $g$ .

The number of **payments in arrears** is a commonly constructed and accountancy-based measure of delinquency wherein the unpaid portion of an instalment is aged into increasingly severe bins as each 30-day calendar month lapses: 30 days, 60 days, 90 days, and so forth, as discussed in the introduction of Cyert et al. (1962). The most severe bin attained finally becomes the delinquency measurement itself, which is referred to as the  $g_0$ -measure in this research. More formally,  $g_0$  is constructed using the arrears amount  $A_t$  (measured at time  $t$ ) and a fixed instalment  $I$ , as assembled in Def. 3.2 with an appropriate rounding function  $f$  that maps the input to the number of payments in arrears. This measure is most sensible within the context of amortising loans or credit facilities with a contingent series of regular payments due, e.g., credit cards or drawn/utilised overdrafts. That said, it can be generalised to other credit commitments without regular instalments becoming overdrawn, by calculating an artificial 'instalment' that settles the overdrawn portion (effectively an 'arrears' amount) over a set period. Regardless, the  $g_0$ -measure has many variations in practice, with the most common variant thereof given in Def. 3.2. Moreover, the present study will later examine some of the flaws of the  $g_0$ -measure<sup>6</sup> in subsection 3.3.1 towards devising a more robust variant thereof called the  $g_1$ -measure.

Banks commonly specified 90 DPD (or approximately whenever  $g_0(t) \geq 3$  payments in arrears) as the point of default, long before the introduction of Basel II when building application scorecards. This particular point, however, is largely informed by managerial discretion and can vary from 30 days up to 180 days, depending on data availability and the particular loan portfolio, as discussed in Thomas et al. (2002, pp. 123–124), Siddiqi (2005, pp. 32–42), Van Gestel and Baesens (2009, pp. 208–212), and Baesens et al. (2016, pp. 90, 115). Factors that influence this default threshold  $d \in \mathbb{Z}^{\geq 0}$  include the type of security (or collateral) underlying credit commitments, e.g., secured mortgages vs. unsecured term loans. Another factor is the contractual loan term, which imposes some reasonable bounds on the chosen default threshold. This is perhaps best

---

<sup>6</sup>The  $g_0$ -measure carries many names, e.g., the number of missed instalments, payments in arrears, *contractual delinquency* (or CD-level), and arrears categories. Though a more robust variant  $g_1$  is later developed in this study, the name 'CD-measure' is retained for both  $g_0$  and  $g_1$ , given their conceptual similarities.

demonstrated when considering a very short term 2-month pay-day loan versus a standard 60-month vehicle loan. In this case, a default threshold of  $d \geq 3$  is clearly nonsensical for the 2-month loan, though it demonstrates the principle. Lastly, the bank's risk appetite is arguably the most important factor when deciding the default threshold. The basic principle hereof is that the higher the appetite for arrears, the more forgiving/greater the default threshold  $d$ , and *vice versa*.

#### Payments in arrears ( $g_0$ -measure)

**Definition 3.2.** If  $A_t$  denotes the accumulated amount in arrears at a given loan time  $t = 0, \dots, T$  with  $T$  being the contractual maturity (or fixed time horizon), and if  $I$  represents the level instalment, then the **payments in arrears** delinquency measure (henceforth called the  $g_0$ -measure) is defined as

$$g_0(t) = f\left(\frac{A_t}{I}\right), \quad (3.1)$$

where  $f$  is a chosen rounding function  $f: \mathbb{R} \rightarrow \mathbb{N}_0$  that maps the real-valued input to the non-negative integer-valued output, which denotes the number of payments in arrears.

The use of a loan delinquency measure such as  $g_0$  is widespread and deeply entrenched across the various divisions and functions of a typical modern bank (especially retail banking). Measuring delinquency is commonplace in almost every piece of analytics or statistical model employed in retail banking, which are called 'exercises' in this work. A basic (but non-exhaustive) taxonomy of these exercises is given in Fig. 3.2; see Finlay (2010) for more detailed discussion. Many of these account-level models focus on predicting the default event itself (e.g., PD models) by way of collapsing the delinquency measure into binary 'default/non-default' outcomes via the chosen default definition. Models of this particular kind are commonly used to support various operational decisions typically found across the first three phases (Marketing, Acquisition, and Customer Management) in Finlay's credit life cycle model, as discussed in Thomas et al. (2002, pp. 169–171) and previously in section 2.2. The first model group includes prescreening and preapproval scorecard models in which the default risk of potentially new customers are predicted using any available data. The subjects are then risk-ordered and perhaps selectively culled, thereby producing a list of 'marketing leads' whom the lender can then solicit during the Marketing-phase. Secondly, application scorecard models are perhaps the best known example wherein default risk is predicted, upon which a credit offer decision is subsequently made by the lender during the Acquisition-phase. Lastly, behavioural scoring and the PD-component within capital and loss provisioning models (see section 2.6) continuously predict default risk throughout loan life.

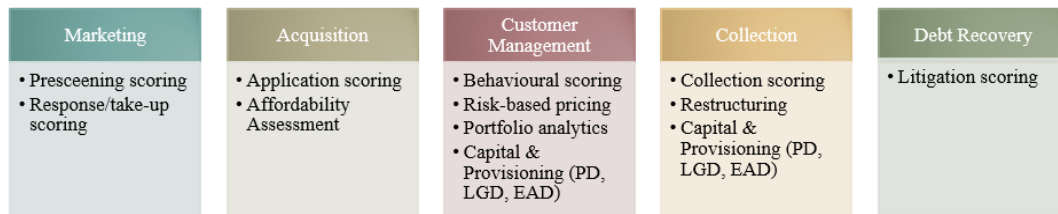


FIG. 3.2: The various credit risk-related models and types of analyses used across the five-phase credit life cycle of retail banking. The common factor amongst these ‘exercises’ is the use of a delinquency measure, used either directly or indirectly alongside a default definition.

The remaining exercises in Fig. 3.2 use this delinquency measure as well, though not necessarily as an outcome variable or within a predictive setting. Instead, collection and litigation scoring commonly use it as an input variable when predicting the likelihood of successful collection or even successful litigation amidst debt recovery proceedings, as discussed in Thomas et al. (2002, pp. 176). Some lenders may deploy advanced default risk models to aid them in restructuring a delinquent account in such a way that the risk of ‘re-default’ is minimised, which naturally requires using a delinquency measure. Furthermore, risk-based pricing commonly uses default risk models to help determine the interest rate (or other contract variables, e.g., the loan amount) of a new loan, as explained in Thomas et al. (2002, pp. 174–175, 227–228) and Thomas (2009a, pp. 152–156). Outside of modelling, a delinquency measure is commonly used in producing various pieces of portfolio analytics within credit reports (e.g., default rates), or to investigate more *ad hoc* phenomena, e.g., finding a new trend of increasing arrears within some portfolio.

Perhaps most noteworthy is the use of a delinquency measure  $g$  within the very structure of capital and loss provisioning models when modelling the expected loss<sup>7</sup> of a portfolio. Specifically, the default definition itself conditions the various modelling samples used in estimating each of the three risk parameters, i.e., PD, LGD, and EAD. Moreover, many PD models use delinquency measurements as an input variable itself, apart from already using it in labelling the default event that PD-models typically seek to predict. In estimating LGD, loan accounts that are at risk of a loss can only feasibly include defaulted loans, which means their histories are observed from the moment of entering the default state, up to their resolution (if resolved). Sensibly, using a different default threshold, e.g.,  $d = 1$ , will likely alter the starting point of the workout period when preparing data, which will likely in turn affect the realised loss itself, as calculated from historically written-off accounts. The statistical distributions of various characteristics of defaulted loans may change when changing  $d$  within a default definition, simply due to sampling

<sup>7</sup>For a thorough treatise on modelling the generic aspects of quantifying credit risk, refer to Thomas (2009a, pp. 289–293), Van Gestel and Baesens (2009, §4, §6), and Baesens et al. (2016, §5–11).

at a different starting point. Lastly, the sample sizes themselves will likely change given each candidate threshold  $d$ , especially when modelling default risk – not unlike the floodgates of a dam changing the water flow as it opens or closes.

In summary, measuring loan delinquency and using these measurements in predictive models seems ubiquitous in modern banking. Choosing a threshold  $d$  on the domain of a delinquency measure as part of defining loan ‘default’, e.g.,  $g_0(t) \geq 3$ , appears to be a deeply embedded practice in many of these modelling exercises. This ubiquity is unsurprising given that the very business model of a bank relies on accurately quantifying credit risk, which implicitly belies some internally agreed default threshold. However, the supposed suitability of any particular default threshold, as used within a wider default definition, still remains questionable – regardless of whether the threshold was decreed by a regulator or chosen arbitrarily by the bank itself for whichever purpose. The particular modelling exercise in which it operates, albeit capital modelling or application scoring, becomes almost superfluous, especially if the underlying principle of any default threshold remains that of approximating the “point of no return” in probability. Intuitively, changing the default point will likely have cascading effects across all exercises built upon the default definition – almost unfathomable when considering the modern regulatory environment.

### 3.1.3 Roll rate analyses as decision-support tools

Practitioners often conduct a statistical analysis called a *roll rate analysis* in providing quantitative assurance on a chosen default threshold, as explained in Siddiqi (2005, pp. 41–42). At its core, a roll rate analysis is a cross-tabulation of observed transition rates amongst various ordinal-valued (and increasingly severe) arrears categories across a length of time, called the outcome period. These categories include the newly-chosen and so-called “default state” as imposed via a pre-selected value for  $d$ . An individual roll rate expresses the proportion of a portfolio (or segment) that transitioned from a particular pre-defined delinquency level to another. This can include transitioning to a worse state (rolling forwards), recovering or *curing* back to a better state (rolling backwards), or staying within a particular state (milling). When analysed together, roll rates summarise the delinquency dynamics amongst loans, from which the overall risk profile of a given loan portfolio can be characterised.

It is worthwhile to discuss the typical structure of most loan datasets before estimating these roll rates, as illustrated in Fig. 3.3. A bank generally disburses approved loans continuously over time as part of actively growing its portfolio, which translates into a loan distribution that is staggered over time in the underlying dataset. For most kinds of risk modelling, the first decision is to choose the epoch in time from which to sample loan accounts, i.e., the *sample window* as shown in Fig. 2.4, with associated trade-offs previously discussed in section 2.2. Part of this decision relates to whether the analysis is solely focused on new loans or the entire portfolio. In

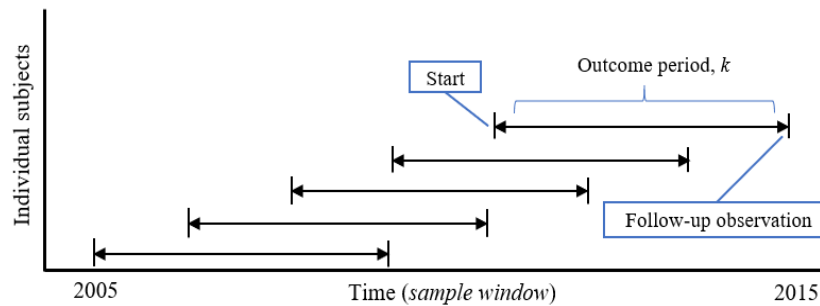


FIG. 3.3: Illustrating the role of an outcome period within roll rate analyses and within most cross-sectional models. A particular subject (e.g., a loan along with its characteristics) is observed at a specific starting point in time, called a ‘snapshot’ or a cross-section, whereupon it is merged with a follow-up observation (e.g., repayment status) after the outcome period has lapsed.

other words, only using new loans implies that the starting time of each loan within the greater sample window will be  $t = 0$ , which is appropriate when building application credit scoring models. Alternatively, the starting time can simply refer to a specific cross-section (or ‘snapshot’) of accounts taken of the portfolio at a particular time, e.g., monthly cohorts, which can include both newly-originated (at the time) and existing accounts. This design has a few extensions of the starting point in practice, though most are more applicable to building behavioural scoring models than credit risk models. From Baesens et al. (2016, pp. 118–120), one such extension is based on an additional *behavioural window*, during which some characteristics are aggregated over time up to the starting point, e.g., the average utilisation of the borrower’s approved credit limit. Naturally, this extension is only relevant for those accounts with sufficient history prior to the starting point.

A secondary set of decisions relate to observing the repayment performance of each loan from the starting time across a given outcome period of  $k$  periods (e.g., 12 months), thereby forming pooled cross-sectional data, as discussed in Thomas (2000) and Siddiqi (2005, pp. 33–36). Other than choosing  $k$  itself, the practitioner has to decide the manner of aggregating loan performances across the outcome period. This can be as simple as recording the latest delinquency value  $g_0(t+k)$  of each loan after the outcome period has lapsed, or by taking the worst delinquency value across the outcome period, i.e.,  $\max g_0(t')$  for  $t' = t, \dots, t+k$ . The latter so-called *worst-ever* approach takes into account a wider span of data and can aid in the detection of distressed debt restructuring during the outcome period, as noted in R. Anderson (2007, pp. 339–340). However, the worst-ever approach can be more punitive towards curing behaviour in that an isolated case of delinquency will persist throughout the remainder of the outcome period, despite subsequent repayments. As such, choosing the more risk-averse worst-off approach can deliberately skew the portfolio’s assessed risk profile away from reality, especially if a portfolio has significant curing

### 3.1. DEFAULT DEFINITIONS: A SERVANT OF MANY MASTERS

behaviour.

Cohort	Outcome period ( $k$ )																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Jan-2018	0.00%	0.13%	0.53%	1.40%	2.07%	2.49%	3.45%	3.59%	3.85%	4.05%	4.31%	4.45%	4.49%	4.53%	4.56%	4.59%	4.62%	4.64%
Feb-2018	0.00%	0.16%	0.58%	1.37%	1.90%	2.87%	3.38%	3.47%	3.89%	4.03%	4.29%	4.42%	4.46%	4.52%	4.54%	4.58%	4.60%	
Mar-2018	0.00%	0.17%	0.51%	1.07%	1.98%	2.37%	3.05%	3.32%	3.67%	3.95%	4.28%	4.37%	4.43%	4.47%	4.51%	4.54%		
Apr-2018	0.00%	0.12%	0.45%	0.97%	1.85%	2.72%	3.35%	3.43%	3.71%	3.98%	4.24%	4.35%	4.39%	4.48%	4.50%			
May-2018	0.00%	0.16%	0.47%	1.10%	1.79%	2.47%	3.08%	3.29%	3.77%	3.92%	4.23%	4.38%	4.42%	4.49%				
Jun-2018	0.00%	0.19%	0.37%	1.30%	2.09%	2.35%	3.17%	3.49%	3.69%	3.95%	4.21%	4.39%	4.43%					
Jul-2018	0.00%	0.11%	0.42%	0.95%	1.89%	2.67%	2.98%	3.33%	3.71%	3.99%	4.18%	4.36%						
Aug-2018	0.00%	0.08%	0.41%	1.15%	2.37%	2.77%	3.07%	3.46%	3.75%	4.01%	4.26%							
Sep-2018	0.00%	0.09%	0.37%	1.20%	1.83%	2.61%	3.16%	3.39%	3.79%	3.91%								
Oct-2018	0.00%	0.17%	0.41%	0.90%	1.88%	2.48%	2.97%	3.55%	3.72%									
Nov-2018	0.00%	0.25%	0.43%	1.31%	2.14%	2.51%	3.02%	3.51%										
Dec-2018	0.00%	0.16%	0.37%	1.32%	2.34%	2.64%	3.32%											
Jan-2019	0.00%	0.14%	0.49%	1.49%	2.17%	2.84%												
Feb-2019	0.00%	0.17%	0.51%	1.43%	2.01%													
Mar-2019	0.00%	0.13%	0.49%	1.27%														
Apr-2019	0.00%	0.09%	0.47%															
May-2019	0.00%	0.13%																
Jun-2019	0.00%																	
Average default rate	0.00%	0.14%	0.46%	1.22%	2.02%	2.60%	3.17%	3.44%	3.76%	3.98%	4.25%	4.39%	4.44%	4.50%	4.53%	4.57%	4.61%	4.64%

FIG. 3.4: An illustrative cohort analysis of cumulative default rates across various candidate outcome periods  $k$ , given a particular default definition. This hypothetical analysis is conducted in retrospect at a particular point in time, e.g., July 2019.

Practitioners often conduct a so-called *cohort* (or *vintage*) analysis to help select an appropriate outcome period  $k$ . This analysis, as explained in Siddiqi (2005, pp. 33–35), considers a particular cohort of loans, e.g., January-2018, whereupon the cumulative default rate for that cohort is iteratively calculated across various candidate outcome periods, e.g.,  $k = 1, \dots, 24$ . Thereafter, these calculations are repeated for subsequent cohorts, thereby forming a triangular matrix as illustrated in Fig. 3.4. As an example, 3.59% of the January-2018 cohort were in default 8 months later. More recent cohorts will likely have less data available than older cohorts, which is testament to the intrinsic right-censoring effect present in most real-world ‘incomplete’ loan portfolios. Furthermore, the resulting cohort analysis is unique to a particular default definition, sample window, loan portfolio (or segment therein), and even product type. By implication, a cohort analysis does not lend itself flexibly to analysing various candidate default thresholds  $d$  themselves, which is problematic when the lender is unsure of a particular criterion within its default definition.

Performing a cohort analysis becomes more valuable once the default rates from Fig. 3.4 are summarised across the various candidate periods under consideration, as shown in Fig. 3.5. Shorter periods usually coincide with default rates that are still in flux, while longer periods typically yield default rates that have stabilised, as exemplified by large (or small) changes in the standard deviation of each period. The practitioner often selects a period from the latter group with more stable default rates simply due to risk-aversion, i.e., default rates should preferably

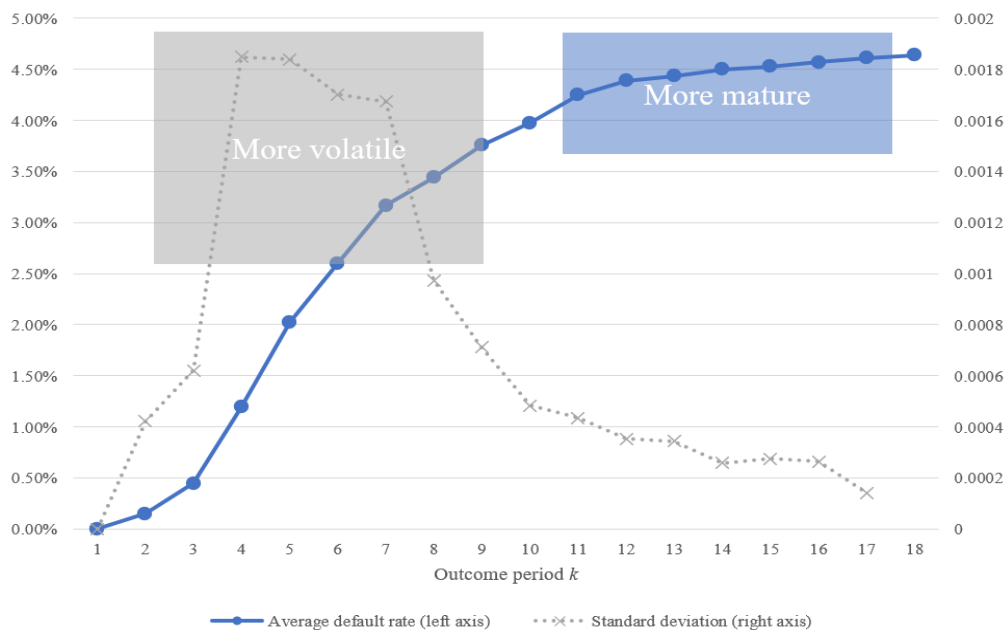


FIG. 3.5: The development of the average default rate across candidate outcome periods as a summary of the cohort analysis from Fig. 3.4. The grey-shaded area shows outcome periods during which the portfolio is still immature, while the blue-shaded area shows outcome periods yielding more mature average default rates.

be overstated rather than understated. From Thomas et al. (2002, pp. 91) and Van Gestel and Baesens (2009, pp. 101–102), the outcome period generally varies from 6–24 months, though literature guiding its selection optimally is fairly limited.

In fact, the work of Kennedy et al. (2013) experimented with various outcome periods in predicting default risk using binary logit models on Irish data. Shorter periods exhibited greater volatility in default/curing rates due to seasonal effects and/or insufficient loan maturity, which may misrepresent the ‘true’ default/curing rates. On the other hand, longer periods may no longer represent current market conditions nor reflect the portfolio’s original risk composition. The authors found that classifier accuracy decreases for increasingly longer outcome periods. Too long a window may also fail to capture unusually rapid delinquency movements, e.g., an oscillating series of defaults and cures, as argued in Kelly and O’Malley (2016). Lastly, Kennedy et al. suggests that using an outcome period between 3–12 months (and a worst-ever approach in assigning the class label) will yield the most accurate model of default risk on average. This result was corroborated in Mushava and Murray (2018) wherein nine different classification algorithms were investigated using South African data across various outcome periods.

Having selected an outcome period  $k$ , an epoch of time, and a method of aggregating loan



### 3.1. DEFAULT DEFINITIONS: A SERVANT OF MANY MASTERS

performance, the practitioner is finally able to conduct a roll rate analysis. As a typical example, consider Fig. 3.6 wherein a large proportion of accounts (80%) remained up to date with their payments following a 12-month outcome period. Similarly, consider the 60% of accounts that remained 90+ DPD (the supposed default state for this hypothetical), and consider the measly 10% of accounts previously in ‘default’ that cured completely. The principle, at least when building application scorecards, is based on back-solving for stability in that accounts identified as ‘lost’ should stay lost at the end of the outcome period. This is to find the "point of no return" at which only a minimum of accounts recover from a supposed default state as imposed via  $d$ , which gives more confidence in the overall stability of the default definition using this  $d$  over time – in this case, 90 DPD. By extension, this instils confidence in any subsequent model that seeks to predict ‘default’, or use the predicted default probability in some way or another.

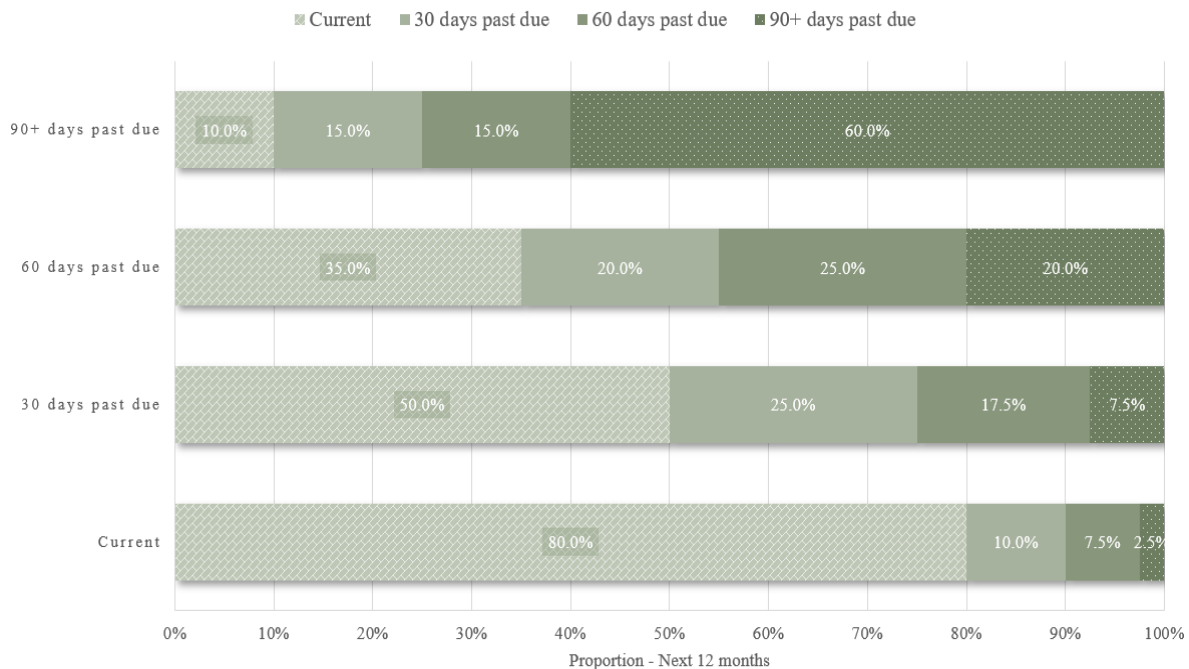


FIG. 3.6: An illustrative roll-rate analysis showing the transitions rates amongst arrears categories across a period of time (12-months).

As a more advanced form of a roll rate analysis, the sophisticated practitioner can use a Markovian approach instead. Specifically, one can model the conditional transition probability itself given several starting and ending delinquency states, in estimating a Markov chain. This was first explored in Cyert et al. (1962) as part of loss provisioning, wherein a time-homogeneous discrete-space Markov chain was built to model the transitions amongst arrears categories. Specifically, the balances of accounts receivable at a particular time point  $t$  were classified into  $n + 1$  delinquency levels (or states) denoted as  $B_0, \dots, B_n$  where  $B_0$  denotes all balances

with  $g_0(t) = 0$  payments in arrears,  $B_{n-1}$  signifies all balances with  $g_0(t) = n - 1$  payments in arrears, and  $B_n$  represents all remaining balances to be written-off with  $g_0(t) \geq n$  payments in arrears. Afterwards, the balances can again be reclassified at time  $t + 1$ , although it becomes necessary to define an additional (absorbing) state  $\bar{0}$  to cater for the balances of settled/prepaid accounts denoted as  $B_{\bar{0}}$ . This leads to calculating an  $n + 2$  square matrix  $B$  wherein the individual element  $B_{ij}$  denotes balances in state  $i$  at time  $t$  that moved to state  $j$  at time  $t + 1$  for all states  $i, j = \bar{0}, 0, 1, \dots, n$ , which is expressed as

$$B = \begin{bmatrix} B_{\bar{0}\bar{0}} & B_{\bar{0}0} & \cdots & B_{\bar{0}j} & \cdots & B_{\bar{0}n} \\ B_{0\bar{0}} & B_{00} & \cdots & B_{0j} & \cdots & B_{0n} \\ \vdots & \vdots & & \vdots & & \vdots \\ B_{i\bar{0}} & B_{i0} & \cdots & B_{ij} & \cdots & B_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ B_{n\bar{0}} & B_{n0} & \cdots & B_{nj} & \cdots & B_{nn} \end{bmatrix}. \quad (3.2)$$

Naturally, one can estimate an  $n + 2$  transition matrix  $P$  from  $B$  that measures the likelihood of transiting amongst the delinquency states over the same time period. The element  $P_{ij}$  is therefore defined as the conditional probability of one rand in state  $i$  at time  $t$  transiting to state  $j$  at time  $t + 1$  for all states  $i, j = \bar{0}, 0, 1, \dots, n$ . The repayment history of each account constitutes a repeated observation of the underlying Markov chain formed by the sequence of random variables  $B(0), B(1), \dots$  over time  $t = 0, 1, \dots$  that can each assume states  $\bar{0}, 0, 1, \dots, n$  respectively, as discussed in T. W. Anderson and Goodman (1957). Assuming stationarity, the *maximum likelihood estimate* (MLE) for each  $P_{ij}$  is defined as

$$P_{ij} = P(B(t+1) = j | B(t) = i) = \frac{B_{ij}}{\sum_{k=\bar{0}}^n B_{ik}} \quad \forall i, j = \bar{0}, 0, 1, \dots, n. \quad (3.3)$$

The work of Cyert et al. (1962) provided useful theorems using Markov theory for directly estimating the allowance for bad debts (state  $n$ ). This is achieved by calculating the particular absorption probability, followed by estimating its variance – as illustrated in Appendix A.1 for  $n = 2$ . Using these theorems, the loss expectancy rates can be estimated for each delinquency state, as well as provide the steady state distribution into which the Markov chain will settle over time. Their work was later extended in Corcoran (1978) wherein account balances were stratified according to size before estimating a transition matrix  $P$  for each stratum, which improved the overall predictive power. The transition matrix itself was exponentially-smoothed each lapsing month to counteract the assumption of time-homogeneity and to incorporate subsequent repayment behaviour. Lastly, Cyert et al. (1962) used an unconventional approach (the so-called "total balance" method) in ageing balances across delinquency states, which is based on retaining the oldest amount due despite partial payments posted subsequently. The ageing procedure

was modified in Van Kuelen et al. (1981) such that collections are no longer understated when estimating the transition matrix.

In general, Markovian approaches are quite prolific in the credit risk modelling literature, even though they are perhaps not commonplace in practice yet. As an example, the bankruptcy process of firms was modelled in Jarrow et al. (1997) as a Markov chain across various credit rating states and an absorbing ‘default’ state. Specifically, the time distribution of the chain first entering the default state was explicitly modelled, from which a probability of default over a certain term structure can be estimated. In agreement with Corcoran (1978) on the Markov chain being time-dependent in reality, a non-stationary forecasting model with a Markovian structure was developed in Smith and Lawrence (1995) using US mortgage loans from the 1970’s and 1980’s. Each transition probability  $P_{ij}(t, t + 1)$  was modelled separately as a multinomial logistic regression explicitly containing the age covariate  $t$ , amongst others. The state space of the eventual Markov chain was limited to four states, which implies that this approach can quickly become extreme for larger though more realistic state spaces.

Alternatively, the Markovian approach followed in Grimshaw and Alexander (2011) proposed modelling only certain transitions within the matrix as time-dependent multinomial logits, which was illustrated using US subprime home loans data. The remainder of the transition matrix can instead be estimated as simple intercept-only models. Additionally, the authors introduced new empirical Bayes estimators that are appropriate for estimating non-stationary and heterogeneous transition matrices. This heterogeneity recognises that certain segments within a loan portfolio may have fundamentally different transition rates. Furthermore, the authors compared estimating the transition matrix based on the actual *dollar* movement amongst states (value-based) versus the *number* of accounts that transited (volume-based). The former value-based approach was originally used in Cyert et al. (1962) and Van Kuelen et al. (1981). While both approaches are statistically unbiased, the value-based approach was shown to be inefficient due to exhibiting greater variance than the volume-based approach.

In modelling the insurance purchase behaviour of customers, the time-homogeneity and first-order assumptions of Markov chains were relaxed in Bozzetto et al. (2005). Portfolio-level credit risk models for corporate exposures were translated into the retail banking context using a Markovian approach, as investigated in Thomas (2009b). The same study argued that the implicit assumption in most Markov models of independent default behaviours amongst borrowers is likely flawed due to the default contagion principle in a given market. Lastly, a four-state non-homogeneous Markov chain was built as part of a larger suite of intensity models in Leow and Crook (2014) for credit card delinquencies. Other examples of Markovian approaches applied to the credit risk context are reviewed in Hao et al. (2010).

Both a roll rate analysis and its more advanced form, a Markov chain, can be iteratively used to analyse the impact of a proposed threshold  $d$  within a larger default definition, based on attaining stability. Whether or not a candidate  $d$  passes managerial muster will likely depend on the particular portfolio, regulatory restrictions, and the associated curing rate implied by  $d$ . Such an analysis, however, often depends on other design parameters that may influence any inference drawn from it, including the sample window and the length of the outcome period. For greater assurance, the practitioner can laboriously repeat the same analysis a few times using slightly different parameter choices each time. Given the bank-wide ubiquity, complexity, and implied impact of varying  $d$  within any default definition, it is perhaps unsurprising that most practitioners (and regulators) simply interpret the default definition as something fixed.

### **3.2 Towards opportune loan recovery: analysing true ‘default’**

Although various regulations exist that constrict the default definition for Basel and IFRS 9 types of credit risk modelling, there are areas in retail banking that can benefit from a new ‘philosophy’ of default – in particular, that of application scoring and optimising the collections process, including related policy decisions. Consider that the original premise of a default definition is reaching a rather probabilistic "point of no return", which may differ for every portfolio (or segment therein) in reality. Therefore, ‘default’ is reinterpreted in this study by simply using  $d$  as a threshold upon the domain of a delinquency measure  $g$ , which is deliberately divorced from current practices and from relevant regulations. This more fundamental meaning becomes useful when attempting to optimise the recovery decision’s timing later in section 3.4, i.e., finding the *best* time at which the lender should forsake the loan and instead pursue debt recovery, including seizing any collateral.

As discussed in subsection 3.1.3, the choice of outcome period and the sample window from which loan performances are drawn can clearly complicate any roll rate-based analysis. This is further exacerbated when trying to choose the ‘best’ threshold  $d$  to approximate the "point of no return" based on these roll rate results, at least without expending a great deal of additional analytic effort. As an example, it would be difficult to decide if a particularly low curing rate is artificially due to an overly short outcome period, testament of the portfolio’s risk profile, or shifting market conditions – without conducting additional analysis. Moreover, a roll rate-based approach ignores the competing financial and opportunity costs that may be in play when varying  $d$  itself, e.g., legal and administration costs, collection staff salaries, loss provision increases, as well as overall effort. While stability in these roll rate analyses is a noble criterion for finding  $d$ , a better alternative may be to consider the direct loss implications associated with any chosen  $d$  instead. For these reasons, a roll rate analysis is deemed unfit as an approach for finding an *ideal* threshold  $d$ , and an alternative becomes necessary.

The work of Harris (2013b) and Harris (2013a) first broached the subject of varying  $d$  within a default definition and studied the subsequent effects thereof on model accuracy<sup>8</sup>. Specifically, support vector machines were used as default-classifiers and it was found that those classifiers trained with increasingly intolerable definitions (or ‘broader’, e.g., 30 DPD) had progressively *higher* accuracy on validation datasets pre-classified with gradually more tolerable definitions (or ‘narrower’, e.g., 120 DPD). In other words, models built with intolerable definitions are seemingly more accurate in predicting defaults that are defined more tolerably (or more severe cases of delinquency). This seems counter-intuitive as one would expect models built with a specific definition to be the best at predicting that very same definition – surely they ought to outperform models built with another definition trying to do the same. However, larger values of  $d$  yield decreasing sample sizes since greater severities of delinquency typically occur less frequently, as noted in Thomas et al. (2002, pp. 124) and Siddiqi (2005, pp. 38), and corroborated by anecdotal experience. These decreasing sample sizes of ‘defaults’ afford less opportunity for the classifier to learn overall patterns of delinquency. This explains the improved accuracy of classifiers trained with smaller values of  $d$ , since they inherently had more training samples.

However, these findings – while proving that  $d$  significantly influences model accuracy – say little about the direct impact on profitability. As originally argued in Hand and Henley (1997) and Hand (2001), a lender is primarily interested in the underlying profitability of a credit decision, with credit risk being but a facet thereof. This is to say that lower default risk borrowers are not necessarily the most profitable since default risk is merely a proxy for profitability, amongst other factors. There is certainly some truth to the idea that borrowers with no arrears (considered as ‘good’ risks) will likely yield a profit for the bank. Conversely, those accounts with sufficient arrears (enough to prompt a default decision) will likely lead to losses and are justifiably ‘bad’ risks. Using 90 DPD as a heuristic rule therefore seems apt and fitting with the historical pragmatism in banking analytics, as discussed in section 2.2. However, the presumption of profitability underlying this particular rule has little objective evidence in literature, with little research effort spent on the topic as a whole.

Loan profitability itself depends on much more than just delinquency, with loan pricing amidst credit market dynamics representing another major factor. As explored in Edelberg (2006), Thomas (2009a, §3), and Phillips (2013), most lenders employ default risk-based pricing by charging higher prices for riskier borrowers, as compensation for the greater credit losses expected in aggregate. This practice, having started with mortgage lenders from the mid-1990s, only became feasible as default risk estimates of the individual borrower became available, which was in turn enabled by lower data storage costs and improvements in both technology and underwriting models. The basis of risk-based pricing is to segment a loan portfolio into a few ordinal-valued risk grades and define a rate factor  $l_k$  specific to the  $k^{\text{th}}$  segment, e.g., a

<sup>8</sup>Accuracy was measured using the *area under the curve* (AUC) and a cross-validation setup.

rate based on the segment's average expected loss. The final customer rate  $r_k$  is then given as  $r_k = r_c + m + l_k$  where  $r_c$  is the overall cost of capital and  $m$  is a desired profit margin targeted by the bank. Moreover, the work of Edelberg (2006) showed that the difference in  $r_k$  (or risk premium) between high- and low-risk borrowers almost doubled for secured loans on average during the 1990s, and increased for most unsecured loans. In turn, lower-risk households had an incentive to increase borrowing given the lower costs, while higher-risk households gained better credit access overall. Risk-based pricing as a practice clearly has demonstrable benefits, not only to borrowers, but also to the base profitability of a bank.

However, risk-based pricing suffers from *adverse selection* wherein riskier borrowers are less price-sensitive than their lower-risk counterparts in accepting a higher-priced loan. Adverse selection, as induced by loan pricing, may lead to banks rationing credit at certain price intervals, which was famously explored in Stiglitz and Weiss (1981). As a secondary market-driven effect, adverse selection implies that the volume of higher-risk borrowers may exceed the lender's original expectation *a posteriori*. It was first empirically evidenced in Ausubel (1999) using the results of several "market experiments" that were conducted by a major US bank. Several pre-approved and varied credit card offers were issued in randomised trials. Evidently, the risk characteristics of respondents were inferior to those of the non-respondents, which confirms the so-called "*Winner's Curse*"<sup>9</sup> phenomenon in the auctions literature. Furthermore, those customers who accepted inferior offers were substantially more likely to default. These inferior offers include higher introductory (or teaser) rates, shorter teaser periods, and higher post-teaser rates. Adverse selection was further demonstrated in Cressy and Toivanen (2001), wherein the variation and trade-offs amongst interest rates, loan sizes, and collateral requirements were empirically modelled given imperfect information. Finally, price-driven adverse selection was incorporated in Phillips and Raffard (2011) by developing a consumer pricing model based on differential price-sensitivity.

Another interesting phenomenon related to profitability is that of *price-based risk* wherein a higher price may itself contribute to higher default rates. One of the many drivers of the 2008 GFC was the prevalence of so-called *adjustable-rate mortgages* (ARMs), wherein the interest rate rose after an introductory period (usually 1-2 years) by contractual design. Perhaps unsurprisingly, the associated default rates rose significantly as rates increased and house prices fell, which was demonstrated in Campbell and Cocco (2015) using rational expectations theory. Moreover, credit-constrained borrowers with uncertain incomes were found to favour ARMs, especially when the offered teaser rates were low. It so happened that this scenario attracted ever riskier borrowers into ARM-portfolios leading up to the crisis, i.e., adverse selection. As the crisis started to unfold, some borrowers exercised a "default decision" to avert negative home equity as house

---

<sup>9</sup>In the absence of adverse selection, there should be no difference in the characteristics between respondents and non-respondents, as explained in Ausubel (1999).

prices started to fall. In tandem, unemployment shocks meant that more borrowers wanted to access home equity, and the resulting oversupply exacerbated the decline in house prices. However, the interest rates declined rapidly during the crisis as well, which would ordinarily imply lower default rates in these ARMs. The authors explain that as home equity became negative, many borrowers rather chose to default strategically than to repay their mortgages; thereby explaining the higher default rates in spite of decreasing interest rates.

The notion of price elasticity can be incorporated into price-based risk as well. In particular, price elasticity of loan demand, or the so-called *price-response* elasticity, is typically expressed in economics literature as a differential that measures the change in the volume of a good/service demanded by consumers, given a change in the price – see Thomas (2009a, §3). In fact, the work of Oliver and Thaker (2013) helped develop a price-risk-response trifecta by defining so-called *price-response* and *price-risk* elasticities. While the former is familiar, the latter measures the change in the PD given changes in the loan price. Taken together, these two elasticities were incorporated into an equation that reveal how risk behaviours and response preferences are simultaneously exchanged as the underlying price is varied. Moreover, it was demonstrated in Phillips and Raffard (2011) and Phillips (2013) that a necessary condition for price-based risk to exist is for higher-risk customers to be less price-sensitive in their take-up decision than lower-risk customers. In summary, both price-driven adverse selection and price-based risk are intricate market-related phenomena that influence overall profitability beyond risk-based pricing. More importantly, both phenomena rely on a chosen threshold  $d$  within a default definition, which implies that their dynamics on profitability are largely unknown when varying  $d$  itself – a rich avenue of future pricing research. Regardless, the pricing literature explored so far suggests that evaluating different values of  $d$ , based on their relative contribution to financial loss, may be a viable approach.

The reasons underlying consumer loan ‘default’ are numerous, though can be crudely grouped into either fraud (“*won’t pay*”) or financial distress (“*can’t pay*”), which are explored in Thomas (2009a, pp. 282) and Bravo et al. (2015). The first group is rather self-explanatory, however, though the pursuit of modelling fraud events is valuable in its own right, it is an unnecessary complication that is best left outside the ambit of credit risk modelling. More interesting is the second group of default reasons, which can include simple financial naivety. As an example hereof, the borrower may not have understood (or appreciated) the financial discipline that is necessary to service the loan, and may therefore become overburdened in its repayment. Financial distress can include life altering-events that are outside of the borrower’s control, e.g., loss of employment or marital breakdown, that compromises the borrower’s ability to repay his loans, despite his best intentions or efforts. However, the exact reasons for delinquency are difficult to substantiate since lenders rarely keep record of them. In this regard, the administrative burden alone may prove too costly for a bank to track (and validate) all possible default reasons. Perhaps a more

tangible avenue is to consider whether this impairment in repayment ability is either persistent or temporary – and the costs of either case.

A sufficiently patient lender may afford some distressed borrowers enough time to recover, which will likely prompt them to resume their repayment schedule. In this case, the borrower's arrears will at first increase over time due to the distress, after which it will decrease again as the borrower regains his ability to repay (e.g., by finding new employment), as shown in Fig. 3.7 for the first loan. The account has recovered from what has clearly been a temporary episode of financial distress, which was enabled by *not* yet writing-off the loan. In effect, the lender trusts that the borrower's 'distress' only reaches an inflection point of sorts, after which arrears will hopefully subside again. Critically, the lender also trusts that this inflection point is below a certain threshold, in line with the bank's risk appetite. On the other hand, being patient for too long a period may prove naive and costly. The borrower may simply never reach this turning point and instead continue deeper into arrears, which may unnecessarily delay debt recovery and incur liquidity and other opportunity costs. Specifying a default threshold may therefore serve as a margin of tolerance towards accruing arrears before spurring a lender into taking recovery action and forsaking the credit relationship.

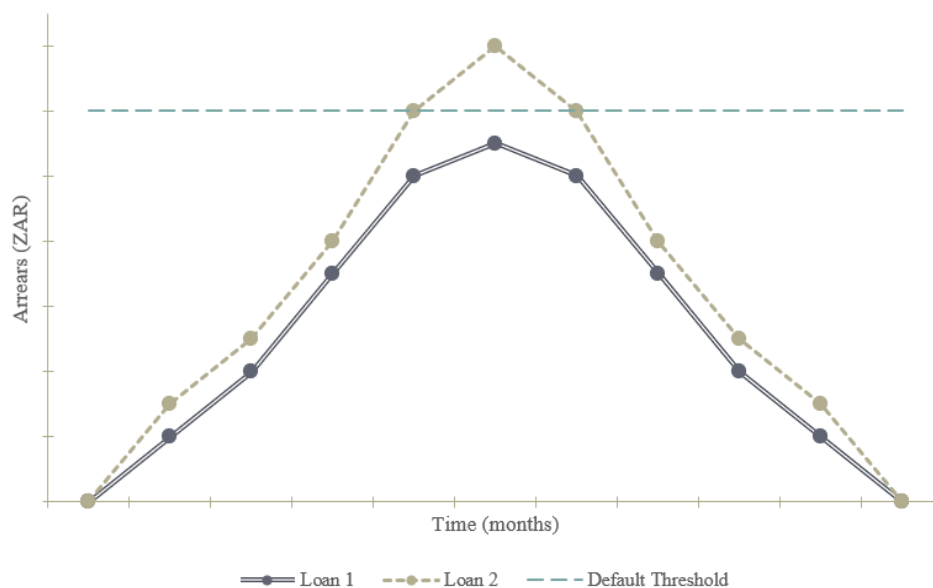


FIG. 3.7: Illustrating the stylised speeds at which two hypothetical delinquent accounts can accumulate arrears over time. Both loans reach an inflection point, after which arrears subside again due to a supposedly recovered ability of a borrower to repay. The arrears of the second loan breached the bank's margin of tolerance while that of the first loan did not. This discrepancy casts doubt on the default threshold as a supposed "point of no return".

As a contrived example, consider the second loan in Fig. 3.7 that breached this tolerance



level at some point, though cured again from default at another point in time. One can argue that the overall tolerance level was perhaps too strict, even punitive if the lender subsequently wrote-off this curable loan. More generally, the mere possibility of curing from 'default' (as fixed via  $d$ ) inherently injects uncertainty into any chosen  $d$  as the supposed "point of no return". Multi-period 'episodes' of delinquency are actually more widespread in practice than one would otherwise believe, corroborated in part by anecdotal experience. In fact, the work of Thomas et al. (2016) demonstrated this oscillating regime-switching effect wherein borrowers have stochastic episodes of payment and non-payment. Specifically, the authors modelled the collections process of defaulted UK loans using both a four-state homogeneous Markov chain and a time-sensitive hazard rate model. Amongst other things, these models demonstrated how LGD modelling can partially depend on the write-off policy of a bank, which typically constrains overall loan recovery. To this point, a write-off policy is similar in its objective to that of imposing a certain "point of no return" via setting  $d$ , which suggests another worthy research avenue to explore in time.

In fact, candidate write-off policies may be indirectly tested by incorporating them as boundary conditions for loan duration within their base models, according to Thomas et al. (2016). However, the authors admitted that this may not really be an appropriate solution for directly trying to find the 'best' write-off policy. Instead, the seminal contribution of Mitchner and Peterson (1957) investigated the optimal pursuit duration of overall loan recovery, which was based on maximising the net profit of a collections department. Using US personal loans, the authors found that pursuing loan recovery should cease (and the remaining balance written-off) whenever the one-period expected repayment equals the cost of pursuing loan recovery itself, which is quite intuitive. However, the authors admitted that their results depend critically on a few simplifying assumptions, which may not realistically hold. One of these assumptions is that a defaulted borrower is forever absorbed into a paying regime, once entered. Regardless, their work offers valuable insights into trying to optimise the write-off decision's timing.

Collections optimisation was explored more explicitly in De Almeida Filho et al. (2010) wherein a dynamic programming model was devised to find both the ideal type of recovery action and its pursuit duration for the "average" debtor over discrete monthly time intervals. Using unsecured European personal loans, the authors maximised the net recovery rate in their dynamic approach by pursuing a particular action for a set period. In order of severity, these actions include telephonic calls, various letters, house visits, legal redress, and write-off. However, the state space formulation excludes any repayment success (or cash flows) from previous periods or indeed those from the future. Their work was later extended in So et al. (2019) wherein a Bayesian approach was instead followed to obtain similarly optimised outputs, though on the individual debtor-level. Within the same context, the work of Liu et al. (2019) devised a Markov-based decision process wherein the optimal collection action is theoretically found across possible delinquency state progressions over time, using designed data. A schedule of optimal

collection actions is then calculated accordingly, which was shown to supersede a static policy based on maximising the expected net present value. However, the authors made some strong simplifying assumptions when designing both their 'data' and elements within their method, which may have implications for their end-results when trying to use their method on real data. In addition, they impose explicit write-off criteria exogenously within the state space of their Markov chain, instead of structuring it as a candidate collection action.

Another extension is the work of Duman et al. (2017) wherein outputs similar to that of De Almeida Filho et al. (2010) were sought using Turkish data, though on a finer daily time scale. The authors posited that each recovery action may have unintended consequences on subsequent customer loyalty, which they incorporated into their optimisation along with collection capacity constraints. Chehrazi et al. (2019) examined the same collection problem and formulated repayments as a complex self-exciting point process in continuous time. As a stochastic control problem, the authors allowed both the size and the timings of receipts to influence each other. Both of these random processes are in turn perturbed by pursuing a particular recovery action. Matuszyk et al. (2010) provided another perspective on optimising collections by developing a decision tree-based approach to inform the overall collection strategy. Candidate strategic choices included keeping collections in-house, outsourcing collections to external agencies, and selling the debt itself. Their work formed part of subsequently modelling the LGD using a two-stage approach and unsecured personal loans. Lastly, Han and Jang (2013) extended this by showing the positive effects of including collection action history within LGD models, using Korean data.

Using a delinquency-based approach for optimising the recovery decision presupposes that a single measure thereof is used. In fact, Rosenberg and Christen (1999) surveyed a few other ways of quantifying loan delinquency, which can serve as important diagnostic tools in managing and modelling credit risk. Three broad classes of so-called 'delinquency ratios' were discussed and compared. The first of these are those assembled from an arrears-basis as overdue amounts divided by total loan amounts (or instalments), e.g., the  $g_0$ -measure from Def. 3.2. The second kind are those that compare amounts actually received from borrowers against amounts that fell due within the same period. These ratios measure delinquency from a cash-flow perspective<sup>10</sup> with variants thereof constructed across different lengths of time periods. The third class of delinquency ratios measures the portion of the portfolio at risk, which is commonly expressed as the remaining loan balances (plus unpaid instalments) divided by the portfolio's outstanding balance. Other than these three types of delinquency ratios, the literature appears limited on more advanced methods of quantifying loan delinquency.

In this regard, the work of Moffatt (2005) is of particular interest since it advocates a two-staged approach (or a so-called "double hurdle" model) for 'measuring' loan delinquency and then

---

<sup>10</sup>From this basis, a more robust delinquency measure is later developed in subsection 3.3.1.

modelling it. Not only was the number of payments in arrears considered (the "first hurdle") in classifying the conventional default event, but also the extent of the arrears amount itself (the "second hurdle"). Using two dimensions in defining 'default' inherently recognises that there is a subset of borrowers who will never let a cent go unpaid in any circumstance. More importantly, it recognises that not all unpaid payments are equal in their cash flow impact and that defining the "point of no return" can be more nuanced than currently espoused by the industry<sup>11</sup>. In fact, the work of Kelly and McCann (2016) demonstrated exactly this heterogeneity using mortgage defaults from the Irish market. A legal peculiarity during 2009–2013 made it extremely difficult for Irish lenders to liquidate troubled mortgages, which led to some borrowers venturing into a disproportionately deep level of arrears. A multinomial logit model was subsequently estimated using different severities of delinquency (measured in DPD) as the levels of the output variable, i.e., 0–89 (or 'early default', also the reference level), 90–359, and 360+ (or 'deep default'). Even when controlling for the time spent in arrears, the overall delinquency and curing experiences were markedly different amongst these severities, which can be interpreted as different default thresholds within a grander default definition. Amongst other things, these results cast doubt on the supposed finality of classical default definitions that ought to serve as the "points of no return".

In conclusion, there are some gaps between the collections and credit risk modelling literatures regarding the principal meaning of 'default' as the supposed point of initiating recovery action. The present study is closest in form to the collection optimisation works of De Almeida Filho et al. (2010) and Liu et al. (2019). However, a different and more general approach is followed in this work based on delinquency measures that leverage the entire portfolio instead of only defaulted loans. Moreover, the focus is more fundamentally on if and when to forsake a loan based on delinquency progression over time, instead of attempting to compile a menu of collection actions. Regarding an optimisation basis, financial loss can serve as a unifying framework into which a wide array of competing costs and opportunities can be incorporated and weighted accordingly. This basis aligns directly with the main profit-based objective of a lender when making decisions. Furthermore, exploring other ways of quantifying loan delinquency may be worthwhile when delinquency is to be used as a central criterion in optimising the timing of the recovery decision.

### 3.3 Measures of loan delinquency

In measuring the severity of eroded trust between bank and borrower for a fixed-term amortising loan contract, three quantities that track the severity of delinquency are presented and discussed, called delinquency measures. Firstly, a variant of the widely-used  $g_0$ -measure from Def. 3.2 is

<sup>11</sup>Arguably, this idea may have inspired Basel II's paragraph 452 on requiring a *material* balance for default considerations, which was later enforced by Article 178(1)(b) of the EU's CRR. In turn, this led to the materiality tests imposed by the UK's PRA (2019), as discussed in subsection 3.1.1.

refined into a more robust measure in subsection 3.3.1, called the  $g_1$ -measure (or *CD*-measure). The  $g_1$ -measure uses a weighting scheme based on specifiable risk-aversion towards accrued arrears, which solves many of the challenges of the original  $g_0$ -measure. Secondly, a more concise algorithm is contributed in subsection 3.3.2 that creates the Macaulay Duration Index from Sah (2015), called the  $g_2$ -measure (or *MD*-measure). This measure is best interpreted as an index of the weighted average time to recover the capital portion of a loan. Lastly, a modified but novel variant of  $g_2$  is introduced in subsection 3.3.3, called the  $g_3$ -measure (or *DoD*-measure). This measure incorporates the extent of disrupted cash flows into the base assessment of delinquency.

### 3.3.1 Contractual Delinquency (*CD*): the $g_1$ -measure

As a common delinquency measure, the  $g_0$ -measure from Def. 3.2 is not without its flaws, predominantly the choice of the rounding function. Many practitioners simply round the ratio between arrears  $A_t$  and the level instalment  $I$  upwards to the nearest integer-valued ceiling, thereby calculating the number of payments in arrears. However, this is quite stringent in that even a small difference  $I_t - R_t = \epsilon < \text{ZAR } 1.00$  will increase the delinquency measurement, purely due to rounding. While the prevalence of this rounding error surely depends on the overall volatility of  $R_t$  over time, it makes little business sense to penalise a borrower when  $\epsilon$  is but a few cents. That said, there must intuitively exist some upper limit on  $\epsilon > 0$  at which point delinquency should be incremented; otherwise, the concept of delinquency (and measures thereof) becomes meaningless if delinquency can never increase. This suggests a boundary of sorts by which  $\epsilon$  ought to be constrained when measuring loan delinquency. Relatedly, should the ratio between  $A_t$  and  $I$  instead be rounded to the nearest integer, then a change in  $g_0(t)$  over time  $[t_1, t_2]$  depends entirely on whether  $A_t/I$  is above or below 50%. This implied ‘threshold’ of 50% seems arbitrary, inflexible, and devoid of any risk consideration, which surely contrasts the risk-sensitive practices of a bank, at least in spirit.

A simple solution to this rounding problem may be to ignore rounding all together and use the arrears ratio  $A_t/I$  directly. This will, however, revert the domain of  $g_0$  back to a real-valued construct and therefore no longer honour its definition. Furthermore, there are still quite a few situations in which arrears categories (constructed from this arrears ratio) can be useful. For example: 1) modelling the transition probabilities amongst these arrears categories; 2) compiling portfolio analytics showing the incidence rates over time within various stages of arrears; or 3) modelling the repayment behaviour within a particular arrears category. Inevitably, these situations require the collapse of the arrears ratio back into some category, e.g.,  $2 \leq A_t/I < 3 \implies 2$  payments in arrears, which again circles back to the original rounding problem. As a consequence of a particular rounding function  $f$  (or rounding ‘threshold’), the  $g_0$ -measure can potentially lag overall measurement by one (or more) periods when a significant overpayment is immediately followed by a severe underpayment the following month, simply due to rounding. In turn,

this "measurement error" has negative implications for the true accuracy of any models using delinquency in some fashion, especially default risk models.

Another flaw of the  $g_0$ -measure is its inherent reliance on the accrued arrears amount  $A_t$ . This dependence is not necessarily a problem when  $A_t$  is simply the sum of partial instalments not yet paid. However,  $A_t$  becomes 'impure' when the same quantity accrues interest on itself or attracts delinquency-related penalty fees of sorts (even if indirectly via the outstanding balance). Specifically, it becomes possible for a  $g_0$ -measured value to change simply due to these exogenous factors, especially when considering the aforementioned rounding problem. In turn, the lender may inadvertently inflate/deflate delinquency measurement; not as a result of the fundamental breakdown of trust, but instead due to its own pricing structures or system constraints at the time. Lastly, the construction of  $g_0$  quickly becomes cumbersome when the instalment has the potential to vary over time (e.g., prime rate-linked), as is common for secured lending. Short of ignoring this feature (and simply using the latest instalment available), the practitioner has little choice but to construct a haphazard and nested variant of  $g_0$ , which may very well present its own set of unstudied risks and flaws as a delinquency measure.

Therefore, a more comprehensive variant, called the Contractual Delinquency ( $CD$ )  $g_1$ -measure, is presented here that circumvents some of these challenges. In particular,  $g_1$  incorporates a boundary parameter  $z$  between the receipt and the instalment beneath which delinquency is increased, which may be set (and later optimised) by the lender. Furthermore, it does not rely on the corruptible arrears amount and instead measures the erosion of trust between bank and borrower purely from a cash flow perspective. Let the receipt vector be  $\mathbf{R} = [R_0, R_1, \dots, R_T]$  with its elements (or receipt amounts)  $R_t \geq 0$ , and let the instalment vector be  $\mathbf{I} = [I_0, I_1, \dots, I_T]$  with its elements  $I_t > 0$  (instalment amounts). Both vectors are defined for a specific loan account across its discrete time periods  $t = 0, \dots, T$ , with  $t = 0$  representing the origination point and  $T$  denoting the tenure (or current loan age). Note that  $T$  may exceed the contractual term  $t_c$ , especially in unsecured lending or cases of extreme delinquency. The repayment ratio  $h_t \in [0, \infty)$  is then defined as

$$h_t = \frac{R_t}{I_t} \quad \forall t = 1, \dots, T \quad \text{and} \quad h_0 = 0. \quad (3.4)$$

A boundary  $z \in [0, 1]$  for  $h_t$  is then specified accordingly. If the  $h_t$ -value of an account equal or exceed this given  $z$  parameter, then the account is considered current at time  $t$ ; otherwise it is considered delinquent for that particular period's expected cash flow. As an illustration, this boundary is set as  $z = 90\%$  in this study, though the lender should certainly adjust (or optimise) this  $z$  parameter for its particular context. Next, an interim Boolean-valued decision function  $d_1(t) \in \{0, 1\}$  is defined for  $t = 1, \dots, T$ , using Iverson brackets  $[a]$  that outputs 1 if the enclosed statement  $a$  is true, and 0 if false, as

$$d_1(t) = [h_t < z]. \quad (3.5)$$

Memory of past delinquency is introduced by defining another integer-valued function  $m(t) \in \{-1, 0, 1, \dots\}$  for  $t = 1, \dots, T$ , which outputs the reduction in accrued delinquency (if any), as

$$\begin{aligned} m(t) &= \left( \left\lfloor \frac{h_t}{z} \right\rfloor - 1 \right) (1 - d_1(t)) - d_1(t) \\ &= \left\lfloor \frac{h_t}{z} \right\rfloor (1 - d_1(t)) - 1. \end{aligned} \quad (3.6)$$

This function  $m(t)$  gives the magnitude by which the measured delinquency at time  $t$  should be reduced (if at all) in catering for past delinquency. When overpaying, i.e.,  $R_t > I_t$ , the ratio between  $h_t$  and  $z$  in Eq. 3.6 signifies the total number of ‘payments’ by which accrued delinquency should be decreased, as weighed by  $z$ . Moreover, the rounding problem is inherently resolved when dividing by the  $z$ -parameter since its value reflects the lender’s tolerance towards underpayment by design. Therefore, taking the floor of this particular ratio  $h_t/z$  does not detract from the previous discussion of the flaws of  $g_0$ . Instead, it is merely intended for  $g_1$  to be an integer-valued delinquency measure, signifying the  $z$ -weighted number of payments in arrears. Furthermore, the currently-owed instalment should be recognised first before reducing any accrued delinquency, which is achieved by subtracting one instalment. For sufficient underpayment, i.e.,  $R_t < zI_t$ , the delinquency is sensibly increased by one payment, which resolves to  $m(t) = -1$  when  $d_1(t) = 1$ .

To indicate previous cases of delinquency using  $g_1$  at time  $t - 1$ , let  $d_2(t) \in \{0, 1\}$  be another Boolean-valued decision function for  $t = 1, \dots, T$ , which is defined using Iverson brackets again, as

$$d_2(t) = [g_1(t - 1) = 0]. \quad (3.7)$$

The reduction in delinquency  $m(t)$  at time  $t$  is subtracted from delinquency as measured at the previous period  $t - 1$ , thereby giving the net delinquency. Finally, the *Contractual Delinquency* (or *CD*)  $g_1$ -measure is then formally defined as in Def. 3.3. Note the necessary starting condition of  $g_1(0) = 0$ , since a newly-disbursed loan account cannot yet be delinquent. In extraordinary cases where a lender has temporarily suspended the expected instalment (perhaps as part of a payment holiday), the practitioner can artificially set the  $I_t$  very close to zero in order to calculate  $g_1$ .

### Contractual Delinquency (CD): the $g_1$ -measure

**Definition 3.3.** Let the Boolean-valued decision functions  $d_1$  and  $d_2$  be defined as in Eq. 3.5 and Eq. 3.7 respectively. Let the memory function  $m$  be as defined in Eq. 3.6 by which accrued delinquency is reduced, respective to the  $z \in [0, 1]$  boundary parameter for the repayment ratio  $h_t$  as defined in Eq. 3.4. The integer-valued *Contractual Delinquency*  $g_1$ -measure is then defined such that  $g_1(t) \geq 0$  for  $t = 1, \dots, T$  and  $g_1(0) = 0$ , recursively expressed as

$$g_1(t) = \max \left[ 0, \quad d_1(t)d_2(t) + (1 - d_2(t))(g_1(t - 1) - m(t)) \right]. \quad (3.8)$$

The output of the  $g_1$ -measure is best interpreted as the  **$z$ -weighted number of payments in arrears**, weighed by the lender's tolerance (or appetite) towards accrued arrears. This appetite level is represented by the boundary parameter  $z$  that constrains the repayment ratio  $h_t$  accordingly when assessing delinquency. Since delinquency only increases if  $h_t < z$  by definition, a higher value of  $z$  effectively translates to greater risk-aversion towards accrued arrears, and *vice versa* for lower  $z$ -values.

### 3.3.2 Macaulay Duration ( $MD$ ): the $g_2$ -measure

The Macaulay Duration Index, recently introduced as a measure of delinquency in Sah (2015), is an index based on the idea of bond duration, i.e., the weighted average time to recover the capital portion of a loan. In its assessment of delinquency, this measure incorporates the loan's interest rate as well as the arrears balance, weighed by the time value of money. It is constructed as a ratio between the actual and the expected loan duration, which is reformulated in this study as the  $g_2$ -measure. The values of  $g_2$  cannot be compared directly to those of the previous  $g_1$ -measure since the domains of both functions – and the meaning of measurements taken within these domains – differ from each other.

Let  $\Delta_t = I_t - R_t$  be the difference between the instalment  $I_t$  and the receipt  $R_t$  at every time point  $t = 0, \dots, T$  during the life of a loan, including at disbursement  $t = 0$  to capture any applicable initiation fees. Considering the time value of money, let  $v_j = (1 + r)^{-j}$  be a discounting function that uses a nominal monthly interest rate  $r$ . In addition, let  $\delta$  be the continuously compounded rate with its nominal variant  $\delta^{(p)} = \delta/p$  and with an annual compounding period  $p = 12$ . Let  $L_P$  denote the loan amount (or principal) that is to be amortised. Ordinarily, the Macaulay Duration is calculated (perhaps once) at origination as the weighted average time to recover sunk capital from future cash flows. However, here it is recursively calculated instead at each subsequent period  $t = 0, \dots, T$  across the remaining  $m$  instalments as at each  $t$ . Naturally, this *expected duration* quantity, denoted as  $f_{ED}$ , tends towards zero over time as it nears the end of loan life, expressed as

$$f_{ED}(t) = \sum_{m=t}^T \left[ \left( \frac{I_m v_{m-t}}{L_P} \right) \left( \frac{m-t}{p} \right) \right] \quad \forall t = 0, \dots, T. \quad (3.9)$$

However, Eq. 3.9 assumes that instalments  $I$  are free of uncertainty. When substituting these instalments with the actual receipts  $R$ , a significant difference is intuitively expected. Moreover, it becomes necessary to track the arrears balance as it develops (if it does) over the loan life. In line with Sah (2015), any arrears at any time are added to the last expected (contractual) instalment at  $t = t_c$ , since it represents the last contractual opportunity to repay any such arrears, short of the lender intervening and restructuring the loan. This last instalment is then recursively updated at each subsequent period  $t$ , using the available arrears information at each  $t$ . Note

that  $R_t$  does not enter the calculation directly but instead via its effect on the arrears at each  $t$ . In turn, accrued arrears will affect the last instalment, which is denoted as  $I'_{(T)}$ . The vector  $I'$  itself equals instalments  $I$  at the outset. Lastly, the *actual duration*  $f_{AD}(t)$  value is recursively calculated, each time starting again at  $t = 0$  up to the 'current' loan age  $T$ . This is illustrated using pseudo-code in Algorithm 1. Finally, the *Macaulay Duration* (or *MD*)  $g_2$ -measure is expressed as the ratio between the actual duration and the expected duration, as formally defined in Def. 3.4.

---

**Algorithm 1** Calculating  $g_2$ 


---

```

1:  $I' := I$ , where  $I = [I_0, \dots, I_T]$  and  $T \leq t_c$ 
2:  $f_{AD}(0) := f_{ED}(0)$ 
3: for  $t = 0, \dots, T$  do ▷ such that  $T \leq t_c$ 
4:    $I'_{(T)} := I'_{(T)} + \Delta_t \left(1 + \frac{\delta^{(p)}}{p}\right)^{T-t}$ ,  $\forall t = 1, \dots, T$  ▷ Add any arrears to  $I'_{(T)}$ 
5:    $f_{AD}(t) := \sum_{m=t}^T \mathbb{1}_{T \leq t_c} \left[ \left( \frac{I'_m v_{(m-t)}}{L_P} \right) \left( \frac{m-t}{p} \right) \right]$ ,  $\forall t = 1, \dots, T$ 
6: end for
    
```

---

**Macaulay Duration (MD): the  $g_2$ -measure**

**Definition 3.4.** Let  $f_{ED}$  be defined as in Eq. 3.9, and let  $f_{AD}(t)$  be calculated recursively at every time period  $t = 0, \dots, T$  for a loan account aged  $T \leq t_c$ , following Algorithm 1. The real-valued *Macaulay Duration*  $g_2$ -measure is then defined such that  $g_2(t) \geq 0$  for  $t = 0, \dots, T - 1$ , which is expressed as

$$g_2(t) = \frac{f_{AD}(t)}{f_{ED}(t)}. \quad (3.10)$$

### 3.3.3 Degree of Delinquency (DoD): the $g_3$ -measure

From a cash flow perspective, an ideal delinquency measurement should penalise the non-payment of a larger loan's instalment to a greater degree than that of a smaller loan's instalment, given the relatively larger impact on a bank's cash flow. Furthermore, the differences in risk concentration between a larger number of small loans versus a small number of larger loans should also be incorporated by the ideal delinquency measure. As a possible solution, the actual duration  $f_{AD}$  from Eq. 3.10 can be altered such that the eventual  $g_2(t)$  value is greater for larger loans than for smaller loans by defining an appropriate multiplier (or function thereof).

Note that  $g_2$  is only defined up to the contractual term  $t_c$ . However, delinquency can continue even past the contractual term  $T \geq t_c$  of a loan, likely due to persisting underpayment. Ignoring loan write-off policies for the moment, let  $d_3(t) \in \{0, 1\}$  be a Boolean-valued decision function that returns 1 if the given time point  $t$  precedes the contractual term  $t_c$ , and 0 if otherwise. Using Iverson brackets, this is expressed as

$$d_3(t) = [t \leq t_c]. \quad (3.11)$$



When  $t > t_c$ , any arrears can clearly no longer be added to the last contractual instalment (since it has lapsed), as is the case with  $I'_{(T)}$  when calculating  $g_2$  in Algorithm 1. Instead, at least one more payment, albeit out-of-contract, can reasonably be expected as time lapses, provided that collection efforts are still actively pursued. Therefore, delinquency can now be computed up to time  $\mathcal{T}$  instead of the previous  $T$ , with  $\mathcal{T}$  either representing the contractual term  $t_c$  when  $t < t_c$ , or a moving target  $\mathcal{T} = t$  when  $t \geq t_c$ . Note that both  $\mathbf{I}$  and  $\mathbf{R}$  will incrementally expand with additional elements for as long as collection efforts<sup>12</sup> continue past the contractual term. A revised algorithm is given in Algorithm 2.

---

**Algorithm 2** Calculating  $g_3$ 


---

```

1:  $\mathbf{I}' := \mathbf{I}$ , where  $\mathbf{I} = [I_0, \dots, I_T]$  and  $0 < t_c \leq T$ 
2:  $\mathcal{T} := t_c$ 
3: for  $t = 0, \dots, T$  do
4:    $\alpha := I'_{(\mathcal{T})}$  ▷ This refers to the element at the  $\mathcal{T}^{\text{th}}$  position of  $\mathbf{I}'$ 
5:    $\mathcal{T} := t_c d_3(t) + t(1 - d_3(t))$  ▷  $\mathcal{T}$  is either equal to  $t_c$  or to  $t \geq t_c$ 
6:    $I'_{(\mathcal{T})} := I'_{(\mathcal{T})} d_3(t) + \Delta_t \left(1 + \frac{\delta^{(p)}}{p}\right)^{\mathcal{T}-t} + \alpha(1 - d_3(t)) \left(1 + \frac{\delta^{(p)}}{p}\right)$ ,  $\forall t = 1, \dots, T$ 
7:    $\beta(m) := m - t + 1 - d_3(t)$ ,  $\forall t = 1, \dots, T$  ▷ Discounting period, see next 3 lines
8:    $f_{ED}(t) := \sum_{m=t}^{\mathcal{T}} \left[ \left( \frac{I_m v^{\beta(m)}}{L_P} \right) \left( \frac{\beta(m)}{p} \right) \right]$ ,  $\forall t = 0, \dots, T$ 
9:    $f_{AD}(t) := f_{ED}(t)$ , for  $t = 0$ 
10:   $f_{AD}(t) := \sum_{m=t}^{\mathcal{T}} \left[ \left( \frac{I'_m v^{\beta(m)}}{L_P} \right) \left( \frac{\beta(m)}{p} \right) \right]$ ,  $\forall t = 1, \dots, T$ 
11: end for
    
```

---

Afterwards, let  $\lambda(L_M, L_P, s)$  denote a multiplier function that inflates the value  $f_{AD}(t)$  at every relevant period  $t$ . Regarding this function's arguments, let  $L_M > 0$  denote the maximum loan size and let  $s \in [0, 1]$  be a real-valued sensitivity that represents the 'strength' at which to apply this inflationary effect. Let  $d_4(t) \in \{0, 1\}$  be another Boolean-valued decision function that returns 1 if there is currently any accrued delinquency at  $t$ , and 0 otherwise, defined using Iverson brackets as

$$d_4(t) = [f_{AD}(t) > f_{ED}(t)]. \quad (3.12)$$

As a simple example, this multiplier is defined as

$$\lambda(L_M, L_P, s) = s \left( 1 - \frac{L_M - L_P}{L_M} \right). \quad (3.13)$$

This particular multiplier function simply measures the percentage deviation between the maximum size allowed by a lender and the size of the particular loan's principal. This base value is then further adjusted by the multiplicative factor  $s$ . The inflated variant of  $f_{AD}$ , denoted as  $\tilde{f}_{AD}$ , is at last expressed as

$$\tilde{f}_{AD}(t) = f_{AD}(t) (d_4(t) \lambda(L_M, L_P, s) + 1). \quad (3.14)$$

---

<sup>12</sup>It is typically the role of the collections department of a bank to negotiate new repayment schedules with the customer (or his debt councillor) once a loan agreement lapses its contractual term or becomes sufficiently delinquent. This includes reduced instalments and interest rates.

By including  $d_4$  into  $\tilde{f}_{AD}$  in Eq. 3.14, accrued delinquency will not be inflated when overpaying at some period  $t$ . Once  $f_{AD}(t) \leq f_{ED}(t)$  as indicated by  $d_4(t) = 0$ , then either significant or persistent overpayment is implied. In this case, the benefit of a depressed actual duration would be lost if  $f_{AD}(t)$  is still inflated when assessing delinquency. Finally, the real-valued *Degree of Delinquency* (or *DoD*)  $g_3$ -measure is defined as in Def. 3.5.

#### Degree of Delinquency (*DoD*): the $g_3$ -measure

**Definition 3.5.** Let  $\tilde{f}_{AD}$  be defined as in Eq. 3.14, and let  $f_{ED}(t)$  and  $f_{AD}(t)$  be calculated recursively at every time period  $t = 0, \dots, T$  for a loan account aged  $T$ , following Algorithm 2. The real-valued *Degree of Delinquency*  $g_3$ -measure is then defined such that  $g_3(t) \geq 0$  for  $t = 0, \dots, T - 1$ , which is expressed as

$$g_3(t) = \frac{\tilde{f}_{AD}(t)}{f_{ED}(t)} = \left( \frac{g_2(t)}{f_{AD}(t)} \right) \tilde{f}_{AD}(t) = g_2(t) (d_4(t) \lambda(L_M, L_P, s) + 1). \quad (3.15)$$

The sensitivity  $s$ , which is fixed in this study at  $s = 100\%$  (though should ideally be optimised), represents a universal and intuitive lever at the lender's disposal. Its adjustment can align with the lender's (or loan portfolio's) particular risk appetite and tolerances. The  $g_3$ -measure still collapses back into  $g_2$  when setting  $s = 0$ , though at values  $s > 0$ , the  $g_3$ -measure purposefully resembles a more risk-averse form of  $g_2$ . Delinquency measurements are more varied than those of  $g_2$  due to the sensitivity to loan principals that is intrinsic to  $g_3$  by design.

In summary, three measures of loan delinquency are formulated in this section, each progressively catering for identifiable weaknesses, albeit at the cost of increasing complexity. The outputs of these measures are illustrated in section A.2 using a simple two-loan case study to aid interpretation. These measures can be applied to the performance history of each loan within a greater portfolio, thereby yielding vectors of delinquency measurements respective to each loan across its history, for each measure  $g \in \{g_1, g_2, g_3\}$ . To find the optimal point where trust between bank and borrower has historically collapsed, is conjectured to be a certain threshold on the domain of each measure  $g$ , after having combined all of these vectors of measured loan delinquencies. Given that these measures cater differently for under- and overpayments, it may very well be that each measure  $g$  signals recovery at different times for the same loan, whenever breaching the threshold (if at all) of each measure.

### 3.4 Optimising loan recovery times: the LROD-procedure

In this section, an expert system is developed, called the *Loss-based Recovery Optimisation across Delinquency* (LROD) procedure. This procedure helps to find the 'best' delinquency-based

threshold for a given delinquency measure  $g \in \{g_1, g_2, g_3\}$  at which the portfolio's recovery decision is loss-optimised, as illustrated in Fig. 3.8. The following two points are noted regarding the procedure's design. Firstly, using loss (instead of profit) as the optimisation basis is intuitively preferable since 'default' implies a continuous stream of non-payments, which can only induce losses in principle. Secondly, construing the threshold choice as a delinquency-based criterion is convenient since delinquency measurements are scale-invariant across the portfolio, while still containing all necessary behavioural information on the borrower.

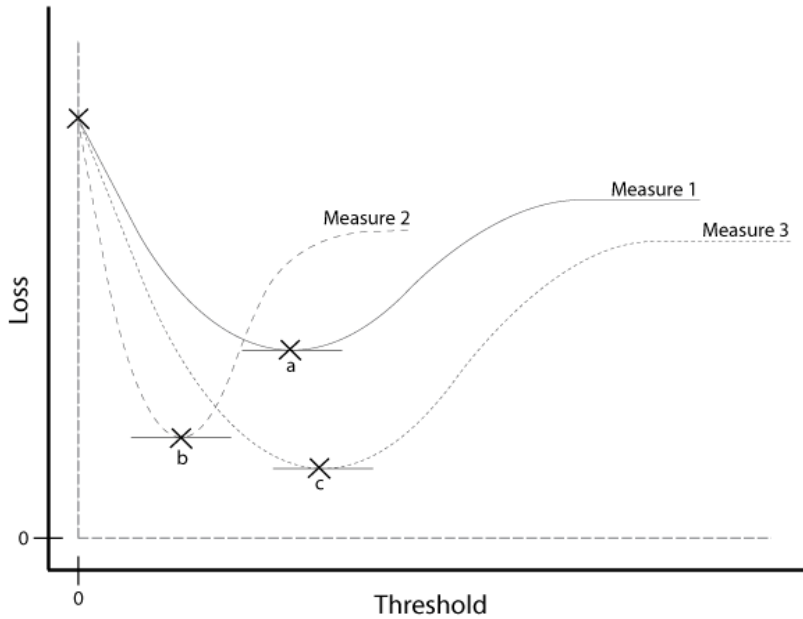


FIG. 3.8: Illustrating the loss optimisation of the recovery decision across several delinquency measures. As a result, Measure 3 is chosen as the best measure with its minimum loss attained at threshold  $c$ .

Consider a portfolio of  $N$  loans, indexed by  $i = 1, \dots, N$ , and let  $g(i, t)$  denote the value of a particular measure  $g \in \{g_1, g_2, g_3\}$  at periods  $t = 0, \dots, t_{c_i}$  with  $t_{c_i}$  representing the contractual term of the  $i^{\text{th}}$  account. Let  $v_t^{(a)}$  and  $v_t^{(b)}$  be standard discounting functions that use an alternative risk-free interest rate and the loan interest rate respectively in discounting back  $t$  periods, both expressed as annual effective rates and parametrised later. Let  $R_t^i$  and  $I_t^i$  represent the receipt and expected instalment respectively at time  $t$  for the  $i^{\text{th}}$  account. Then, let  $R(i, t)$  be the present value of the sum of discounted receipts from origination up to given time  $t = 0, \dots, t_{c_i}$  of the  $i^{\text{th}}$  account, expressed as

$$R(i, t) = \sum_{l=0}^t R_l^i v_l^{(a)}. \tag{3.16}$$

For the expected outstanding balance, let  $O(i, t)$  denote the present value of the sum of discounted

instalments from given time  $t = 0, \dots, t_{c_i}$  up to maturity  $t_{c_i}$  of the  $i^{\text{th}}$  account, defined as

$$O(i, t) = v_t^{(a)} \sum_{l=t+1}^{t_{c_i}} I_l^i v_{l-t}^{(b)}, \quad O(i, t) = 0 \quad \text{for } t = t_{c_i}. \quad (3.17)$$

Finally, to cater for instalments (or portions thereof) in arrears, let  $A(i, t)$  represent the present value of the sum of discounted differences (or shortfalls) between instalments and receipts from origination up to given time  $t = 1, \dots, t_{c_i}$  of the  $i^{\text{th}}$  account, which is given by

$$A(i, t) = \sum_{l=0}^t \left( I_l^i - R_l^i \right) v_l^{(a)}. \quad (3.18)$$

Financial loss can only be realised when the lender disposes of the impaired asset, regardless of the extent of the impairment. This implies that a loan must first become non-performing before sensibly writing-off the non-recoverable part as a loss. For now, all provisioning-related technicalities and regulatory requirements are discarded when defining default; see Novotny-Farkas (2016), Xu (2016), Cohen, Edwards Jr et al. (2017), and Skoglund (2017) for a comparison between IAS 39 and IFRS 9 accounting frameworks regarding loss provisioning. In this study, 'default' is interpreted as a variable state, which will become useful for optimising the eventual recovery decision. Having breached some threshold based on some  $g$  (thereby signifying broken trust), the lender's objective immediately changes to collecting the most it can from the distressed borrower within the shortest amount of time possible. As a simplifying assumption, a fixed portion of the loan is immediately written-off upon entering 'default'. In reality, this portion will likely depend on many factors, including the workout period itself. This assumption can certainly be relaxed in future research when refining this optimisation procedure and what is essentially its LGD-component.

Accordingly, let  $r_E \in [0, 1]$  be a loss rate levied on the expected balance  $O(i, t)$  to help reflect any underlying opportunity costs of forsaking future revenue. Moreover, assume that any amount in arrears  $A(i, t)$  is partly written-off though at a different loss rate  $r_A \in [0, 1]$  to account for impairment. Using two different loss rates recognises that the recovery success may differ between these two components (expected balance and arrears). Simultaneously, it unifies the loss calculation on the arrears portion for both performing and non-performing loans. This setup accounts for implicit trade-offs between forsaking future revenue versus accruing arrears for a given  $t$  respective to Eq. 3.17 and Eq. 3.18. Ideally, these loss rates serve as placeholders for the output from more sophisticated loss models (or expert knowledge of the loss experience), presumably including all other costs. Furthermore, let  $l(i, t)$  be the discounted "blended loss" for the  $i^{\text{th}}$  account assessed at the given time  $t = 0, \dots, t_{c_i}$ , expressed as

$$l(i, t) = O(i, t)r_E + A(i, t)r_A. \quad (3.19)$$

The notation  $(g, d)$  is adopted wherein a particular threshold  $d$  is coupled with the domain of a particular delinquency measure  $g$ , as formalised in Def. 3.6.

**The notion of  $(g, d)$ -defaulting accounts**

**Definition 3.6.** Let  $g(i, t)$  denote the value of a delinquency measure  $g \in \{g_1, g_2, g_3\}$  applied on a particular loan  $i = 1, \dots, N$  at every loan period  $t = 0, \dots, t_{c_i}$  where  $t_{c_i}$  is its particular contractual term. Then, let  $d \geq 0$  be a delinquency threshold applicable to the domain of a particular delinquency measure  $g$  such that the  $i^{\text{th}}$  account is considered as  $(g, d)$ -defaulting if and only if  $g(i, t) \geq d$  at any particular time  $t = 0, \dots, t_{c_i}$  during its lifetime. Accordingly, let  $S_D$  be the subset of all loan accounts within the portfolio that may be considered as  $(g, d)$ -defaulting at a particular point in time such that

$$S_D = \{i \mid \exists t \in [0, t_{c_i}] : g(i, t) \geq d\} . \quad (3.20)$$

Since an account may enter and leave the  $(g, d)$ -default state multiple times in reality, let  $t_i^{(g, d)}$  be the earliest moment of ‘default’ for the  $i^{\text{th}}$   $(g, d)$ -defaulting account, defined as

$$t_i^{(g, d)} = \min(t : g(i, t) \geq d), \quad \forall i \in S_D . \quad (3.21)$$

Similarly, let  $S_P$  be the subset of all loan accounts within the portfolio that may be considered as  $(g, d)$ -performing such that

$$S_P = \{i : g(i, t) < d \quad \forall t \in [0, t_{c_i}]\} . \quad (3.22)$$

The main difference in assessing the loss between a  $(g, d)$ -defaulting and a  $(g, d)$ -performing account is simply the time of assessment, which is set at either  $t = t_i^{(g, d)}$  or  $t = t_{c_i}$  respectively within  $l(i, t)$  from Eq. 3.19. At each time  $t$ , the lender effectively decides an account’s membership between  $S_D$  or  $S_P$ , based on accrued delinquency  $g(i, t)$  and a particular  $(g, d)$ -configuration. The latter is to be adopted as a portfolio-wide delinquency-based collection policy at the outset  $t = 0$ . In a sense, accrued delinquency forms the time-invariant *action* space of a Markov Decision Process (MDP) in choosing  $d$ , whereas accrued delinquency formed the *state* space in Liu et al. (2019). Accordingly, the state space in this work is set membership itself, i.e., either  $S_P$  or  $S_D$ . A classical MDP framework is not employed, instead opting for a simpler approach that facilitates choosing a static  $(g, d)$ -policy. Both ‘payoff’ and the element of time is already accounted for in Eq. 3.19 by having discounted the associated loss to  $t = 0$  for a given  $(g, d)$ -policy. As such, the objective function is simply the total portfolio loss  $L_g(d)$  for a particular  $(g, d)$ -configuration, as formalised in Def. 3.7 and used throughout this study.

**A simple portfolio loss model for a given  $(g, d)$ -configuration**

**Definition 3.7.** Given a  $(g, d)$ -configuration as defined in Def. 3.6 and Eq. 3.22, and given a loss function  $l(i, t)$  as defined in Eq. 3.19, let the argument  $t$  denote the time of loss assessment either at the earliest moment of  $(g, d)$ -default  $t_i^{(g, d)}$  for the  $i^{\text{th}}$   $(g, d)$ -defaulting account or at the contractual maturity  $t_{c_i}$  for the  $i^{\text{th}}$   $(g, d)$ -performing account. The discounted portfolio loss for a  $(g, d)$ -configuration is then given by the value of the objective function  $L_g(d)$ , expressed as

$$L_g(d) = \sum_{i \in S_D} l\left(i, t_i^{(g, d)}\right) + \sum_{i \in S_P} l\left(i, t_{c_i}\right). \quad (3.23)$$

Losses are iteratively calculated across a range of thresholds  $d \in \mathcal{D}_g$  using  $L_g$  from Eq. 3.23 in Def. 3.7 with a particular measure  $g \in \{g_1, g_2, g_3\}$ . To summarise then, the practitioner should complete three preparatory steps before conducting loss optimisation:

1. Delinquency must be measured for every account and across its history using  $g \in \{g_1, g_2, g_3\}$ ;
2. Select appropriate thresholds  $d \in \mathcal{D}_g$  on the domain of a particular  $g$  for optimisation;
3. A portfolio loss model  $L_g$  must be applied for every chosen threshold  $d \in \mathcal{D}_g$  of each  $g$ .

The procedure now becomes a search for a minimum in  $L_g$  and output the associated threshold  $d$  from the search space  $\mathcal{D}_g$  for a given measure  $g$ . By dividing the main optimisation problem into smaller  $(g, d)$ -based sub-problems, the ideal  $(g, d)$ -policy can be found. More specifically, for each  $(g, d)$ -iteration, the associated portfolio loss  $L_g(d)$  is calculated and stored centrally as an element within a wider collection. As formalised in Def. 3.8, there may exist a global minimum loss  $m^{(g)}$  at threshold  $d^{(g)}$  respective to each  $g$ . Delinquency measures themselves become indirectly comparable on the portfolio-level by minimising across the set formed by  $m^{(g)}$  for  $g \in \{g_1, g_2, g_3\}$ . The optimal measure  $g^*$  is then the  $g$  that yielded the lowest loss at its corresponding threshold, as illustrated in Fig. 3.8, though expressed more formally as

$$g^* = \arg_g \min_{g \in \{g_1, g_2, g_3\}} \left[ m^{(g_1)}, m^{(g_2)}, m^{(g_3)} \right]. \quad (3.24)$$

Note that this procedure can also be used with a single measure, e.g.,  $g_1$ . In this case, the optimisation problem from Eq. 3.26 simply resolves to finding  $d^{(g_1)} \in \mathcal{D}_{g_1}$ , that is, finding the optimal input  $\arg_d \min L_{g_1}(d)$  that minimises  $L_{g_1}$ .

**Loss-optimising the recovery decision: the LROD-procedure**

**Definition 3.8.** Assume delinquency thresholds  $d \in \mathcal{D}_g$  respective to a delinquency measure  $g \in \{g_1, g_2, g_3\}$  are adequately chosen, and that portfolio losses are calculated at each chosen threshold  $d$  using a loss model  $L_g$ , e.g., as in Def. 3.7. The function  $L_g : \mathcal{D}_g \rightarrow \mathbb{R}$  is adopted as the objective function with its smaller search space  $\mathcal{D}_g$  within the domain of  $g$ . Using an iterative approach, the optimisation problem then seeks a threshold  $d' \in \mathcal{D}_g$  such that  $L_g(d') \leq L_g(d)$  for all chosen  $d \in \mathcal{D}_g$ . More generally, the minimum loss  $m^{(g)}$  and optimal input argument  $d^{(g)}$  for a particular measure  $g$  are respectively expressed as

$$m^{(g)} = \min_{d \in \mathcal{D}_g} L_g(d) \quad \text{and} \quad (3.25)$$

$$d^{(g)} = \arg_d \min_{d \in \mathcal{D}_g} L_g(d). \quad (3.26)$$

The optimisation's feasibility relies on having an adequately populated search space  $\mathcal{D}_g$ . However, the choice of  $g$  affects  $\mathcal{D}_g$ , which can complicate choosing thresholds  $d$  in practice. Imagine that an underlying real-valued loss curve  $\mathcal{L}_g$  exists but cannot be specified in closed-form. Its functional shape only becomes apparent by iteratively calculating  $L_g(d)$  across a sufficiently wide range  $d \geq 0$ . This is trivial for the integer-valued  $g_1$ -measure since any finite integer interval  $[d_1, d_2] \in \mathcal{D}_g$  is countable. However, this is not the case for  $g_2$  and  $g_3$ , with an infinite number of possible real-valued thresholds that may be chosen between the now real-valued points  $d_1$  and  $d_2$ . In this case, one may calculate the derivative of  $L_g$  with respect to  $d$  at certain points  $d^*$  within  $[d_1, d_2]$ , e.g., using finite difference methods. The approximated derivatives  $L'_g(d^*)$  can then be inspected across the chosen points  $d^*$  for sign-changes, i.e., negative to positive, to help isolate neighbourhoods containing minima. Upon finding such a neighbourhood, this scheme can be repeated across a smaller range of points located closer to where the sign changed.

However, the practical assembly (or approximation) of  $\mathcal{L}_g$  remains challenging since these thresholds are still selected manually, even if using numerical differentiation. There are two competing interests when populating  $\mathcal{D}_g$  with thresholds. Firstly, inadequate threshold choices may lead to failure in materialising the true shape of  $\mathcal{L}_g$ , which can obscure hidden optima and ruin the optimisation. Secondly, too large a set of chosen thresholds can become computationally burdensome in evaluating  $L_g$ , especially as loans increase in either contractual term or number. As a practical expedient,  $\mathcal{D}_g$  for  $g_1$  is simply populated in this study by choosing  $d = 0, \dots, d_N$  where  $d_N$  corresponds to a reasonable (but admittedly arbitrary) proportion of the maximum contractual term, e.g., 60%. In addition to lowering computation time, a lower value for  $d_N$  is sensible since it would be unintuitive to search for optimal thresholds that are near the contractual term in value, especially for longer term loan portfolios. To populate  $\mathcal{D}_g$  for the real-valued measures  $g_2$  and  $g_3$ , the output of these functions are simply binned using a combination

of equal-width discretisation, some numerical differentiation, and a healthy dose of discretion.

### 3.5 Concluding remarks

Having reviewed the international standards relating to default definitions, lenders are clearly afforded some discretion in defining 'default'. Individual regulators may, however, prescribe the definition thereof to varying degrees, though it seems that many of these prescriptions still coalesce around "*unlikeliness to repay*" as a central tenet. Regardless, prescribing an admittedly probabilistic idea by fiat may render the default event into little more than a standardised static hurdle. In principle, reaching 'default' is impetus for the lender to forsake the credit agreement in having reached a certain "point of no return". This implies that there must surely exist different consequences when varying the timing of the recovery decision, lest the idea of 'default' itself becomes meaningless.

The notion of loan delinquency is widely used in many other (less regulated) contexts outside of modelling capital and loss provisions, most notably that of application and collection scoring. It is perhaps this level of ubiquity that exacerbates the fact that little scientific evidence exist for decreeing any threshold (e.g., 90 DPD) as an optimised absolute, regardless of context. That said, using a roll rate analysis (instead of regulation) to inform a delinquency threshold  $d$  is common practice in some modelling contexts. However, such an analysis can lead to spurious results simply due to arbitrary settings in its design, e.g., the choice of outcome period and the particular sample window. Moreover, roll rate-based approaches are oblivious to any competing financial and opportunity costs when varying the threshold  $d$ , in addition to using stability instead of financial loss as a base criterion. All of these factors complicate finding the *ideal* threshold  $d$  when using a roll rate-based approach, which calls for devising a more appropriate and optimisation-friendly method. A chosen threshold  $d$  can serve as a time-sensitive margin of tolerance towards accruing arrears, beyond which the associated costs for keeping the loan would trump any benefit thereof. Financial loss can therefore be used as an optimisation basis for finding the *ideal* threshold  $d$  on a given measure  $g$ . Accordingly, this work has explored and refined a few ways of quantifying delinquency itself, thereby enumerating the choice of  $g$  with a few functional forms. These alternative measures can potentially enhance any subsequent optimisation given  $g$ -measurable delinquency.

Finally, the philosophy underlying the recovery decision is framed anew as a loss-based nonlinear optimisation problem. The LROD-procedure manifests this problem by trying to find an ideal threshold  $d$  (the main decision variable, apart from the choice of  $g$ ) such that loan recovery occurs neither too early nor too late, if at all. Too strict a threshold will surely marginalise some loan accounts that would otherwise have resumed payment, had the bank been more trusting. Yet too forgiving a threshold will naively tolerate increasing arrears amounts at the cost of greater



liquidity risk, increased capital buffers, and overall higher levels of credit risk. Ultimately, the "net cost" of each candidate threshold is assessed and collated into a *g*-specific collection, thereby forming a loss curve that may be inspected for optima. As such, debt ought to be recovered at this optimum (or 'default' point) whereupon the portfolio loss is minimised in aggregate. Pursuing loan collection any further beyond this point would be sub-optimal, which agrees more fundamentally with 'default' translating into a *variable* "point of no return" on the banker's imagined gauge of eroded trust.



*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## OPTIMISING LOAN RECOVERY TIMING: A COMPUTATIONAL STUDY

The prevailing philosophy when timing loan recovery is framed as a hypothetical delinquency-based optimisation problem, such that loans are forsaken neither too early nor too late, if at all. In this chapter, the LROD-procedure is tested from "first principles" using portfolios that are meaningfully generated across the entire range of credit risk scenarios. In particular, a few probabilistic techniques and their associated parameters are described in section 4.1 by which the cash flows of any portfolio can be generated, guided by expert judgement and industry experience. Accordingly, various parameter searches are conducted in section 4.2 to identify a set of ideal parametrisations for which recovery optimisation becomes feasible, i.e., yielding optima. Finally, this chapter is concluded in section 4.3, which includes a discussion on directions for future research. Overall, the results demonstrate that the timing of loan recovery depends greatly on the inherent risk level of a given loan portfolio and its composition. A research article titled "*Simulation-based optimisation of the timing of loan recovery across different portfolios*" is associated with this chapter, published in the journal *Expert Systems with Applications*; see Botha et al. (2021). The associated source code is available in Botha (2020a).

### 4.1 Portfolio generation: a testbed for the LROD-procedure

A real-world portfolio inherently suffers from censoring insofar that delinquent loans are only kept on the balance sheet up to a certain point, as controlled by the bank's write-off policies. Although eventually optimising the recovery decision of a real-world portfolio would be ideal, it is arguably prudent first to demonstrate the efficacy hereof from "first principles" on designed data. In this section, a broad but simple simulation-based setup is described, guided by expert

judgement and industry experience. Using this setup as a testbed, replicable loan portfolios of varying risk levels are iteratively generated in testing the LROD-procedure. This testbed is subsequently used to identify a certain range of credit risk profiles for which optima are found, simply by varying the simulation parameters.

Some delinquent accounts will simply never recover in reality, which implies a continuous stream of zeros in their receipts  $\mathbf{R} = [R_1, R_2, \dots, R_{t_c}]$  after some point. Given a measure  $g \in \{g_1, g_2, g_3\}$  and a so-called *truncation parameter*  $k \geq 0$ , this effect is simulated from a certain starting point  $t' = \min(j : g(j) \geq k)$  that only exists when delinquency has accrued sufficiently, i.e., the earliest period  $j \in [0, t_c]$  at which  $g(j) \geq k$  is potentially triggered. A process, called  $(k, g)$ -truncation, then changes  $\mathbf{R}$  to  $\mathbf{R}'$  by

$$\mathbf{R}' = \begin{cases} [R_1, R_2, \dots, R_{t'}, 0, \dots, 0] & \text{if } t' \text{ exists} \\ \mathbf{R} & \text{otherwise} \end{cases} . \quad (4.1)$$

Consider  $N = 10,000$  standard amortising loan accounts that are indexed by  $i = 1, \dots, N$ , with a fixed contractual term of  $t_c = 60$  months, a fixed effective annual interest rate of 20%, and a fixed principal amount such that the level instalment is  $I_t = 100$  at every period  $t = 1, \dots, t_c$ . Admittedly, these quantities are oversimplified and will typically vary in a real portfolio based on the level of credit risk and loan demand. However, sampling them instead from stylised<sup>1</sup> distributions did not have nearly the same effect as that of credit risk in the optimisation itself. These simplifications are therefore justified for the time being. Furthermore, an effective annual risk-free rate of 7% is used in discounting, which is realistic for the South African market. Let the maximum loan size be  $L_M = 5,000$  and let  $r_E = 40\%$  and  $r_A = 70\%$  with the rationale that losses on arrears ought to be penalised more than losses on expected balances. The latter is a decreasing quantity while the former increases over time for a persistently delinquent loan. All of these parameter values represent expert knowledge though can certainly be varied in practice, which will be demonstrated later for some of these parameters.

In simulating the receipt vector  $\mathbf{R}$  of each loan account, three probabilistic techniques are now described. As a basic technique (called *random defaults*), let  $u_t \in [0, 1]$  be a randomly generated number at every period  $t = 1, \dots, t_c$  and let  $b$  be the probability of payment, i.e.,  $\mathbb{P}(R_t = I) = b$  with  $I$  denoting the level instalment. Note that  $b = 80\%$  is chosen as a default value, though this is later varied. Each element  $R_t$  within  $\mathbf{R}$  is then populated with either  $I$  or 0, expressed as

$$R_t = \begin{cases} I & \text{if } u_t < b \\ 0 & \text{otherwise} \end{cases} . \quad (4.2)$$

<sup>1</sup>In particular, a beta distribution was first parametrised to resemble the typically right-skewed distributional shape of unsecured retail loan rates in the South African market, thereby reflecting expert knowledge and risk-based pricing practises. Secondly, the loan amount was also sampled from a similarly parametrised distribution, again based on the authors' experience in the industry.

Despite its simplicity, random defaults do not feasibly generate periods of consecutive non-payments followed by resumed payment, which frequently occurs in practice as "episodic delinquency". Therefore, the so-called *episodic defaults* technique is devised wherein  $p_D = 50\%$  represents the overall probability of default, i.e., half the portfolio is bound to have a default episode by design. Let  $l_j$  be the number of consecutive non-payments to be simulated for the  $j^{\text{th}}$  delinquent account within the defaulting-segment. This episode length  $l_j \in [1, k]$  is sampled from the uniform distribution up to  $k$ , coinciding with  $(k, g_1)$ -truncation. When applying  $(k, g_1)$ -truncation, accounts will only cure if they had less than  $k$  consecutive non-payments, as a limiting condition. Thereafter, the starting point  $o_j \in [1, t_c - l_j]$  of the episode is sampled from the uniform distribution up to  $t_c - l_j$ , which is to say the entire episode must fit within the remaining loan life. Finally, each element  $R_t$  within  $\mathbf{R}$  of the  $j^{\text{th}}$  delinquent account is then simulated as

$$R_t = \begin{cases} 0 & \text{if } o_j \leq t \leq (o_j + l_j) \\ I & \text{otherwise} \end{cases} . \quad (4.3)$$

Realistically, an account may experience multiple default episodes during its life, though the previous episodic technique only produces one such episode. Therefore, and similar to Thomas et al. (2016), the *Markovian defaults* technique is defined where  $X_t \in \{P, D, W\}$  denotes a random variable that can assume one of three states at each period  $t$ ; the paying state  $P : R_t = I$ , the delinquent state  $D : R_t = 0$ , and the absorbing write-off state  $W : R_{t \geq t'} = 0$  from a certain point  $t'$  onwards. Then, let  $X_1, X_2, \dots$  be a sequence of random variables that form a discrete-time first-order Markov chain. One can reasonably assume that every account starts off as non-delinquent, i.e.,  $\mathbb{P}(X_1 = P) = 1$  while  $\mathbb{P}(X_1 \in \{D, W\}) = 0$ . Subsequently, the one-period transition probability from the current state  $i$  at  $t$  to the future state  $j$  at  $t + 1$  is denoted as  $P_{ij}$ . However, let the write-off probabilities be sensibly set to 0.1% and 1% respective to the starting states P and D. These values agree with general industry experience of an unsecured portfolio, though can certainly be tweaked to the individual portfolio in practice. The remaining elements in the transition matrix can now be derived from but two probabilities,  $P_{PP}$  and  $P_{DD}$ . In turn, both of these can be systematically varied to generate a portfolio's cash flows according to a certain level (or profile) of credit risk. The transition matrix is accordingly expressed in Table 4.1.

		To		
		P	D	W
From	P	$P_{PP}$	$1 - P_{PP} - 0.1\%$	0.1%
	D	$1 - P_{DD} - 1\%$	$P_{DD}$	1%
	W	0%	0%	100%

TABLE 4.1: A conceptual transition matrix for the Markovian defaults technique, wherein the rates  $P_{PP}$  and  $P_{DD}$  are to be systematically varied.

## 4.2 Computational results of recovery optimisation

In this section, the LROD-procedure is demonstrated and tested across a wide array of credit risk scenarios, generated using the testbed described in section 4.1. The computational results are grouped below by technique, followed by suggestions given in subsection 4.2.4 for applying the LROD-procedure on real-world data.

### 4.2.1 Random defaults

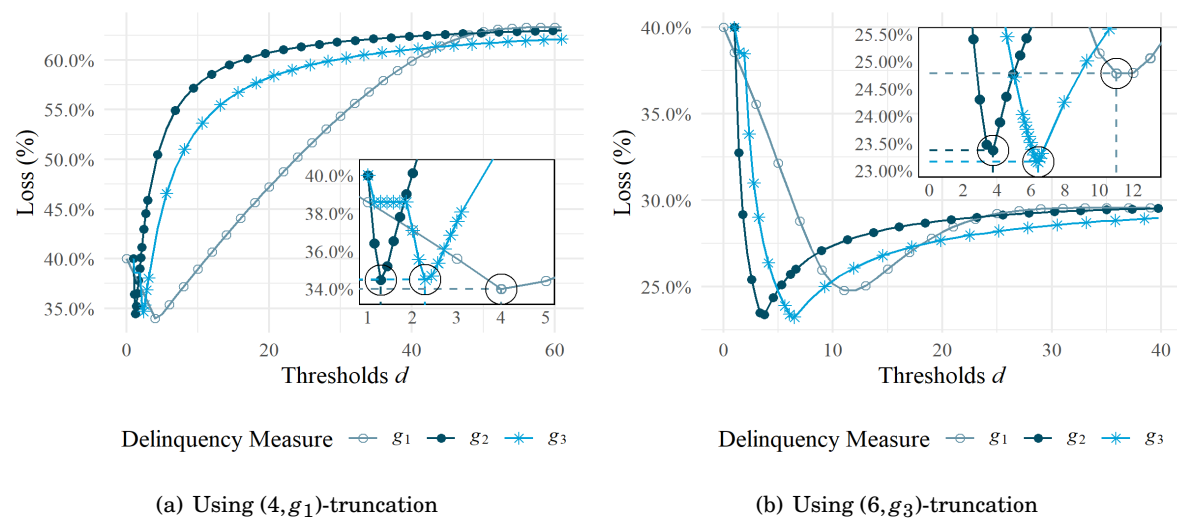


FIG. 4.1: Losses (as a proportion of summed principals) across thresholds  $d$  by measure  $g \in \{g_1, g_2, g_3\}$  using the random defaults technique. In (a), loans are  $(4, g_1)$ -truncated, while they are  $(6, g_3)$ -truncated in (b). In both cases, the zoomed plots show that global minima occur at or near the truncation point,  $d = k$ .

This technique leverages  $(k, g)$ -truncation to control the portfolio generation itself, thereby serving as a sanity check when testing the optimisation results and its underlying logic. Intuitively, the lowest loss should be at threshold  $d = k$ , since receipts are zeroed after having breached  $k$  by design. As an illustration,  $(4, g_1)$ -truncation is applied in Fig. 4.1(a), which shows the lowest loss occurs at  $d = 4$  for  $g_1$  as expected. However, the choice of  $g \in \{g_1, g_2, g_3\}$  when truncating introduces bias in the timing of cash flows, such that this  $g$  will likely contain the lowest loss as well. This is demonstrated in Fig. 4.1(b) when using  $(6, g_3)$ -truncation instead, where the minimum loss now occurs approximately at  $d = k = 6$  for  $g_3$ . Whilst seemingly artificial, truncation is merely used as an intuitive testing tool. However, the notion of truncation is plausibly similar to default contagion during a real-world economic downturn, during which borrowers may default systematically at some level of accrued delinquency  $k$  on average.

Minimum losses ought to occur wherever  $d = k$  in applying  $(k, g)$ -truncation on receipts. This intuition is largely confirmed in Fig. 4.2 wherein truncation parameters  $k = 1, \dots, 10$  are applied

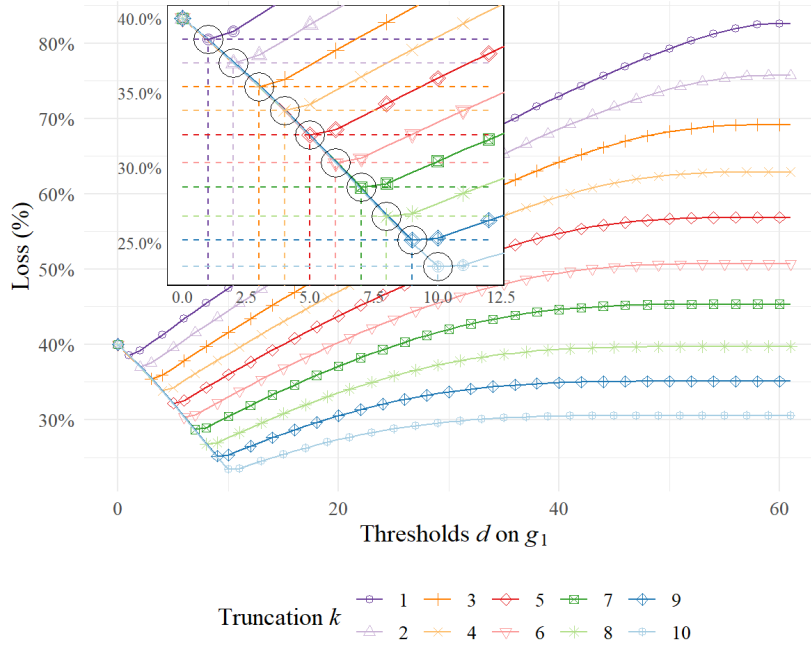


FIG. 4.2: Losses (as a proportion of summed principals) across thresholds  $d$  for the  $g_1$ -measure with  $(k, g_1)$ -truncation, using the random defaults technique. Several truncation points  $k = 1, \dots, 10$  are used, with the zoomed plot confirming that global minima in losses occur at each truncation point  $d = k$ .

during portfolio generation. As a result, loss minima occur consistently at the truncation point  $d = k$  as expected, while holding other factors constant. Each increasing value of  $k$  also yielded a smaller minimum loss as a result of the overall lessening truncation effect. Since receipts are truncated less frequently as  $k$  increases, generated portfolios exhibit overall less delinquency (or credit risk), which explains both lower loss curves and lower loss minima. Although not shown, this result holds similarly for  $g_2$  and  $g_3$  when used in truncation. Therefore, the optimisation is deemed sensitive to systematic defaults and can react accordingly should the defaulting behaviour of borrowers converge, as simulated by truncation.

Besides truncation, this technique has another parameter that is arguably more relevant: that of the one-period repayment probability  $b$ . Each value of  $b$  corresponds to a particular level of credit risk during portfolio generation. By varying  $b$ , the effect of credit risk can be broadly tested when optimising loan recovery, as shown in Fig. 4.3. Applying  $(6, g_1)$ -truncation as a benchmark, loss minima still occur at  $d = k = 6$  as expected, though only for a certain range of  $0.5 < b < 0.94$ . This suggests that optimising loan recovery in practice is infeasible for either very risky loan portfolios or near riskless portfolios. In particular, the two boundary cases of  $b = 0$  and  $b = 1$  in Fig. 4.3 support this idea in that loans should be forsaken at the outset when  $b = 0$ , as evidenced by the loss minimum at  $d = 0$ , since all receipts will be zero-valued by design.

Conversely, if there is no credit risk, i.e.,  $b = 1$ , then no loss is made at any  $d > 0$  and loan recovery itself becomes a moot point. These computational results can directly translate into practical value when estimating the parameter  $b$  from a real-world portfolio, as well as estimating the extent of any underlying truncation effect.

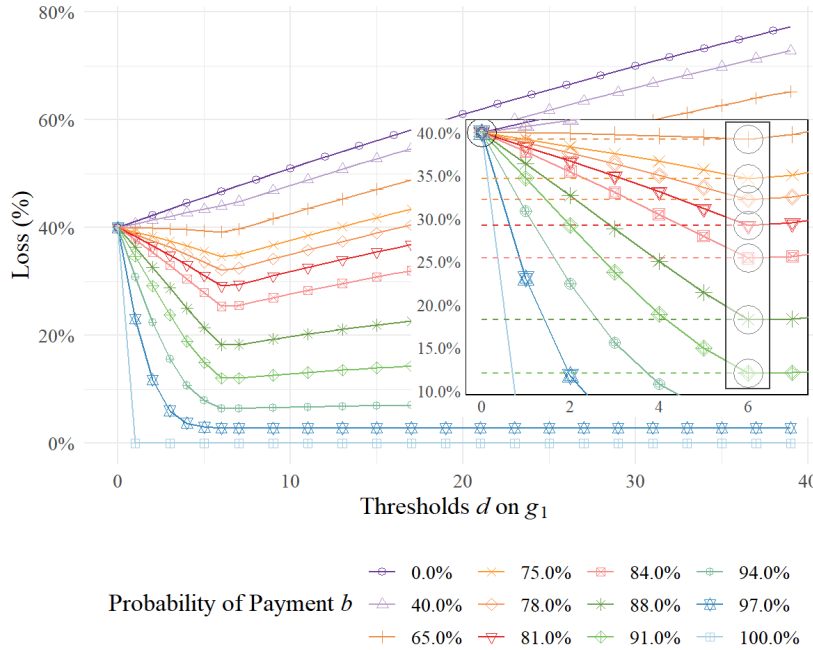


FIG. 4.3: Losses (as a proportion of summed principals) across thresholds  $d$  for the  $g_1$ -measure with  $(6, g_1)$ -truncation, using the random defaults technique and several probabilities of payment  $b \in [0, 1]$ . The zoomed plot shows a smaller range of  $0.65 \leq b \leq 0.91$  where loss minima occur at the chosen truncation point.

Intuitively, the loss experience (or LGD) of a particular portfolio ought to affect the results of recovery optimisation as well, especially when considering loan security in the event of default. This is testable by varying the loss rate  $r_A$  during portfolio generation while holding other factors constant, as illustrated in Fig. 4.4 using  $g_1$  (though similar results hold for  $g_2$  and  $g_3$ ). As a proxy for more secure portfolios, smaller values of  $r_A$  lead to flatter loss curves, until reaching a point where recovery optimisation becomes infeasible. Conversely, larger values of  $r_A$  yield loss curves with a greater ‘bend’ at the chosen truncation point, which signifies the greater risk involved with more unsecured portfolios. Since  $b$  is held constant, one can conclude that once default does occur, the viability of recovery optimisation only increases with the risk of loss, which is intuitively sensible. This is to say that unsecured portfolios will likely benefit even more from recovery optimisation than secured portfolios.



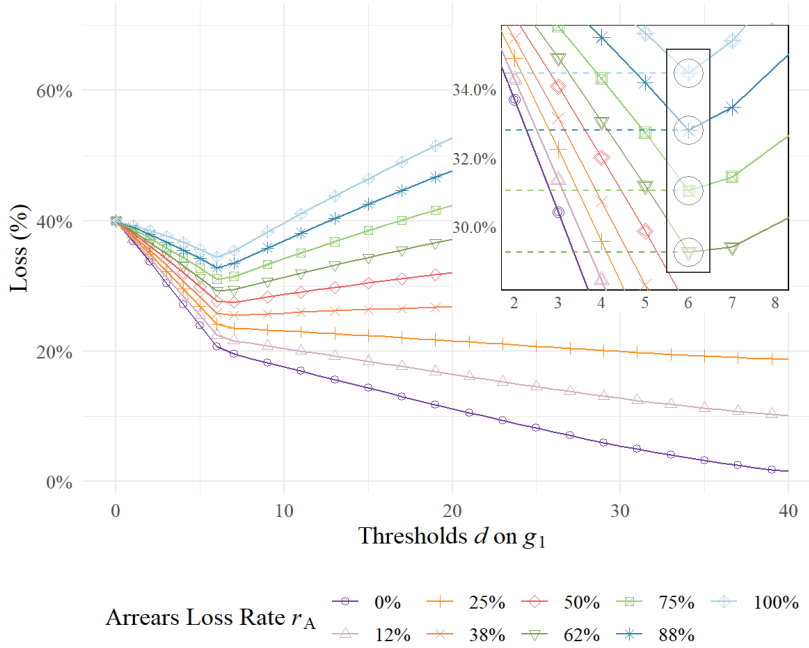


FIG. 4.4: Losses (as a proportion of summed principals) across thresholds  $d$  for the  $g_1$ -measure with  $(6, g_1)$ -truncation, using the random defaults technique and several arrears loss rates  $r_A \in [0, 1]$ . The zoomed plot shows a smaller range of loss rates  $0.62 \leq r_A \leq 1$  where loss minima occur at the chosen truncation point.

### 4.2.2 Episodic defaults

Since this technique is tightly coupled with  $(k, g_1)$ -truncation by design, portfolios are generated accordingly for  $k = 1, \dots, 10$ , as shown in Fig. 4.5 for  $g_1$  (with similar results for  $g_2$  and  $g_3$ ). Clearly, the shapes of loss curves are different, even though loss minima are still found at each successive truncation point. Furthermore, accounts only resume payment provided that the length of the default episode is less than  $k$ . Longer episodes (higher  $k$ ) seem to absorb the loss that is specifically introduced by truncation itself, which is signified by flattening loss curves for  $d \geq k$ . Since higher  $k$  implies that truncation will occur less frequently, a greater proportion of accounts with an episode length less than  $k$  will resume payment. In turn, arrears stabilise given less frequent truncation, which explains the flattening slopes of loss curves for greater  $k$ .

In general then, small  $k$  implies shorter episode lengths but more truncated accounts, while large  $k$  means longer episode lengths but fewer truncated accounts. It seems this technique generates portfolios with an interesting trade-off between default episode length and truncation frequency, regarding their credit risk compositions. Since loss minima are still obtained at each truncation point as hoped, it suggests that recovery optimisation will likely remain viable in practice when facing portfolios with these more interesting characteristics. This includes

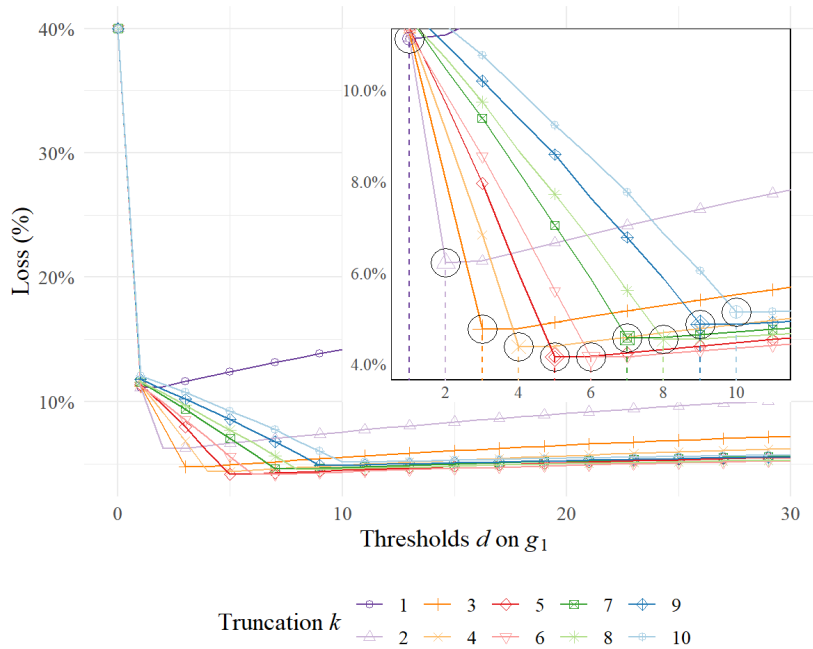


FIG. 4.5: Losses (as a proportion of summed principals) across thresholds  $d$  for the  $CD$ -measure  $g_1$  with  $(k, g_1)$ -truncation, using the episodic defaults technique with  $p_D = 50\%$  and several truncation points  $k = 1, \dots, 10$ . The zoomed plot shows that loss minima occur at each successive truncation point.

portfolios wherein truly delinquent accounts (as proxy for truncation) occur more frequently (e.g., unsecured lending), but are also more prone to recover from shorter bouts of delinquency, and *vice versa*.

### 4.2.3 Markovian defaults

This technique affords greater flexibility in generating portfolios with more sporadic repayment histories. Accordingly, the LROD-procedure is demonstrated in Fig. 4.6 using some of the parametrisations of the underlying Markov chain that yield optima across all delinquency measures. Evidently, the  $g_1$ -measure appears to outperform the other measures since it yields the lowest loss within each of these settings, including a number of other parametrisations not shown. However, summarily concluding the supremacy of  $g_1$  across *all* portfolios would be disingenuous. It is still possible that some real-world portfolios may be better served using measures other than  $g_1$  within the LROD-procedure (or more broadly in risk management). The current objective is not to determine the best measure conclusively. Indeed, conducting such an empirical study would require expansive real-world data on all types of portfolios across the risk spectrum, which is prohibitively impractical at this stage. That said, the  $g_1$ -measure is henceforth used in this section given its supremacy in this instance.

## 4.2. COMPUTATIONAL RESULTS OF RECOVERY OPTIMISATION

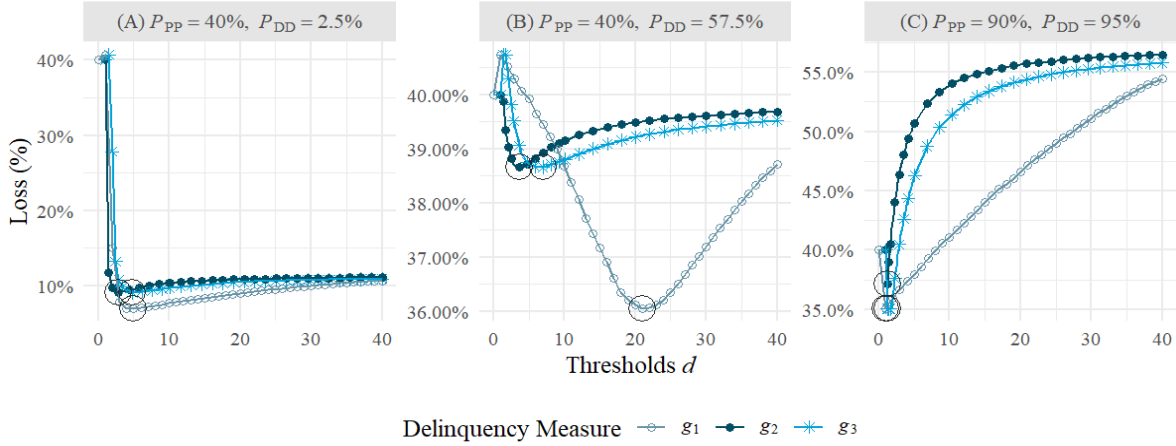


FIG. 4.6: Losses across thresholds  $d$  by measure  $g \in \{g_1, g_2, g_3\}$  using the Markovian defaults technique to generate three different loan portfolios. Each panel explores a specific setting of the transition matrix, using the titular probabilities within the matrix defined in Table 4.1. Encircled points indicate loss minima at associated thresholds  $d^{(g)}$ .

Using this technique, a broad iterative scheme is devised to generate portfolios systematically across the entire credit risk spectrum, as measured with  $g_1$ . In particular,  $P_{DD}$  is held constant at a certain value while varying  $P_{PP}$ , followed by fixing  $P_{DD}$  to a different value and varying  $P_{PP}$  again, and so on. This scheme allows for suitably varying the transition matrix in Table 4.1 using fixed intervals, with some of the resulting loss curves and associated loss minima presented in Fig. 4.7. The subplots in both panels (A) and (I) represent boundary cases that confirm intuition. Specifically, panel (A) demonstrates recovery optimisation for portfolios with highly transitive delinquency states such that accounts immediately exit this state in the next period, once entered. Accordingly, the loss curves increasingly resemble a near risk-less case as the value of  $P_{PP}$  tends towards 1, which is similar to setting  $b = 1$  in Fig. 4.3 when using random defaults. In turn, recovery optimisation itself becomes progressively infeasible in tandem with  $P_{PP}$  approaching 1. Conversely, panel (I) showcases the loss curves of extremely risky portfolios, which are again similar to setting  $b = 0$  in Fig. 4.3 as  $P_{PP}$  approaches 0. More importantly, the fact that loss minima occur at very small thresholds agrees intuitively with cutting losses sooner rather than later, especially for extreme default risk.

The remaining panels in Fig. 4.7 are perhaps the most instructive. As the delinquency state becomes more absorbing (or less transient), i.e., moving from panel (B) to (F), the loss-optimal thresholds  $d^{(g_1)}$  become increasingly staggered across both axes. This is to say that  $d^{(g_1)}$  becomes progressively more sensitive to both the threshold  $d$  and the value of  $P_{PP}$ . Moreover, it is sensible that ever greater losses (at  $d^{(g_1)}$ ) are associated with lower values of  $P_{PP}$  since the latter implies

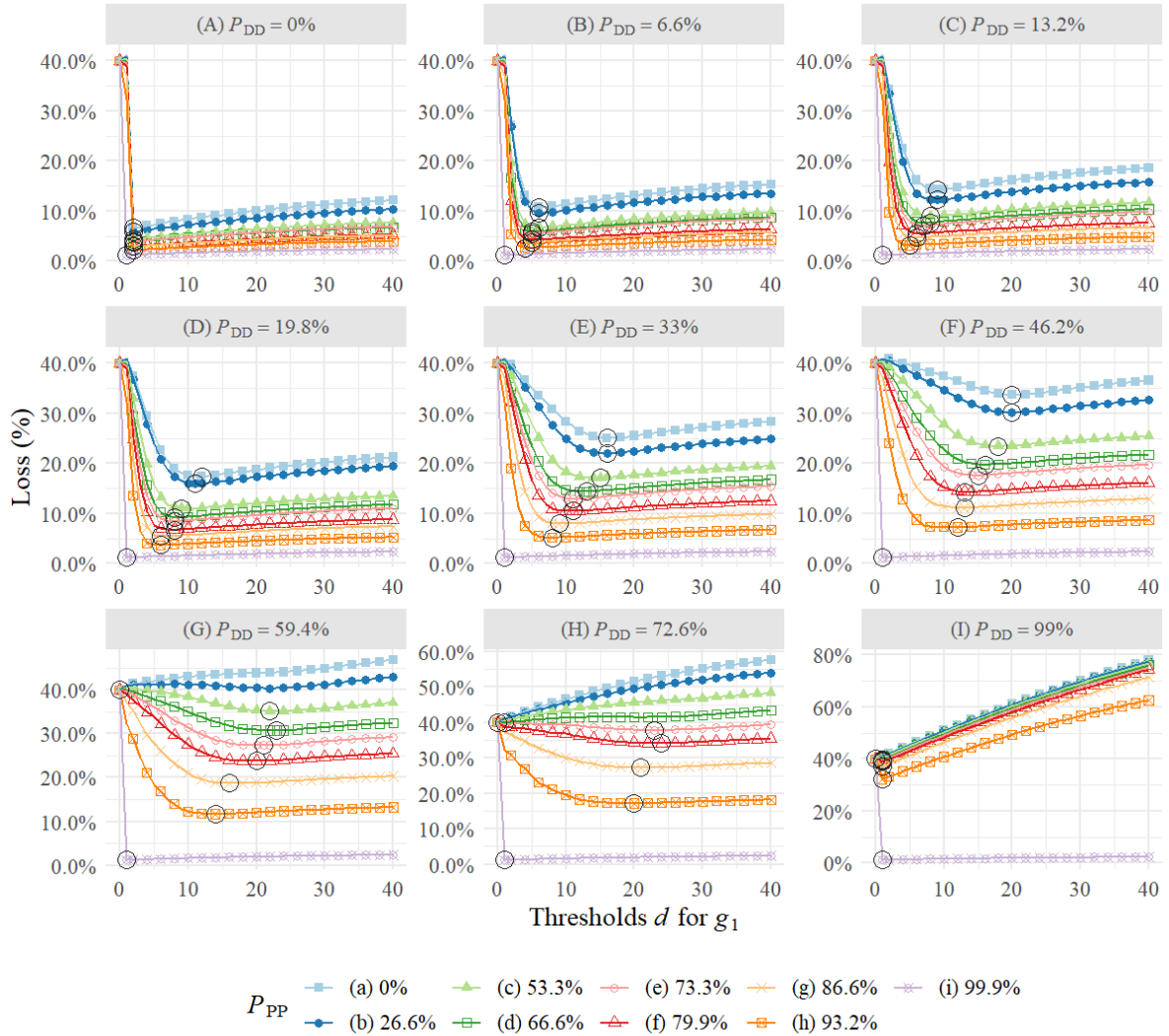


FIG. 4.7: Losses across thresholds  $d$  for the  $g_1$ -measure using the Markovian defaults technique with several transition rates  $P_{PP} \in [0, 1]$  and  $P_{DD} \in [0, 1]$ . Encircled points indicate loss minima at associated thresholds  $d^{(g_1)}$ .

less time being spent in the paying state, even as the delinquency state becomes less transient. Furthermore, consider that  $d^{(g_1)}$  increases in threshold-value when  $P_{PP}$  decreases and  $P_{DD}$  increases, i.e., moving from curve (i) down to curve (a) whilst moving across panels (B) to (F). This suggests that gradually postponing loan recovery is the better strategy even as delinquency becomes more likely, at least up until a certain point, in this case, panel (G). However, this suggestion is counter-intuitive since one would rather cut losses sooner than later when risk supposedly increases, which implies selecting lower thresholds instead.

Two factors help explain this phenomenon. Firstly, the relevant portfolios are increasingly turbulent by design when  $P_{PP}$  changes from higher to lower values in each successive panel. The effect hereof is that loans start to oscillate quite rapidly between the paying and delinquent states as  $P_{PP}$  decreases. The slightly increased rate of absorption into the delinquent state (when moving across panels) is not sufficient to support earlier loan recovery as intuition would otherwise suggest, especially so when an account still frequently exits the delinquent state. This has the side-effect of muting the severity of ‘default’, which is plausible when curing from ‘default’ itself becomes increasingly likely due to the same turbulence. Therefore, the associated opportunity cost of forsaking the loan earlier is too high when future repayments are still likely to be received over the longer run, albeit sporadic. Accordingly, greater turbulence in a portfolio requires greater patience to collect upon these repayments, which is why postponing loan recovery (by virtue of  $d^{(g_1)}$  increasing) would be loss-efficient. Secondly, even if  $d^{(g_1)}$  increases in value, the associated loss minimum reassuringly increases alongside  $P_{DD}$ , as expected from more turbulent and riskier portfolios.

There is little need for applying  $(k, g)$ -truncation on these results since the Markovian technique already has a realistically-set write-off state that achieves the same effect. While additional truncation will surely confound the results,  $(12, g_1)$ -truncation is experimentally applied in the interest of completeness. The results (not shown) are largely similar to that of random defaults in that loss minima still occur at or near  $k = 12$  across most portfolios. The exceptions are the two boundary cases, i.e., at or close to panels (A) and (I). Furthermore, the Markovian technique is especially geared towards generating "regime-switching" portfolios where accounts suffer from episodes of delinquency that vary in length, as controlled by the state probabilities. In this regard, episodic delinquency is more common a phenomenon in practice than one would think, which is why investigating recovery optimisation for these cases is more valuable than exploring explicit truncation/write-off any further in this section.

#### **4.2.4 Applying the LROD-procedure on real-world data**

The steps in section 3.4 require data to be in a longitudinal-format, having measured delinquency in retrospect across all accounts and time (usually monthly), based on expected instalments and actual receipts. Letting the contractual term, loan and risk-free rates, and even the loss rates vary across the portfolio ought not to impede the practical use of the LROD-procedure. However, the portfolio is assumed to be fully observed (or ‘completed’) in this study, with little consideration given to any right-censoring and its effect on the receipt history of an account. This particular avenue is further explored in Botha et al. (2020), thereby demonstrating the empirical use of the procedure on real-world data. That said, simply excluding incomplete accounts from the dataset can sidestep this possible issue, though at the cost of a reduced sample size. The effect hereof will likely vary based on the typical tenure of the loan product.

The results, particularly those from subsection 4.2.3, can easily translate into practical value with relatively little analytical effort. For example, one can fit the same three-state Markov chain on a real portfolio's delinquency progressions, just to obtain the associated transition rate estimates. In turn, these estimates can be used as a rough guide in finding a corresponding loss curve amongst all those presented in Fig. 4.7, i.e., a look-up exercise. The associated optimised threshold can provide a high-level idea of recovery optimisation, provided the assumptions are reasonably met. That said, applying the LROD-procedure remains the imperative in order to capture all idiosyncrasies of a particular portfolio and the prevailing market conditions.

### 4.3 Concluding remarks

The results demonstrate that optimising the timing of the recovery decision using the LROD-procedure is sensitive to the level of credit risk of a portfolio and its particular composition. This rather intuitive result rests upon weighing two competing interests against each other: the prospect of reaping future revenue from troubled loans versus the cost of retaining these loans any further. In addition, the LROD-procedure is formulated in such a way that it can be used with multiple loan delinquency measures. This facilitates the objective testing of alternative measures, e.g., those provided in the appendix, that may better suit the recovery optimisation (or even broader risk management) of a portfolio. That said, the study objective is not to establish the best measure conclusively, which would likely be a data-intensive and costly endeavour.

Regarding results, a simple simulation-based setup is first described in which the LROD-procedure (and its goal of recovery optimisation) is closely examined from "first principles". Using this setup as a testbed, a broad computational study is conducted wherein basic amortising loan portfolios are systematically generated by varying the simulation parameters, though still constrained by expert judgement. Having spanned the entire credit risk spectrum (as measured with the payment probability  $b$ ), the computational results show that optimising the recovery decision's timing is viable across most risk levels, except at the extremes. The results further indicate that optimised recovery times are sensitive to systematic defaults that may structurally affect a portfolio during an economic downturn, as approximated by the notion of  $(k, g)$ -truncation in the testbed. Another factor is that of collateral and the portfolio's loss experience (or LGD), insofar that optima were successfully found across most of the loss spectrum (as measured with the loss rate  $r_A$ ). Moreover, recovery optimisation seems to become an increasingly viable practice as the risk of loss increases.

In addition, recovery optimisation is tested on more turbulent portfolios wherein borrowers repay intermittently, thereby causing episodic delinquency. Once accounts oscillate rapidly between paying and nonpayment, 'default' itself diminishes in severity, especially when curing also becomes more likely as a result of the very same turbulence. Accordingly, optimised

thresholds are shown to increase in value as turbulence develops, though only up to a point. Postponing loan recovery in tandem with greater turbulence is therefore strategically optimal since it allows greater scope to collect upon these repayments, albeit sporadic. As a secondary contribution, the testbed itself can serve as a valuable tool in exploring the strategic viability of the LROD-procedure. Once appropriately parametrised, the testbed can generate a wide variety of portfolios, which allows a bank to investigate (at least preliminarily) the prospects of recovery optimisation for a certain type of portfolio. Ultimately, the LROD-procedure can be used to tweak existing collection policies and, perhaps in time, default definitions themselves.

Future studies can focus on refining the LROD-procedure using real-world portfolio data. So-called ‘incomplete’ portfolios, i.e., those wherein many loans have not yet reached contractual maturity, may prove a challenge for recovery optimisation at this stage. The simplest solution would be to exclude the incomplete accounts, though unfortunately reducing the sample size as well. Alternatively, one can perhaps explore an appropriate forecasting approach in future work. Furthermore, homogeneity is currently assumed in that the optimised threshold is a portfolio-wide criterion. However, exploring segmentation schemes may be worthwhile such that the LROD-procedure yields an ideal threshold for each identified segment within the portfolio. Lastly, the current loss model  $L_g$  can be refined by incorporating historical loss experiences and transforming it into a more dynamic component. As an example, calculating the realised LGD generally requires a specific point of entering ‘default’. From this point, cash flows are observed during its workout up to the applicable write-off point. By introducing  $d$  as the  $(g, d)$ -default state, the starting points of cash flows will naturally vary with  $d$ , thereby impacting the LGD calculation itself for each  $(g, d)$ -policy. Intuitively, longer or shorter workout periods will affect the loss experience, which will influence recovery optimisation based on the study results. This particular refinement will likely intersect with the existing literature on credit loss modelling and IFRS 9, which as a field is currently quite in vogue.



*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*



## RECOVERY OPTIMISATION USING REAL-WORLD DATA WITH FORECASTING

The LROD-procedure from section 3.4 allows for the objective comparison, evaluation, and optimisation of a bank's recovery decision (i.e, foreclosure) across competing delinquency measures. This procedure was demonstrated in chapter 4 as an optimisation problem using toy portfolios in such a way that loans are forsaken neither too early nor too late, if at all. In this chapter, the same LROD-procedure is extended and refined using a rich real-world portfolio of 20-year residential mortgages, provided by a large South African bank. However, real-world amortising loan portfolios typically suffer from right-censoring in that the many of its constituent loan accounts have not yet reached contractual maturity, barring those that were settled early or written-off historically. Older monthly cohorts will have more observable history available than those cohort originated more recently. Consequentially, the cash inflows (or receipts) of these active loan accounts can only be observed up to the most recent time point  $t_0$  and no further. More formally, a loan account has the receipt vector  $\mathbf{R} = [R_1, R_2, \dots, R_{t_0}, R_{t_1}, \dots, R_{t_c}]$  of which elements are observed from data only up to time  $t_0 \leq t_c$  with  $t_c$  denoting the contractual term. The remaining future elements at times  $t_1, \dots, t_c$  (where  $t_0 < t_1 \leq t_c$ ) are right-censored and therefore unobservable.

Accordingly, censored portfolios present a secondary challenge since the LROD-procedure was originally developed within the context of uncensored 'completed' portfolios. It is also quite challenging to procure a real-world uncensored though still data-rich portfolio from a willing lender. However, even if procured, its use may become questionable since the completed portfolio

may no longer reflect current market conditions, which may adversely affect recovery optimisation. Moreover, the unsurprising reality is that most loan portfolios will be censored to some degree since they are being actively grown every month, which suggests that censoring will remain a prevalent problem whenever seeking to apply the LROD-procedure in practice. Failure to treat for this censoring leads to unusable and counter-intuitive results<sup>1</sup>. Arguably, it would be more feasible to use available data and try to forecast the remaining cash flows of each censored account up to its contractual term, as originally depicted in Fig. 1.3. Performing this necessary step will then in turn enable the empirical use of the LROD-procedure.

This chapter is begun by outlining two candidate forecasting techniques in section 5.1 by which a censored portfolio can be completed. Each technique is parametrised from 20-year residential mortgage data and assessed on its forecasting quality in section 5.2. The LROD-procedure is then applied on the now-completed portfolio in section 5.3, accompanied by a discussion of the ensuing results. A Monte Carlo-based refinement to the procedure is demonstrated in subsection 5.3.2 by which the variance of the underlying forecasts can be analysed, thereby granting additional assurance on the stability of the optimisation results. Finally, this chapter is concluded in section 5.4, which includes a discussion on limitations and avenues for future research. Overall, the timing of a bank's recovery decision is successfully illustrated as a delinquency-based optimisation problem using real-world data. This work can therefore facilitate the revision of relevant bank policies or strategies towards optimising the loan collections process. A research article titled "*The loss optimisation of loan recovery decision times using forecast cash flows*" is associated with this chapter, accepted for publication in the *Journal of Credit Risk*. A preprint is published in Botha et al. (2020) while the associated source code is available in Botha (2020b).

## 5.1 Two techniques to forecast future loan receipts

Two techniques are presented in this section to forecast the future cash flows  $R_{t_1}, \dots, R_{t_c}$  up to the contractual term  $t_c$  of each censored loan account, using its observed receipt history  $R_1, \dots, R_{t_0}$ . These techniques include a simple probabilistic technique called *random defaults* as well as a more sophisticated eight-state Markov chain-based technique called *Markovian defaults*. Note that both techniques are data-driven extensions of those used in chapter 4.

### 5.1.1 Random defaults with empirical truncation

Let  $u_t \in [0, 1]$  be a randomly generated number from the uniform distribution at every loan period  $t = t_1, \dots, t_c$  that is to be forecast. Let  $b$  be an estimable probability of payment, i.e.,  $P(R_t = I_c) = b$  with  $I_c$  being the calculated level instalment. This instalment is calculated such that it amortises

<sup>1</sup>Using an untreated real-world portfolio in the procedure did not yield loss-optimised thresholds on any measure, regardless of the chosen loss rate – see Appendix A.3.

the most recent outstanding balance as observed at time  $t_0$  to zero at time  $t_c$ , using the most recent client interest rate observed from data. The receipt is then initially forecast as

$$R_t = \begin{cases} I_c & \text{if } u_t < b \\ 0 & \text{otherwise} \end{cases} . \quad (5.1)$$

A truncation effect is introduced (similar to section 4.1) via a structural break in the forecast receipt vector at a certain point (if at all) and replacing elements thereafter with zeros. Tempering the predicted receipts in this way mimics the fact that some loan accounts will simply never resume payment in reality. This is similar to Thomas et al. (2016) wherein the parameters controlling the payment and non-payment sequences were fixed after reaching some point in the process. More formally, consider all periods  $j = t_0, \dots, t_c$  within the now-forecast receipt vector  $\mathbf{R}$  of a particular account, with the measure  $g_1$  applied accordingly across all periods. Let  $k \geq 0$  be a truncation parameter above which the receipts are truncated. The starting period of this truncation, denoted as  $t_k \geq 0$ , may then exist if the account has experienced sufficient delinquency  $g_1(j) \geq k$  at some  $j$ , i.e.,  $t_k = \min(j : g_1(j) \geq k)$ . Conversely, if delinquency has not breached  $k$ , then this time point  $t_k$  does not exist. A process called  $(k, g_1)$ -truncation then changes  $\mathbf{R}$  to  $\mathbf{R}'$  by

$$\mathbf{R}' = \begin{cases} [R_{t_1}, \dots, R_{t_k}, 0, \dots, 0] & \text{if } t_k \text{ exists} \\ \mathbf{R} & \text{otherwise} \end{cases} . \quad (5.2)$$

In estimating this truncation parameter  $k$  (as opposed to fixing it previously in section 4.2), consider that the maximum delinquency across time can be obtained for each account in a loan portfolio, using  $g_1$  for simplicity's sake. In turn, the histogram of these maxima is plotted, followed by fitting statistical distributions to these maxima. One can then draw a random sample  $\hat{k}_i$  from an appropriately fitted distribution for each account and finally  $(\hat{k}_i, g_1)$ -truncate the initially forecast receipt vector. This introduces some realistic variance to the overall truncation effect.

Lastly, consider an indicator function  $\mathcal{I}_t^{(i)}$  that signals payment using the receipt  $R_t^i$  and instalment  $I_t^i$  of the  $i^{\text{th}}$  account at its historical periods  $t = 1, \dots, t_0(i)$ . This  $\mathcal{I}_t^{(i)}$  is then formally defined as

$$\mathcal{I}_t^{(i)} = \begin{cases} 1 & \text{if } R_t^i \geq I_t^i \\ 0 & \text{otherwise} \end{cases} \quad t = 1, \dots, t_0(i) . \quad (5.3)$$

The probability of payment  $b$  can be estimated by  $\hat{b}$ , which is defined as

$$\hat{b} = \frac{1}{N} \sum_i \frac{1}{t_0(i)} \sum_t \mathcal{I}_t^{(i)} \quad \forall i = 1, \dots, N \text{ and } t = 1, \dots, t_0(i) . \quad (5.4)$$

### 5.1.2 Markovian defaults

Let  $X_t \in \{x_0, \dots, x_7\}$  be a random vector that can assume one of eight increasingly-severe delinquency states derived from  $g_1$ , across all historical periods  $t = 1, \dots, t_0$  of an account. The

states  $x_0, \dots, x_5$  correspond to  $g_1(t)$  having the respective values  $0, \dots, 5$  at any  $t$ . State  $x_6$  is semi-absorbing such that  $g_1(t) \geq 6$  at any  $t$  and the state  $x_7$  denotes write-off (fully-absorbing). The sequence  $X_1, \dots, X_{t_0}$  then forms a discrete-time first-order Markov chain from which receipts can be forecast, based on the predicted states  $X_{t_1}, \dots, X_{t_c}$  at future periods  $t = t_1, \dots, t_c$ . Note that  $g_1$  can only ever increase in value by one delinquency level, while it can decrease by several levels depending on the magnitude of the overpayment  $R_t > I_t$ . Lastly, previous studies used Markov chains with fewer states to characterise the delinquency process (see subsection 3.1.3). However, more delinquency states should theoretically translate into better capturing a portfolio's delinquency dynamics over time. While this claim is not explicitly tested, an eight-state Markov chain is deemed a reasonable compromise between greater sophistication and simplicity.

To generate receipts at these future periods, temporarily ignore write-off ( $x_7$ ) and consider the one-period delinquency difference  $\delta_t$ , defined as  $\delta_t = g_1(t) - g_1(t-1)$ . A positive difference  $\delta_t > 0$  implies  $R_t < h_t I_c$  since delinquency has increased and  $R_t$  is therefore simply zeroed. Secondly,  $\delta_t = 0$  implies  $R_t = I_c$  since the delinquency level remained unchanged. Finally,  $\delta_t < 0$  implies  $R_t \geq 2I_c$  since  $\delta - 1$  extra payments are needed to decrease the delinquency level beyond the instalment normally due at the time. When  $X_t = x_6$ , the account remains semi-absorbed as long as  $g_1(t) \geq 6$ , which implies either increasing or constant delinquency. For the sake of prudence, the former case is assumed (i.e.,  $\delta_t > 0$ ) and  $R_t$  is zeroed accordingly. These ideas (barring  $x_7$ ) are combined into forecasting the receipt as

$$R_t = \begin{cases} -I_c(\delta_t - 1) & \text{if } \delta_t < 0 \\ I_c & \text{if } \delta_t = 0 \\ 0 & \text{if } \delta_t > 0 \end{cases} . \quad (5.5)$$

Note that truncation is effectively incorporated whenever an account transitions to the absorbing write-off state  $x_7$  at a supposed time point  $t_w$  that only exists when  $X_{t_w} = x_7$  with  $t_1 \leq t_w \leq t_c$ . This inherently implies zeroed receipts from that point forward, i.e.,  $R_t = 0$  for  $t = t_w, \dots, t_c$  if  $t_w$  exists.

In estimating the transition matrix of this Markov chain, note that the receipt history of each loan account effectively signifies a repeated observation of the underlying chain, as discussed in T. W. Anderson and Goodman (1957). Assuming stationarity, the maximum likelihood estimates (MLEs) for the transition probabilities  $p_{ij}$  from state  $i$  to state  $j$  are then  $\hat{p}_{ij} = n_{ij}/n_i^*$  where  $n_{ij}$  is the number of observed transitions across all time periods from state  $i$  to  $j$  and  $n_i^*$  is the observed number of total transitions starting in state  $i$ . In this context, it is not necessary to estimate initial state probabilities since the starting delinquency state is simply observed from the last available time point  $t_0$  of an account.

## 5.2 Calibrating the forecasting techniques to mortgage data

The aforementioned forecasting techniques are calibrated in this section using credit data on a rich portfolio of ordinary home loans granted to the lower-income segment of the South African market. Although shorter-term matured loans would be ideal for this study, only mortgage data was available. This longitudinal dataset has monthly loan performance observations over time  $t = 1, \dots, t_0(i)$  for account  $i = 1, \dots, N$  with  $N = 61,648$  single-advance 20-year mortgage accounts. These mortgages were originated from April 2004 (and beyond) and observed throughout up to December 2017, thereby yielding 3,271,534 raw monthly observations of loan performance. This data includes actual net cash flows (receipts), expected instalments (including credit life insurance add-ons and fees, or special arrangements), variable interest rates, original loan principals, month-end balances, write-off amounts and indicators, asset sale proceeds, and early settlement indicators.

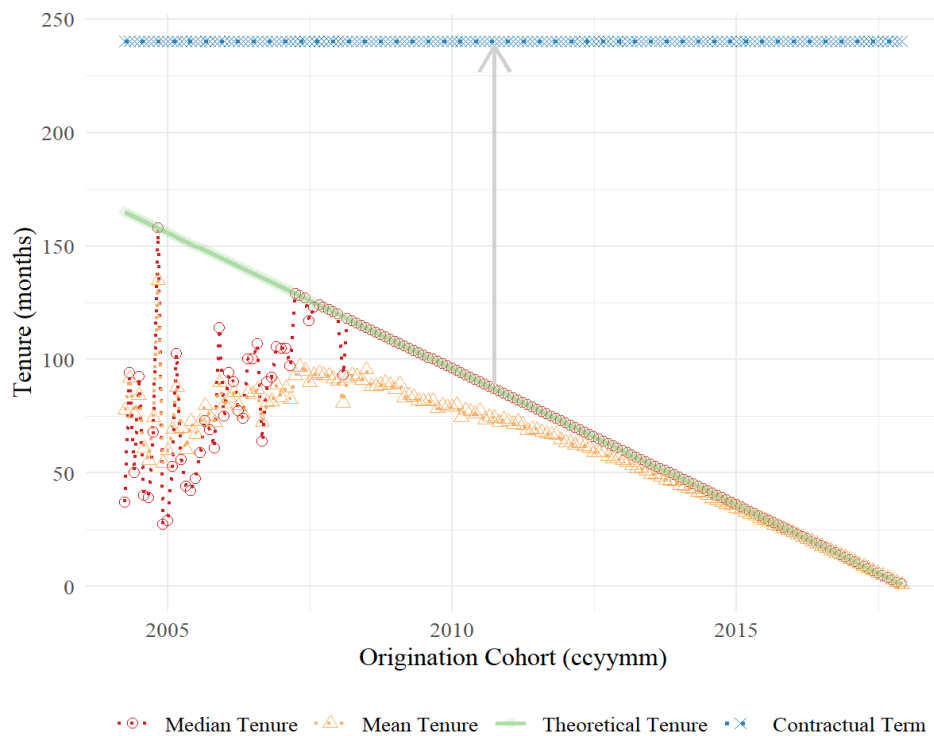


FIG. 5.1: The difference between the theoretically observable loan tenure (as measured at December 2017 in retrospect) and the remaining contractual term across monthly loan cohorts. The mean and median loan ages per monthly cohort are overlaid.

As measured at December 2017, the difference between the maximum theoretical loan tenure and the remainder of the contractual term is shown in Fig. 5.1 at every historical monthly loan cohort, with aggregates overlaid. Clearly, these aggregates are below the theoretical maximum

for most cohorts, which demonstrates some additional right-censoring. To this point, mortgage loans can exit the portfolio pre-maturely either via write-off or via early settlement, e.g., private sales, bond cancellations, or transfers. Moreover, the volatility in both of these aggregates at earlier times attests to low sample sizes, which is unsurprising for a fledgling loan portfolio at the time. This volatility, however, gradually subsides until both aggregates approach the theoretical maximum. This is sensible since more recently originated mortgages have less time available to develop write-off or early settlement outcomes than their older counterparts.

In estimating the various parameters of the forecasting techniques, the data is partitioned to form three specific samples:  $S_1$  as the full dataset,  $S_2$  as the delinquents-only sample (all accounts that had at least one payment in arrears historically, or were eventually written-off), and  $S_3$  as the write-offs sample. These samples and the relationships amongst them are illustrated with a Venn diagram shown in Fig. 5.2. Some accounts will simply never experience any delinquency and their exclusion in  $S_2$  and  $S_3$  removes an optimism bias during model training. There is little practical benefit to finding the best time for loan recovery on a near risk-less portfolio. Furthermore, recovery optimisation is only sensible for loans likely to become delinquent in the first place, which is predicated upon forecasting them as such. Likewise, it would be pointless to forecast cash flows of closed accounts, though their repayment histories are retained for model training purposes. Lastly, this particular partitioning scheme is an experimental proxy for risk compositions that differ across both product and risk appetites in reality. As an example, mortgages typically have a much lower default rate than unsecured personal loans, which is catered for in the current setup.

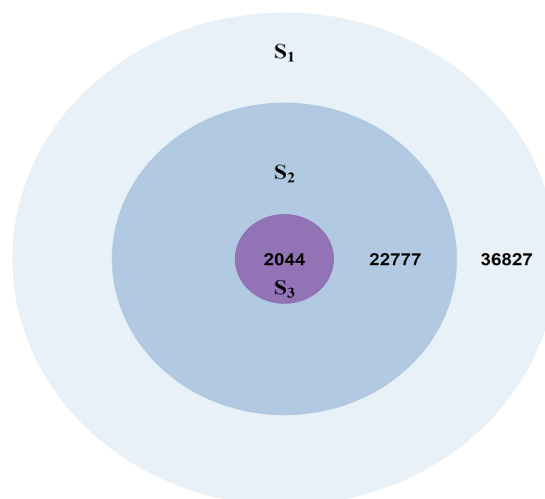


FIG. 5.2: A Venn diagram showing the relative sizes and overlaps amongst the three main samples of mortgage accounts:  $S_1$  (full sample),  $S_2$  (delinquents), and  $S_3$  (write-offs). These samples are used both in training the forecasting techniques and during the subsequent loss optimisation.

### 5.2.1 Calibrating the random defaults technique

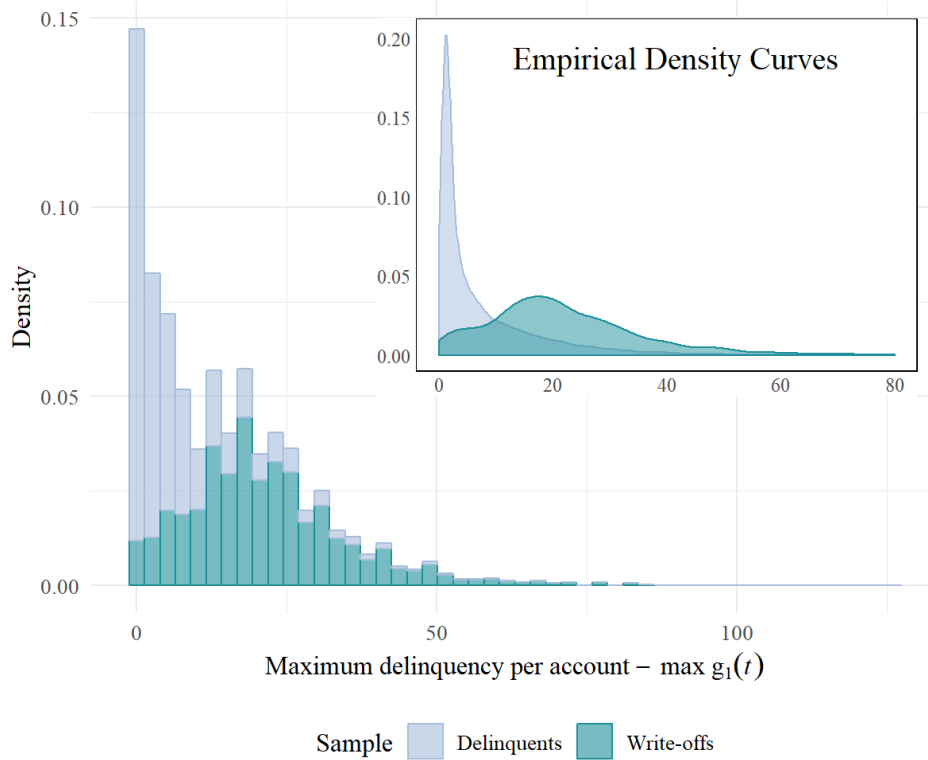


FIG. 5.3: A histogram and empirical density curve of the maximum delinquency level observed per account, drawn for the samples  $S_2$  (delinquents) and  $S_3$  (write-offs). A theoretical distribution is then fit on each sample (see Appendix A.4), from which the truncation parameter  $k$  is drawn randomly for each loan account prior forecasting.

The aforementioned probability of payment  $b$  used in this technique is estimated from samples  $\{S_1, S_2, S_3\}$  respectively as  $\hat{b}_1 = 87\%$ ,  $\hat{b}_2 = 81\%$ , and  $\hat{b}_3 = 45\%$ . The descending values are plausible given that each successive sample contains a greater proportion of delinquency by design. Note that the random truncation of forecasts that accompanies this technique is only sensibly performed for delinquent cases, which implies  $k > 0$ . Therefore, ignoring  $S_1$ , the distribution of the maximum delinquency level per account, i.e.,  $\max g_1(t)$  across all historically observed periods  $t = 1, \dots, t_0$ , is given in Fig. 5.3 for both samples  $S_2$  and  $S_3$ . A few statistical distributions were tested against the data (see Appendix A.4), though the exponential and two-parameter Weibull distributions are chosen for  $S_2$  and  $S_3$  respectively, denoted as  $\text{Exp}(\lambda)$  and  $\text{Weibull}(\lambda, \phi)$ . The MLEs of these parameters are  $\lambda = 0.1378555$  for the exponential distribution, scale  $\lambda = 24.449566$  and shape  $\phi = 1.688026$  for the Weibull distribution. The truncation parameter then follows either one of these distributions, i.e.,  $k \sim \text{Exp}(\lambda)$  for both  $S_1$  and  $S_2$ , as well as  $k \sim \text{Weibull}(\lambda, \phi)$  for  $S_3$ , as part of a comparative study. Note that the exponentially-distributed  $k$  has a stronger

truncation effect since it generally yields lower values of  $k$  than those yielded by its Weibull-distributed counterpart. This is also evidenced by the sample mean of  $k$  estimated from  $S_2$  being 7.25 versus that from  $S_3$  being 21.58.

### 5.2.2 Calibrating the Markovian defaults technique

		Ending state							
		$x_0$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
Starting state	$x_0$	0.9477	0.0521	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	$x_1$	0.0942	0.8074	0.0980	0.0000	0.0000	0.0000	0.0000	0.0004
	$x_2$	0.0138	0.0502	0.7735	0.1621	0.0000	0.0000	0.0000	0.0004
	$x_3$	0.0064	0.0084	0.0481	0.7372	0.1993	0.0000	0.0000	0.0006
	$x_4$	0.0064	0.0030	0.0082	0.0488	0.6957	0.2371	0.0000	0.0007
	$x_5$	0.0051	0.0020	0.0029	0.0081	0.0469	0.6846	0.2496	0.0009
	$x_6$	0.0044	0.0006	0.0007	0.0009	0.0021	0.0095	0.9756	0.0061
	$x_7$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

TABLE 5.1: Maximum likelihood estimates for the transition matrix of the multi-state Markov chain, estimated from the delinquents sample  $S_2$ . States  $x_0, \dots, x_5$  correspond to  $g_1$  having the respective values  $0, \dots, 5$  (weighted payments in arrears). States  $x_6$  (semi-absorbing) and  $x_7$  (absorbing) indicate  $g_1 \geq 6$  and write-off respectively.

		Ending state							
		$x_0$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
Starting state	$x_0$	0.8820	0.1126	0.0000	0.0000	0.0000	0.0000	0.0000	0.0054
	$x_1$	0.0962	0.5387	0.3534	0.0000	0.0000	0.0000	0.0000	0.0117
	$x_2$	0.0254	0.0453	0.4607	0.4600	0.0000	0.0000	0.0000	0.0086
	$x_3$	0.0136	0.0103	0.0430	0.3824	0.5393	0.0000	0.0000	0.0114
	$x_4$	0.0117	0.0032	0.0093	0.0412	0.3187	0.6048	0.0000	0.0112
	$x_5$	0.0079	0.0037	0.0037	0.0053	0.0293	0.3181	0.6194	0.0127
	$x_6$	0.0076	0.0006	0.0005	0.0007	0.0012	0.0035	0.9474	0.0385
	$x_7$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

TABLE 5.2: Maximum likelihood estimates for the transition matrix of the multi-state Markov chain, estimated from the write-offs sample  $S_3$ . States  $x_0, \dots, x_5$  correspond to  $g_1$  having the respective values  $0, \dots, 5$  (weighted payments in arrears). States  $x_6$  (semi-absorbing) and  $x_7$  (absorbing) indicate  $g_1 \geq 6$  and write-off respectively.

The MLEs for the transition matrices, as used in this technique, are estimated only from the samples  $S_2$  and  $S_3$ , shown respectively in Tables 5.1–5.2. Note that the estimates using  $S_1$  differ from those yielded by using  $S_2$  only in the first row, which is sensible since  $S_1$  contains the same delinquent accounts (and therefore the same transitions) as  $S_2$  by design. The estimates are realistic in that an account in any particular delinquency state (barring write-off) can increase its delinquency level only by one level within a monthly period. Additionally, these estimates reflect the fact that an account can significantly overpay and thereby recover either partially or



entirely from distress. The probability of staying within a particular starting state is greatest, though it decreases gradually as the delinquency level increases, at least for states  $x_0, \dots, x_5$ . Simultaneously, the probability of becoming even more delinquent increases as the starting delinquency level increases, which agrees with anecdotal experience in the industry. This is corroborated by the increasing probability of write-off, effectively representing an increasing probability of truncation.

### 5.2.3 Assessing the quality of forecasts

Although forecasts are trained specifically on  $\{S_1, S_2, S_3\}$ , the forecast quality itself is examined in this section by following a more general  $k$ -fold cross-validation approach as additional assurance. However, the available mortgage portfolio did not have a single completed 20-year loan, against which the receipt forecasts could be validated across *all* periods. Nonetheless, available loan data up to  $t_0(i)$  is still used within a  $k = 5$  cross-validation setup, despite the censoring-related bias this likely introduces into measuring the forecast error. Moreover, the main objective is not to produce the most accurate or robust forecasting model on the account-level, although that is certainly a worthwhile endeavour. Instead, the present focus is more fundamental: using different forecasts (regarding accuracy) when optimising the timing of loan recovery, which suggests using multiple forecasting techniques.

Metric	Random defaults ( $T_a$ )	Markovian defaults ( $T_b$ )
Mean Absolute Error (MAE) of cash flows	3,233.19	1,414.03
Instalment-scaled MAE	103.7%	45.3%
Delinquency Forecast Error (DFE): Mean	30.3	5.5
DFE: Median	22.2	0
Portfolio Arrears Rate (PAR)	6.715% (-1.64%)	0.695% (-1.64%)
Mean parameter %-difference	0.00012%	-0.0068%

TABLE 5.3: The results of various measures, calculated and averaged across a 5-fold cross-validation setup. The receipt forecasts are validated using MAE against the actual receipts within the  $k^{\text{th}}$  subset per technique, having trained the technique on the rest of the data. The DFE-metric compares the  $g_1$ -based delinquency levels underlying forecasts against the actual values by taking the difference thereof at each period and averaging. The PAR-metric expresses the sum of discounted shortfalls (essentially ‘arrears’) between instalments and forecasts as a proportion of all gross advances, using 7% as the discounting rate. The actual PAR-value is -1.64% on average, which is negatively signed due to large historical overpayments at earlier periods.

A few measures are used that span forecast error, portfolio impact, and overall parameter stability, with the results thereof given in Table 5.3. These results reflect the significant differences between each technique’s performance, which is unsurprising given that the simpler technique

( $T_a$ ) deliberately ignores the possibility of curing. Another factor that affects the forecast accuracy is the ability of the Markovian technique ( $T_b$ ) to produce forecasts based on the level of accrued delinquency. This advantage gave a *Mean Absolute Error* (MAE) of less than half that of  $T_a$ , which is perhaps made more contextual when expressing the error as a proportion of the overall mean instalment. The receipt forecasts are perhaps less important themselves than the delinquency calculations that they enable, as eventually used during recovery optimisation. Accordingly, the *Delinquency Forecast Error* (DFE) quantifies the ‘error’ in measured delinquency levels when replacing historical cash flows with their forecast-counterparts. As a result, the DFE also suggests a clear preference for  $T_b$  with its much lower error, regardless of taking the mean or median of these account-level errors. On the other hand, the *Portfolio Arrears Rate* (PAR) reflects the portfolio-wide arrears rate as implied when using each technique’s forecasts of historical receipts. Again, the PAR of  $T_b$  is much closer to the actual rate than that of  $T_a$ . Regarding parameter stability, the mean %-difference in parameter estimates is reassuringly close to 0, as calculated between using all data versus using each training fold.

### 5.3 Optimising the recovery decision: an empirical illustration

The parameters of each forecasting technique were previously estimated three times, each from a progressively worse sample, thereby recognising that a portfolio’s historical risk composition itself will bias the forecast receipts. Naturally, the LROD-procedure itself can be applied on each of these subsequent samples. The locations of the recovery thresholds at which losses are minimised (if found) are expected to differ significantly, given the different risk profiles. That said, this procedure is imagined to be applied on the entire loan portfolio when loss-optimising a bank’s recovery decision in practice. It is, however, iteratively applied in this study as an experimental and artificial proxy for various risk compositions found in reality, as if each sample is a stand-alone portfolio. Moreover, the sample  $S_i$  from which a forecasting technique is parametrised (or trained) may differ from the sample  $S_j$  on which the LROD-procedure is applied, where both  $i$  and  $j$  are indexes that signify samples  $\{S_1, S_2, S_3\}$ . Apart from using data more efficiently, this approach approximates the reality of a portfolio’s historical risk composition changing in the future. As an example, parametrising a forecasting technique from  $S_3$  but optimising recovery thresholds on  $S_1$  simulates the context of a proportionally lower-risk portfolio ( $S_1$ ) undergoing heavy financial strain in the future (by using forecasts trained from  $S_3$ ). Additionally, this proposed setup aligns with the IFRS 9 accounting standard, which requires expected credit losses to be estimated based on various macroeconomic scenarios, as stated in IFRS 9 (2014, §5.5). Therefore, the experimental setup is illustrated as a  $3 \times 3$  matrix in Table 5.4 wherein each cell  $s_{ij}$  represents the results from a specific scenario. Greater values of  $j$  denote riskier portfolios, while greater values of  $i$  represent more pessimistic forecasts.

On interpreting the following results, the LROD-procedure’s particular loss model from

		$j$		
		$S_1$	$S_2$	$S_3$
$i$	$S_1$	$s_{11}$	$s_{12}$	$s_{13}$
	$S_2$	$s_{21}$	$s_{22}$	$s_{23}$
	$S_3$	$s_{31}$	$s_{32}$	$s_{33}$

TABLE 5.4: The experimental setup containing nine scenarios wherein row  $i$  represents the sample used for parametrising a forecasting technique, and column  $j$  denotes the sample on which optimisation is performed.

section 3.4 assumes that a portion of the expected balance and arrears amount are immediately lost upon entering  $(g, d)$ -default. In effect, this equates the actual default and write-off events to a single point, which implies that a loss-optimised threshold is not necessarily the suggested starting point of legal proceedings, but rather the optimal ending point. For that matter, finding the optimal starting point of legal proceedings will likely lead to a workout period that varies in its length, depending on the starting point. Data collection is likely to be challenging since a lender would have to delay debt recovery deliberately for enabling such an experiment. Other than these data challenges, there may be a few operational and legal factors (e.g., jurisdictions differing in their legal processes) that influence the workout length, which could require a more sophisticated portfolio loss model than the one used in this study. Therefore, pursuing the best starting point of debt and/or legal proceedings is left as an avenue of future investigation.

### 5.3.1 Optimisation results using $S_1$ , $S_2$ , and $S_3$ respectively

The first set of results are presented in Fig. 5.4 wherein all loss curves exhibit minima at certain thresholds  $d^*$  when loss-optimising the recovery decision on the full sample  $S_1$ , i.e., results based on the first column in Table 5.4. In this case, optimising on  $S_1$  represents a historically lower-risk portfolio, while training forecasts from  $\{S_1, S_2, S_3\}$  represents increasingly dire credit risk scenarios in future. Specifically, the loss minimum increases both in value as well as occur at decreasing thresholds as the forecast scenario worsens, i.e., progressing from  $s_{11} \rightarrow s_{21} \rightarrow s_{31}$  when parametrising the forecasting technique. Furthermore, overall losses across all thresholds increase as the forecast scenario deteriorates, which is evidenced by the steeper slope of the loss curve after having reached its minimum at  $d^*$ . This agrees with the intuition of cutting losses sooner rather than later when facing increasingly higher credit risk on future cash flows. Moreover, the  $s_{31}$ -results yielded the lowest thresholds  $d^* = \{4, 6\}$  respective to each technique, whose values seem close to the current practice of using  $d = 3$  with the  $g_0$ -measure as a default definition. Therefore, training forecast models from  $S_3$  may serve as a conservative ‘boundary’ case, thereby deliberately introducing risk aversion when optimising the recovery decision itself.

Regarding the techniques, the base scenario  $s_{11}$  clearly gives two very different loss minima at  $d^* = 5$  for random defaults versus  $d^* = 13$  for Markovian defaults. Incidentally, the latter also

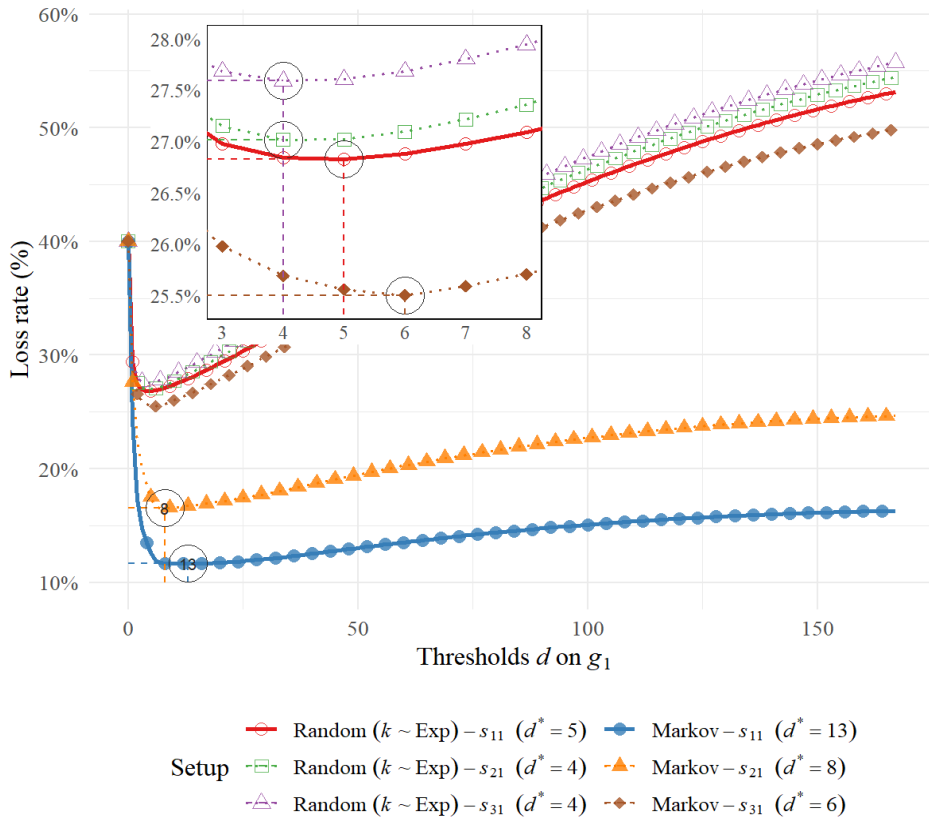


FIG. 5.4: Loss rates across recovery thresholds  $d$  for measure  $g_1$  on the full sample  $S_1$  across various forecasting scenarios, using the random defaults technique with  $k \sim \text{Exp}(\lambda)$  truncation, and using the Markovian defaults technique independently. Solid lines indicate base scenarios wherein both optimisation and training forecasts use the same sample. Zoomed plot and encircled points show global minima for each loss curve, also bracketed in the legend.

yielded lower loss rates at less stringent (higher) values of  $d^*$  in general, when compared to those given by random defaults across all scenarios  $s_{i1}$ . Moreover, the difference between  $d^*$  yielded by each technique becomes smaller as the forecast scenario worsens. In fact, Markovian defaults for  $s_{31}$  gives a loss curve that is similarly shaped to those produced by random defaults irrespective of forecast scenario, which suggests some connection. Consider that the underlying transition matrix in Table 5.2 is generally much more transient than the one in Table 5.1, with far greater conditional probabilities of transiting to worse states. As the possibility of curing back to a better state declines, the Markovian technique increasingly resembles the simpler technique in effect. The latter provides inherently less realistic forecasts since it deliberately ignores curing, which means the larger cash flows associated with curing events are not generated. Since the Markovian forecasts are demonstrably more accurate, they are therefore clearly preferable, though the random forecasts are kept for expositional purposes, including model risk.

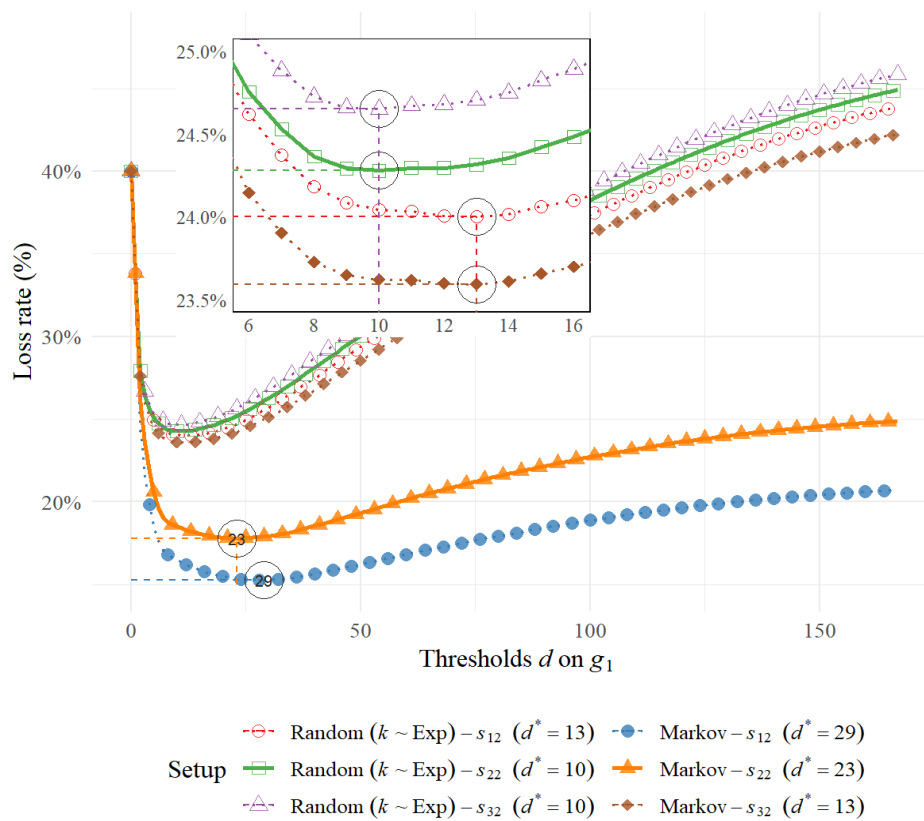


FIG. 5.5: Loss rates across recovery thresholds  $d$  for measure  $g_1$  on the delinquents sample  $S_2$  across various forecasting scenarios, using the random defaults technique with  $k \sim \text{Exp}(\lambda)$  truncation, and using the Markovian defaults technique independently. Graphical formatting follows that of Fig. 5.4.

The same trends hold true in Figs. 5.5–5.6 when optimising on samples  $S_2$  and  $S_3$  instead, i.e., scenarios from the second/third columns in Table 5.4. As the main result, optima are still obtained (though at different locations) across all techniques and forecast scenarios, thereby demonstrating the LROD-procedure’s sensitivity to the inherent risk profile of a portfolio. In particular, optimising across increasingly riskier portfolios ( $S_1 \rightarrow S_2 \rightarrow S_3$ ) remains viable, even if the loss curves become somewhat vertically compressed relative to lower-risk samples. Moreover,  $d^*$  seem to *increase* across riskier samples. The base scenarios  $\{s_{11}, s_{22}, s_{33}\}$ , i.e., the diagonal in Table 5.4, demonstrate this phenomenon with loss minima found at  $d^* = \{5, 10, 35\}$  for random defaults and  $d^* = \{13, 23, 35\}$  for Markovian defaults. That said, the recoveries realised from selling the underlying asset largely explains this phenomenon. Since these recoveries are generally recognised only at the write-off point after a typically long workout period, the suddenly large receipt will dramatically decrease the delinquency level at the last period. Therefore, when optimising on  $S_3$ , it is indeed statistically better to wait strategically and collect some of these

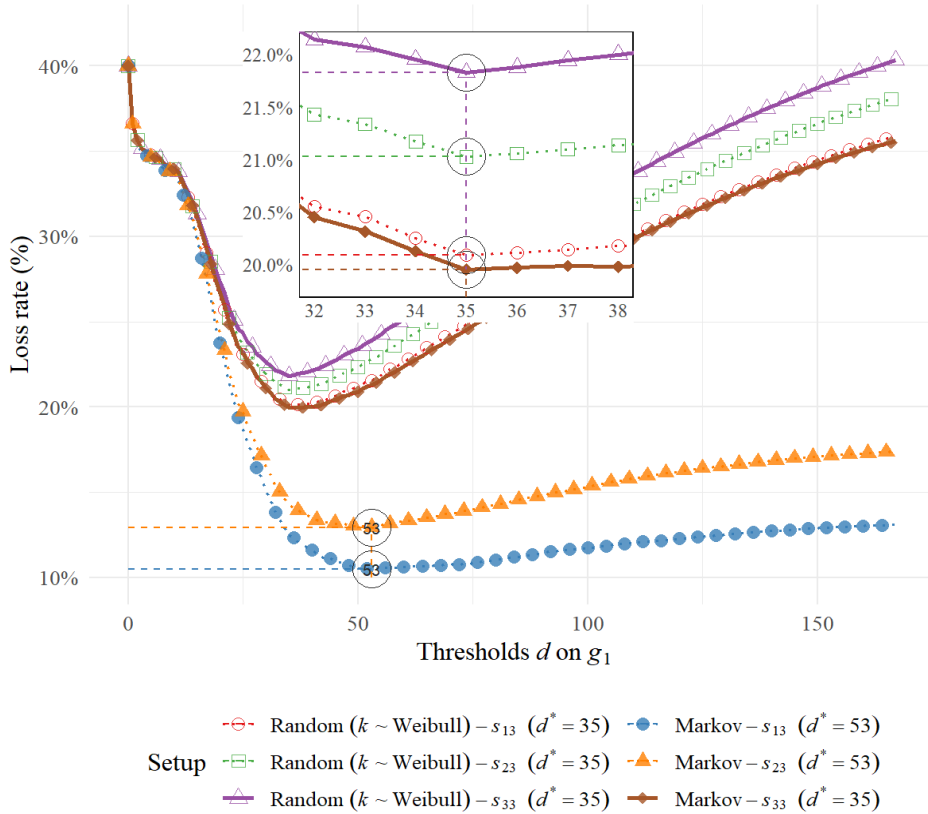


FIG. 5.6: Loss rates across recovery thresholds  $d$  for measure  $g_1$  on the write-offs sample  $S_3$  across various forecasting scenarios, using the random defaults technique with  $k \sim \text{Weibull}(\lambda, \phi)$  truncation, and using the Markovian defaults technique independently. Graphical formatting follows that of Fig. 5.4.

large cash flows. This is evidenced by the relatively high threshold  $d^* = 35$ , which indicates the optimal *ending* point of legal proceedings. The fact that both loss minima and their thresholds change when optimising on  $S_3 \rightarrow S_2 \rightarrow S_1$  merely attests to the dilution of written-off cases as a proportion of the overall portfolio.

While the  $CD$ -measure  $g_1$  is primarily used in this study, loss-optimality was also found for the other two measures  $g_2$  and  $g_3$  across the experimental setup for both forecasting techniques. Interestingly, the  $d^*$  yielded by  $g_2$  and  $g_3$  are much less varied than those yielded by  $g_1$ , which suggests these measures are not as sensitive as  $g_1$  to the choice of technique, risk level, or forecast scenario. Specifically, the loss minima for  $g_2$  and  $g_3$  occur within the threshold ranges  $[1.2, 1.9]$ ,  $[1.3, 2.3]$ , and  $[3.2, 6]$  when optimising respectively across  $\{S_1, S_2, S_3\}$ . However, the loss minima themselves are greater than those yielded by  $g_1$  with the percentage difference thereof averaging at 3.6% (excluding  $S_3$ ). Evidently, the LROD-procedure suggests that the  $g_1$ -measure is objectively the best delinquency measure for signalling loan recovery – at least for this particular

mortgage portfolio.

### 5.3.2 Monte Carlo simulations for analysing the variance of optima

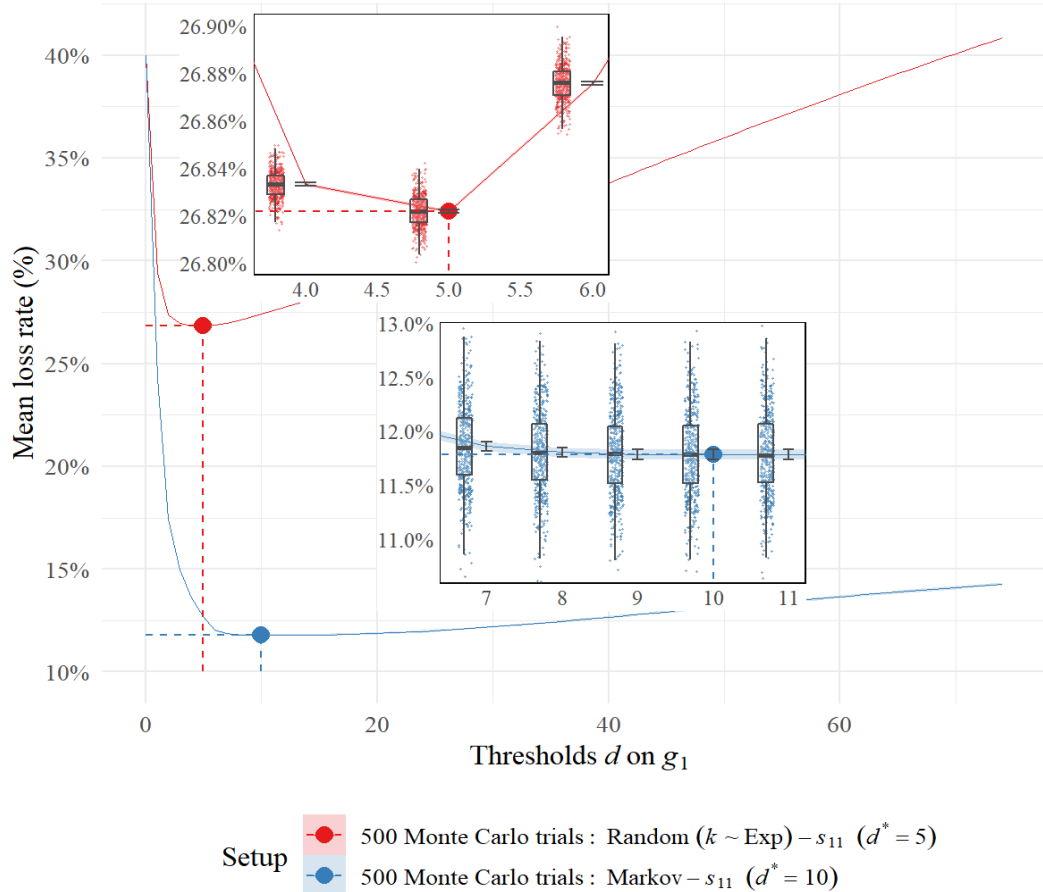


FIG. 5.7: Average loss rates (solid lines) across thresholds  $d$  for measure  $g_1$ , estimated from  $n = 500$  Monte Carlo trials, for scenario  $s_{11}$ . Forecasts are iteratively and independently made using the random defaults technique with  $k \sim \text{Exp}(\lambda)$  truncation, and the Markovian defaults technique. The averages are accompanied by a 99% shaded confidence band with error bars. Zoomed plots show global minima for each loss curve, also bracketed in the legend. Box-and-whiskers mini-plots within the zoomed plots summarise the overlaid loss estimates at each  $d$ .

Given that forecast receipts are inherently probabilistic, their subsequent use within a delinquency measure injects uncertainty into the latter's output as well as into the optimisation itself. Therefore, any loss minimum that is found at a certain threshold may, in fact, be spurious. As an example, a random but systemic perturbation at some time point in the underlying forecasts can produce an alternative minimum at an entirely different threshold, which has implications for the overall precision of the optimisation. Confidence in this supposed minimum can be enhanced by conducting a variance study on the loss curve. One approach to this problem is to produce

multiple sets of forecasts of the portfolio's future cash flows using simple Monte Carlo simulation and the laws of large numbers. Each iteration thereof will have its own independent loss curve using a particular set of random forecasts generated from a specific technique. As an example, consider  $n$  such Monte Carlo trials, thereby resulting in  $n$  loss rate estimates at each threshold  $d$ , from which a sample mean  $\mu_d$  is calculated at each  $d$ . The corresponding sample variance  $s_d^2$  is estimated, which is finally used in constructing a standard 99% confidence interval for the mean as  $\mu_d \pm 2.58 s_d/\sqrt{n}$ .

Monte Carlo simulation is illustrated in Fig. 5.7 for both forecasting techniques using the  $s_{11}$  base scenario after 500 runs. The forecasts yielded by the simpler technique appear to be quite robust since the resulting loss rates had relatively little variation and retained the overall shape of the original loss curve in Fig. 5.4. These results, particularly the difference in the widths of the confidence intervals per technique, attest to the *bias-variance* trade-off phenomenon in statistical learning, as the model's complexity varies. More specifically, the simpler technique with its relatively invariant forecasts is also much less accurate than the Markovian technique. Reassuringly, the lowest sample mean still occurs at  $d^* = 5$  as it did previously. However, the same cannot be said for the Markovian forecasts since the minimum now occurs at  $d^* = 10$  (down from the previous  $d^* = 13$ ). While the overall shape of the Markovian loss curve is still the same, the loss rate estimates exhibit greater variance than those of the simpler technique. Moreover, the loss curve is relatively flat in the region near  $d^* = 10$  (as in Fig. 5.4), which helps explain the 'ease' at which the minimum shifted in this particular scenario.

Conducting these Monte Carlo simulations clearly refines the LROD-procedure one step further by controlling for the uncertainty within forecasts. That said, it is not necessarily true that the average minimum loss will *always* occur at a different threshold, as it did in Fig. 5.7. In fact, the average minima remained at the same thresholds as they did in Figs. 5.5–5.6, when running these Monte Carlo simulations for the other base scenarios  $s_{22}$  and  $s_{33}$  in Fig. 5.8, regardless of forecasting technique. Moreover, the general shape of each loss curve in Fig. 5.8 remained the same, all of which provides combined assurance on the precision of the optimisation results. Lastly, the practitioner may consider a smaller and more focused range of thresholds, especially within the general region of optima, when conducting these Monte Carlo simulations in practice. In contrast, a larger range is chosen in this study simply to demonstrate the LROD-procedure (and its viability) as a "proof of concept".



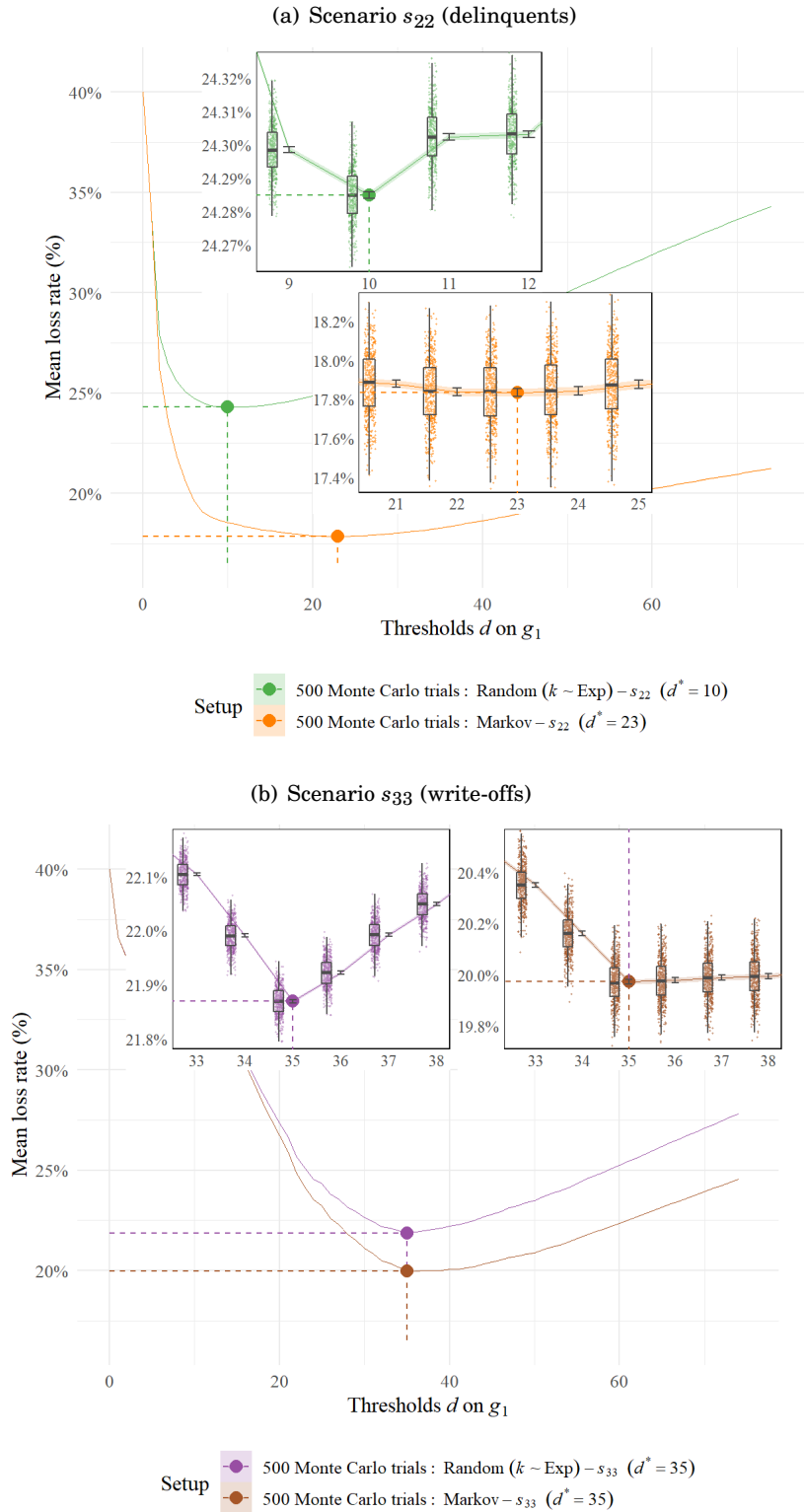


FIG. 5.8: Average loss rates (solid lines) across thresholds  $d$  for measure  $g_1$ , estimated from  $n = 500$  Monte Carlo trials, for scenarios  $s_{22}$  and  $s_{33}$ . Forecasts are iteratively and independently made using the random defaults technique with  $k \sim \text{Exp}(\lambda)$  truncation in (a) and with  $k \sim \text{Weibull}(\lambda, \phi)$  truncation in (b), with Markovian forecasts provided in both panels. Graphical formatting follows that of Fig. 5.7.

## 5.4 Concluding remarks

The timing of the recovery decision is empirically illustrated as a delinquency-based optimisation problem, such that loans are forsaken neither too early nor too late, if at all. This empiricism, however, has to contend with the non-trivial practicalities of a real-world portfolio, most notably that of extensive right-censoring wherein most loan accounts have not yet reached their full contractual maturity. The LROD-procedure, previously presented in section 3.4, only caters for ‘completed’ uncensored loan portfolios, which poses an additional challenge<sup>2</sup>. While an uncensored portfolio would be ideal, the paucity of both data and lenders willing to avail sufficiently rich data makes this difficult. Moreover, most portfolios are actively grown by banks, which causes right-censoring and implies that recovery optimisation will likely remain problematic in practice. A more feasible remedy is demonstrated wherein available data is first used to forecast the residual cash flows of each account up to its contractual maturity. This step ‘completes’ the portfolio and enables the practical use of the LROD-procedure for optimising the bank’s recovery decision.

As a secondary contribution, two forecasting techniques are proposed, parametrised, and applied on the portfolio before optimisation. This includes a simple probabilistic technique and a more sophisticated eight-state Markov chain, both of which are subsequently used in forecasting cash flows independently. However, the manner in which receipts are forecast will greatly affect the portfolio’s subsequent credit risk profile, which influences the timing of loan recovery at which the minimum loss is subsequently attained. Accordingly, forecasts are artificially differentiated by training them from different account subsets (or samples), where each sample contains a progressively greater proportion of delinquent accounts by design. Effectively, each sample approximates a different risk composition typically found in reality, e.g., mortgages vs. unsecured loans, as if each sample is a stand-alone portfolio. Furthermore, the sample that is optimised may differ from the sample from which forecasts are trained. This simulates the reality of a portfolio’s historical risk composition changing in the future by forecasting receipts accordingly, while also making more efficient use of data. Additionally, this setup aligns with IFRS 9 by using various macroeconomic scenarios when estimating expected losses.

Within each scenario of this experimental setup, a so-called ‘Goldilocks’-region is found that contains an *ideal* delinquency threshold at which the portfolio loss is minimised. This setup demonstrates that the LROD-procedure is sensitive to the historical risk profile of a portfolio. Moreover, riskier forecasts yield smaller (or more stringent) optimal delinquency thresholds, which agrees intuitively with cutting losses sooner rather than later as risk expectations deteriorate. Another contribution is that of a Monte Carlo-based refinement to the procedure that can provide additional assurance on the stability of optima, especially given the uncertainty

---

<sup>2</sup>Ignoring the censoring is ill-advised since it leads to unusable results, see Appendix A.3

underlying all forecasts. To this point, the choice of forecasting technique itself affects recovery optimisation, which is demonstrated by the significant differences between each technique's optima. By design, the Markovian technique is much more realistic since it allows for curing backwards to lower delinquency levels, which affects the size of forecasts. Conducting a 5-fold cross-validation further verified the superior quality of Markovian forecasts. However, the simpler technique is retained for expositional purposes since it clearly demonstrates the dangers of model risk when forecasting. That said, an ensemble of forecasting techniques suggests that a meta-learning approach may be viable, which can certainly be further examined in future work. For example, optima can be averaged across technique and forecast scenario using a weighting scheme of sorts. Besides optimisation, there is also practical value for developing these forecast models within an IFRS 9-compliant loss provisioning context.

Regarding limitations, historical cash flows are surely affected by past collection strategies (and their subsequent success or failure) that were employed by the bank at the time. Therefore, training a forecast model from the same data carries the unavoidable risk of embedding the effects of previous strategies into the optimisation, as additional data 'noise'. Future research can perhaps focus on controlling for the bank's strategic influence on these cash flows over time when forecasting receipts. Another avenue of future study is to explore a finer-grained segmentation scheme during the optimisation step. Partitioning data into three increasingly riskier samples correctly assumes homogeneity within each sample. However, recovery decision times can surely be further optimised within certain segments of the portfolio, instead of yielding a portfolio-wide criterion. This may attenuate the LROD-procedure further to the idiosyncrasies of a portfolio, though one will have balance greater segmentation against too little data within a segment. Furthermore, future work can certainly explore a less censored (and therefore richer) portfolio of shorter-term loans, perhaps from different epochs of time. Doing so can reduce the necessary forecasting extent as well as improve the forecasting ability. Lastly, future studies can expand upon the current loss model by incorporating dynamic cost components more explicitly, e.g., funding costs. The static loss rates  $r_E$  and  $r_A$  may be converted into proper LGD-models instead such that loss rates are estimated from the time of entering the  $(g, d)$ -default state. Pursuing this particular avenue will likely intersect with the literature on credit risk modelling and IFRS 9, which can enhance model sophistication given that the field itself is currently in vogue.



*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## CONCLUSION

**B**ankers and merchants have plied their trades for millennia in what is ostensibly a commercial symbiosis. The tradesman's wealth invariably gravitates to the security and convenience of the banker's vaults, which initiates the cycle of trust. In fact, the banker's integrity is what attracts deposits in the first place; exemplified by the ancient temples and their divine sanctity, the state-owned grain banks spread across Ptolemaic Egypt, the Bank of Delos within Apollo's temple, and the Catholic castles of the Knights Templar. With capital pooled together, the banker is undoubtedly better placed to facilitate transactions amongst his depositors. Merchants soon trusted their bankers to conduct exchanges amongst the many currencies in which they traded. However, flourishing trade also increases the overall demand for money, which is precisely when bankers transform stored wealth into loan assets, thereby easing the money demand in what is perhaps their greatest feat. It is this last role that establishes the banker more as a catalyst for trade than its enabler, e.g., granting business loans for risky sea-faring ventures. Lending completes the cycle of trust wherein credit flows freely between depositor and banker and again between bank and borrower. This symbiosis continues *in perpetuum* unless trust erodes sufficiently between any two agents in this triad.

Although the history of banking dates back to antiquity, the use of statistical and mathematical models in driving decision-making has only entered the discourse during the last few decades. Credit risk still poses the single largest source of bank risk, especially since the credit losses from even a few defaulting borrowers may trigger a liquidity crisis – or even an outright bank failure. Accordingly, reserving capital and raising write-off provisions are perhaps the greatest examples of model-driven decision-making in modern-day banking. Another example is that of application

scorecards, which have revolutionised and largely automated the credit granting decision in the 1960s. This advent has both spurred and was made necessary by the extreme growth of consumer credit since that time. The recent introduction of IFRS 9 requires even greater sophistication when modelling expected losses, which only further embeds the use of statistical models beyond the level already required by Basel II.

The breakdown of trust between bank and borrower has remained a largely unavoidable risk, despite the increasing prevalence of statistical models. While there are certainly model-based strategies for managing credit risk, most of these strategies arguably depend on first tolerating a certain level of eroded trust before their activation. It is at this level (or threshold) at which confidence in loan repayment is supposedly lost entirely. As a consequence, the bank renounces the relationship and proceeds with debt recovery, presumably safe in its assumption that retaining the impaired loan asset any longer will inevitably lead to deeper delinquency. The idea of reaching a so-called "point of no return" is believed to be the historical basis of a "default definition", predating the prescripts of Basel II and most related regulations. However, it is argued that the various contexts and jurisdictions in which 'default' is used (or decreed) today has made the very concept thereof utterly incoherent in trying to serve so many 'masters' at once. Even if the 'default' point becomes uniform across all contexts and jurisdictions, the optimality thereof remains questionable and without objective evidence.

A more fundamental meaning of 'default' is explored in this thesis; posited as the risk-based "point of no return" beyond which loan recovery becomes sub-optimal, thereby answering Question 1 in part. Any impairment in a borrower's repayment ability may either persist indefinitely or simply turn out to be a momentary weakness. Deciding exactly where this dichotomy fractures, however, is a non-trivial decision. On the one hand, sufficient patience may afford some distressed borrowers enough time to recover and ultimately to resume their repayments. However, too much patience may prove naive and costly, especially since the inevitably greater arrears will require more capital, thereby crowding out new loans. To specify a 'default' threshold therefore serves as a margin of tolerance towards accumulating arrears. In this study, the admittedly abstract notion of trust and its subsequent erosion is first made concrete using a mathematical delinquency measure  $g$ . In this regard (and that of Question 3), a few measures are re-examined in this work, followed by presenting three refined measures. Secondly, the 'default' point is reinterpreted as simply exceeding a variable threshold  $d$  upon the domain of  $g$ , which better aligns with the rather probabilistic idea of breaching the aforementioned "point of no return". As such, a novel optimisation procedure (LROD) is contributed that yields the *ideal* time (or delinquency threshold) for recovering debt, thereby answering Question 2.

The LROD-procedure weighs two competing interests against each other: extracting the residual revenue from troubled loans versus the risk-adjusted cost thereof. Each distinct ( $g, d$ )-

---

pair serves as a candidate collection policy that carries a "net cost" if applied to a portfolio. Keeping the choice of  $g$  constant, the procedure iterates across values of  $d$  (or 'policies') and calculates the overall portfolio loss of each  $d$ . Doing so produces a loss curve for each  $g$ , which is then inspected for a certain threshold at which the lowest loss occurs, thereby concluding the optimisation. By way of its formulation, the LROD-procedure indirectly facilitates the objective comparison and evaluation of competing delinquency measures, which satisfies Question 4. Alternative measures may better suit a given portfolio's recovery optimisation, or even enhance risk modelling more broadly. However, establishing the *best* measure conclusively would be data-intensive and costly in its own right, and therefore left as future work.

The LROD-procedure is tested by conducting a comprehensive computational study to examine its optimisation ability. To this end, a simulation-based system (or testbed) is devised wherein loan portfolios are meaningfully generated by systematically varying the underlying parameters. Using this testbed, the different types of portfolios wherein threshold optima are found can be broadly determined across the entire credit risk spectrum, in line with Question 5. In fact, the results demonstrate that optimising the recovery decision's timing is viable across most levels of default risk<sup>1</sup>, though to different degrees. Furthermore, systematic defaults<sup>2</sup> occurring during an economic downturn are shown to affect recovery optimisation; itself affected by the extent of a portfolio's loss experience (or LGD). Lastly, recovery optimisation is explored on more turbulent portfolios wherein borrowers repay intermittently, thereby causing episodic delinquency. In this case, postponing loan recovery in response to greater turbulence is demonstrably the strategic optimum (though only up to a point) since it affords greater scope to collect upon sporadic repayments.

Real-world loan portfolios are often right-censored in that many loan accounts have not yet reached contractual maturity. This is unsurprising given that most portfolios are being actively grown by banks, thereby causing 'perpetual' right-censoring (unless loan origination stops entirely). However, the LROD-procedure itself was designed for 'completed' portfolios and *not* treating the inherent right-censoring yields unintuitive results, as explored in section A.3. Furthermore, guiding subsequent policy design based on such flawed results is downright dangerous. As a remedy, a forecasting step is introduced into the procedure, wherein the residual cash flows of each censored loan are first forecast up to its contractual maturity. Accordingly, this step enables the practical and feasible application of the LROD-procedure on a now-completed portfolio, thereby answering Question 6.

Armed with forecasts trained from real-world data, the results show that riskier forecasts

---

<sup>1</sup>In this case, default risk is measured by the payment probability  $b$ , as defined in section 4.1.

<sup>2</sup>The notion of  $(k, g)$ -truncation approximates the idea of systematic defaults in the testbed; see section 4.1.

lead to earlier recovery times<sup>3</sup> during optimisation. This finding agrees intuitively with cutting losses sooner rather than later, especially when risk expectations deteriorate. Another factor that affects the optimisation is that of choosing (and calibrating) a forecasting technique. A more sophisticated technique (like the Markovian method) produces more accurate forecasts, thereby increasing the credibility of the LROD-procedure's optima. However, having compared a few forecasting techniques by design, an ensemble of forecasting techniques may actually be more useful than a single technique; a worthy avenue of future research. Regarding Question 7, the factors that influence optimisation include the portfolio's historical risk composition, its subsequent forecast (i.e., macroeconomic or other strategic factors), and the level of modelling sophistication. Lastly, the uncertainty underlying any forecast may destabilise subsequent threshold optima. As a solution (that satisfies Question 7), the censored cash flows are repeatedly forecast in an additional Monte Carlo-based step in the LROD-procedure. This approach allows one to analyse the variance of both forecasts and resulting optima, thereby inspiring greater confidence.

The thesis of the recovery decision and its timing is successfully demonstrated as a nonlinear optimisation problem, both theoretically and empirically. Balancing risk against reward is at the core of the LROD-procedure such that loans are forsaken neither too early nor too late, if at all. This study has significant implications for the policy design of most banks, especially in tweaking their collection policies. In particular, the quantitative aspects thereof can be better informed using the LROD-procedure than relying on arbitrary discretion alone. Moreover, related business strategies and certain default-driven models (e.g., application credit scorecards, pricing and collection models) can be similarly enhanced. Another more fundamental implication is that a default definition's quantum (e.g., 90+ DPD) can itself be optimised within capital modelling or broader portfolio management. That said, optimising 'default' in this way undoubtedly raises legal questions about the contractual design of credit agreements, not to mention the existing regulatory prescriptions relating to 'default'. However, there remains little scientific evidence for the supposed optimality of using 90+ DPD as a default criterion in the first place. In this regard, my work attempts to bridge the gap in literature between credit risk modelling and collection optimisation. Using the LROD-procedure may become standard practice in time, which is certainly preferable to regulators prescribing 'default' by fiat alone. In turn, optimising a definition in this way would likely have a tremendous impact on credit risk modelling and its improvement.

Regarding limitations, the LROD-procedure currently assumes homogeneity in that a single threshold is sought that serves as a portfolio-wide criterion. This assumption may be relaxed in future work by exploring a simple segmentation scheme, thereby yielding a segment-specific optimised threshold. However, finer segmentation must be balanced against having too little

---

<sup>3</sup>When slotting in riskier forecasts, the optimised thresholds decreased in value; see subsection 5.3.1.



---

data (as a result) when attenuating the procedure to a portfolio. Furthermore, one can refine the procedure's current loss model by converting its static components into dynamic probabilistic models, conditioned on a given  $(g, d)$ -default state, e.g., converting the static loss rate  $r_E$  into a model  $r_E(g, d)$ . Calculating the realised LGD requires that cash flows be observed from a certain starting point, which will naturally vary with the value of  $d$ . Regarding calibration, a future study can further validate the LROD-procedure by exploring a less censored portfolio of shorter-term loans. While doing so will surely lessen the forecasting burden, it will not resolve another challenge, i.e., controlling for a bank's own influence in past data. Since historical cash flows are inevitably affected by past collection strategies, any model trained on this data will likely embed the effects thereof into the subsequent optimisation, at least to some degree. Therefore, future work can devise and examine remedies for what basically amounts to a bit of data noise or 'contamination'. Lastly, future researchers can compare the output from the LROD-procedure to that of a typical roll rate analysis. However, the latter cannot directly compete with the former given the considerable differences in sophistication.

In conclusion, there is no doubt that the erosion of trust will remain fixed in the banker's mind. As the economy waxes and wanes over its many cycles, so too will its influence on the borrower's disposable income; itself subject to other life-altering events. The borrower's measured ability to service his debt cannot be an absolute nor can it be time-invariant. Moreover, future defaults may soon be predicted even better, given the current era of machine learning and overall greater sophistication in risk modelling. For these reasons, isolating the ideal default point every so often should henceforth be structured as a modelling exercise in its own right, especially when modelling credit risk. Such an exercise certainly fits the banker's supposed prerogative in defining what is essentially a dynamic and context-sensitive 'true' default state.



*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*



## ANCILLARY MATERIAL ON VARIOUS UNRELATED SUBTOPICS

The following sections are ancillary to this study. This includes an illustration of using Markov theory in base loss reservation in section A.1. A few delinquency measures are compared in section A.2 using a simple two-loan case study. The dangers of *not* treating a real-world portfolio for right-censoring during recovery optimisation are demonstrated in section A.3. In section A.4, various statistical distributions are fit to the maximum delinquency observed at the account-level using a real-world portfolio.

### A.1 An example of loss reservation using Markov theory

The theorems developed in Cyert et al. (1962) using Markov theory can be simplistically applied to loss reservation as an illustration of using a Markov chain in practice. In a nutshell, loss reservation involves calculating the relevant absorption probability and applying it to a balance vector, which is then modified by adding/subtracting a few standard deviations as needed. This is achieved firstly by re-ordering the states in the transition matrix  $P$  (as estimated in Eq. 3.3) so that the two absorbing states are put first, i.e., the settled state  $\bar{0}$  and the write-off state  $n$ , followed by the remaining transient delinquency states  $0, 1, \dots, n - 1$ . As a simple though unrealistic example, suppose  $n = 2$  with the correspondingly re-ordered matrix  $P$  given as

$$P = \begin{matrix} & \bar{0} & 2 & 0 & 1 \\ \bar{0} & \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .2 & 0 & .6 & .2 \\ .4 & .2 & .3 & .1 \end{pmatrix} \end{matrix} \quad (\text{A.1})$$

Following Kemeny-Snell notation, the matrix  $P$  can be partitioned as

$$P = \left[ \begin{array}{c|c} I & O \\ \hline R & Q \end{array} \right], \quad (\text{A.2})$$

where  $I$  is the  $2 \times 2$  identity matrix;  $O$  is a  $2 \times n$  zero matrix;  $R$  is an  $n \times 2$  absorption matrix; and  $Q$  is an  $n \times n$  matrix of the remaining transient states. The inverted matrix  $N = (I - Q)^{-1}$  is the so-called *fundamental* matrix of the absorbing Markov chain. Finally, the entries of the resulting  $n \times 2$  matrix  $NR$ , yield the absorption probabilities into either state  $\bar{0}$  or  $n$ . Using the example matrix  $P$  from Eq. A.1, the corresponding partition matrices are

$$R = \begin{bmatrix} .2 & 0 \\ .4 & .2 \end{bmatrix}, \quad Q = \begin{bmatrix} .6 & .2 \\ .3 & .1 \end{bmatrix}, \quad \text{and} \quad N = (I - Q)^{-1} = \begin{bmatrix} .4 & -.2 \\ -.3 & .9 \end{bmatrix}^{-1} = \begin{bmatrix} 3 & .67 \\ 1 & 1.33 \end{bmatrix}. \quad (\text{A.3})$$

The absorption probabilities in  $NR$  are then calculated as

$$NR = \begin{bmatrix} 3 & .67 \\ 1 & 1.33 \end{bmatrix} \begin{bmatrix} .2 & 0 \\ .4 & .2 \end{bmatrix} = \begin{bmatrix} .87 & .13 \\ .73 & .27 \end{bmatrix}. \quad (\text{A.4})$$

If  $\boldsymbol{\beta} = (B_0, B_1, \dots, B_{n-1})$  represents a balance vector of balances at time  $t$  classified across all  $0, 1, \dots, n - 1$  transient delinquency states, then Cyert's first theorem states that the 2-component vector  $\boldsymbol{\beta}NR$  will output the expected settled and write-off amounts respectively. As an example, suppose that  $\boldsymbol{\beta} = [70 \quad 30]$  is the balance vector observed at  $t$  for the transient states  $B_0$  and  $B_1$ , in which case,  $\boldsymbol{\beta}NR$  will yield 82.67 and 17.33 as the expected settled and write-off amounts respectively. Furthermore, if  $b$  is the sum of all elements in  $\boldsymbol{\beta}$  and  $\boldsymbol{\beta}' = \frac{1}{b}\boldsymbol{\beta}$  is the probability vector denoting the fraction of balances observed across all transient delinquency states, then Cyert's first theorem caters for the variance as follows. Let  $A$  be a two-component vector that gives the variances of settlements and write-offs (per example), which is formally defined as

$$A = b(\boldsymbol{\beta}'NR - (\boldsymbol{\beta}'NR)_{sq}), \quad (\text{A.5})$$

where  $(\bullet)_{sq}$  is a matrix operation that outputs the square of each element. If  $\boldsymbol{\beta}' = [.7 \quad .3]$  as an example, then the corresponding variances using Eq. A.5 are calculated as

$$\begin{aligned} A &= 100 \left( [.7 \quad .3] \begin{bmatrix} .87 & .13 \\ .73 & .27 \end{bmatrix} - \left( [.7 \quad .3] \begin{bmatrix} .87 & .13 \\ .73 & .27 \end{bmatrix} \right)_{sq} \right) \\ &= 100 \left( [.8267 \quad .1733] - \left( [.8267 \quad .1733] \right)_{sq} \right) \\ &= [14.33 \quad 14.33]. \end{aligned} \quad (\text{A.6})$$

This particular theorem allows for customising the loss reserve beyond its expectation at 17.33 in a simple way, e.g., raising it by one standard deviation  $\sqrt{14.33} = 3.79$  as  $17.33 + 3.79 = 21.12$ .

Loan age	Small Loan History 1	Large Loan History 1	Small Loan History 2	Large Loan History 2
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	265	4,000	265	4,000
4	795	12,000	397.50	6,000
5	300	4,528.30	132.50	2,000
6	0	0	0	0
7	0	0	0	0
8	265	4,000	265	4,000
9	795	12,000	397.50	6,000
10	300	4,528.30	132.50	2,000
11	0	0	0	0
12	0	0	0	0
13	265	4,000	265	4,000
14	795	12,000	397.50	6,000
15	300	4,528.30	132.50	2,000
16	0	0	0	0
17	0	0	0	0
18	265	4,000	265	4,000
19	795	12,000	397.50	6,000
20	300	4,528.30	132.50	2,000

TABLE A.1: Hypothetical repayment histories (denominated in ZAR) for two loan accounts: a case study.

## A.2 Illustrating three delinquency measures: a case study

Consider two 20-month amortising loans with the same interest rate as in Sah (2015), but with significantly different instalments: a small loan and a large loan. Assume a fixed instalment of ZAR 265 (with principal ZAR 5,000) and ZAR 4,000 (with principal ZAR 75,471.54), both using a continuously compounded interest rate  $\delta = 6.7\%$  expressed per annum. Further assume that the receipt elements  $R_1, R_2, \dots, R_T$  coinciding with loan periods  $1, 2, \dots, T$  of each loan account are generated by one of two possible vectors containing a specific cyclic pattern, with  $I \in \{265, 4000\}$  denoting the instalment of each loan:  $[0 \ 0 \ I \ 3(I) \ 1.13(I)]$  for a particular receipt history;  $[0 \ 0 \ I \ 1.5(I) \ 0.5(I)]$  for another particular receipt history. Applying these two patterns repeatedly will give two hypothetical receipt histories for each loan in populating the elements of  $\mathbf{R}$ , as detailed in Table A.1.

Delinquency is then calculated using a few distinct delinquency measures on these two loans as a case study. Fig. A.1 illustrates the  $g_1$ -measure using a repayment ratio  $z = 90\%$  that was set within reason. Clearly, the  $g_1$ -measure cannot differentiate in delinquency between loans

of different sizes but with proportionally equal receipts, as evidenced by the curves of both the small and large loan being equal to each other. Moreover,  $g_1$  cannot account for the effects of overpayments when assessing delinquency beyond its zero-valued lower bound, i.e., amounts paid in advance. Lastly, the impact of unpaid interest amounts when delinquent is completely disregarded insofar that  $g_1$  considers the incurred delinquency to be independent of its timing within the life of the loan. From a cash flow perspective of an amortising loan, non-payment at larger  $t$ , i.e., closer to contractual maturity, should reasonably be considered with greater urgency than at smaller  $t$ . The rationale for this is that there is simply less contractual time available to repay any accumulated arrears, without restructuring such a loan.

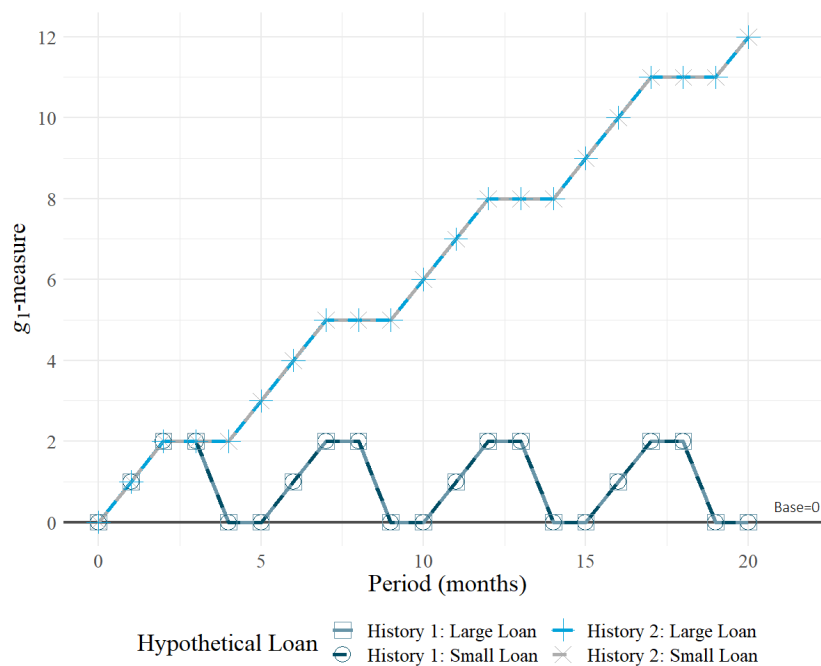


FIG. A.1: The  $g_1$ -measure for a two-loan case study (different loan sizes) with two particular receipt histories over time, as detailed in Table A.1. A reasonable repayment ratio threshold of  $z = 90\%$  is used for illustration purposes.

To this last point, the inability of  $g_1$  to capture the timing of delinquency relative to residual maturity is perhaps best seen when studying Fig. A.2. Specifically, the  $g_2$ -measure is showcased using the same two-loan setup, clearly showing that  $g_2$  can assign greater (or lesser) delinquency-values than  $g_1$  as their particular receipt patterns manifest cyclically over loan life  $t \rightarrow T$  (or  $t \rightarrow 0$ ). This apparent sensitivity of the  $g_2$ -measure to loan life seems sensible when considering it from a cash flow perspective. However, the relative insensitivity of delinquency to earlier loan life may be naive from a risk perspective. Whether this sensitivity is either a boon or a distinct disadvantage is unknown, especially in the absence of a comparative framework. Factors that may be considered within such a framework include loan contract variables (e.g., loan amount, or

interest rate), portfolio and borrower characteristics, and the macroeconomic reality.

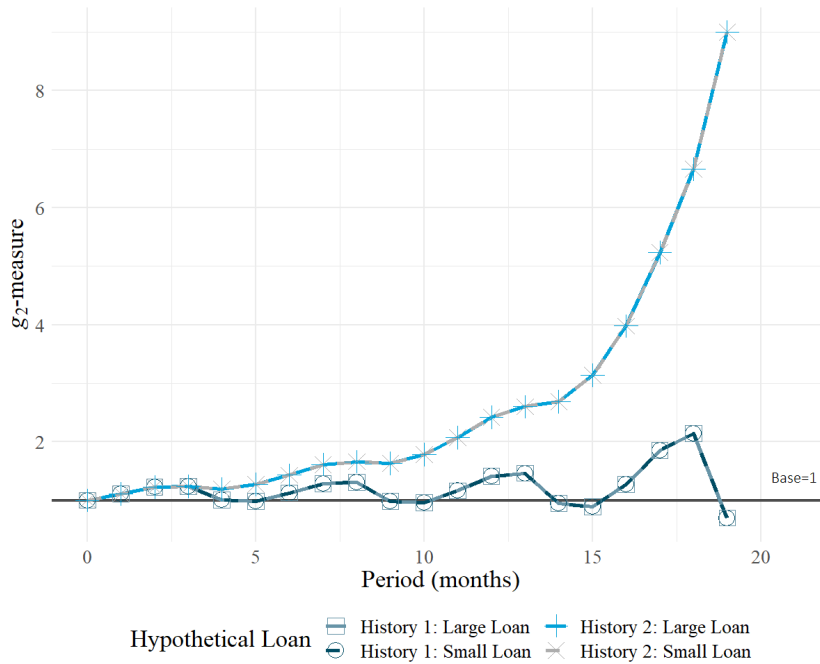


FIG. A.2: The  $g_2$ -measure for a two-loan case study (different loan sizes) with two particular receipt histories over time, as detailed in Table A.1.

In addition, the  $g_2$ -measure accounts for the effects of overpayments beyond the case of zero-valued arrears, which is evident whenever  $g_2(t) < 1$ . Overpayment itself, i.e.,  $R_t > I_t$  at any  $t \in [0, T]$ , may very well be considered positively in that the borrower has beaten the monthly expectation, which leaves the bank with marginally more cash and with a slightly better liquidity position. On the other hand, overpayment can become a profit-related concern since habitual overpayment can cause overall loan life to become shortened. This will in turn inhibit the force of interest and therefore compromise profit margins. Regardless, the  $g_2$ -measure incorporates both under- and overpayments as apparent polar opposites in its assessment of delinquency, which seems more flexible when compared to the  $g_1$ -measure.

Despite these improvements over  $g_1$ , the  $g_2$ -measure still cannot differentiate in delinquency between the proportionally equal receipts of the small and the large loan. This is clear from Fig. A.2 where  $g_2(t)$  of either loan is equal to each other at every time period  $t$ . On the other hand, the  $g_3$ -measure remedies this and is sensitive to loan principal sizes when assessing delinquency. This is evident in Fig. A.3 using the first<sup>1</sup> receipt history across three sensitivities  $s \in [0\%; 50\%; 100\%]$ , wherein  $g_3(t)$  differs at every period  $t$ , for both loan sizes. This result strongly suggests that  $g_3$  can uniquely pivot the delinquency assessment itself in accordance to the

<sup>1</sup>The second receipt history is excluded since it has very similar results.

severity of the unpaid (or underpaid) instalment, relative to other instalments. Amongst other factors, the disruption in cash flow of a particular loan is therefore assessed on its magnitude relative to other differently-sized loans.

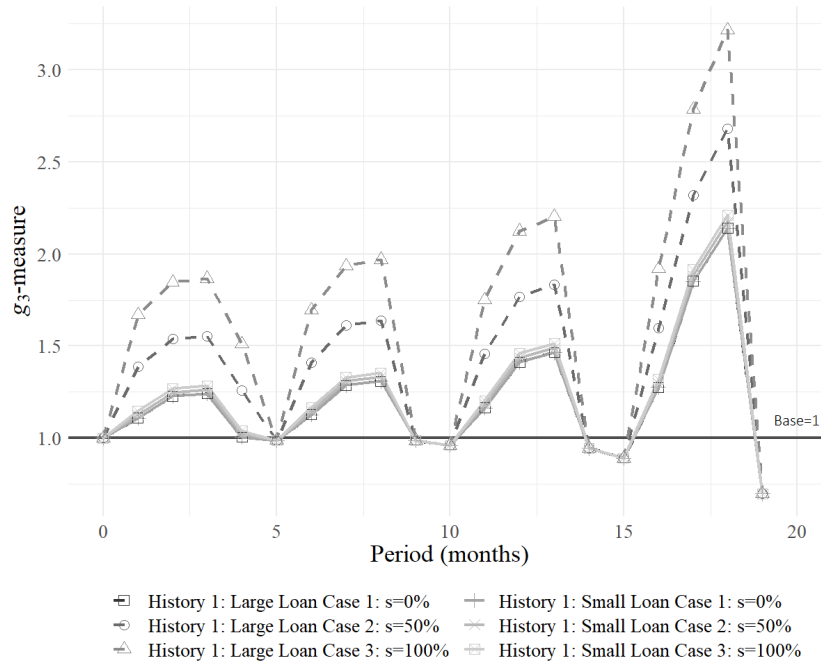


FIG. A.3: The  $g_3$ -measure for a two-loan case study (different loan sizes) using the first receipt history as detailed in Table A.1, for sensitivities  $s \in [0; 50\%; 100\%]$ .

On interpretation, the  $g_1$ -measure is best interpreted as the  $z$ -weighted number of payments in arrears, as weighed by the risk appetite of a lender towards accrued arrears. The  $g_2$ -measure is the *extent* at which the expected duration is increased (or penalised/contorted). As a simple example, the value  $g_2(t) = 1.5$  at a particular  $t$  represents a 50% penalty in the expected time to recover the loan capital originally lent by the bank, accounting for the residual maturity, arrears balance, and the force of interest. Furthermore, the  $g_3$ -measure carries a very similar meaning in its output, albeit merely nuanced/inflated given a discretionary sensitivity  $s$  towards loan principal differences. Lastly, it is quite clear that these delinquency measures  $g_1$  vs.  $g_2$  and  $g_3$  represent different things and even operate on different measurement scales (interval vs. ratio scales respectively). As such, their measurements cannot be compared *directly* to one another, especially so at the account-level, without becoming meaningless.

### A.3 Failing to forecast before recovery time optimisation

As an illustration, the LROD-procedure devised in section 3.4 is applied on a real-world loan portfolio that is left untreated. This is to say that the receipt vector  $\mathbf{R} = [R_1, R_2, \dots, R_{t_0}]$  only



contains elements as observed from data up to the most recent time point  $t_0 < t_c$ . The remaining future elements  $t_1, \dots, t_c$  are unobservable and deliberately ignored in this illustration to demonstrate the effect of foregoing any forecasting on the results of the LROD-procedure. Furthermore, the balances of each account are observed at relevant time periods and simply multiplied with a static loss rate  $l_\alpha \in [0, 1]$ , as a simpler loss model. More specifically, the most recent balance at time  $t_0$  is used for a  $(g, d)$ -performing account whilst the balance at the default time  $\tau \leq t_0$  is used for a  $(g, d)$ -defaulting account, as signalled by a particular  $(g, d)$ -configuration. In both cases, the observed balance is simply discounted back to time  $t = 1$  (loan origination) using the same 7% risk-free rate. Selecting a range of loss rates at will, the LROD-procedure is then iteratively applied on the entire portfolio. The resulting loss curves are presented in Fig. A.4 using the  $CD$ -measure  $g_1$ . There are no significant differences in the shapes of loss curves for loss rates exceeding 50%.

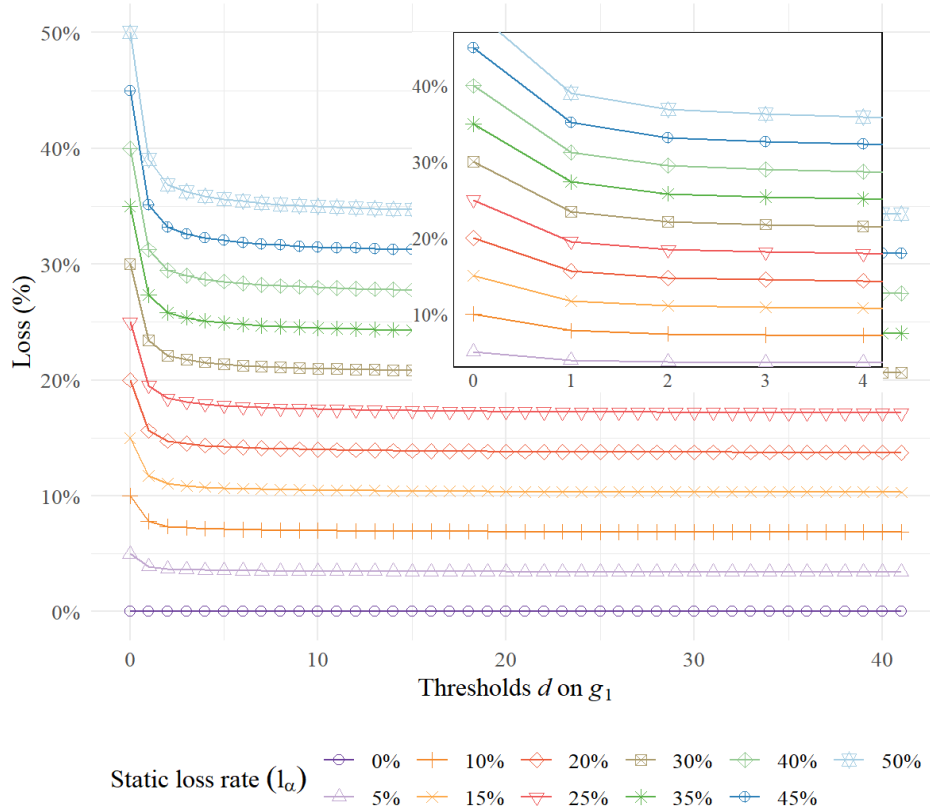


FIG. A.4: Total losses (expressed as a % of the summed principals) across default thresholds  $d$  for the  $CD$ -measure  $g_1$ , using a range of static loss rates  $l_\alpha \in [0, 1]$  and an untreated real-world loan portfolio.

No global minima in losses exist at any particular threshold, regardless of the chosen loss rate. Instead, all losses tend toward a certain asymptote that is influenced by the loss rate,

which renders the optimisation of the recovery threshold a moot point. Moreover, the resulting loss curves suggest that one should simply ignore any accrued delinquency, except for very low thresholds  $d \leq 2$ , which coincide with the greatest losses. Although unusual, consider that the majority of the portfolio's receipts are still pending. The LROD-procedure's particular loss model recognises this and logically suggests never to recover a single account. Regardless, ignoring accrued delinquency at large is intuitively false and ill-advised for a credit risk-based business like a bank. Instead, this result rather attests to the breakdown of the LROD-procedure itself when foregoing the necessary forecasting of a loan portfolio's cash flows.

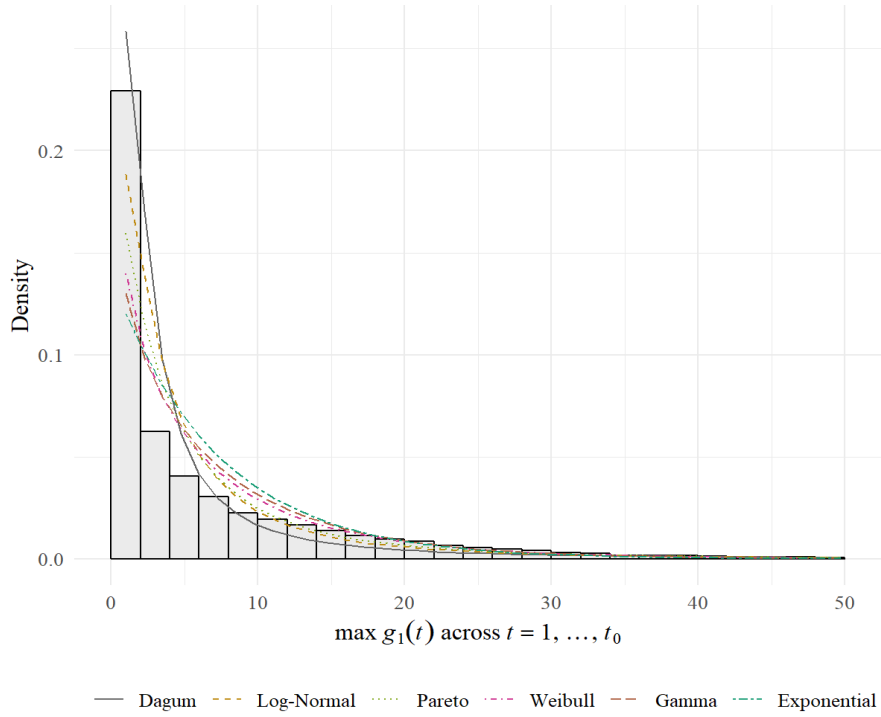
#### **A.4 Fitting statistical distributions to the truncation parameter $k$**

In calibrating the random defaults technique to forecast cash flows, a truncation effect simulates the reality that some accounts will simply never resume payment. This is achieved when the forecast receipts are zero-valued after a certain point  $t_k$  coinciding with a certain  $k$ -threshold in measured delinquency. In estimating this  $k$  truncation parameter, the maximum observed delinquency (using  $g_1$ ) per account is calculated with the resulting empirical distributions of these maxima shown in Fig. 5.3, respective to each sample  $S_2$  (delinquents) and  $S_3$  (write-offs). Several candidate statistical distributions are then fit using maximum likelihood on each respective sample, with the fitted probability density function overlaid on the histogram, as shown in Fig. A.5 for some of these candidates.

In selecting the best fit, both Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit tests are conducted for each candidate distribution against the standard 5% significance level. However, all of the null hypotheses are rejected for both  $S_2$  and  $S_3$ , presumably due to the heavily right-skewed distributions of maxima in both cases. Secondly, the Akaike Information Criterion (AIC) reveals that the Dagum, log-normal, Pareto, Weibull, exponential, and gamma distributions were amongst the best fitting candidates for  $S_2$ . Though the Dagum distribution had the best AIC, the exponential distribution is chosen owing to its simplicity and its somewhat greater popularity in statistical literature. Furthermore, the exponential distribution is strictly decreasing for  $x$ , which is deemed more appropriate given the histogram's shape. Similarly, the AIC for  $S_3$  suggests that the Dagum, Burr (Type 12), Weibull, Gumbel, Gamma, and Logistic distributions were the better-fitting candidates. Of these, the Weibull distribution is chosen since it best approximates the histogram visually without lending too much credence to the left-tail though still yielding a sufficiently heavy right-tail.



(a) Using the delinquents sample  $S_2$



(b) Using the write-offs sample  $S_3$

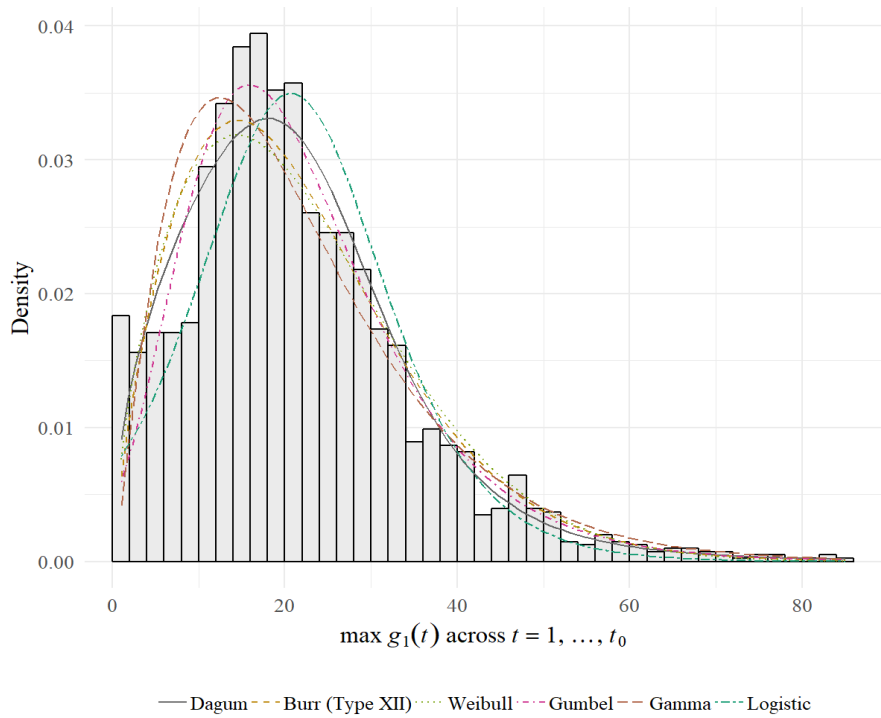


FIG. A.5: Candidate statistical distributions fit on the maximum of the weighted payments in arrears  $\max g_1(t)$  observed per account across historical periods  $t = 1, \dots, t_0$ . These maxima are respectively calculated from the  $S_2$  sample (delinquents) in (a) and  $S_3$  (write-offs) in (b), with a histogram of maxima given in each case.

*This page is intentionally left blank so that all front-matter sections and chapters start on the right-most page when printing double-sided*

## BIBLIOGRAPHY

1. Alghazo, J. M., Kazmi, Z. & Latif, G. (2017). Cyber security analysis of internet banking in emerging countries: User and bank perspectives. *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 1–6. <https://doi.org/10.1109/ICETAS.2017.8277910>
2. Allen, F. & Gale, D. M. (2017). How should bank liquidity be regulated? *Achieving financial stability* (pp. 135–157). [https://doi.org/10.1142/9789813223400\\_0011](https://doi.org/10.1142/9789813223400_0011)
3. Allen, F. & Santomero, A. M. (2001). What do financial intermediaries do? *Journal of Banking & Finance*, 25(2), 271–294. [https://doi.org/10.1016/S0378-4266\(99\)00129-6](https://doi.org/10.1016/S0378-4266(99)00129-6)
4. Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press.
5. Anderson, T. W. & Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, 89–110. <https://doi.org/10.1214/aoms/1177707039>
6. Anginer, D., Demirguc-Kunt, A. & Zhu, M. (2014). How does deposit insurance affect bank risk? evidence from the recent crisis. *Journal of Banking & Finance*, 48, 312–321. <https://doi.org/10.1016/j.jbankfin.2013.09.013>
7. Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3), 203–228. <https://doi.org/10.1111/1467-9965.00068>
8. Ausubel, L. M. (1999). *Adverse selection in the credit card market* (tech. rep.). Working paper, University of Maryland, Department of Economics. <https://pages.ucsd.edu/~aronatas/conference/adverse.pdf>
9. Baesens, B., Rösch, D. & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons.
10. Baltensperger, E. (1980). Alternative approaches to the theory of the banking firm. *Journal of Monetary Economics*, 6(1), 1–37. [https://doi.org/10.1016/0304-3932\(80\)90016-1](https://doi.org/10.1016/0304-3932(80)90016-1)

## BIBLIOGRAPHY

---

11. Basel Committee on Banking Supervision. (2006a). *International convergence of capital measurement and capital standards: A revised framework, comprehensive version* (tech. rep.). Bank for International Settlements. <https://www.bis.org/publ/bcbs128.pdf>
12. Basel Committee on Banking Supervision. (2006b). *The joint forum – the management of liquidity risk in financial groups* (tech. rep.). Bank for International Settlements. <https://www.bis.org/publ/joint16.pdf>
13. Basel Committee on Banking Supervision. (2017). *D403: Prudential treatment of problem assets – definitions of non-performing exposures and forbearance* (tech. rep.). Bank for International Settlements. <https://www.bis.org/bcbs/publ/d403.pdf>
14. Berger, S. C. & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *BuR – Business Research*, 2(1), 39–65. <https://doi.org/10.1007/BF03343528>
15. Besanko, D. & Thakor, A. V. (1992). Banking deregulation: Allocational consequences of relaxing entry barriers. *Journal of Banking & Finance*, 16(5), 909–932. [https://doi.org/10.1016/0378-4266\(92\)90032-U](https://doi.org/10.1016/0378-4266(92)90032-U)
16. Bhattacharya, S. & Thakor, A. V. (1993). Contemporary banking theory. *Journal of Financial Intermediation*, 3, 2–50. <https://doi.org/10.1006/jfin.1993.1001>
17. Board of Governors of the Federal Reserve System. (2020). *Households and Nonprofit Organisations; One-to-Four-Family Residential Mortgages; Liability, Level [HHMSDODNS]* (tech. rep.). Federal Reserve Bank of St. Louis. Retrieved October 24, 2020, from <https://fred.stlouisfed.org/series/HHMSDODNS>
18. Botha, A. (2020a). Simulation-based optimisation of the timing of loan recovery across different portfolios: the LROD-procedure [Source Code]. <https://doi.org/10.5281/zenodo.4005703>
19. Botha, A. (2020b). The loss optimisation of loan recovery decision times using forecast cash flows: the empirical use of the LROD-procedure [Source Code]. <https://doi.org/10.5281/zenodo.4006113>
20. Botha, A., Beyers, C. & De Villiers, P. (2020). The loss optimisation of loan recovery decision times using forecast cash flows. *Accepted by the Journal of Credit Risk*. Retrieved October 12, 2020, from <https://arxiv.org/abs/2010.05601>
21. Botha, A., Beyers, C. & De Villiers, P. (2021). Simulation-based optimisation of the timing of loan recovery across different portfolios. *Expert Systems with Applications*, 177. <https://doi.org/10.1016/j.eswa.2021.114878>

22. Bozzetto, J.-F., Tang, L., Thomas, L. C. & Thomas, S. (2005). *Modelling the purchase dynamics of insurance customers using Markov chains* (tech. rep.). University of Southampton. <https://eprints.soton.ac.uk/36179/1/CORMSIS-05-02.pdf>
23. Bravo, C., Thomas, L. C. & Weber, R. (2015). Improving credit scoring by differentiating defaulter behaviour. *Journal of the Operational Research Society*, 66(5), 771–781. <https://doi.org/10.1057/jors.2014.50>
24. Campbell, J. Y. & Cocco, J. F. (2015). A model of mortgage default. *The Journal of Finance*, 70(4), 1495–1554. <https://doi.org/10.1111/jofi.12252>
25. Chan, Y.-S., Greenbaum, S. I. & Thakor, A. V. (1992). Is fairly priced deposit insurance possible? *The Journal of Finance*, 47(1), 227–245. <https://doi.org/10.1111/j.1540-6261.1992.tb03984.x>
26. Chehrazi, N., Glynn, P. W. & Weber, T. A. (2019). Dynamic credit-collections optimization. *Management Science*, 65(6), 2737–2769. <https://doi.org/10.1287/mnsc.2018.3070>
27. Cohen, B. H., Edwards Jr, G. A. et al. (2017). The new era of expected credit loss provisioning. *BIS Quarterly Review*. <https://EconPapers.repec.org/RePEc:bis:bisqtr:1703f>
28. Corcoran, A. W. (1978). The use of exponentially-smoothed transition matrices to improve forecasting of cash flows from accounts receivable. *Management Science*, 24(7), 732–739. <https://doi.org/10.1287/mnsc.24.7.732>
29. Cressy, R. & Toivanen, O. (2001). Is there adverse selection in the credit market? *Venture Capital: An International Journal of Entrepreneurial Finance*, 3(3), 215–238. <https://doi.org/10.1080/13691060110052104>
30. Crook, J. N., Edelman, D. B. & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465. <https://doi.org/10.1016/j.ejor.2006.09.100>
31. Cyert, R. M., Davidson, H. J. & Thompson, G. L. (1962). Estimation of the allowance for doubtful accounts by Markov chains. *Management Science*, 8(3), 287–303. <http://www.jstor.org/stable/2627386>
32. Davies, G. (2002). *A history of money: From ancient times to the present day* (3rd ed.). University of Wales Press.
33. De Almeida Filho, A. T., Mues, C. & Thomas, L. C. (2010). Optimizing the collections process in consumer credit. *Production and Operations Management*, 19(6), 698–708. <https://doi.org/10.1111/j.1937-5956.2010.01152.x>

## BIBLIOGRAPHY

---

34. De Roover, R. (1963). *The rise and decline of the Medici Bank: 1397-1494*. Harvard University Press.
35. de Jongh, R., Verster, T., Reynolds, E., Joubert, M. & Raubenheimer, H. (2017). A critical review of the Basel margin of conservatism requirement in a retail credit context. *The International Business & Economics Research Journal*, 16(4), 257–274. <https://doi.org/10.19030/iber.v16i4.10041>
36. Dermine, J. (2007). ALM in banking. In S. Zenios & W. Ziemba (Eds.), *Handbook of asset and liability management: Applications and case studies* (pp. 490–541). North-Holland. [https://doi.org/10.1016/S1872-0978\(06\)02011-4](https://doi.org/10.1016/S1872-0978(06)02011-4)
37. Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *Review of Economic Studies*, 51(3), 393–414. <https://doi.org/10.2307/2297430>
38. Diamond, D. W. & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3), 401–419. <https://doi.org/10.1086/261155>
39. Duman, E., Ecevit, F., Çakır, Ç. & Altan, O. (2017). A novel collection optimisation solution maximising long-term profits: A case study in an international bank. *Journal of Decision systems*, 26(4), 328–340. <https://doi.org/10.1080/12460125.2017.1422318>
40. Durand, D. (1941). *Risk elements in consumer instalment financing*. National Bureau of Economic Research.
41. EBA. (2016). *Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013* (tech. rep.). European Banking Authority (EBA). <https://eba.europa.eu/documents/10180/1597103/Final+Report+on+Guidelines+on+default+definition+%28EBA-GL-2016-07%29.pdf/004d3356-a9dc-49d1-aab1-3591f4d42cbb>
42. EBA. (2017a). *Annex to the EBA opinion EBA-OP-2017-17: Report on the use of the 180 days past due criterion* (tech. rep.). European Banking Authority (EBA). <https://eba.europa.eu/documents/10180/2071742/EBA+Op+2017+17+%28Annex+to+the+EBA+Opinion+on+180+DPD%29.pdf>
43. EBA. (2017b). *Op/2017/17: Opinion of the European Banking Authority on the use of the 180 days past due criterion* (tech. rep.). European Banking Authority (EBA). <https://eba.europa.eu/documents/10180/2071742/EBA+BS+2017+17+%28Opinion+on+the+use+of+180+DPD%29.pdf>



44. EBA. (2018). *Commission Delegated Regulation (EU) 2018/171 of 19 October 2017 on supplementing Regulation (EU) No 575/2013 of the European Parliament and of the Council with regard to regulatory technical standards for the materiality threshold for credit obligations past due (Text with EEA relevance)* (tech. rep.). European Banking Authority (EBA). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R0171&from=EN>
45. Edelberg, W. (2006). Risk-based pricing of interest rates for consumer loans. *Journal of Monetary Economics*, 53(8), 2283–2298. <https://doi.org/10.1016/j.jmoneco.2005.09.001>
46. Edgeworth, F. Y. (1888). The mathematical theory of banking. *Journal of the Royal Statistical Society*, 51(1), 113–127. <http://www.jstor.org/stable/2979084>
47. European Parliament. (2013). Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 Text with EEA relevance. *Official Journal of the European Union*. Retrieved October 8, 2019, from <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02013R0575-20190627&qid=1570522454700&from=EN>
48. Finlay, S. (2010). *The management of consumer credit: Theory and practice* (2nd ed.). Palgrave Macmillan.
49. FirstRand Bank Limited. (2019). *Analysis of financial results for the year ended 30 June 2019* (tech. rep.). Johannesburg, South Africa. Retrieved September 6, 2019, from <https://www.firststrand.co.za/media/investors/financial-results/frb-analysis-of-results-june-2019.pdf>
50. Freedman, S. & Jin, G. Z. (2008). *Do social networks solve information problems for peer-to-peer lending? Evidence from Prosper.com* (tech. rep.). NET Institute Working Paper No. 08-43. Indiana University, Bloomington, School of Public & Environment Affairs Research. Retrieved August 20, 2019, from <https://ssrn.com/abstract=1936057>
51. Gordy, M. B. (2003). A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation*, 12(3), 199–232. [https://doi.org/10.1016/S1042-9573\(03\)00040-8](https://doi.org/10.1016/S1042-9573(03)00040-8)
52. Grimshaw, S. D. & Alexander, W. P. (2011). Markov chain models for delinquency: Transition matrix estimation and forecasting. *Applied Stochastic Models in Business and Industry*, 27(3), 267–279. <https://doi.org/10.1002/asmb.827>

## BIBLIOGRAPHY

---

53. Han, C. & Jang, Y. (2013). Effects of debt collection practices on loss given default. *Journal of Banking & Finance*, 37(1), 21–31. <https://doi.org/10.1016/j.jbankfin.2012.08.009>
54. Hand, D. J. (2001). Modelling consumer credit risk. *IMA Journal of Management Mathematics*, 12(2), 139–155. <https://doi.org/10.1093/imaman/12.2.139>
55. Hand, D. J. & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523–541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
56. Hao, C., Alam, M. & Carling, K. (2010). Review of the literature on credit risk modeling: Development of the past 10 years. *Banks and Bank Systems*, 5(3), 43–60. <http://urn.kb.se/resolve?urn=urn:nbn:se:du-4687>
57. Harris, T. (2013a). Default definition selection for credit scoring. *Artificial Intelligence Research*, 2(4), 49–62. <https://doi.org/10.5430/air.v2n4p49>
58. Harris, T. (2013b). Quantitative credit risk assessment using support vector machines: Broad versus narrow default definitions. *Expert Systems with Applications*, 40(11), 4404–4413. <https://doi.org/10.1016/j.eswa.2013.01.044>
59. Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
60. Heider, F., Hoerova, M. & Holthausen, C. (2015). Liquidity hoarding and interbank market rates: The role of counterparty risk. *Journal of Financial Economics*, 118(2), 336–354. <https://doi.org/10.1016/j.jfineco.2015.07.002>
61. Howgego, C. (1992). The supply and use of money in the Roman world 200 BC to AD 300. *The Journal of Roman Studies*, 82, 1–31. <https://doi.org/10.2307/301282>
62. Hudson, M. (2010). Entrepreneurs: From the near eastern takeoff to the Roman collapse. *The invention of enterprise: Entrepreneurship from ancient Mesopotamia to modern times* (pp. 8–39). Princeton University Press. <http://www.jstor.org/stable/j.ctt7t7h2.7>
63. IFRS 9. (2014). *International financial reporting standard (IFRS) 9: Financial instruments* (tech. rep.). International Accounting Standards Board. London, IFRS Foundation. <https://www.ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments/>
64. Jankowitsch, R., Pichler, S. & Schwaiger, W. S. (2007). Modelling the economic value of credit rating systems. *Journal of Banking & Finance*, 31(1), 181–198. <https://doi.org/10.1016/j.jbankfin.2006.01.003>

65. Jarrow, R. A., Lando, D. & Turnbull, S. M. (1997). A Markov model for the term structure of credit risk spreads. *The review of financial studies*, 10(2), 481–523. <https://doi.org/10.1093/rfs/10.2.481>
66. Jung, K. M., Thomas, L. C. & So, M. M. (2013). Time varying or static cut-offs for credit scorecards. *Journal of the Operational Research Society*, 64(9), 1299–1306. <https://doi.org/10.1057/jors.2012.118>
67. Kelly, R. & McCann, F. (2016). Some defaults are deeper than others: Understanding long-term mortgage arrears. *Journal of Banking & Finance*, 72, 15–27. <https://doi.org/10.1016/j.jbankfin.2016.07.006>
68. Kelly, R. & O'Malley, T. (2016). The good, the bad and the impaired: A credit risk model of the Irish mortgage market. *Journal of Financial Stability*, 22, 1–9. <https://doi.org/10.1016/j.jfs.2015.09.005>
69. Kennedy, K., Mac Namee, B., Delany, S. J., O'Sullivan, M. & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications*, 40(4), 1372–1380. <https://doi.org/10.1016/j.eswa.2012.08.052>
70. Leland, H. E. & Pyle, D. H. (1977). Informational asymmetries, financial structure, and financial intermediation. *The Journal of Finance*, 32(2), 371–387. <https://doi.org/10.2307/2326770>
71. Leow, M. & Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research*, 236(2), 685–694. <https://doi.org/10.1016/j.ejor.2013.12.026>
72. Liu, Z., He, P. & Chen, B. (2019). A Markov decision model for consumer term-loan collections. *Review of Quantitative Finance and Accounting*, 52(4), 1043–1064. <https://doi.org/10.1007/s11156-018-0735-4>
73. Longhofer, S. D., Carlstrom, C. T. et al. (1995). Absolute priority rule violations in bankruptcy. *Economic Review-Federal Reserve Bank of Cleveland*, 31, 21–30. <https://www.clevelandfed.org/~media/content/newsroom%20and%20events/publications/discontinued%20publications/economic%20review/1995/er%201995q4%20absolute%20priority%20rule%20violations%20pdf.pdf?la=en>
74. Louzada, F., Ara, A. & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21, 117–134. <https://doi.org/10.1016/j.sorms.2016.10.001>
75. MacDonald, S. B. & Gastmann, A. L. (2001). *A history of credit and power in the western world*. Transaction Publishers.

## BIBLIOGRAPHY

---

76. Matuszyk, A., Mues, C. & Thomas, L. C. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society*, 61(3), 393–398. <https://doi.org/10.1057/jors.2009.67>
77. Merton, R. C. (1977). An analytic derivation of the cost of deposit insurance and loan guarantees an application of modern option pricing theory. *Journal of Banking & Finance*, 1(1), 3–11. [https://doi.org/10.1016/0378-4266\(77\)90015-2](https://doi.org/10.1016/0378-4266(77)90015-2)
78. Miller, S. M. (1975). A theory of the banking firm: Comment. *Journal of Monetary Economics*, 1(1), 123–128. [https://doi.org/10.1016/0304-3932\(75\)90012-4](https://doi.org/10.1016/0304-3932(75)90012-4)
79. Mitchner, M. & Peterson, R. P. (1957). An operations-research study of the collection of defaulted loans. *Operations Research*, 5(4), 522–545. <https://doi.org/10.1287/opre.5.4.522>
80. Modigliani, F. (1986). Life cycle, individual thrift, and the wealth of nations. *Science*, 234(4777), 704–712. <https://doi.org/10.1126/science.234.4777.704>
81. Moffatt, P. G. (2005). Hurdle models of loan default. *Journal of the Operational Research Society*, 56(9), 1063–1071. <https://doi.org/10.1057/palgrave.jors.2601922>
82. Mushava, J. & Murray, M. (2018). An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market. *Expert Systems with Applications*, 111, 35–50. <https://doi.org/10.1016/j.eswa.2018.02.030>
83. Nichols, G. O. (1971). English government borrowing, 1660-1688. *Journal of British Studies*, 10(2), 83–104. <https://www.jstor.org/stable/175350>
84. Novotny-Farkas, Z. (2016). The interaction of the IFRS 9 expected loss approach with supervisory rules and implications for financial stability. *Accounting in Europe*, 13(2), 197–227. <https://doi.org/10.1080/17449480.2016.1210180>
85. OECD. (2020). *Household debt (indicator)* (tech. rep.). The Organisation for Economic Co-operation and Development. Retrieved October 24, 2020, from <https://doi.org/10.1787/f03b6469-en>
86. Oliver, R. M. & Thaker, A. (2013). Adverse selection and non-take inference with coherent risk and response scoring. *Journal of the Operational Research Society*, 64(1), 70–85. <https://doi.org/10.1057/jors.2012.3>
87. Phillips, R. (2013). Optimizing prices for consumer credit. *Journal of Revenue & Pricing Management*, 12(4), 360–377. <https://doi.org/10.1057/rpm.2013.9>

88. Phillips, R. & Raffard, R. (2011). *Price-driven adverse selection in consumer lending* (tech. rep.). Center for Pricing and Revenue Management Working Paper 2011-3, Columbia University. <https://www8.gsb.columbia.edu/cprm/sites/cprm/files/files/Working%20Papers/2011-3-Price-Driven-Adverse-Selection.pdf>
89. Poole, W. (1968). Commercial bank reserve management in a stochastic model: Implications for monetary policy. *The Journal of Finance*, 23(5), 769–791. <https://doi.org/10.1111/j.1540-6261.1968.tb00316.x>
90. PRA. (2019). *PS7/19: Credit risk: The definition of default* (tech. rep.). Bank of England, Prudential Regulation Authority (PRA). <https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/policy-statement/2019/ps719.pdf?la=en&hash=AB7A1E33FD2ED033A2CC488F75087AF477A31409>
91. Quinn, S. & Roberds, W. (2005). The big problem of large bills: The Bank of Amsterdam and the origins of central banking. *Federal Reserve Bank of Atlanta: Working Paper Series*. <https://www.frbatlanta.org/-/media/documents/research/publications/wp/2005/wp0516.pdf>
92. Rigas, C. A. & Riga, E. (2003). Banks in ancient Greece. *Archives of Economic History*, 15(1), 55–70. [http://archivesofeconomichistory.com/webdata/magaz/040113165111\\_Volume%20XV\\_No1\\_2003.pdf](http://archivesofeconomichistory.com/webdata/magaz/040113165111_Volume%20XV_No1_2003.pdf)
93. Rosenberg, R. & Christen, R. (1999). *Measuring microcredit delinquency: Ratios can be harmful to your health* (tech. rep.). World Bank. CGAP Occasional paper. Retrieved October 27, 2018, from <http://documents.worldbank.org/curated/en/986221468138281838/Measuring-microcredit-delinquency-ratios-can-be-harmful-to-your-health>
94. Sah, R. (2015). Loan recovery monitoring mechanism. *International Journal of Trade, Economics and Finance*, 6(1), 62. <https://doi.org/10.7763/IJTEF.2015.V6.444>
95. Santomero, A. M. (1984). Modeling the banking firm: A survey. *Journal of Money, Credit and Banking*, 16(4), 576–602. <https://doi.org/10.2307/1992092>
96. Santos, J. A. (2006). Insuring banks against liquidity shocks: The role of deposit insurance and lending of last resort. *Journal of Economic Surveys*, 20(3), 459–482. <https://doi.org/10.1111/j.0950-0804.2006.00286.x>
97. SARB. (2014). C2/2014: Interpretation of definition of default as outlined in Regulation 67 of the Regulations relating to Banks. <https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6116/C2%20of%202014.pdf>
98. SARB. (2015). D7/2015: Restructured credit exposures. <https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6716/D7%20of%202015.pdf>

## BIBLIOGRAPHY

---

99. Sarkar, M., Butler, B. & Steinfield, C. (1998). Cybermediaries in electronic marketplace: Toward theory building. *Journal of Business Research*, 41(3), 215–221. [https://doi.org/10.1016/S0148-2963\(97\)00064-7](https://doi.org/10.1016/S0148-2963(97)00064-7)
100. Schuermann, T. (2004a). What do we know about loss given default? *Wharton Financial Institutions Center Working Paper*, (04-01). Available at SSRN: <https://doi.org/10.2139/ssrn.525702>
101. Schuermann, T. (2004b). Why were banks better off in the 2001 recession? *Current Issues in Economics and Finance*, 10(1). Federal Reserve Bank of New York. [https://www.newyorkfed.org/medialibrary/media/research/current\\_issues/ci10-1.pdf](https://www.newyorkfed.org/medialibrary/media/research/current_issues/ci10-1.pdf)
102. Sekoni, A. (2015). The basic concepts and feature of bank liquidity and its risk. *Institute of Islamic Banking and Finance*. Available at MPRA: <https://mpra.ub.uni-muenchen.de/id/eprint/67389>
103. Siddiqi, N. (2005). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. John Wiley & Sons.
104. Skoglund, J. (2017). Credit risk term-structures for lifetime impairment forecasting: A practical guide. *Journal of Risk Management in Financial Institutions*, 10(2), 177–195. <https://www.econbiz.de/Record/credit-risk-term-structures-for-lifetime-impairment-forecasting-a-practical-guide-skoglund-jimmy/10011670671>
105. Smith, L. D. & Lawrence, E. C. (1995). Forecasting losses on a liquidating long-term loan portfolio. *Journal of Banking & Finance*, 19(6), 959–985. [https://doi.org/10.1016/0378-4266\(94\)00065-B](https://doi.org/10.1016/0378-4266(94)00065-B)
106. So, M. C., Mues, C., de Almeida Filho, A. T. & Thomas, L. C. (2019). Debtor level collection operations using Bayesian dynamic programming. *Journal of the Operational Research Society*, 70(8), 1332–1348. <https://doi.org/10.1080/01605682.2018.1506557>
107. South Africa. (2012). *Banks Act (94/1990): Regulations relating to Banks (regulation gazette no. 35950)* (tech. rep.). Government Gazette. Retrieved May 3, 2018, from [https://www.gov.za/sites/default/files/gcis\\_document/201409/35950rg9872gon1029.pdf](https://www.gov.za/sites/default/files/gcis_document/201409/35950rg9872gon1029.pdf)
108. Standard Bank Group. (2018). *Analysis of financial results for the year ended 31 December 2018* (tech. rep.). Johannesburg, South Africa. Retrieved September 6, 2019, from [https://thevault.exchange/?get\\_group\\_doc=18/1551936555-SBKFY18Resultsanalysis.pdf](https://thevault.exchange/?get_group_doc=18/1551936555-SBKFY18Resultsanalysis.pdf)
109. Stewart, R. (2011). A profit-based scoring system in consumer credit: Making acquisition decisions for credit cards. *Journal of the Operational Research Society*, 62(9), 1719–1725. <https://doi.org/10.1057/jors.2010.135>

110. Stiglitz, J. E. & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, 393–410. <https://www.jstor.org/stable/1802787>
111. Taggart, R. A. & Greenbaum, S. I. (1978). Bank capital and public regulation. *Journal of Money, Credit and Banking*, 10(2), 158–169. <https://doi.org/10.2307/1991868>
112. Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172. [https://doi.org/10.1016/S0169-2070\(00\)00034-0](https://doi.org/10.1016/S0169-2070(00)00034-0)
113. Thomas, L. C. (2009a). *Consumer credit models: Pricing, profit and portfolios*. Oxford University Press.
114. Thomas, L. C. (2009b). Modelling the credit risk for portfolios of consumer loans: Analogies with corporate loan models. *Mathematics and Computers in Simulation (MATCOM)*, 79(8), 2525–2534. <https://doi.org/10.1016/j.matcom.2008.12.006>
115. Thomas, L. C. (2010). Consumer finance: Challenges for operational research. *Journal of the Operational Research Society*, 61(1), 41–52. <https://doi.org/10.1057/jors.2009.104>
116. Thomas, L. C., Edelman, D. B. & Crook, J. N. (2002). *Credit scoring and its applications*. SIAM: Monographs on Mathematical Modeling and Computation.
117. Thomas, L. C., Matuszyk, A., So, M. C., Mues, C. & Moore, A. (2016). Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research*, 249(2), 476–486. <https://doi.org/10.1016/j.ejor.2015.09.013>
118. Van Gestel, T. & Baesens, B. (2009). *Credit risk management: Basic concepts*. Oxford University Press.
119. Van Kuelen, J. A., Spronk, J. & Corcoran, A. W. (1981). Note—On the Cyert-Davidson-Thompson doubtful accounts model. *Management science*, 27(1), 108–112. <https://doi.org/10.1287/mnsc.27.1.108>
120. Vasicek, O. (2002). The distribution of loan portfolio value. *Risk*, 15(12), 160–162. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.8181&rep=rep1&type=pdf>
121. Vento, G. A. & La Ganga, P. (2009). Bank liquidity risk management and supervision: Which lessons from recent market turmoil. *Journal of Money, Investment and Banking*, 10(10), 78–125.
122. Witzany, J. et al. (2013). A note on the vasicek's model with the logistic distribution. *Ekonomický časopis*, 61(10), 1053–1066. <https://doi.org/10.2139/ssrn.2132583>

## BIBLIOGRAPHY

---

123. Xu, X. (2016). Estimating lifetime expected credit losses under IFRS 9. *Unisys Machine Learning and Advanced Analytics Services*. Available at SSRN: <https://doi.org/10.2139/ssrn.2758513>