Visualising Basins of Attraction for the Cross-Entropy and the Squared Error Neural Network Loss Functions

Anna Sergeevna Bosman[a,*], Andries Engelbrecht[b,c], Mardé Helbig[d]

[a] *Department of Computer Science, University of Pretoria, Pretoria, South Africa*
[b] *Department of Industrial Engineering, Stellenbosch University, Stellenbosch, South Africa*
[c] *Computer Science Division, Stellenbosch University, Stellenbosch, South Africa*
[d] *School of Information and Communication Technology, Griffith University, Southport, Australia*

Abstract

Quantification of the stationary points and the associated basins of attraction of neural network loss surfaces is an important step towards a better understanding of neural network loss surfaces at large. This work proposes a novel method to visualise basins of attraction together with the associated stationary points via gradient-based stochastic sampling. The proposed technique is used to perform an empirical study of the loss surfaces generated by two different error metrics: quadratic loss and entropic loss. The empirical observations confirm the theoretical hypothesis regarding the nature of neural network attraction basins. Entropic loss is shown to exhibit stronger gradients and fewer stationary points than quadratic loss, indicating that entropic loss has a more searchable landscape. Quadratic loss is shown to be more resilient to overfitting than entropic loss. Both losses are shown to exhibit local minima, but the number of local minima is shown to decrease with an increase in dimensionality. Thus, the proposed visualisation technique successfully captures the local minima properties exhibited by the neural network loss surfaces, and can be used for the purpose of fitness landscape analysis of neural networks.

Keywords: fitness landscape analysis, neural networks, cross-entropy, squared error, local minima, loss functions

[*] Corresponding author
URL: `anna.bosman@up.ac.za` (Anna Sergeevna Bosman)

1. Introduction

In the wake of the deep learning research explosion in the artificial neural network (NN) research community [1, 2], it becomes increasingly important to develop a better general understanding of NN training as a non-convex optimisation problem. Lack of understanding causes practitioners to make arbitrary choices for various hyperparameters, yielding potentially subpar performance. Failure or success of a particular combination of NN architecture and training algorithm parameters is hard to predict. Specifically, the nature of the error landscapes associated with the NN loss functions is still poorly understood [3, 4, 5]. There are on-going debates and theories regarding the presence or absence of local minima in NN error landscapes, as well as the properties of stationary points and the associated basins of attraction in the search space [6, 7, 8]. Such lack of understanding hinders the development of new training algorithms that would take the discovered properties of the search space into consideration.

One of the main reasons for this lack of insight is the high dimensionality inherent to NN problems. High-dimensional spaces are not intuitively visualisable, thus other means of analysis have to be employed. Theoretical analysis, however, often relies on unrealistic assumptions, sometimes causing erroneous conclusions. For example, papers were published claiming that XOR has no local minima [9], to be subsequently followed by other publications that explicitly listed all local minima of the XOR problem [10]. Sprinkhuizen et al. [10] have also stated that the listed local minima are in fact saddle points [10]. More recent studies confirm that local optima are indeed present in the NN error landscapes [11], although saddle points are likely to become more prevalent as the dimensionality of the problem increases [6, 12]. Similarly to local minima, the properties of the NN basins of attraction are being actively studied and questioned [8, 13].

The number of local minima, as well as the properties of local minima, were theoretically shown to depend on the chosen error metric [14], among other parameters. Solla et al. [14] analysed two common NN loss functions, quadratic loss and entropic loss, and came to the conclusion that quadratic loss exhibits a higher density of local minima, and entropic loss has steeper gradients, which is likely to benefit gradient-based training. Entropic loss has gained popularity in the deep learning community due to speeding up gradient descent convergence, and providing more robust results than squared loss [15, 16]. However, studies were published advocating the hybrid use of both entropic and squared loss, as squared loss was shown to be able to refine the solution discovered with entropic loss [16].

This study aims to explore the properties of the stationary points and the associated basins of attraction exhibited by the NN loss functions by means of proposing a low-dimensional visualisation. The stationary points of the NN error surfaces are visualised using sampling-based techniques developed for fitness landscape analysis (FLA). Hessian matrix analysis is further employed to classify the discovered stationary points into minima, maxima, and saddles.

The novel contributions of this paper are summarised as follows:

- A 2-dimensional visualisation of the NN stationary points is proposed.
- A simple numerical metric to quantify the number and extent of the basins of attraction is proposed.
- An empirical comparison of the basins of attraction associated with squared loss and entropic loss is carried out using the proposed techniques.

The rest of the paper is structured as follows: Section 2 reviews the previously published literature on local minima, stationary points, and attraction basins in the NN error landscapes. Section 3 discusses the two loss functions considered in this study. Section 4 describes FLA in the context of NN training problems, discusses the sampling technique used, and proposes: (1) a novel method to visualise stationary points and the associated basins of attraction of NN loss surfaces in 2-dimensional space, and (2) two metrics to numerically quantify the discovered basins of attraction. Section 5 details the experimental procedure. Section 6 presents a visual and numerical analysis of stationary points and basins of attraction of the quadratic and the entropic error landscapes. Finally, Section 7 concludes the paper and proposes some topics for future research.

2. Local Minima and Basins of Attraction in Neural Networks

Many studies of local minima in NNs were carried out on the XOR (exclusive-or) problem. XOR is a simple, but linearly non-separable problem that can be solved by a feedforward NN with at least two hidden neurons. As such, XOR is often used to analyse the basic properties of NNs. Studies of the XOR error landscape are especially interesting, because researchers have arrived at somewhat contradictory conclusions. Hamey [9] claimed that the NN error surface associated with XOR has no local minima. A year later, Sprinkhuizen-Kuyper et al. [10, 17] showed that stationary points are present in the XOR NN search space, but that the stationary points are in fact saddle points. A more recent study of the XOR error surface was published by Mehta et al. [18], where techniques developed for potential energy landscapes were used to quantify local minima of the XOR problem under a varied number of hidden neurons and regularisation coefficient values. Mehta et al. [18] showed that the XOR problem exhibits local minima, and that the number of local minima grows with an increase in the size of the hidden layer.

Further theoretical analysis performed for more complex problems than XOR highlighted the fact that saddle points are more prevalent in high-dimensional spaces than local minima, and that the number of local minima decreases with an increase in dimensionality [6, 12]. Counterexamples have also been published, artificially constructing problems with difficult local minima that can potentially trap the training algorithm [11]. Current understanding of the stationary points in NN error surfaces remains incomplete, partially due to the lack of empirical evidence and intuitive visualisations.

The discovery of the prevalence of saddle points in NN error landscapes has led researchers to question the nature of the basins of attraction associated with the stationary points [8]. It has been observed that NN error landscapes are comprised of wide and narrow valleys, and that the solutions discovered at the bottom of such valleys may have different generalisation behaviour [19, 20, 21]. It has also been observed that it may be possible to find a path of non-increasing error value that connects any two valleys, thus indicating that the valleys may all be part of a single manifold, or attraction basin [22]. This study estimates the properties of the basins of attraction associated with two different loss functions, namely quadratic and entropic, discussed in the next section.

3. Loss Functions

The modality of an NN search space, i.e., the number of local minima, as well as the properties of local minima and the associated basins of attraction, were theoretically shown to depend on the chosen error metric [14], among other parameters. The two most widely used error metrics are the quadratic loss function and the entropic loss function, discussed in this section.

Quadratic loss, also referred to as the sum squared error (SSE), simply calculates the sum of squared errors produced by the NN:

$$E_{sse} = \sum_{p=1}^{P} \sum_{k=1}^{K} (t_{k,p} - o_{k,p})^2 \tag{1}$$

where P is the number of data points, K is the number of outputs, $t_{k,p}$ is the k'th target value for data point p, and $o_{k,p}$ is the k'th output obtained for data point p. Minimisation of the SSE minimises the overall error produced by the NN.

If the outputs of the NN can be interpreted as probabilities, then the cross-entropy between two distributions can be calculated, i.e., the distribution of the desired outputs (targets), and the distribution of the actual outputs. Entropic loss, also referred to as log loss, or as the cross-entropy (CE) error, is formulated as follows:

$$E_{ce} = -\sum_{p=1}^{P} \sum_{k=1}^{K} t_{k,p} \log o_{k,p}. \tag{2}$$

Minimisation of the cross-entropy leads to convergence of the two distributions, i.e., the actual output distribution resembles the target distribution more and more, thus minimising the NN error.

Solla et al. [14] analysed quadratic loss and entropic loss theoretically, and came to the conclusion that quadratic loss exhibits a higher density of local minima. Solla et al. [14] further showed that entropic loss must generate a "steeper" landscape with stronger gradients, which may be the reason for the observed faster convergence of gradient descent on CE compared to SSE. Faster convergence of entropic loss has led to entropic loss becoming more popular than quadratic loss in the deep learning community [15, 16]. In addition to faster convergence, entropic loss was shown to exhibit better statistical properties, such as more precise estimation of the true posterior probability on average [23].

From a theoretical standpoint, however, the global minima of both SSE and CE will correspond to the true posterior probability derived from the given dataset [24]. Thus, if a global minimum is found on either of the error landscapes, the quality of either minimum will be equally good. A study by Golik et al. [16] showed that, although squared loss may cause the training algorithm to converge to a poor minimum, this behaviour is only exhibited if the algorithm was initialised poorly. Golik et al. [16] demonstrated the benefit of applying gradient descent to the error landscape generated by entropic loss at first, and then "switching" to quadratic loss to further refine the solution discovered on the entropic loss surface. Such a training scheme may be successful due to the fact that entropic loss is known to turn flat around the global minimum [25].

This paper aims to study the difference between the loss landscapes of the quadratic loss and the entropic loss by applying fitness landscape analysis techniques, discussed in the next section.

4. Fitness Landscape Analysis

The concept of fitness landscape analysis (FLA) comes from the evolutionary context, where quantitative metrics have been developed to study the landscapes of combinatorial problems [26, 27]. FLA was successfully adapted to continuous fitness landscapes at a later stage [28, 29, 30, 31]. Various fitness landscape properties, such as ruggedness, neutrality, modality, and searchability, can be estimated by taking multiple samples of the search space, calculating the objective function value for every point in each sample, and analysing the relationship between the spatial and the qualitative characteristics of the sampled points. If the samples cover the search space in a meaningful way, the characteristics of the fitness landscape captured by the sampling will apply to the fitness landscape at large.

The NN search space is defined as all possible real-valued weight combinations. Thus, samples of the weight space can be taken and analysed to approximate the search space properties. Several studies were conducted showing FLA to be a useful tool for analysis and visualisation of the NN error surfaces [32, 33, 34, 35]. However, none of the previous FLA studies have attempted to quantify the modality of the NN error landscapes, i.e., the presence and characteristics of local minima.

This study uses NN error landscape samples to quantify the loss surface modality. The progressive gradient walk algorithm used to obtain the samples is discussed in Section 4.1. Further, this study proposes two novel FLA techniques to visualise and quantify the stationary points and the associated basins of attraction exhibited by NN loss surfaces. Section 4.2 introduces the loss-gradient clouds, which offer a 2-dimensional visualisation of the stationary points discovered by the gradient walks. Section 4.3 proposes two metrics to quantify the properties of the basins of attraction encountered by the progressive gradient walks.

4.1. Progressive Gradient Walk

One of the simplest FLA approaches to estimate the presence of local minima is to take a uniform random sample of the search space, and then to calculate the proportion of local minima within the sample [36]. To identify minima, stationary points need to be identified first. Since the loss functions are differentiable, the gradient can be calculated for each point in the sample. Points with a gradient of zero are stationary points. Stationary points can be further categorised into local minima, local maxima, and saddle points by calculating the eigenvalues of the corresponding Hessian matrix. A positive-definite Hessian is indicative of a local minimum [37].

However, an earlier study by Bosman et al. [38] demonstrated that random samples capture very few points of high fitness even for such a simple problem as XOR, and thus are unlikely to discover local or global minima. Additionally, random samples do not capture the neighbourhood relationship between individual sample points, which is crucial to the analysis of the basins of attraction. Besides simply identifying the presence or absence of local minima, the possibility of escaping the minima, as well as the structure of the minima, should also be quantified.

An alternative to a uniform random sample is a sample generated by a random walk. To perform a random walk, a random point is chosen within range, and consecutive steps in randomised directions are taken to generate the sample. This way, the sampled points will be related to each other topographically. However, a random walk faces the same problem as the uniform random sample, in that random traversal of the search space provides no guarantee of locating areas of good fitness [38].

Instead of analysing random walks, the trajectory of a training algorithm can be analysed. However, such an approach will bias the observations towards the performance of the specific algorithm under specific hyperparameter settings. Convergent behaviour typical of training algorithms will also prevent sufficient exploration of the attraction basins. To address this problem, Bosman et al. [38] proposed a sampling method called a *progressive gradient walk*. This approach is an adaptation of the progressive random walk, proposed by Malan and Engelbrecht [39], where random persistent direction bias was first applied to a random walk. A progressive gradient walk uses the numeric gradient of the loss function to determine the direction of each step. The size of the step is randomised per dimension within predefined bounds. The progressive gradient walk algorithm is summarised as follows:

1. At each iteration, gradient vector $\vec{g}_l$ is calculated for a point $\vec{x}_l$, where $l \in \{1, \ldots, L\}$, and L is the length of the walk.
2. A binary direction mask $\vec{b}_l$ is extracted from $\vec{g}_l$ as follows:
$$b_{lj} = \begin{cases} 0 & \text{if } g_{lj} < 0, \\ 1 & \text{otherwise,} \end{cases}$$
where $j \in \{1, \ldots, m\}$ for the m-dimensional vector $\vec{g}_l$.
3. The progressive random walk algorithm, proposed by Malan and Engelbrecht [39], is used to generate the next step $\vec{x}_{l+1}$. A single step of a progressive random walk can be defined as randomly generating an m-dimensional step vector $\Delta\vec{x}_l$, such that $\Delta x_{lj} \in [0, \varepsilon]\ \forall j \in \{1, \ldots, m\}$, and setting the sign of each Δx_{lj} according to the corresponding b_{lj}:
$$\Delta x_{lj} := \begin{cases} -\Delta x_{lj} & \text{if } b_{lj} = 0, \\ \Delta x_{lj} & \text{otherwise.} \end{cases}$$
To generate the next step, $\vec{x}_{l+1}$, the current step $\vec{x}_l$ is modified by adding $\Delta\vec{x}_l$:
$$\vec{x}_{l+1} = \vec{x}_l + \Delta\vec{x}_l.$$

The progressive gradient walk algorithm requires one parameter to be set: the maximum dimension-wise step size, ε. The main advantage of this sampling approach is that gradient information is combined with stochasticity, preventing convergence, yet guiding the walk towards areas of higher fitness.

The next section proposes a visual way to study the NN loss surface samples obtained by the progressive gradient walk.

4.2. Loss-Gradient Clouds

Stationary points in the search space are identified by the absence of gradient, i.e., gradient of zero. Therefore, for each sampled point, the magnitude of the gradient vector can be calculated in order to determine whether the point is stationary. Further, stationary points of non-zero loss can be either local minima, local maxima, or saddle points. To determine if a particular stationary point is a local minimum, local maximum, or a saddle point, local curvature information can be derived from the eigenvalues of the corresponding Hessian matrix [37]. If the eigenvalues of the Hessian are positive, the point is a maximum. If the eigenvalues are negative, the point is a minimum. If the eigenvalues are positive as well as negative, the point is a saddle. If any of the eigenvalues are zero, i.e., if the Hessian is indefinite, the test is considered inconclusive.

Thus, three metrics need to be calculated to identify local minima and other stationary points: (1) gradient magnitude, (2) loss value, and (3) local curvature. To study the properties of the attraction basins that surround the discovered stationary points, the same metrics can be calculated for the points sampled in the vicinity of the stationary points. To avoid making assumptions regarding the size and shape of the attraction basins, the three metrics can be calculated for all sampled points.

Thus, high-dimensional NN search spaces can be projected onto three dimensions: gradient magnitude, loss value, and local curvature. To study the interactions between the gradient magnitude and the loss value, a 2-dimensional scatterplot [40] projection is proposed, referred to as the *loss-gradient cloud*, or l-g cloud. Further, curvature can be represented on the same plot by assigning a unique colour to convex, concave, saddle, and indefinite curvatures. A scatterplot is a common statistical tool designed to visualise the relations between two variables measured on the same observational units [40]. Scatterplots can be generated for any given multivariate data using statistical visualisation tools such as ggplot [41]. To the best of authors' knowledge, this study is the first to use scatterplots for visualisation of the NN loss landscape modality.

An example l-g cloud is shown in Figure 1. The loss value is shown on the x-axis, and the gradient magnitude is shown on the y-axis. The sampled points are split into four panes according to the four curvature types. Figure 1 shows that convex stationary points of non-zero loss were sampled. Therefore, the sampled search space exhibited local minima.

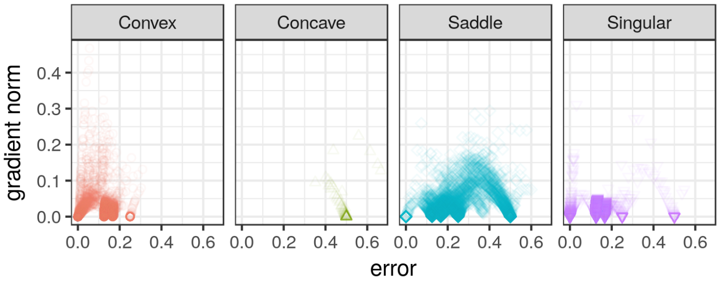


Figure 1: Example l-g cloud generated for the XOR problem.

The main benefit of l-g clouds is the 2-dimensional, interpretable representation of the high-dimensional search space which is otherwise very hard to study. Studying the discovered stationary points in 2-dimensional space allows the identification of the total number of attractors, both local and global, corresponding to unique loss values. The gradient behaviour of the attractors is also visualised by the l-g clouds, and can provide useful insights into the structure of the attraction basins, such as the steepness of the basins, and the connectedness of the basins, i.e., the ability of a sampling algorithm to make a transition from a local attractor to the global attractor. L-g clouds allow for empirical studies of the loss surface modality properties, and enable comparisons between the loss landscapes yielded by different NN configurations. Since the distance between sampled points is not represented in the l-g clouds, the actual number of distinct local minima and other attractors cannot be estimated using this technique.

L-g clouds provide information about the total number of stationary attractors of non-zero loss. Next section proposes two additional metrics to quantify the corresponding basins of attraction.

4.3. Quantifying Basins of Attraction

A progressive gradient walk samples the search space by taking stochastic steps of consistent magnitude in the general direction of the steepest gradient descent. If a step taken in the direction of the negative gradient is too large, the step may miss an area of low error, and result in an area of higher error. Thus, a progressive gradient walk will not necessarily produce a sequence of points with strictly non-increasing error values. In fact, any gradient-based sample or algorithm trajectory is likely to exhibit oscillatory behaviour if the gradient in some dimensions is significantly steeper than in others [42].

Even though the gradient step sequence will not necessarily be strictly decreasing in error, the sample is nonetheless expected to travel in the general direction of the global minimum. The areas of the landscape where a gradient-based walk oscillates or otherwise fails to reduce the error for a number of steps are the stationary areas of the search space that may hinder the optimisation process. Quantification of the number and extent of such areas will provide an indication of the "difficulty" of the search space, as well as an empirical estimate of the landscape modality. Thus, an important error landscape property to estimate is the number of times that the sampling algorithm will become "stuck" along the way.

To smooth out the potential oscillations of the sample, an exponential moving average of the sample can be calculated. An exponentially weighted moving average (EWMA) [43] is a smoothing filter commonly used for time series prediction. EWMA calculates the moving average for each step in the time series by taking all previous steps into account, and assigning exponentially decaying weights to the previous steps, such that the weight for each older step in the series decreases exponentially, never reaching zero. Given a sequence $T = \{T_i\}_{i=1}^{Z}$ of length Z, the EWMA-smoothed sequence T' is given by:

$$T'_i = \begin{cases} T_i & \text{if } i = 1, \\ \alpha \cdot T_i + (1 - \alpha) \cdot T'_{i-1} & \text{if } i > 1. \end{cases} \tag{3}$$

The decay coefficient $\alpha \in [0, 1]$ determines the degree of smoothing, where larger values of α facilitate faster decay and weaker smoothing, and smaller values of α facilitate slower decay and thus stronger smoothing.

To identify the sections of the sample where the behaviour is stagnant, the standard deviation of the smoothed sample is calculated first. Then, a sliding window approach is used to generate a sequence of the moving standard deviations of the sample. If the standard deviation of the values in the current window is less than the standard deviation of the entire sample for a number of steps, then these steps can be said to form a stagnant sequence. The average number of stagnant regions encountered per sample, n_{stag}, and the average length of the stagnant regions, l_{stag}, can be used to quantify the number and size of the basins of attraction present in the search space.

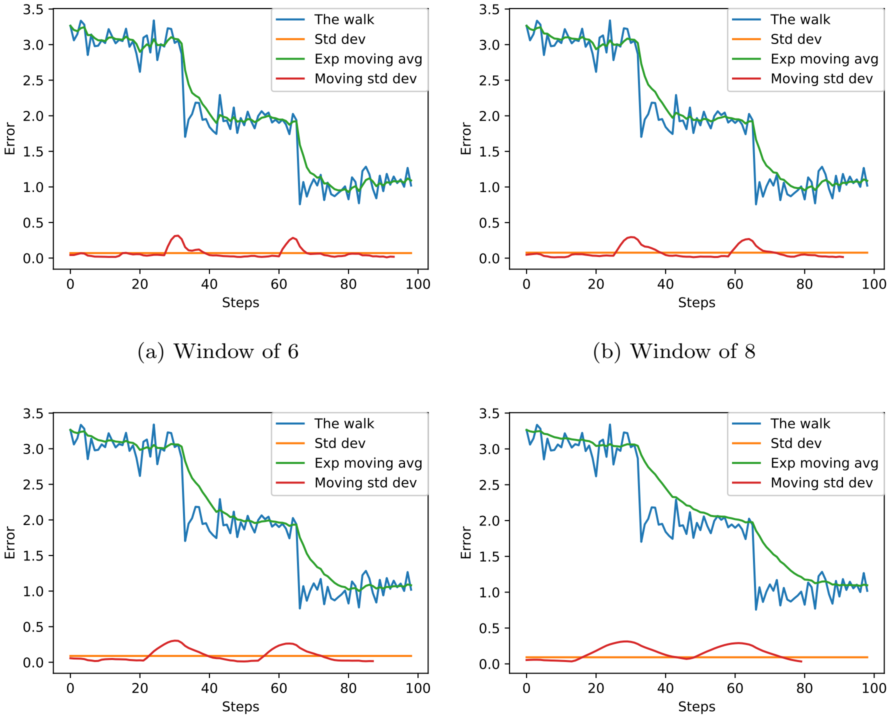
The proposed approach is illustrated in Figure 2. The simulated walk oscillates around three different error values. The moving standard deviation line dips below the all-sample standard deviation threshold three times, which corresponds to the three simulated stagnant areas.

Figure 2 illustrates that the window size has a significant effect on the attraction basin estimates: too little smoothing (Figure 2a) may cause fluctuations to be perceived as stationary regions. Excessive smoothing, on the other hand (Figure 2d), may fail to detect all stationary regions. Therefore, the window size has to be optimised per sample. If the sequence contains oscillations, then too little smoothing will cause multiple “spikes” in the walk to be regarded as areas of stagnation. These short bursts of “stagnation” will yield a small average basin length, l_{stag}. If the sequence is smoothed excessively, the sample will start to resemble a wave more and more, perceiving flat areas as areas with an incline, which will once again cause the l_{stag} to decrease. Thus, too little as well as too much smoothing will shrink the l_{stag}. Therefore, the window size w can be optimised by maximising the l_{stag} value. Table 1 lists l_{stag} values obtained on the simulated walk shown in Figure 2 under various values of w. Table 1 shows that l_{stag} reaches its maximum for $w = 8$, and decreases for smaller, as well as larger, values of w.

Table 1: Effect of window size w on l_{stag}

w	6	8	10	12	14	16	18	20
l_{stag}	18.75	**22.0**	20.67	18.0	16.33	14.33	14.0	12.0

The window size w can therefore be automatically optimised by calculat-



(a) Window of 6

(b) Window of 8

(c) Window of 12

(d) Window of 20

Figure 2: Illustration of the proposed technique to estimate the number and extent of the basins of attraction. Figures 2a to 2d show the effect of window size on the sample smoothing. The EWMA-smoothed sample is shown in green (Exp moving avg).

ing l_{stag} over a range of w values, and picking the value of w that yields the highest l_{stag} value. In this study, w is optimised by successively applying $w \in \{6, 8, \ldots, 18, 20\}$. Given a window of size w, the EWMA value of α is calculated as $\alpha = 2/(w + 1)$. The w value yielding the largest l_{stag} is subsequently used for the final l_{stag} and n_{stag} estimates.

Thus, two estimates to quantify the basins of attraction are proposed:

1. The average number of times that stagnation was observed, n_{stag}.
2. The average length of the stagnant sequence, l_{stag}.

A pseudocode to calculate n_{stag} and l_{stag} is provided in Appendix A.

It is important to note that n_{stag} and l_{stag} are approximations, and may produce misleading results in some scenarios. Specifically, if the observed sequence is chaotic, i.e., does not exhibit convergence or stagnant areas, the estimates provided by n_{stag} and l_{stag} are likely to be overly optimistic. In order to maximise l_{stag}, the algorithm will apply excessive smoothing to the chaotic sequence, potentially interpreting chaotic fluctuations as multiple stagnation regions. In general, because the algorithm is designed to maximise the stagnation length of the estimate, erroneous results are expected for sequences that do not exhibit any form of stagnation.

Experiments conducted to empirically test the proposed modality visualisation and quantification techniques are discussed in the next section.

5. Experimental Procedure

The aim of the study was to visually and numerically investigate the local minima and basins of attraction exhibited by quadratic and entropic loss functions. This section discusses the experimental set-up of the study, and is structured as follows: Section 5.1 lists the benchmark problems used and the NN hyperparameters employed in the experiments; Section 5.2 outlines the sampling algorithm parameters, and the data recorded for each sampled point.

5.1. Benchmark problems

A selection of well-known classification problems of varied dimensionality were used in this study. Table 2 summarises the NN architecture parameters used for each dataset, as well as the total dimensionality of the resulting weight space. The specified sources point to publications from which each dataset and/or NN architectures were adopted.

The properties of each dataset are briefly discussed below:

1. **XOR:** exclusive-or (XOR) is a simple, but linearly non-separable problem that can be solved by a feedforward NN with at least two hidden neurons. As such, XOR is often used to analyse basic properties of artificial neural networks. The dataset consists of 4 binary patterns.

Table 2: Benchmark Problems

Problem	Input	Hidden	Output	Dimension	Source
XOR	2	2	1	9	[9]
Iris	4	4	3	35	[44]
Diabetes	8	8	1	81	[45]
Glass	9	9	6	150	[45]
Cancer	30	10	1	321	[45]
Heart	32	10	1	341	[45]
MNIST	784	10	10	7960	[46]

2. **Iris:** The famous Iris flower data set [44] contains 50 specimens from each of the three species of iris flowers, i.e., *Iris Setosa*, *Iris Versicolor*, and *Iris Virginica*. There are 150 patterns in the dataset.
3. **Diabetes:** The diabetes dataset [45] captures personal data of 768 Pima Indian patients, classified as diabetes positive or diabetes negative.
4. **Glass:** The glass dataset [45] captures chemical components of glass shards. Each glass shard belongs to one of six classes: float processed or non-float processed building windows, vehicle windows, containers, tableware, or head lamps. There are 214 patterns in the dataset.
5. **Cancer:** The breast cancer Wisconsin (diagnostic) dataset [45] consists of 699 patterns, each containing tumor descriptors, and a binary classification into benign or malignant.
6. **Heart:** The heart disease prediction dataset [45] contains 920 patterns, each describing various patient descriptors.
7. **MNIST:** The MNIST dataset of handwritten digits [46] contains 70,000 examples of grey scale handwritten digits from 0 to 9. For the purpose of this study, the 2-dimensional input is treated as a 1-dimensional vector.

Input values for all problems except XOR were standardised by subtracting the mean per input dimension, and scaling every input variable to unit variance. All outputs were binary encoded for problems with two output classes, and one-hot binary encoded for problems with more than two output classes.

All experiments employed feed-forward NNs with a single hidden layer. The sigmoid activation function was used in the experiments, given by $f_{NN}(net) = 1/(1 + e^{-net})$, where net is the sum of weighted inputs. While the choice of activation function has an effect on the resulting error landscape, the aim of this study was to investigate the difference between quadratic and entropic loss.

5.2. Sampling parameters

For the purpose of sampling the areas of low error, a progressive gradient walk, discussed in Section 4, was used as the sampling mechanism. To allow for adequate coverage of the search space, the number of independent walks was set to be one order of magnitude higher than the dimensionality of the problem, i.e., for a problem of d dimensions, $10 \times d$ independent progressive gradient walks were performed. The walks were not restricted by search space bounds, however, two different initialisation ranges were considered, namely $[-1, 1]$ and $[-10, 10]$. The smaller range is typically used for NN weight initialisation. The larger range is likely to contain high fitness solutions [33]. Since the granularity of the walk, i.e., the average step size, has a bearing on the resulting FLA metrics [28], two granularity settings were used throughout the experiments: micro, where the maximum step size was set to 1% of the initialisation range, and macro, where the maximum step size was set to 10% of the initialisation range. Micro walks performed 1000 steps each, and macro walks performed 100 steps each.

For all problems except the XOR problem, the dataset was split into 80% training and 20% test subsets. The training set was used to calculate the direction of the gradient, as well as the error of the current point on the walk. The test set was used to evaluate the generalisation ability of each point in the walk. To calculate the training and the generalisation errors, the entire train/test subsets were used for all problems except MNIST. For MNIST, random batches of 100 patterns were sampled from the respective training and test sets.

In order to identify stationary points discovered by the gradient walks, the magnitude of the gradient vector was recorded for each step together with the loss value. Additionally, the eigenvalues of the Hessian matrix were calculated for each step, and used to classify each step as convex, concave, saddle, or singular.

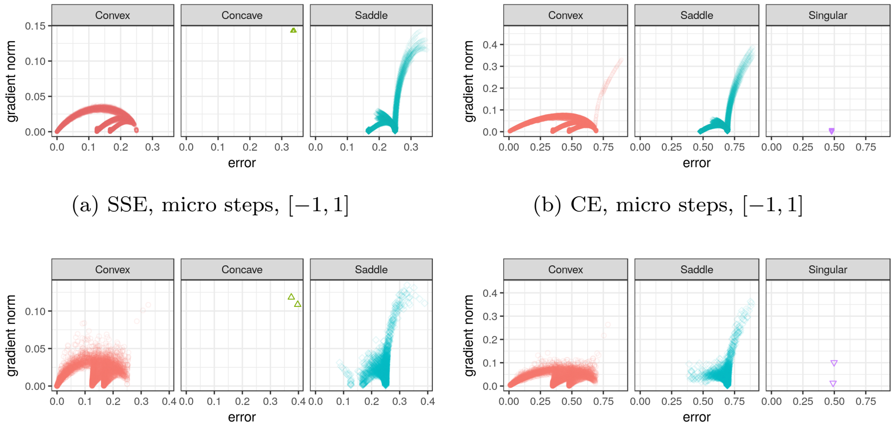
6. Empirical Results

This section presents the analysis of observed local minima and the corresponding basins of attraction as captured by the progressive gradient walks. For each problem, l-g clouds were generated and analysed. Then, n_{stag} and l_{stag} values were studied. The results obtained for each problem are discussed below.

6.1. XOR

Figure 3 shows the l-g clouds obtained for the XOR problem for gradient walks initialised in the $[-1, 1]$ range, separated into panes according to the curvature. The first observation that can immediately be made from Figure 3 is that both SSE and CE yielded exactly four unique stationary attractors. Furthermore, these four attractors were classified as convex according to the Hessian eigenvalues, indicating that the points can be classified as local minima rather than saddle points. A transition from saddle curvature to convex curvature was observed for both SSE and CE. Points further away from a global optimum were classified as exhibiting saddle curvature. Points in the two stationary attractors furthest away from the global attractor were sometimes classified as saddles, indicating that both saddles and local minima of equal loss value were discovered. Under the macro setting (larger steps), a few singular points were sampled in the same apparent basin, indicating that the area was flat (no curvature) in some dimensions. However, the global minima discovered by the gradient walks initialised in the $[-1, 1]$ range appeared perfectly convex. The area surrounding the global minima, as well as the two adjacent local minima, also exhibited convexity. Thus, the XOR problem definitely exhibits convex local minima.

Another interesting observation can be made by observing the trajectories connecting the apparent local minima: It is evident from Figure 3 that most



(a) SSE, micro steps, $[-1, 1]$

(b) CE, micro steps, $[-1, 1]$

(c) SSE, macro steps, $[-1, 1]$

(d) CE, macro steps, $[-1, 1]$

Figure 3: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range for the XOR problem. Micro (maximum 1% of the initialisation range) and macro (maximum 10% of the initialisation range) steps were considered.

high loss, high gradient points first descended to the local minimum furthest away from the global minimum, and from thereon proceeded to one of the three better minima. The three convex minima, however, were not connected by trajectories. In other words, once the gradient walk descended into one of the basins, escape from the basin became unlikely, given the limited step size. To further support this claim, n_{stag} and l_{stag} values calculated for the various XOR gradient walks are reported in Table 3. According to Table 3, the average number of basins visited by the $[-1, 1]$ micro-step walks was 1.88889 for SSE, and 2.04444 for CE. Thus, the walks visited two or fewer basins. The n_{stag} values are even smaller for macro-step walks initialised in the same range, i.e., 1.33333 for SSE, and 1.35556 for CE. Figures 3c and 3d illustrate that larger step sizes allowed some of the walk trajectories to skip the poor loss area, while the smaller steps consistently became stuck, and proceeded directly to one of the better minima. Small n_{stag} values indicate that transition between adjacent minima was still unlikely for the given step size.

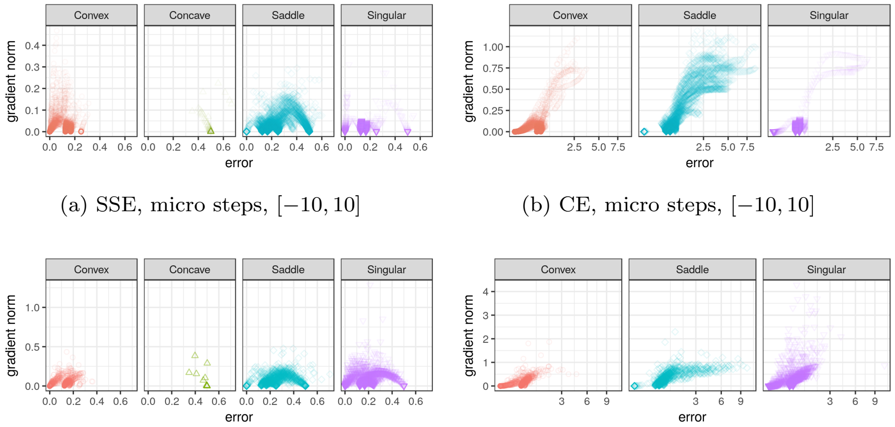
CE and SSE thus exhibited very similar properties when sampled with $[-1, 1]$

Table 3: Basin of attraction estimates calculated for the XOR problem. Standard deviation shown in parenthesis.

	SSE		CE	
	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$, micro	1.88889	367.04444	2.04444	313.32130
	(0.31427)	(134.84453)	(0.44500)	(148.75671)
$[-1, 1]$, macro	1.33333	37.14815	1.35556	30.35000
	(0.49441)	(16.68220)	(0.50136)	(13.32863)
$[-10, 10]$, micro	1.63333	684.77778	1.16667	870.87222
	(0.72188)	(263.72374)	(0.37268)	(180.17149)
$[-10, 10]$, macro	1.10000	57.98889	1.03333	74.79444
	(0.39581)	(24.91864)	(0.23333)	(20.49253)

gradient walks. The same number of local minima was observed, and the basins of attraction exhibited similar behaviour in terms of basin-to-basin transitions. According to Figure 3, CE exhibited stronger gradients. This corresponds to the theoretical predictions made in [14]. A comparison of Figures 3c and 3d shows that SSE exhibited more non-convex behaviour around the apparent local minima, which indicates that SSE would be harder to search for an optimisation algorithm than CE.

Figure 4 shows the l-g clouds obtained for gradient walks initialised in the $[-10, 10]$ range. Figures 4a and 4c indicate that initialisation in a wider range caused the gradient walks to discover more stationary points on the SSE loss surface: Instead of four attractors of zero gradient, six can be seen in the figures. Out of these six, only four exhibited convexity. Even the points that exhibited convexity were surrounded by points with saddle curvature or no curvature. Such overlap between convex and non-convex structure indicates that the surface around the minima was not smooth. Overlap of convexity and non-convexity can also indicate that multiple minima of the same loss value exist that exhibit



(a) SSE, micro steps, $[-10, 10]$

(b) CE, micro steps, $[-10, 10]$

(c) SSE, macro steps, $[-10, 10]$

(d) CE, macro steps, $[-10, 10]$

Figure 4: L-g clouds for the gradient walks initialised in the $[-10, 10]$ range for the XOR problem. Micro (maximum 1% of the initialisation range) and macro (maximum 10% of the initialisation range) steps were considered.

different landscape curvature properties.

Figures 4b and 4d show that the loss surface of CE exhibited noticeably different properties when probed in a larger range. The horizontal axis is shown in square root scale for clarity. While non-convex curvature remained prevalent, CE, as opposed to SSE, did not exhibit additional stationary attractors. Instead, points of high loss exhibited high gradient, leading the gradient walks towards the same basins as discovered with the $[-1, 1]$ walks. Four stationary attractors can be observed, only three of which exhibited convexity. Thus, CE exhibited fewer local minima than SSE. This observation corresponds with the theoretical predictions made in [14].

Once again, the convex minima observed in Figure 4 were disconnected from one another. No convex trajectory has been captured that visited all the stationary points present. Figure 4c shows that the only transition between the global optima and the adjacent local optima corresponded to the indefinite Hessians. Thus, to make a transition from one convex minimum to another one, the algorithm had to traverse a flat area with little to no convexity. With reference to Table 3, the n_{stag} values were smaller for the $[-10, 10]$ initialisation range, and the l_{stag} values were larger than those yielded by the $[-1, 1]$ walks. Thus, the walks were more likely to stagnate once, and to remain in the stagnated state for the entire walk.

A comparison of Figures 4a and 4b shows that CE demonstrated a smoother, more consistent relationship between the gradient and the loss values than SSE. Together with evidently fewer stationary points, this property makes CE an easier loss surface to minimise.

Figures 4c and 4d indicate that gradient walks with a macro step size, initialised in a larger area, still managed to find the global optima for both SSE and CE, but on fewer occasions than the micro walks. A large portion of the points yielded indefinite Hessians, indicating flatness. This is to be expected, as the loss surfaces of NNs with sigmoidal activation functions are known to exhibit increasing hidden neuron saturation with an increased distance from the origin [35].

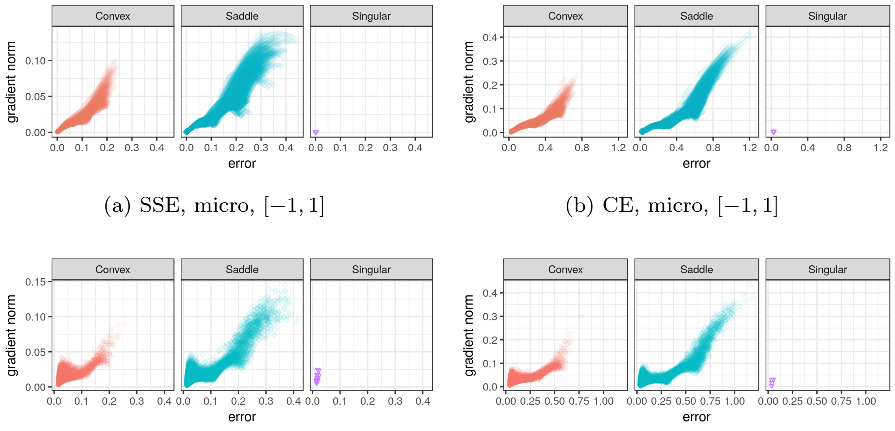
6.2. Iris

The Iris classification problem is one of the most trivial and most commonly used real-world classification datasets. The benefit of studying the Iris problem compared to the XOR problem is that the Iris dataset is large enough to be split into the training and testing subsets. The training subset can then be used to sample the loss surface, and the testing set can be used to evaluate the discovered minima and stationary points for their ability to generalise. For the rest of the paper, the training set loss values are referred to as E_t, and the test set loss values are referred to as E_g.

Figure 5 shows the l-g clouds obtained for the Iris problem using the $[-1, 1]$ initialisation interval. According to Figure 5, only one attractor with zero gradient has been discovered on both the SSE and CE loss surfaces by gradient walks initialised in the $[-1, 1]$ range. Two more attractors of non-zero gradient can also be observed, however, these attractors do not constitute local minima. Transition from non-convex space to convex space was still present, but was less distinct than for XOR. Points around the global minima exhibited convex as well as saddle behaviour, and saddle behaviour was prevalent. Both the SSE and CE surfaces exhibited flatness (indicated by the singular Hessians) around the global optima. This corresponds to theoretical claims that the loss surface around the global minima is flat [25]. However, the flatness was not prevalent.

A comparison of the micro and macro steps in Figure 5 indicates that the macro steps discovered the same landscape characteristics as the micro steps. In the macro setting, a wider range of gradient values around the global minima was discovered. This is explained by the fact that NN loss surfaces are known to contain ravines and valleys [19], and optima are typically found at the bottom of such structures. The macro step size caused the gradient walks to oscillate and to sample points on the sides of the valley where the global minima were discovered.

To further analyse the landscape properties sampled by the gradient walks, the n_{stag} and l_{stag} values were calculated for the E_t values sampled by the gradient walks, as well as for the corresponding E_g values. Table 4 lists the E_t and E_g values obtained. According to Table 4, both SSE and CE yielded



(a) SSE, micro, $[-1, 1]$

(b) CE, micro, $[-1, 1]$

(c) SSE, macro, $[-1, 1]$

(d) CE, macro, $[-1, 1]$

Figure 5: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range for the Iris problem.

an average n_{stag} very close or equal to 1 for all gradient walks initialised in the $[-1, 1]$ range. Thus, a single basin of attraction was discovered by each individual walk. This correlates well with the results shown in Figure 5. For the macro setting in the $[-1, 1]$ range, both SSE and CE produced an n_{stag} average of 1, with a standard deviation of zero. This observation indicates that the macro steps in the $[-1, 1]$ range were sufficient to prevent stagnation in suboptimal areas, yet convergence in an attraction basin still took place. The generalisation error exhibited similar behaviour, as shown in Table 4. The presence of a single global attractor makes the loss surface associated with the Iris problem trivial to search using a gradient-based method.

Figure 6 shows the l-g clouds obtained for the gradient walks initialised in the $[-10, 10]$ interval. According to Figures 6a and 6c, multiple stationary points were discovered on the SSE loss surface. Two of the discovered stationary points, including the global minima, exhibited convexity. Thus, there is at least one local minimum attractor on the SSE loss surface associated with the Iris problem. Additionally, the discovered stationary points were disjoint in the convex and singular (flat) space. The saddle space was more connected; however, the n_{stag} values presented in Table 4 indicate that the gradient walks did not generally become stuck more than twice. Thus, the multiple stationary points discovered were not trivial to escape from.

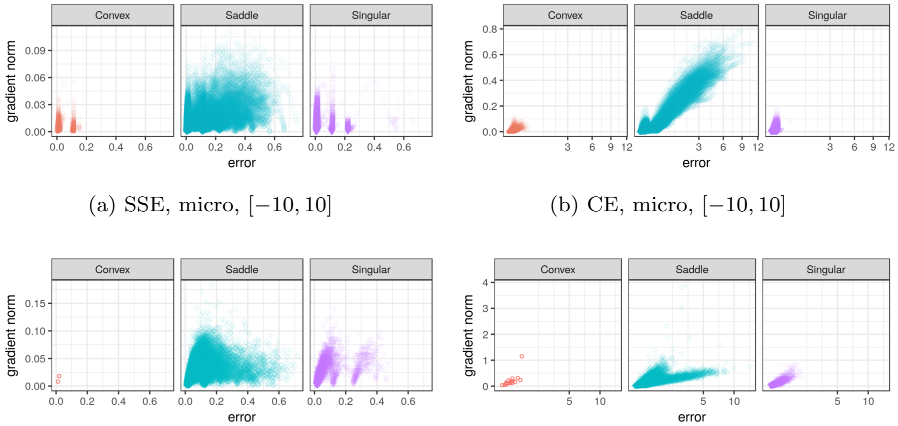
CE, on the other hand, exhibited only one attractor at the global minimum, as illustrated in Figure 6. Even though all points belong to the same global attraction basin, two distinct clusters can be observed in Figure 6b: points that lie in the low error region, and exhibit higher gradients, and points that lie in the higher error region, and exhibit lower gradients. The same tendency can be observed in Figure 6d. These observations indicate that the gradient walks have explored wide (higher error, lower gradient) as well as narrow (higher gradient, lower error) valleys, which the NN error landscapes are known to exhibit [21].

Thus, the CE loss surface again exhibited fewer local minima than SSE. The quality of the discovered minima can also be evaluated in terms of the generalisation capabilities. Figure 7 shows the l-g clouds colourised according

Table 4: Basin of attraction estimates calculated for the Iris problem on the E_t and E_g walks. Standard deviation shown in parenthesis.

	SSE		CE	
E_t	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00857	848.82048	1.00571	820.20429
micro	(0.11922)	(52.82897)	(0.07538)	(54.28522)
$[-1, 1]$,	1.00000	76.46571	1.00000	73.40857
macro	(0.00000)	(4.03382)	(0.00000)	(4.72216)
$[-10, 10]$,	1.28571	796.07000	1.00000	953.26286
micro	(0.48234)	(212.33922)	(0.00000)	(10.64382)
$[-10, 10]$,	1.02571	73.77000	1.00571	84.55857
macro	(0.15828)	(13.50309)	(0.07538)	(6.27217)
E_g	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.10571	820.07167	1.02286	818.33905
micro	(0.38206)	(143.30209)	(0.18375)	(77.64190)
$[-1, 1]$,	1.00000	74.69429	1.00286	67.71143
macro	(0.00000)	(4.52494)	(0.05338)	(7.26987)
$[-10, 10]$,	1.36000	770.39048	1.12286	917.17541
micro	(0.57231)	(229.39260)	(0.75160)	(138.19499)
$[-10, 10]$,	1.03143	75.41714	1.01429	83.85857
macro	(0.17447)	(13.86152)	(0.11867)	(8.78615)

to the corresponding E_g values. It is evident from Figure 7 that CE yielded poor generalisation performance in the area of the global minima: all E_g values reported were an order of magnitude larger than the corresponding E_t values. This observation is to be expected: achieving 100% accuracy on the training can lead to overfitting. SSE also exhibited overfitting at the global minima, but not as strongly as CE. CE exhibited stronger gradients around the global optima,



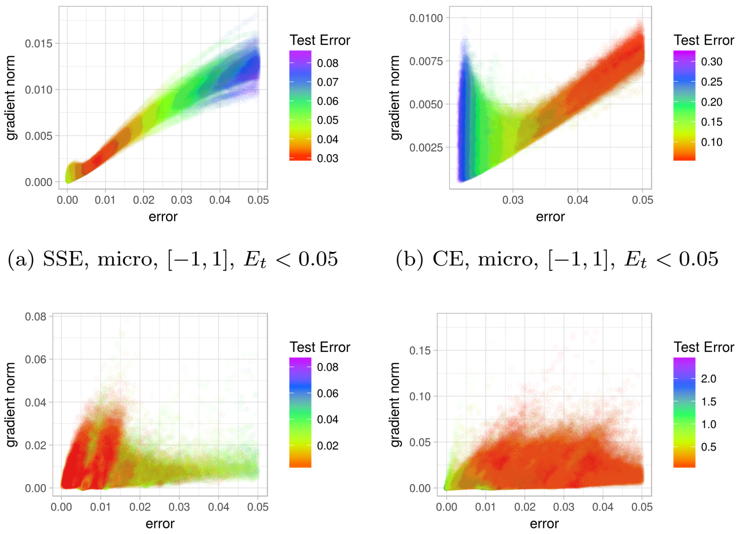
(a) SSE, micro, $[-10, 10]$

(b) CE, micro, $[-10, 10]$

(c) SSE, macro, $[-10, 10]$

(d) CE, macro, $[-10, 10]$

Figure 6: L-g clouds for the gradient walks initialised in the $[-10, 10]$ range for the Iris problem.



(a) SSE, micro, $[-1, 1]$, $E_t < 0.05$

(b) CE, micro, $[-1, 1]$, $E_t < 0.05$

(c) SSE, micro, $[-10, 10]$, $E_t < 0.05$

(d) CE, micro, $[-10, 10]$, $E_t < 0.05$

Figure 7: L-g clouds colourised according to the corresponding E_g values for the Iris problem.

which can promote overfitting when using gradient-based methods.

Appendix B lists all classification errors obtained by the gradient walks on the various problems. Table B.10 indicates for the Iris problem that SSE has indeed yielded better generalisation in most scenarios.

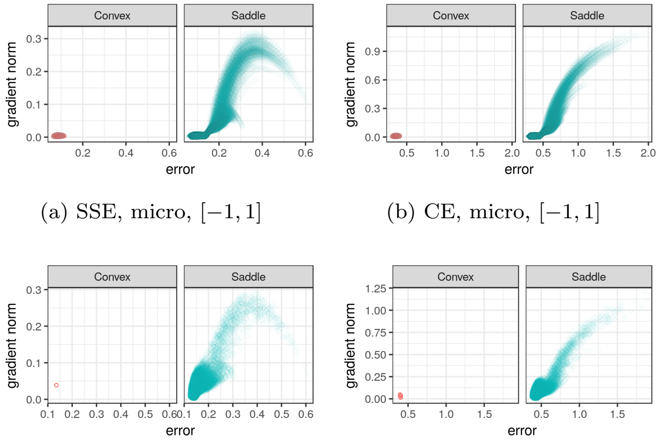
Thus, CE exhibited better global structure than SSE on the Iris problem, and was more searchable from the gradient descent perspective. However, stronger gradients around the global optima indicate that CE exhibited sharper minima, causing stronger overfitting on the CE loss surface.

6.3. Diabetes

Figure 8 shows the l-g clouds obtained for the Diabetes problem using the $[-1, 1]$ initialisation range. According to Figure 8, both SSE and CE exhibited a single attractor of near-zero gradient, and that attractor constituted a wide area of low gradients around the loss of zero. Both SSE and CE exhibited convexity around zero loss, especially when sampled with micro steps. The majority of the sampled points, however, were once again classified as saddles according to their Hessians. This corresponds well with the observations made by Dauphin et al. [6], where the prevalence of saddle points in non-convex optimisation was studied.

An arch-like curve can be observed in Figures 8a and 8c, indicating that higher errors were associated with weaker gradients on the SSE loss surface. A transition to the area of higher fitness was associated with a gradient signal that became stronger for some time, and then began to weaken again as a global optimum was approached. The CE l-g clouds in Figure 8 indicate that the CE loss surface did not have the tendency to exhibit weaker gradients for higher errors, which makes CE favourable from the gradient descent perspective. This corresponds well with the theoretical properties of both loss functions, which indicate that SSE is expected to exhibit weaker gradients for higher errors, as opposed to CE [16].

Figure 9 shows the l-g clouds obtained for the points sampled by the gradient walks initialised in the $[-10, 10]$ interval. SSE loss once again exhibited multiple near-zero gradient attractors (three), and CE loss exhibited only one attractor.



(a) SSE, micro, $[-1, 1]$

(b) CE, micro, $[-1, 1]$

(c) SSE, macro, $[-1, 1]$

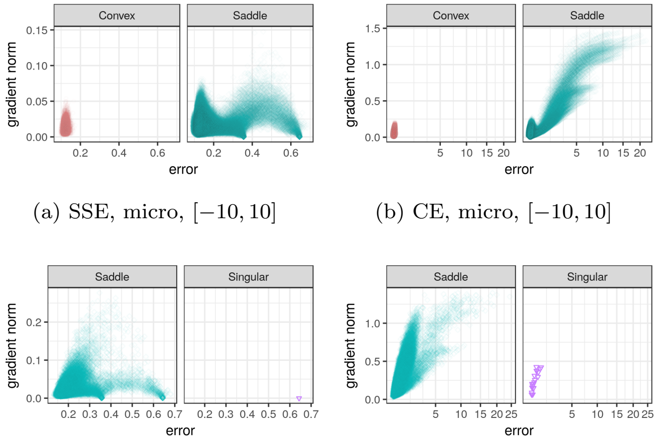
(d) CE, macro, $[-1, 1]$

Figure 8: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range for the Diabetes problem.

The majority of the points sampled by larger steps in a larger area had a saddle curvature. The convex attractors sampled by the $[-10, 10]$ walks exhibited more variation in gradient than the corresponding attractors discovered by the $[-1, 1]$ walks. This observation can be attributed to the valley structure of the optima: Larger steps induced oscillations around the walls of the valley.

The n_{stag} and l_{stag} values reported in Table 5 indicate that most walks discovered a single attractor only, which correlates well with Figures 8 and 9, and also indicates that the two suboptimal attractors discovered on the SSE loss surface were not easy to escape from. Table 5 also shows that the generalisation performance of the points discovered on the SSE loss surface was somewhat volatile when sampled using micro walks. Micro walks took smaller steps, and thus were more likely to exploit a particular attractor, causing overfitting.

Figure 10 shows a close-up depiction of the convex attractors, colourised according to their generalisation performance. Both SSE and CE exhibited deteriorating generalisation performance as the walks sampled points closer to the zero loss, which is to be expected. For micro $[-1, 1]$ walks, both SSE and



(a) SSE, micro, $[-10, 10]$

(b) CE, micro, $[-10, 10]$

(c) SSE, macro, $[-10, 10]$

(d) CE, macro, $[-10, 10]$

Figure 9: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range for the Diabetes problem.

CE exhibited a sudden drop in gradient magnitudes, and the points of low gradient with the highest error exhibited the best generalisation performance. As previously noted by Choromanska et al. [12], finding the global minimum may be unnecessary, as the global minimum is likely to overfit the problem. Figures 10c and 10d indicate for the $[-10, 10]$ walks that points around the global minima have exhibited various degrees of generalisation performance, with a significant overlap between good and poor generalisation. This indicates that the discovered minima had the same training error values, but different test error values. The Diabetes problem is known to contain noisy data, and noise is a common cause of overfitting. Table B.11 lists the classification errors obtained for the Diabetes problem, and shows that CE loss yielded better generalisation when sampled with the $[-1, 1]$ walks, and SSE generalised better when larger step sizes were used.

6.4. Glass

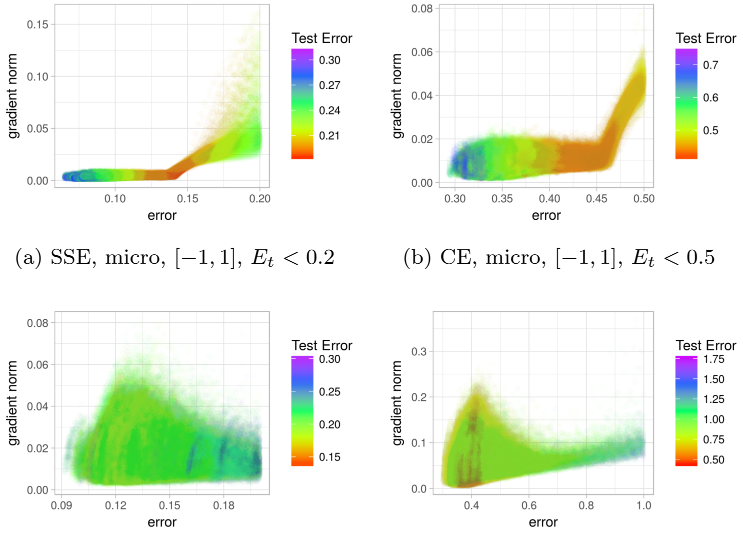
Figure 11 shows the l-g clouds obtained for the Glass problem. According to Figure 11, convexity was found around the global minima only, and only by the

Table 5: Basin of attraction estimates calculated for the Diabetes problem on the E_t and E_g walks. Standard deviation shown in parenthesis.

	SSE		CE	
E_t	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00123	938.66728	1.00000	935.27160
micro	(0.03511)	(22.51936)	(0.00000)	(12.49791)
$[-1, 1]$,	1.00000	85.19012	1.00000	84.84691
macro	(0.00000)	(1.54389)	(0.00000)	(1.97922)
$[-10, 10]$,	1.09259	905.05504	1.00000	962.31235
micro	(0.37525)	(138.96827)	(0.00000)	(5.28613)
$[-10, 10]$,	1.03580	77.06975	1.02716	78.23086
macro	(0.18580)	(13.75685)	(0.18393)	(16.45928)
E_g	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.51852	794.78363	1.04938	925.83735
micro	(1.21727)	(270.89365)	(0.31822)	(100.01013)
$[-1, 1]$,	1.00494	85.80988	1.00123	85.16543
macro	(0.07010)	(4.91134)	(0.03511)	(2.61851)
$[-10, 10]$,	2.76420	703.52152	1.00617	958.96852
micro	(3.96060)	(343.41982)	(0.07832)	(39.60395)
$[-10, 10]$,	1.08148	53.88477	1.08272	70.88848
macro	(0.52189)	(33.59152)	(0.35041)	(24.84558)

micro walks. Macro walks discovered exclusively saddle curvature points. This observation once again confirms that the search space for both SSE and CE is dominated by saddle curvature points. Convexity could only be discovered by the smallest steps tested, indicating that the convex area was sharp, and could easily be "overstepped" by a larger step size.

From Figure 11, the attractor dynamics exhibited by CE and SSE were



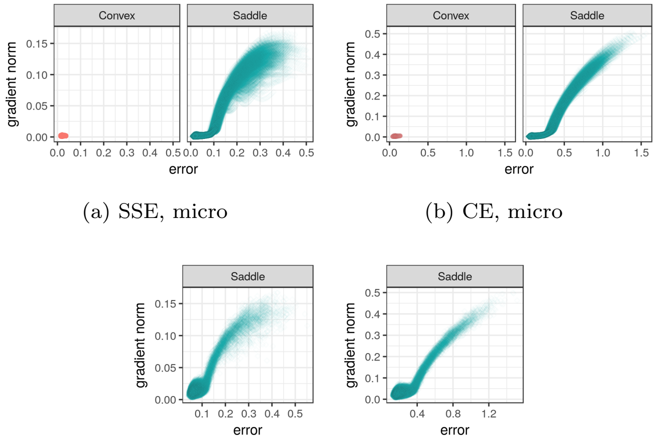
(a) SSE, micro, $[-1, 1]$, $E_t < 0.2$

(b) CE, micro, $[-1, 1]$, $E_t < 0.5$

(c) SSE, micro, $[-10, 10]$, $E_t < 0.2$

(d) CE, micro, $[-10, 10]$, $E_t < 1$

Figure 10: L-g clouds colourised according to the corresponding E_g values for the Diabetes problem.



(a) SSE, micro

(b) CE, micro

(c) SSE, macro

(d) CE, macro

Figure 11: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range for the Glass problem.

quite similar: both losses yielded a general near-linear decline in gradient associated with a decline in error. Once the error became low enough, the gradients flattened, and a further decrease in error towards zero was performed with near-zero gradients. Both CE and SSE exhibited a single major attractor around the global minima, indicating that all near-stationary points discovered by the walks had a similar error value. The macro steps discovered higher gradients around zero error than the micro steps, but the separation into flat and non-flat areas was still evident. This behaviour is likely to be caused by the gradient walks descending to the bottom of a valley first, and then travelling down the bottom of the valley towards a global minimum.

Table 6 reports the n_{stag} and l_{stag} values obtained by the various walks on the glass problem. All walks consistently discovered only one attractor. The l_{stag} values indicate that the attractor was found within the first 10% to 20% of the steps, and from thereon the walks proceeded to explore the discovered attractor. Thus, all walks quickly descended into a valley, and then travelled at the bottom of the valley for the majority of the steps. It was clearly quite easy to find a valley, and the error values at the bottom of all discovered valleys were rather similar. No inter-valley transition was observed.

The corresponding n_{stag} and l_{stag} values obtained for E_g indicate that E_g also yielded a single attractor per walk. Standard deviations of n_{stag} and l_{stag} are higher for E_g than for E_t, indicating that a steady decrease in E_t was not always associated with a steady decrease in E_g.

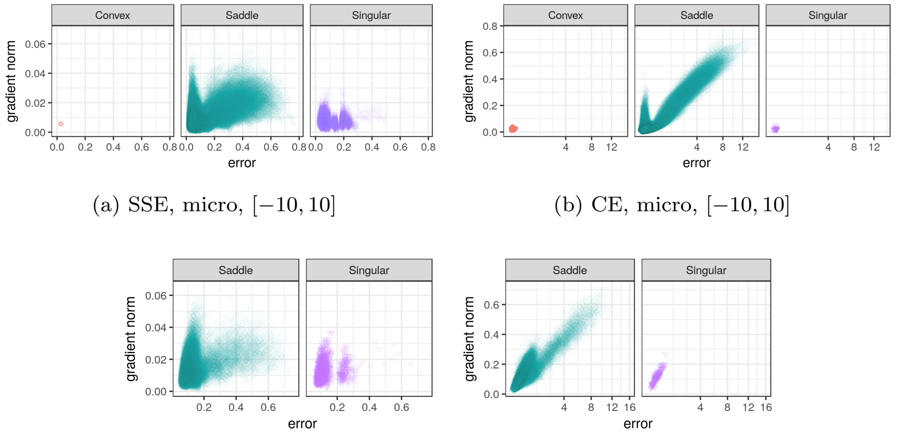
Figure 12 shows the l-g clouds obtained by the micro and macro walks initialised in the $[-10, 10]$ range. According to Figure 12, a larger initialisation range yielded indefinite Hessians, indicating that points of little to no curvature were discovered. A larger initialisation range is more likely to yield exploration of areas further away from the origin. Since the NNs in this study employed the sigmoid activation, the observed flatness is attributed to the saturation of the activation signals. Multiple flat attractors were observed for SSE, while CE exhibited a single major attractor. While this single attractor was at the global minima, the sampled points clustered around two “paths”: lower errors associ-

Table 6: Basin of attraction estimates calculated for the Glass problem. Standard deviation shown in parenthesis.

	SSE		CE	
E_t	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00000	947.12067	1.00000	939.70667
micro	(0.00000)	(7.81525)	(0.00000)	(7.50773)
$[-1, 1]$,	1.00000	86.13867	1.00000	85.04333
macro	(0.00000)	(0.77595)	(0.00000)	(1.03672)
$[-10, 10]$,	1.04133	927.42956	1.00000	961.23867
micro	(0.20238)	(96.74308)	(0.00000)	(5.18617)
$[-10, 10]$,	1.00400	85.29156	1.00133	87.17956
macro	(0.08155)	(4.99437)	(0.05162)	(2.28182)
E_g	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00000	951.66867	1.00800	941.96300
micro	(0.00000)	(8.18235)	(0.10925)	(42.59343)
$[-1, 1]$,	1.00000	86.67000	1.00000	86.01267
macro	(0.00000)	(0.57715)	(0.00000)	(0.71683)
$[-10, 10]$,	1.11400	902.05250	1.02733	950.75153
micro	(0.44084)	(148.80139)	(0.26568)	(73.87091)
$[-10, 10]$,	1.00533	85.01433	1.00400	86.62400
macro	(0.07283)	(6.10341)	(0.06312)	(4.83090)

ated with higher gradients, and higher errors associated with lower gradients. This indicates the presence of two structures: narrow as well as wide valleys.

It was previously observed that wide valleys are likely to yield better generalisation performance [13, 21]. There was also a counter-argument presented, where a sharp minimum with good generalisation properties was artificially created [47]. To study the generalisation performance of the sampled points, the



(a) SSE, micro, $[-10, 10]$

(b) CE, micro, $[-10, 10]$

(c) SSE, macro, $[-10, 10]$

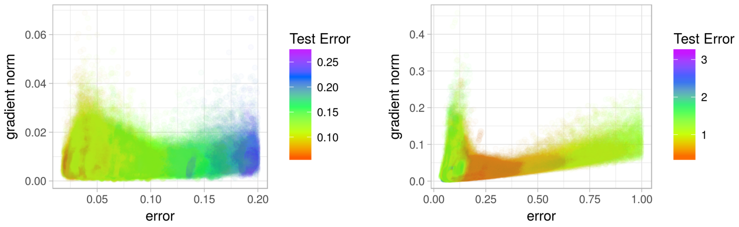
(d) CE, macro, $[-10, 10]$

Figure 12: L-g clouds for the gradient walks initialised in the $[-10, 10]$ range for the Glass problem.

l-g clouds obtained for the $[-10, 10]$ micro walks, colourised according to the E_g values, are presented in Figures 13a and 13b. Figure 13b confirms that points of large gradient and low error generalised poorly for CE, while points of higher error and lower gradient generalised better. Thus, points of low error exhibited overfitting for CE loss on the glass problem, and the wide valleys exhibited better generalisation properties. Interestingly, the same did not hold for SSE loss: according to Figure 13a, the smallest E_g was observed for the points of the lowest E_t. Thus, SSE loss was less prone to overfitting when sampled at the given resolution. Therefore, despite exhibiting more low gradient attractors, SSE exhibits better generalisation properties in some scenarios. The classification error values reported in Table B.12 indicate that SSE and CE have in fact performed very similarly, and have both generalised poorly. The glass dataset is rather small, and small datasets lead to overfitting.

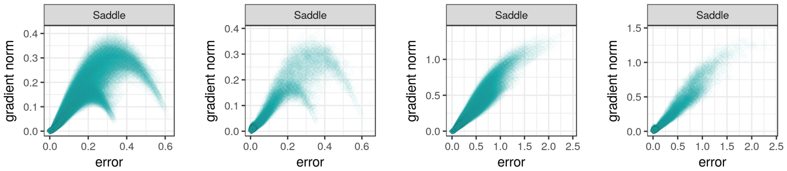
6.5. Cancer

Figure 14 shows the l-g clouds obtained for the Cancer problem. According to Figure 14, all points sampled by micro and macro walks initialised in the



(a) SSE, micro, $[-10, 10]$, $E_t < 0.2$ (b) CE, micro, $[-10, 10]$, $E_t < 1$

Figure 13: L-g clouds colourised according to the corresponding E_g values for the Glass problem.



(a) SSE, micro (b) SSE, macro (c) CE, micro (d) CE, macro

Figure 14: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range on the Cancer problem.

$[-1, 1]$ range exhibited saddle curvature. Total dimensionality of the cancer problem is 321, which is noticeably higher than that of the previous problems considered. Saddle curvature is expected to become more and more prevalent as the dimensionality increases [6].

According to Figure 14, both SSE and CE exhibited a single attractor at the global minimum. In addition to the global attractor, SSE exhibited two more attractors of low, but non-zero gradient. Trajectories can be observed leading to the global attractor from either of the two high error attractors. However, there is no trajectory connecting the attractors to one another. The n_{stag} and l_{stag} values reported in Table 7 confirm that all walks discovered a single attractor only, thus no transition between the attractors took place.

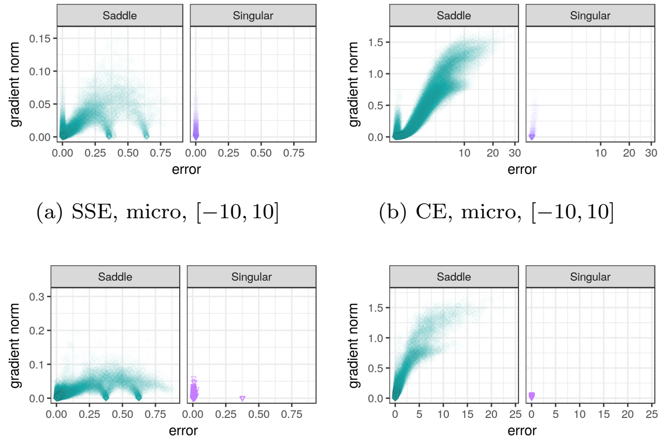
CE, as shown in Figure 14, exhibited almost linear correlation between the

Table 7: Basin of attraction estimates calculated for the Cancer problem. Standard deviation shown in parenthesis.

	SSE		CE	
E_t	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00000	962.09844	1.00000	953.89307
micro	(0.00000)	(5.39294)	(0.00000)	(5.37201)
$[-1, 1]$,	1.00000	87.77788	1.00000	87.18816
macro	(0.00000)	(0.44464)	(0.00000)	(0.51044)
$[-10, 10]$,	1.00000	972.77778	1.00000	975.43836
micro	(0.00000)	(7.45025)	(0.00000)	(3.01133)
$[-10, 10]$,	1.00000	87.44517	1.00000	87.80498
macro	(0.00000)	(0.89423)	(0.00000)	(1.12240)
E_g	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00000	959.38629	1.00725	953.80766
micro	(0.00000)	(5.62718)	(0.09002)	(41.40420)
$[-1, 1]$,	1.00000	87.81838	1.00000	87.25514
macro	(0.00000)	(0.41360)	(0.00000)	(0.54000)
$[-10, 10]$,	1.11111	932.44444	4.13699	541.93665
micro	(0.45812)	(154.23691)	(4.50666)	(384.89318)
$[-10, 10]$,	1.00125	87.16246	1.00903	86.55711
macro	(0.03528)	(2.95538)	(0.09786)	(7.86466)

gradient and the error. Such simple correlation implies that the CE loss surface is likely to be more searchable than the SSE loss surface from the perspective of a gradient-based optimisation algorithm. The cancer problem is known to be an easy classification problem, which must have contributed to the simplicity of the observed attractor.

Figure 15 shows the l-g clouds for the micro and macro walks initialised in



(a) SSE, micro, $[-10, 10]$

(b) CE, micro, $[-10, 10]$

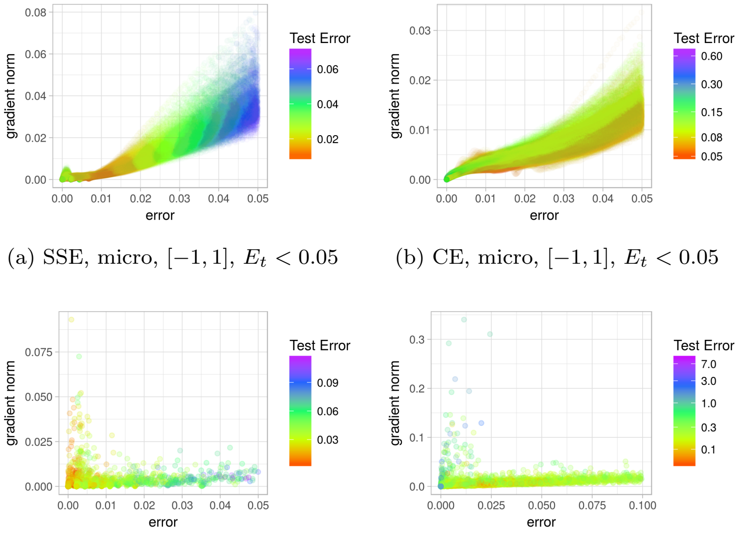
(c) SSE, macro, $[-10, 10]$

(d) CE, macro, $[-10, 10]$

Figure 15: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range on the Cancer problem.

the $[-10, 10]$ range. The larger initialisation range once again exposed points with indefinite Hessians for both SSE and CE, i.e., points with little to no curvature. For CE, the points of no curvature aligned with the global minimum attractor. For SSE, the global minimum, as well as the other two attractors, exhibited flatness. The majority of the points exhibited saddle curvature. Two zero-gradient local minimum attractors were observed for the SSE loss surface. The CE loss surface did not exhibit multiple attractors. However, multiple points of high gradient close to the global minimum were sampled. This once again indicates that CE is more prone to sharp minima (narrow valleys) than SSE.

The n_{stag} and l_{stag} values yielded by E_g (Table 7) are inconsistent with the corresponding n_{stag} and l_{stag} values obtained for E_t. To further study this inconsistency, Figure 16 presents the l-g clouds colourised according to the E_g values for the points around the global attractor. Due to high disparity in the E_g values obtained for CE, the CE l-g clouds were colourised on logarithmic scale. Similar to the previous problems considered, the generalisation performance at



(a) SSE, micro, $[-1, 1]$, $E_t < 0.05$

(b) CE, micro, $[-1, 1]$, $E_t < 0.05$

(c) SSE, micro, $[-10, 10]$, $E_t < 0.05$

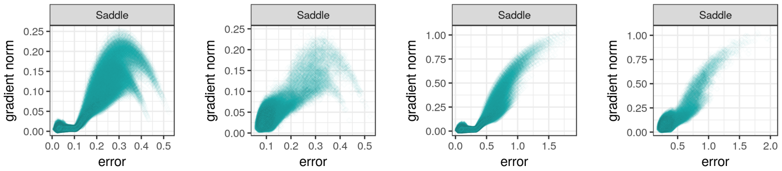
(d) CE, micro, $[-10, 10]$, $E_t < 0.1$

Figure 16: L-g clouds colourised according to the corresponding E_g values for the Cancer problem.

the global optimum was poor for both SSE and CE. However, it is evident from Figures 16c and 16d that low error, high gradient points around the global attractor generalised well for SSE, and poorly for CE. SSE in general produced weaker gradients than CE, indicating that SSE was less prone to sharp minima. Figure 16 also shows that SSE exhibited points of zero gradient for non-zero error, while CE did not. However, the observed local minima, as well as the global optimum of SSE, can yield better generalisation performance than the global minimum exhibited by CE.

6.6. Heart

Figures 17 shows the l-g clouds obtained for the Heart problem. Similar to the cancer problem, all points sampled by the $[-1, 1]$ walks were classified as saddle points. The total dimensionality of the heart problem is 341, which is similar to the dimensionality of the Cancer problem. Figure 17 illustrates that both SSE and CE had a single flat attractor in the general area of the global



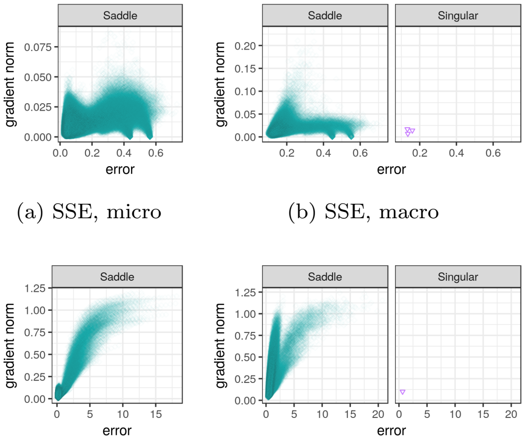
(a) SSE, micro (b) SSE, macro (c) CE, micro (d) CE, macro

Figure 17: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range on the Heart problem.

minima. In addition to this attractor, SSE exhibited two more attractors of much higher error. However, the $[-1, 1]$ walks did not sample any zero-gradient (stationary) points around the high error attractors.

Larger steps and a larger initialisation range, however, allowed gradient walks to discover the stationary points of high error on the SSE loss surface, as illustrated in Figure 18. The CE loss surface sampled by the same walks did not reveal any additional attractors, but was again visibly split into two clusters leading towards the global minima: points of high gradient and low error, and points of lower gradient and higher error. This is once again indicative of narrow and wide valleys, which appears to be a common characteristic of the CE loss surface.

The n_{stag} and l_{stag} values reported in Table 8 confirm that the walks generally did not make transitions between the discovered attractors. The n_{stag} and l_{stag} values calculated over the E_g values were again less stable than the corresponding E_t values, indicating that exploiting an E_t attractor does not necessarily coincide with exploiting a corresponding E_g attractor. Figure 19 illustrates the generalisation behaviour of the flat attractor discovered on both the SSE and CE loss surfaces by the micro $[-1, 1]$ walks: the smallest E_g values were observed on the rightmost side of the attractor, closest to the points of higher error and higher gradient. Exploitation around the global minima yielded superior E_t values, but inferior E_g values. This again illustrates that discovering the global optimum may be unnecessary. The success of techniques



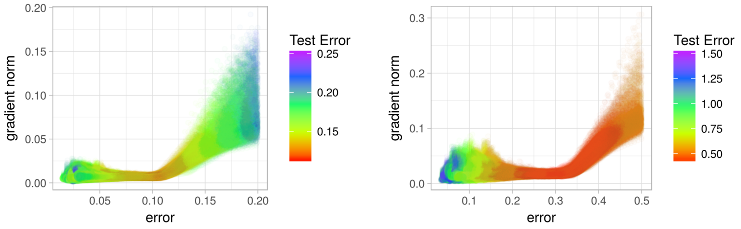
(a) SSE, micro

(b) SSE, macro

(c) CE, micro

(d) CE, macro

Figure 18: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range on the Heart problem.



(a) SSE, micro, $[-1, 1]$, $E_t < 0.2$

(b) CE, micro, $[-1, 1]$, $E_t < 0.5$

Figure 19: L-g clouds colourised according to the corresponding E_g values for the Heart problem.

such as early stopping [48] comes precisely from preventing the algorithm from exploiting a global minimum unnecessarily.

6.7. MNIST

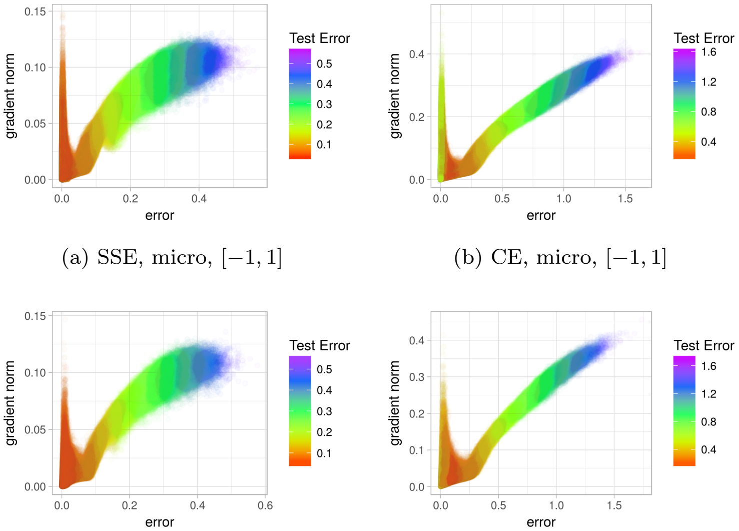
Figures 20 and 21 show the l-g clouds for the MNIST problem. Due to the prohibitively expensive memory requirements, the Hessian matrices were not computed for the MNIST dataset. Thus, the curvature of the loss functions for

Table 8: Basin of attraction estimates calculated for the Heart problem. Standard deviation shown in parenthesis.

	SSE		CE	
E_t	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00000	952.36276	1.00000	947.81432
micro	(0.00000)	(6.49124)	(0.00000)	(7.04222)
$[-1, 1]$,	1.00000	86.70645	1.00000	86.40587
macro	(0.00000)	(0.66921)	(0.00000)	(0.97088)
$[-10, 10]$,	1.02493	937.01486	1.00000	966.72036
micro	(0.15962)	(76.63092)	(0.00000)	(3.70200)
$[-10, 10]$,	1.00176	84.85293	1.00411	84.43886
macro	(0.04191)	(3.68841)	(0.06394)	(7.00978)
E_g	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00733	957.87269	2.14920	710.77930
micro	(0.10669)	(40.88520)	(2.07157)	(342.24282)
$[-1, 1]$,	1.00000	87.70880	1.00088	87.54971
macro	(0.00000)	(0.60395)	(0.02965)	(1.97717)
$[-10, 10]$,	1.54927	821.66135	1.00298	965.68084
micro	(1.78543)	(251.33524)	(0.05453)	(27.96970)
$[-10, 10]$,	1.00440	84.55381	1.04956	79.23624
macro	(0.06618)	(5.61240)	(0.24735)	(17.04847)

the MNIST dataset is not reported in this study. The reader is referred to the previous studies of the MNIST Hessians [8] for a discussion of curvature characteristics, where it was shown that the gradient descent algorithm discovered points of saddle and singular curvature only.

775 Figure 20 shows that both SSE and CE exhibited one attractor around the global minimum. Additionally, SSE exhibited two more attractors of non-zero



(a) SSE, micro, $[-1, 1]$

(b) CE, micro, $[-1, 1]$

(c) SSE, macro, $[-1, 1]$

(d) CE, macro, $[-1, 1]$

Figure 20: L-g clouds for the gradient walks initialised in the $[-1, 1]$ range on the MNIST problem.

gradient. Thus, the error landscape of CE was more searchable than the error landscape of SSE. The n_{stag} and l_{stag} results reported in Table 9 indicate that most walks have discovered a single attractor only, which corresponds to the results in Figures 20 and 21.

A cluster of values of high gradient and low error can be observed for both SSE and CE, indicating that both exhibited sharp minima. SSE, however, exhibited lower gradients overall. Figure 20 illustrates that the generalisation performance improved as the error approached zero. Figure 22 shows the generalisation performance of the points sampled around the global minima. SSE once again exhibited a better generalisation performance around the global minima than CE, confirming the earlier made hypothesis that SSE is less prone to overfitting due to weaker gradients. The classification error results reported in Table B.15, however, indicate that, although SSE yielded a smaller disparity between the E_t and E_g values, both loss functions performed similarly in terms

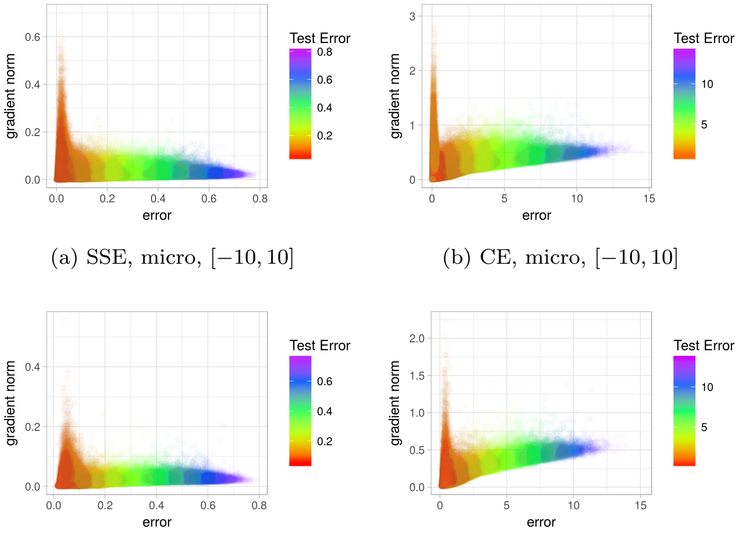
Table 9: Basin of attraction estimates calculated for the MNIST problem. Standard deviation shown in parenthesis.

	SSE		CE	
E_t	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00003	948.96269	1.00020	943.65445
micro	(0.00557)	(7.73257)	(0.01418)	(10.31210)
$[-1, 1]$,	1.00000	89.59761	1.00000	88.94583
macro	(0.00000)	(0.62686)	(0.00000)	(0.79698)
$[-10, 10]$,	1.00338	944.23606	1.00020	955.48716
micro	(0.06420)	(29.25325)	(0.01418)	(9.14026)
$[-10, 10]$,	1.00004	90.19884	1.00001	90.24536
macro	(0.00614)	(1.02171)	(0.00354)	(1.09101)
E_g	n_{stag}	l_{stag}	n_{stag}	l_{stag}
$[-1, 1]$,	1.00028	944.25561	2.84430	570.85913
micro	(0.01762)	(10.01749)	(2.71734)	(334.02547)
$[-1, 1]$,	1.00000	90.04197	1.01201	85.19988
macro	(0.00000)	(0.63536)	(0.11234)	(7.54223)
$[-10, 10]$,	1.00408	943.53542	1.26670	878.55561
micro	(0.11775)	(29.40333)	(1.10976)	(191.85100)
$[-10, 10]$,	1.00005	90.31260	1.00881	88.80587
macro	(0.00709)	(1.00828)	(0.10216)	(6.14361)

of final classification.

Figure 21 shows that SSE exhibited a much weaker correlation between the gradient and the error when sampled by gradient walks initialised in the $[-10, 10]$ interval. For CE, the positive correlation was still clearly manifested. Thus, CE exhibited a more searchable landscape when sampled by the $[-10, 10]$ walks.

The landscape properties exhibited by the MNIST problem were thus very



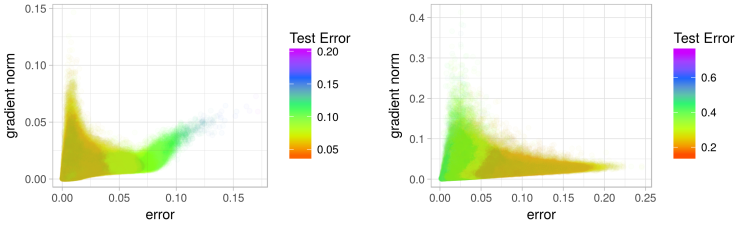
(a) SSE, micro, $[-10, 10]$

(b) CE, micro, $[-10, 10]$

(c) SSE, macro, $[-10, 10]$

(d) CE, macro, $[-10, 10]$

Figure 21: L-g clouds for the gradient walks initialised in the $[-10, 10]$ range on the MNIST problem.



(a) SSE, macro, $[-10, 10]$, $E_t < 0.2$

(b) CE, macro, $[-10, 10]$, $E_t < 0.25$

Figure 22: L-g clouds colourised according to the corresponding E_g values for the MNIST problem.

similar to the landscape properties exhibited by the problems of lower dimensionality. The CE loss surface was more searchable for all problems considered, and exhibited fewer non-global attractors. SSE, however, exhibited somewhat better generalisation capabilities under some of the considered scenarios. Perhaps the two loss functions should be combined to construct an error landscape that is both searchable and robust to overfitting.

7. Conclusions

This study presented a visual and numerical analysis of local minima and the associated basins of attraction for two common NN loss functions, i.e., quadratic loss and entropic loss. The study was performed by analysing the samples obtained by a number of progressive gradient walks proportionate to the dimensionality of the problems. The gradient walks were not restricted to any specific search space bounds, but were initialised in two distinct intervals, i.e., $[-1, 1]$ and $[-10, 10]$. Additionally, two granularity settings were considered for the gradient walks, namely micro and macro.

This study proposed an intuitive visualisation of the local minima and the associated basins of attraction, namely the loss-gradient clouds. By plotting the sampled loss values against the corresponding gradient vector magnitudes, stationary points could be easily identified. To classify the identified stationary points as minima, maxima, or saddles, Hessian matrix information was used to identify the curvature of each sampled point.

Additionally, this study proposed two simple metrics to quantify the number and extent of attraction basins as sampled by the walks. Calculation of statistical metrics over a number of walks provides an idea of the connectedness of the various basins, as well as the likelihood of escaping from the basins.

Both loss functions exhibited convex local minima for the XOR problem. The amount of observed convexity decreased with the increase in problem dimensionality. Saddle curvature was the most prevalent curvature observed, and some higher-dimensional problems considered exhibited only saddle curvature for all sampled points.

SSE consistently exhibited more local stationary points and associated attractors than CE. Analysis of the individual walks further revealed that transition between different attractors was unlikely, and that the paths connecting different attractors exhibited singular Hessian matrices, indicative of flatness.

Thus, CE exhibited a more consistent and searchable structure across the selection of problems considered in this study.

With an increase in problem dimensionality, the number of zero or low gradient attractors decreased. The majority of the problems exhibited a single main attractor around the global optimum. For CE, the gradient was for the most part positively correlated to the error value, indicating that the CE loss surface is highly searchable from the perspective of gradient-based methods. This study did not attempt to quantify the number of optima, but the results obtained clearly indicated that the majority of the optima exhibited similar loss values.

The results confirmed previously made observations of the presence of valley-shaped optima in NN error landscapes. For the majority of the problems, descending into a valley was easily accomplished by the walks. Travelling down the bottom of the valley towards the global minimum yielded a decrease in generalisation performance for both SSE and CE. CE exhibited stronger gradients than SSE in all experiments conducted, which promoted overfitting in CE. For some of the problems, SSE exhibited a better generalisation performance. It can be speculated that the CE loss surface is more prone to sharp minima (narrow valleys) than SSE; thus, CE is more easily overfitted. The experiments revealed the tendency for the points sampled on CE to fall into two major clusters: points of low error and high gradients, and points of higher error and low gradients. These are hypothesised to represent narrow and wide valleys, respectively. The results of this study confirmed that superior generalisation performance was exhibited by the points in the wide valleys.

An analysis of the progressive gradient samples thus illustrated a number of current theories regarding the shape of NN error surfaces, and highlighted the differences between SSE and CE loss surfaces, confirming that FLA is a viable method for visualisation and analysis of NN error landscapes. Future research will apply FLA to analyse the influence of various activation functions, as well as NN architectures, on the resulting stationary points and attraction basins.

The observation that the SSE landscape may have superior generalisation properties suggests that a hybrid of SSE and CE may produce a landscape that combines the searchability of CE with the robustness of SSE. Additionally, the presence of a single attractor in the majority of the problems considered suggests that an exploitative rather than an exploratory approach should be taken for the purpose of NN training. This observation has strong implications for population-based training algorithms, which so far failed to be effectively applied to high-dimensioal NN training problems. A population-based approach designed with exploitation rather than exploration in mind may perform competitively, especially if gradient information is used as one of the guides for the population. This hypothesis is further supported by a recent study of particle swarm optimisation in high-dimensional spaces [49], where the efficacy of exploitation over exploration in high-dimensional spaces was observed. Investigation of exploitative population-based techniques applied to NNs is an interesting topic for future research.

Another interesting observation is the impressive ability of a randomised algorithm to find the global optima, when guided by nothing besides the direction of the gradient. As Appendix B indicates, the average classification error calculated at the last step of the gradient walks approached 100% accuracy on most problems under at least one of the granularity settings. Perhaps gradient-guided stochastic training algorithms should be considered for deeper, more complex problems.

Acknowledgements

The authors would like to thank the Centre for High Performance Computing (CHPC) (http://www.chpc.ac.za) for the use of their cluster to obtain the data for this study.

Declarations of interest: none.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[2] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[3] A. Choromanska, Y. LeCun, G. B. Arous, Open problem: The landscape of the loss surfaces of multilayer networks, in: Proceedings of The 28th Conference on Learning Theory, 2015, pp. 1756–1760.

[4] M. Kordos, W. Duch, A survey of factors influencing MLP error surface, Control and Cybernetics 33 (4) (2004) 611–631.

[5] H. Shen, Towards a mathematical understanding of the difficulty in learning with feedforward neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 811–820.

[6] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, in: Advances in Neural Information Processing Systems, 2014, pp. 2933–2941.

[7] K. Kawaguchi, Deep learning without poor local minima, in: Advances in Neural Information Processing Systems, 2016, pp. 586–594.

[8] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, L. Bottou, Empirical analysis of the Hessian of over-parametrized neural networks, in: Proceedings of the International Conference on Learning Representations, 2018, pp. 1–15.

[9] L. G. Hamey, XOR has no local minima: A case study in neural network error surface analysis, Neural Networks 11 (4) (1998) 669–681.

[10] I. G. Sprinkhuizen-Kuyper, E. J. Boers, The local minima of the error surface of the 2-2-1 XOR network, Annals of Mathematics and Artificial Intelligence 25 (1-2) (1999) 107.

[11] G. Swirszcz, W. M. Czarnecki, R. Pascanu, Local minima in training of neural networks, arXiv e-printsarXiv:1611.06310.

[12] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. LeCun, The loss surfaces of multilayer networks, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, 2015, pp. 192–204.

[13] C. Xing, D. Arpit, C. Tsirigotis, Y. Bengio, A Walk with SGD, arXiv e-printsarXiv:1802.08770.

[14] S. Solla, E. Levin, M. Fleisher, Accelerated learning in layered neural networks, Complex Systems 2 (1988) 625–640.

[15] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[16] P. Golik, P. Doetsch, H. Ney, Cross-entropy vs. squared error training: a theoretical and experimental comparison., in: Interspeech, Vol. 13, 2013, pp. 1756–1760.

[17] I. G. Sprinkhuizen-Kuyper, E. J. Boers, A local minimum for the 2-3-1 XOR network, IEEE Transactions on Neural Networks 10 (4) (1999) 968–971.

[18] D. Mehta, X. Zhao, E. A. Bernal, D. J. Wales, Loss surface of XOR artificial neural networks, Physical Review E 97 (5) (2018) 052307.

[19] M. R. Gallagher, Multi-layer perceptron error surfaces: Visualization, structure and modelling, Ph.D. Thesis, University of Queensland, St Lucia 4072, Australia (2000).

[20] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. T. P. Tang, On large-batch training for deep learning: Generalization gap and sharp minima, in: Proceedings of the International Conference for Learning Representations, 2017.

[21] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, R. Zecchina, Entropy-SGD: Biasing gradient descent into wide valleys, in: Proceedings of The International Conference on Learning Representations, 2017, pp. 1–19.

[22] F. Draxler, K. Veschgini, M. Salmhofer, F. Hamprecht, Essentially no barriers in neural network energy landscape, in: J. Dy, A. Krause (Eds.), Proceedings of the International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 1309–1318.
URL `http://proceedings.mlr.press/v80/draxler18a.html`

[23] D. M. Kline, V. L. Berardi, Revisiting squared-error and cross-entropy functions for training neural network classifiers, Neural Computing & Applications 14 (4) (2005) 310–318.

[24] H. Bourlard, N. Morgan, Connectionist speech recognition: a hybrid approach, Vol. 247 of The Kluwer international series in engineering and computer science, Boston: Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[25] P. Auer, M. Herbster, M. K. Warmuth, Exponentially many local minima for single neurons, in: Advances in Neural Information Processing Systems, 1996, pp. 316–322.

[26] T. Jones, Evolutionary algorithms, fitness landscapes and search, Ph.D. Thesis, The University of New Mexico (1995).

[27] P. Merz, B. Freisleben, Fitness landscape analysis and memetic algorithms for the quadratic assignment problem, IEEE Transactions on Evolutionary Computation 4 (4) (2000) 337–352.

[28] K. M. Malan, A. P. Engelbrecht, Quantifying ruggedness of continuous landscapes using entropy, in: IEEE Congress on Evolutionary Computation, IEEE, 2009, pp. 1440–1447.

[29] K. M. Malan, Characterising continuous optimisation problems for particle swarm optimisation performance prediction, Ph.D. Thesis, University of Pretoria (2014).

[30] M. A. Muñoz, Y. Sun, M. Kirley, S. K. Halgamuge, Algorithm selection for black-box continuous optimization problems: A survey on methods and challenges, Information Sciences 317 (2015) 224–245.

[31] Y. Sun, S. K. Halgamuge, M. Kirley, M. A. Muñoz, On the selection of fitness landscape analysis metrics for continuous optimization problems, in: Proceedings of the International Conference on Information and Automation for Sustainability, IEEE, 2014, pp. 1–6.

[32] A. Rakitianskaia, E. Bekker, K. M. Malan, A. Engelbrecht, Analysis of error landscapes in multi-layered neural networks for classification, in: Proceedings of the IEEE Congress on Evolutionary Computation, IEEE, 2016, pp. 5270–5277.

[33] A. S. Bosman, A. Engelbrecht, M. Helbig, Search space boundaries in neural network error landscape analysis, in: Proceedings of the IEEE Symposium Series on Computational Intelligence, IEEE, 2016, pp. 1–8.

[34] A. Bosman, A. Engelbrecht, M. Helbig, Fitness landscape analysis of weight-elimination neural networks, Neural Processing Letters 48 (1) (2018) 353–373.

[35] W. A. van Aardt, A. S. Bosman, K. M. Malan, Characterising neutrality in neural network error landscapes, in: Proceeding of the IEEE Congress on Evolutionary Computation, IEEE, 2017, pp. 1374–1381.

[36] K. Alyahya, J. E. Rowe, Simple random sampling estimation of the number of local optima, in: International Conference on Parallel Problem Solving from Nature, Springer, 2016, pp. 932–941.

[37] C. Edwards, Advanced Calculus of Several Variables, Academic Press, 1973.

[38] A. S. Bosman, A. P. Engelbrecht, M. Helbig, Progressive gradient walk for neural network fitness landscape analysis, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, ACM, 2018, pp. 1473–1480.

[39] K. M. Malan, A. P. Engelbrecht, A progressive random walk algorithm for sampling continuous fitness landscapes, in: Proceedings of the IEEE Congress on Evolutionary Computation, IEEE, 2014, pp. 2507–2514.

[40] M. Friendly, D. Denis, The early origins and development of the scatterplot, Journal of the History of the Behavioral Sciences 41 (2) (2005) 103–130.

[41] H. Wickham, ggplot2: Elegant graphics for data analysis, software available from `https://ggplot2.tidyverse.org` (2016).
URL `https://ggplot2.tidyverse.org`

[42] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, Tech. rep., California University, San Diego La Jolla Institute for Cognitive Science (1985).

[43] P. R. Winters, Forecasting sales by exponentially weighted moving averages, Management science 6 (3) (1960) 324–342.

[44] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188.

[45] L. Prechelt, Proben1 – a set of neural network benchmark problems and benchmarking rules, Tech. rep., Universität Karlsruhe, Karlsruhe, Germany (Sep. 1994).

[46] Y. LeCun, C. Cortes, C. Burges, MNIST handwritten digit database, AT&T Labs [Online]. Available: `http://yann.lecun.com/exdb/mnist`.

[47] L. Dinh, R. Pascanu, S. Bengio, Y. Bengio, Sharp minima can generalize for deep nets, in: Proceedings of the International Conference on Machine Learning, 2017, pp. 1019–1028.

[48] N. Morgan, H. Bourlard, Generalization and parameter estimation in feedforward nets: Some experiments, in: Advances in Neural Information Processing Systems, 1990, pp. 630–637.

[49] E. T. Oldewage, The perils of particle swarm optimization in high dimensional problem spaces, Master's Thesis, University of Pretoria (2018).

Appendix A. Pseudocode for basin of attraction estimates

Two estimates to quantify the basins of attraction are proposed in this study:

1. The average number of times stagnation observed, n_{stag}.
2. The average length of the stagnant sequence, l_{stag}.

The pseudocode given in Algorithms 1 and 2 summarises the proposed method to obtain both metrics.

Appendix B. Classification errors

The average classification errors arrived at by the gradient walks are reported in this appendix. Averages are calculated across the error values as observed at the last step of each walk. The classification error of the training set is referred to as C_t, and the classification error of the test set is referred to as C_g. Tables B.10, B.11, B.12, B.13, B.14, and B.15 list the average C_t and C_g values obtained for the iris, diabetes, glass, cancer, and MNIST problems, respectively. Standard deviation is shown in parenthesis.

Table B.10: Iris, classification errors.

	micro				macro			
	SSE		CE		SSE		CE	
	C_t	C_g	C_t	C_g	C_t	C_g	C_t	C_g
$[-1, 1]$	1.00000	0.91819	0.98352	0.93333	0.96638	1.00000	0.97557	0.96667
	(0.00000)	(0.01660)	(0.00125)	(0.00000)	(0.00881)	(0.00000)	(0.00646)	(0.00000)
$[-10, 10]$	0.97252	0.97105	0.99245	0.90581	0.92155	0.92829	0.92857	0.92457
	(0.07622)	(0.08790)	(0.00734)	(0.05097)	(0.09578)	(0.10521)	(0.05844)	(0.05806)

Table B.11: Diabetes, classification errors.

	micro				macro			
	SSE		CE		SSE		CE	
	C_t	C_g	C_t	C_g	C_t	C_g	C_t	C_g
$[-1, 1]$	0.91094	0.66913	0.85453	0.73725	0.81141	0.73586	0.81187	0.74165
	(0.01002)	(0.02400)	(0.00991)	(0.02684)	(0.00959)	(0.01712)	(0.01017)	(0.02307)
$[-10, 10]$	0.85434	0.74521	0.83494	0.69911	0.79669	0.68657	0.71915	0.66424
	(0.01441)	(0.02648)	(0.01480)	(0.03019)	(0.02970)	(0.03580)	(0.06485)	(0.05338)

Table B.12: Glass, classification errors.

	micro				macro			
	SSE		CE		SSE		CE	
	C_t	C_g	C_t	C_g	C_t	C_g	C_t	C_g
$[-1, 1]$	0.94400	0.55738	0.93769	0.62740	0.79600	0.68398	0.79793	0.67744
	(0.01636)	(0.04912)	(0.01551)	(0.04766)	(0.02738)	(0.04128)	(0.02285)	(0.05012)
$[-10, 10]$	0.79578	0.60513	0.90388	0.62657	0.71373	0.58626	0.69585	0.55828
	(0.08762)	(0.08170)	(0.02785)	(0.05711)	(0.06541)	(0.06897)	(0.07784)	(0.08112)

Table B.13: Cancer, classification errors.

	micro				macro			
	SSE		CE		SSE		CE	
	C_t	C_g	C_t	C_g	C_t	C_g	C_t	C_g
$[-1, 1]$	0.99944	0.97298	1.00000	0.97322	0.99451	0.97656	0.99633	0.97487
	(0.00096)	(0.00788)	(0.00000)	(0.00861)	(0.00227)	(0.00759)	(0.00275)	(0.00887)
$[-10, 10]$	0.99813	0.96206	1.00000	0.96408	0.99539	0.96574	0.99357	0.97335
	(0.00150)	(0.01170)	(0.00000)	(0.00685)	(0.00279)	(0.01006)	(0.00612)	(0.00961)

Table B.14: Heart, classification errors.

	micro				macro			
	SSE		CE		SSE		CE	
	C_t	C_g	C_t	C_g	C_t	C_g	C_t	C_g
$[-1, 1]$	0.97447	0.78274	0.97918	0.77466	0.91038	0.83086	0.90601	0.82772
	(0.00477)	(0.02250)	(0.00648)	(0.02138)	(0.00925)	(0.01538)	(0.00915)	(0.01743)
$[-10, 10]$	0.95409	0.76148	0.93496	0.80585	0.85821	0.83363	0.80135	0.74857
	(0.00910)	(0.02149)	(0.01096)	(0.02425)	(0.01829)	(0.02063)	(0.05700)	(0.05340)

Table B.15: MNIST, classification errors.

	micro				macro			
	SSE		CE		SSE		CE	
	C_t	C_g	C_t	C_g	C_t	C_g	C_t	C_g
$[-1, 1]$	0.98922	0.56534	0.99846	0.57332	0.96611	0.60834	0.98404	0.61831
	(0.01097)	(0.04874)	(0.00395)	(0.04828)	(0.01829)	(0.04600)	(0.01334)	(0.04547)
$[-10, 10]$	0.87408	0.49537	0.95444	0.52401	0.74988	0.46981	0.70077	0.44259
	(0.05040)	(0.05590)	(0.02668)	(0.05063)	(0.06487)	(0.06232)	(0.07736)	(0.06666)

Algorithm 1 Basins of attraction estimates

```
Initialise $n_{stag}$, average number of basins, to 0;
Initialise $l_{stag}$, average basin size, to 0;
Initialise $n_w$ to the number of walks to perform;
Initialise *walk*, the sample, to $\emptyset$;
**for** $\forall i \in \{1, ..., n_w\}$ **do**
    *walk* $\leftarrow$ sample the input problem using a progressive gradient walk [38];
    Normalise the sample fitness range in *walk* to $[0, 1]$;
    Initialise $n_{stag,i}$ and $l_{stag,i}$ to 0 for walk $i$;
    **for** $\forall j \in \{6, 8, ..., 18, 20\}$ **do**
        *walk* $\leftarrow$ calculate the EWMA of *walk* using Equation (3), $\alpha = 2/(j+1)$
        $\varsigma \leftarrow$ calculate the standard deviation of *walk*
        $\sigma \leftarrow$ calculate the sequence of moving standard deviations of *walk*
        Get a list of stagnant regions, $list$, using Algorithm 2 with inputs $\sigma$, $\varsigma$.
        **if** average length of regions in $list > l_{stag,i}$ **then**
            $l_{stag,i} \leftarrow$ average($list$)
            $n_{stag,i} \leftarrow$ number of regions in $l$
        **end if**
    **end for**
    $n_{stag} \leftarrow n_{stag} + n_{stag,i}$
    $l_{stag} \leftarrow l_{stag} + l_{stag,i}$
**end for**
return $n_{stag}/n_w$, $l_{stag}/n_w$
```

Algorithm 2 Basins of attraction identification

```
Inputs: $\sigma$, $\varsigma$;
Initialise $l_{stag}$, average basin size, to 0;
Initialise stuck to false
Initialise len, length of a stagnant region, to 0;
Initialise list, the list of stagnant regions, to $\emptyset$;
for each step $s_i$ in $\sigma$ do
  if stuck then
    if $s_i < \varsigma$ then
      $len \leftarrow len + 1$
    else
      stuck ←false
      list ← add len to list
      $len \leftarrow 0$
    end if
  else
    if $s_i < \varsigma$ then
      $len \leftarrow len + 1$
      stuck ←true
    end if
  end if
end for
if $len > 0$ then
  list ← add len to list
end if
return list
```