Recent Advances in Landscape Analysis of Deep Learning Loss Functions

Authors: Liang Ruobing, Liu Bo, Sun Yuehong, Liu Bo

Date: 2021-11-29T00:00:00+00:00

Abstract

In the research fields of machine learning and mathematical optimization, providing a mathematical explanation for the ease of optimization in deep learning problems presents a significant challenge. Loss functions exhibit characteristics of high dimensionality, non-convexity, and non-smoothness, yet global optima can be successfully located through gradient descent methods. Loss landscape analysis has emerged as a crucial research direction for uncovering the fundamental nature of this optimization ease in deep learning. To advance the deployment of interpretable and trustworthy deep learning in more critical domains, this paper reviews the research progress and challenges concerning loss landscape characteristics (the number and spatial distribution of local minima, connectivity between optima, optimality of critical points), the convergence properties of gradient descent methods, and loss landscape visualization.

Full Text

Loss Landscape Analysis for Deep Learning: A Survey

Ruobing Liang[1],[2], Bo Liu[1]*, Yuehong Sun[3]
[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Mathematical Sciences, Nanjing Normal University, Nanjing 210046, China

Abstract

In the fields of machine learning and mathematical optimization, providing a mathematical explanation for the ease of optimization in deep learning remains highly challenging. Despite the high-dimensional, non-convex, and non-smooth characteristics of loss functions, gradient descent methods consistently succeed

in finding global optima. Loss landscape analysis has emerged as a crucial research direction for revealing the fundamental reasons behind the optimization tractability of deep learning problems. To promote the application of interpretable and trustworthy deep learning in critical domains, this paper reviews research progress and challenges in three key areas: loss landscape characteristics (including the number and spatial distribution of local minima, connectivity between optima, and optimality of critical points), convergence properties of gradient descent methods, and visualization techniques for loss landscapes.

Keywords: deep learning; loss function; landscape analysis

1. Introduction

Deep learning employs multi-layer network architectures composed of neurons, connection weights, biases, and activation functions to progressively transform low-level feature representations into high-level ones, thereby achieving optimal or near-optimal input-output mappings for complex representation learning tasks [1, 2]. For instance, in lung cancer detection, deep learning utilizes supervised, semi-supervised, or even unsupervised learning to extract features—identifying lesion edges and their combinations at lower layers while performing conceptual recognition at higher layers [3, 4]. After overcoming critical challenges such as vanishing gradients [5], training data scarcity, and limited computational power, various deep learning models have been successfully applied to computer vision, business intelligence, and medical image analysis, surpassing the performance of experienced human experts on specific tasks [6].

The loss function, a non-negative function of network parameters, measures the discrepancy between neural network predictions and ground truth values. Deep learning employs optimization algorithms such as gradient descent to adjust network parameters, minimizing the loss function until network outputs align with or approximate the true values, thereby achieving optimal input-output mapping [1].

Loss functions exhibit challenging properties—including high dimensionality, non-convexity, and non-smoothness—that pose significant difficulties for optimization theory and algorithm design. A mysterious phenomenon emerges when training deep neural networks: despite the highly non-convex nature of the loss function, first-order gradient descent with random initialization can converge to global optima, achieving zero training loss [7, 8]. This empirical observation, which contradicts conventional optimization theory, has motivated extensive research into its underlying causes. Notably, overparameterization [9-11] and universal approximation theorems [12-14], while explaining the representational capacity of deep networks, fail to account for gradient descent' s ability to achieve zero training loss.

Mathematically explaining the optimality of deep learning optimization prob-

lems represents an extremely challenging research endeavor. Since 2015, researchers in machine learning and mathematical optimization have begun analyzing deep learning loss landscape characteristics to provide theoretical explanations for optimization tractability. Over the past five years, substantial progress has been made in three areas: loss landscape feature analysis, gradient descent convergence analysis, and loss landscape visualization. Loss landscape analysis has become an effective mathematical tool for revealing the fundamental nature of deep learning optimization, characterizing properties of optimal solutions, and analyzing optimization algorithm convergence. This survey reviews research progress in deep learning loss landscape analysis, identifies remaining challenges, and outlines future research directions.

The remainder of this paper is organized as follows: Section 2 provides preliminaries on loss functions and gradient descent methods; Section 3 reviews progress in loss landscape feature analysis; Section 4 discusses advances in gradient descent convergence analysis; Section 5 covers loss landscape visualization techniques; and Section 6 presents challenges and future research directions.

---

2. Preliminaries

2.1 Loss Function

The loss function is defined as the error of a neural network on a training sample set, expressed as a function of network parameters $\theta = \{W_i, b_i\}_{i=1}^{L}$, where W_i represents the weight matrix of layer i and b_i denotes the bias vector of layer i. The loss function in equation (1) is commonly used for continuous representation learning tasks, while the binary cross-entropy loss function in equation (2) is typically employed for discrete representation learning tasks.

For a loss function $L(\theta)$, let x and y denote the actual input and label, respectively, where $x \in \mathbb{R}^d$ is a d-dimensional vector and $y \in \mathbb{R}^c$ is a c-dimensional vector. Here $\hat{y}(\theta, x)$ represents the network' s predicted label given parameters θ and input x. The matrix $X \in \mathbb{R}^{d\times n}$ consists of n input vectors as columns, $Y \in \mathbb{R}^{c\times n}$ comprises n label vectors as columns, and $\hat{Y} \in \mathbb{R}^{c\times n}$ contains n network output predictions as columns.

Minimizing the expected loss over the data generation distribution p_{data} yields the population risk: $R(\theta) = \mathbb{E}_{(x,y)\sim p_{\text{data}}}[L(\theta; x, y)]$. Given a finite training set, we can only compute the empirical risk $\hat{R}(\theta)$, which represents the average loss over the training set: $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(\theta; x_i, y_i)$. Following the principle of empirical risk minimization, we seek optimal parameters θ^* that minimize the average training loss: $\theta^* = \arg\min_\theta \hat{R}(\theta)$.

Loss functions exhibit extremely high parameter dimensionality. For example, LeNet-5 [15] contains approximately 60,000 parameters, ResNet-10 [16] about 10 million parameters, AlexNet [17] approximately 60 million parameters, and

VGG16 [18] roughly 140 million parameters.

Loss functions are non-convex. Kawaguchi [19] proved that deep neural network loss functions possess critical points without negative eigenvalues, demonstrating their non-convex nature. Dauphin et al. [20] discovered that the ratio of saddle points to local minima increases exponentially with function dimensionality. Saddle points exhibit large loss values and occupy extensive flat regions, contributing significantly to non-convexity.

Loss functions are not strictly smooth. For instance, the ReLU activation function is non-differentiable at zero. For backpropagation convenience, its derivative at zero is typically defined as zero.

These challenging properties—high dimensionality, non-convexity, and non-smoothness—pose substantial obstacles to mathematical analysis of optimality.

2.2 Gradient Descent

Gradient descent is a fundamental algorithm for neural network parameter updates [1]. Based on the amount of data required for each parameter update, gradient descent methods can be categorized into batch gradient descent (BGD), stochastic gradient descent (SGD), and mini-batch gradient descent (MGD). BGD updates parameters after computing gradients over all training samples. SGD randomly selects one sample per iteration, offering fast learning but unstable convergence. MGD uses m samples per update ($m \leq n$), reducing convergence oscillation while requiring careful selection of batch size.

---

3. Loss Landscape Feature Analysis

This section reviews progress in analyzing deep learning loss landscape features using mathematical tools including surrogate models, matrix analysis, stochastic analysis, and differential geometry. We examine three aspects: the number and spatial distribution of local minima, connectivity between global optima, and optimality of critical points.

3.1 Number and Spatial Distribution of Local Minima

Understanding the number and spatial distribution of local minima helps characterize optimization difficulty. Due to the high-dimensional nature of loss functions, direct acquisition of such information is infeasible, prompting research into surrogate model approximations [21].

Choromanska et al. [22] discovered that deep linear neural network loss functions share properties with the Hamiltonian of spin glass models. Low-index critical points in high-dimensional loss functions form a hierarchical structure where these critical points are also local minima located within a bounded region whose lower bound is defined by global optima. Outside this region, the probability

of finding low-index critical points decreases exponentially with loss function dimensionality. Building on this work, Becker and Zhang [23] employed random matrix theory and algebraic geometry to establish that loss functions follow the same distribution as spherical spin glass model Hamiltonians, expressing the loss function as a function of network depth. When the number of parameters remains constant, increasing network depth reduces the number of critical points and concentrates optima in parameter space, making the loss function easier to optimize. Cooper [24] utilized Sard' s theorem from differential geometry to prove that for fully-connected ReLU networks, the set of critical points forms a non-empty submanifold whose dimension equals the difference between the number of parameters and the number of samples.

3.2 Connectivity Between Optima

Exploring connectivity or reachability between optima helps characterize their spatial distribution and explains optimization tractability. Garipov et al. [25] discovered that optima can be connected by simple curves, and using any point on such a curve as new network parameters yields training loss nearly identical to the original network. This finding provides a novel geometric interpretation for deep networks with different parameters but equivalent representational capacity. Nguyen [26] leveraged properties such as linear independence of hidden layer outputs and connected sets to characterize the solution set for convex loss functions in deep fully-connected networks with piecewise linear activations. When a hidden layer width exceeds the training set size and subsequent layers decrease in width, all weight matrices become full-rank and connected, forming a single connected set containing all global optima.

3.3 Optimality of Critical Points

Critical points—where the loss function gradient vanishes—include saddle points, local minima, and global optima. Investigating properties of critical points, particularly their optimality, helps explain loss landscape tractability.

3.3.1 Deep Linear Neural Networks Kawaguchi [19] established sufficient conditions for the Hessian matrix at critical points of deep linear networks to be positive semidefinite, ensuring no local minima exist in the critical point set. When T_{XX} and T_{XY} are full-rank and the number of distinct eigenvalues equals the label dimension, critical points are either global optima or saddle points, with each saddle point' s Hessian possessing at least one negative eigenvalue. Lu and Kawaguchi [27] relaxed Kawaguchi' s conditions to require only that X and Y be full-rank, using singular value decomposition and continuity of singular subspaces. Zhou and Liang [28] proved that all local minima in the critical point set are global optima without any assumptions on network parameters or data matrices, relying solely on weight matrix factorization. Yun et al. [29] showed that when the product of weight matrices is full-rank, loss function critical points can only be global optima or saddle points, providing a criterion for distinction:

critical points within the set where the weight matrix product is full-rank are global optima, while those outside are saddle points. For deep linear residual networks, Hardt and Ma [30] leveraged the identity mapping structure to prove that when weight matrix spectral norms are uniformly small, all critical points are global optima.

3.3.2 Deep Nonlinear Neural Networks Research has examined how different nonlinear activation functions—such as ReLU, analytic, and smooth functions—affect critical point optimality.

For deep ReLU networks, Kawaguchi [19] adopted assumptions from Choromanska et al. [22] and introduced random vectors following Bernoulli distributions to represent original predictions as products of data matrices, weight matrices, and random vectors. By constructing the expected loss and leveraging averaging properties to cancel activation nonlinearities, they extended deep linear network conclusions to nonlinear cases, showing critical points are either global optima or saddle points.

For deep fully-connected networks with analytic activations and ℓ_2 loss, Nguyen and Hein [31] derived sufficient conditions for critical points to be global optima. Using full-rank matrices and analytic function properties, they proved that when one hidden layer width exceeds input data dimension, weight matrices from layer k onward are row full-rank, and network width decreases layer-by-layer after k, non-degenerate critical points are global optima with no low-rank local minima. Nguyen and Hein [32] extended this analysis to convolutional neural networks by reformulating convolutional operations as fully-connected computations, discovering infinitely many critical points in parameter sets where layer k outputs are linearly independent and weights from k to l are full-rank—all of which are global optima.

For smooth activation functions, Yun et al. [29] defined function spaces for each network layer and formulated the loss function in terms of these layer mappings. Critical points were defined as points where the Fréchet derivative of the loss function vanishes for any mapping. While they provided sufficient conditions for global optimality in function space, these results do not extend to parameter space. No suboptimal points exist in function space—any such point has a descent direction. However, when mapping suboptimal points to parameter space, their descent direction in function space may be orthogonal to the parameter space, potentially creating local minima in parameter space.

4. Gradient Descent Convergence Analysis

This section reviews research on gradient descent convergence analysis using neural tangent kernels, conjugate kernels, and Gram matrices.

4.1 Convergence

The Neural Tangent Kernel (NTK) provides a novel analytical tool for studying gradient descent convergence [33]. Jacot et al. [33] discovered that infinitely wide neural networks are equivalent to Gaussian processes and proposed the NTK to describe network training dynamics. For a depth-L network, the NTK is defined as $K(\theta) = \sum_{i=1}^{L} \langle \frac{\partial F(\theta)}{\partial \theta_i}, \frac{\partial F(\theta)}{\partial \theta_i} \rangle$, where $F(\theta)$ represents network output and $\langle \cdot, \cdot \rangle$ denotes the inner product. They proved that gradient descent in parameter space is equivalent to kernel gradient descent in function space using the NTK. The positive definiteness of the NTK guarantees gradient descent convergence, with infinitely wide networks converging fastest along the principal components of the kernel matrix. Lee et al. [34] showed that gradient descent dynamics in infinitely wide networks can be approximated by first-order Taylor expansion of the network output at initialization, determined by the NTK and initial network output. Using Gronwall' s inequality, they proved that the distance between actual network output and its first-order Taylor approximation is bounded by $\mathcal{O}(1/\sqrt{m})$, where m is network width, establishing global convergence of NTK gradient descent. Chen et al. [35] employed NTK random feature functions [36] to approximate network functions within a neighborhood of initial parameters, proving global convergence of both batch and stochastic gradient descent for ReLU-activated binary classification networks when width is a polynomial function of input dimension logarithm.

4.2 Convergence Rate

Sankararaman et al. [37] analyzed how network architecture affects mini-batch SGD convergence speed. Mini-batch SGD selects small subsets of training samples per iteration, but obtained gradients may be negatively correlated, preventing determination of the loss descent direction—a phenomenon termed "gradient confusion." The study found that increasing network width reduces gradient confusion and accelerates convergence. However, their results only guarantee convergence to a stable point, not necessarily a global optimum.

4.3 Convergence and Convergence Rate

For simpler deep linear networks, Arora et al. [38] constructed weight matrix constraints and proved linear convergence to global optima with appropriate learning rates. Du et al. [39] analyzed batch gradient descent convergence using Gram matrices. Network dynamics depend on the minimum eigenvalue of the Gram matrix; controlling this eigenvalue from initialization effectively bounds it. Additionally, in overparameterized networks, weight matrices remain close to their initial values. Assuming the Gram matrix' s minimum eigenvalue stays positive, they proved that for fully-connected feedforward networks with exponentially growing layer widths, gradient descent converges linearly to zero loss. For deep residual and convolutional residual networks, polynomial width growth suffices. These theoretical results apply only to smooth or Lipschitz continuous

activations with ℓ_2 loss using batch gradient descent.

Zou and Gu [40] proved that when randomly initialized weights follow Gaussian distributions and hidden layer widths are polynomial functions of input data quantity, data point distances, and network depth, both batch and stochastic gradient descent generate sequences with small perturbations near initial weights. For ReLU-activated binary classification networks, this ensures good local properties and global convergence, with convergence rates depending on assumptions.

Under non-degenerate data and polynomial hidden layer width assumptions (in terms of depth and sample size), Allen-Zhu et al. [11] proved two key loss function properties: (1) No saddle points exist—if not at a global optimum, the gradient norm exceeds zero and increases with loss value; (2) Semi-smoothness—the distance between the loss function and its first-order approximation is small. Based on these properties, they proved linear convergence to zero loss for both batch and stochastic gradient descent in deep ReLU networks. With smooth activations, width requirements can be further relaxed, maintaining conclusions even with extremely small hidden layers. Allen-Zhu et al. [41] extended these results to ReLU-activated recurrent neural networks, similarly proving absence of saddle points, semi-smoothness, and linear convergence under polynomial width conditions. Under identical assumptions, Zou and Gu [42] reduced requirements on network width magnitude and iteration count, obtaining tighter gradient lower bounds and more precise convergence rates. Daniely [43] used SGD to learn functions in conjugate kernel space, proving polynomial-time convergence to zero loss for networks with depth between 2 and $\log(n)$ when network size and iterations are polynomial in input-output dimensions.

5. Loss Landscape Visualization

Visualizing high-dimensional loss landscapes using two- or three-dimensional representations presents a significant challenge. This section reviews dimensionality reduction methods that preserve key information while minimizing information loss.

5.1 Filter Normalization

Network scale invariance—the property that scaling weight parameters does not affect predictions—hinders comparative visualization of loss functions across different parameters. Preprocessing is required to remove this effect.

Filter normalization effectively eliminates scale invariance [44]. Direction vectors in the 2D coordinate system are normalized according to filter norms, enhancing correlation between sharpness and generalization error while facilitating observation of local convexity. Specifically, random direction vectors matching network parameter dimensions are generated and normalized per filter, producing scale-invariant directions. Li et al. [44] applied filter normalization to visualize ResNet loss landscapes, plotting contour maps around minima along

two random directions. They observed that deeper networks exhibit increasingly chaotic landscapes with greater non-convexity and larger test errors at minima, indicating reduced generalization capacity. Notably, skip connections in residual networks effectively increase flatness around minima, preventing landscape chaos.

5.2 Principal Component Analysis

Principal Component Analysis (PCA) transforms variables into linearly uncorrelated principal components via orthogonal transformation. Retaining a subset of components preserves most information while achieving dimensionality reduction [45].

Li et al. [44] constructed a matrix from network parameters across iterations, applied PCA to identify the two most informative linearly independent principal components, and projected parameters onto these components. Using projection coefficients as coordinates, they plotted SGD convergence paths with loss contours, effectively visualizing optimization dynamics.

5.3 Multidimensional Scaling

Multidimensional Scaling (MDS) is a technique for displaying distance structures in low-dimensional space while preserving pairwise distances between high-dimensional points [46]. By constructing an inner product matrix from distance matrices and selecting dominant eigenvalues and eigenvectors, MDS creates low-dimensional projections.

Liao and Poggio [47] demonstrated MDS on deep convolutional networks, using distance matrices as similarity measures to maintain consistent relative distances between parameters in 2D space. They visualized convergence paths for both SGD and batch gradient descent, revealing multiple zero-loss points. Despite varying convergence paths from small perturbations, all initializations converge to global optima.

5.4 PHATE Method

PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding) is a diffusion-manifold learning method for dimensionality reduction [48]. It encodes local information via local similarities, represents pairwise similarities as diffusion probabilities, captures global information through diffusion processes, generates distance matrices, and applies MDS for low-dimensional embedding.

Using PHATE, Horoi et al. [49] studied the ruggedness of loss landscapes around ResNet minima, characterizing relationships between surface features and generalization. They found multiple zero-loss points in ResNet loss functions, where generalization capacity depends on local landscape geometry: minima in flatter neighborhoods exhibit smaller generalization errors than those in rugged regions.

6. Challenges and Future Directions

6.1 Loss Landscape Analysis

1) **Relaxing constraints on training data and network parameters:** Historical analyses impose restrictive conditions for theoretical tractability, such as independent inputs [19, 22, 31], analytic activations [24, 31, 32], polynomial network width [24, 29, 31], and differentiable loss functions [19]. Future work should focus on discrete loss functions, non-smooth activations, and arbitrary training distributions.

2) **Investigating complex network architectures:** Prior research primarily addresses linear [19, 27-30] and fully-connected networks [22-24, 26, 31], deriving optimality conditions via matrix factorization [27, 28] and full-rank conditions [29, 31]. Future studies should examine convolutional and recurrent network architectures.

6.2 Gradient Descent Convergence Analysis

1) **Relaxing width requirements for global convergence:** Existing convergence analyses assume exponential or polynomial width [11, 39-42]. Research should address optimality with reduced width magnitudes.

2) **Developing linearized convergence theory:** The NTK characterizes gradient descent dynamics through first-order Taylor expansion, providing a novel tool for infinite-width networks [33, 34, 36, 43]. Extending linearized approximations to nonlinear networks and estimating generalization errors remain open problems.

3) **Accelerating gradient descent:** Techniques for escaping saddle points [50] and balanced initialization [38] have shown promise. Further algorithmic acceleration is needed.

6.3 Generalization Error Analysis via Loss Landscapes

Visualization reveals qualitative relationships between landscape ruggedness and generalization—flatter zero-loss neighborhoods correlate with better generalization [49]. Beyond theoretical generalization bounds [35, 36], visualization-based explanations of network generalization represent a promising research direction.

Acknowledgments

We thank Professors Shouyang Wang and Xiaoshan Gao from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences; Professors Yihui Jin and Ling Wang from Tsinghua University; Professor Jikun Huang

from Peking University; and Professor Jianxing He from the National Clinical Research Center for Respiratory Disease for their valuable suggestions and assistance.

References

[1] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.
[2] Schmidhuber, Jürgen. Deep Learning in Neural Networks: An Overview[J]. Neural Netw, 2015, 61: 85-117.
[3] Liu B, Chi W, Li X, et al. Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect[J]. Journal of cancer research and clinical oncology, 2020, 146(1): 153-185.
[4] Yang Y, Feng X, Chi W, et al. Deep learning aided decision support for pulmonary nodules diagnosing: a review[J]. Journal of thoracic disease, 2018, 10(Suppl 7): S867.
[5] Hochreiter S. Untersuchungen zu dynamischen neuronalen Netzen[J]. Diploma, Technische Universität München, 1991, 91(1).
[6] Fukushima K, Miyake S, Ito T. Neocognitron: A neural network model for a mechanism of visual pattern recognition[J]. IEEE transactions on systems, man, and cybernetics, 1983, (5): 826-834.
[7] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[J]. arXiv preprint arXiv:1611.03530, 2016.
[8] Goodfellow I J, Vinyals O, Saxe A M. Qualitatively characterizing neural network optimization problems[J]. arXiv preprint arXiv:1412.6544, 2014.
[9] Telgarsky M. Representation benefits of deep feedforward networks[J]. arXiv preprint arXiv:1509.08101, 2015.
[10] Li D, Ding T, Sun R. On the benefit of width for neural networks: Disappearance of bad basins[J]. arXiv, 2018: arXiv: 1812.11039.
[11] Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization[C]// International Conference on Machine Learning, 2019: 242-252.
[12] Csáji B C. Approximation with artificial neural networks[J]. Faculty of Sciences, Etvs Lornd University, Hungary, 2001, 24(48): 7.
[13] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural networks, 1989, 2(5): 359-366.
[14] Lu Z, Pu H, Wang F, et al. The expressive power of neural networks: A view from the width[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6231-6239.
[15] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.
[17] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing

Systems, 2012: 1097-1105.
[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[19] Kawaguchi K. Deep learning without poor local minima[C]// Advances in Neural Information Processing Systems, 2016: 586-594.
[20] Dauphin Y N, Pascanu R, Gulcehre C, et al. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization[C]// Advances in Neural Information Processing Systems, 2014: 2933-2941.
[21] Ballard A J, Das R, Martiniani S, et al. Energy landscapes for machine learning[J]. Physical Chemistry Chemical Physics, 2017, 19(20): 12585-12603.
[22] Choromanska A, Henaff M, Mathieu M, et al. The loss surfaces of multilayer networks[C]// Artificial intelligence and statistics: PMLR, 2015: 192-204.
[23] Becker S, Zhang Y. Geometry of energy landscapes and the optimizability of deep neural networks[J]. Physical review letters, 2020, 124(10): 108301.
[24] Cooper Y. The loss landscape of overparameterized neural networks[J]. arXiv preprint arXiv:1804.10200, 2018.
[25] Garipov T, Izmailov P, Podoprikhin D, et al. Loss surfaces, mode connectivity, and fast ensembling of dnns[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 8803-8812.
[26] Nguyen Q. On connected sublevel sets in deep learning[C]// International Conference on Machine Learning: PMLR, 2019: 4790-4799.
[27] Lu H, Kawaguchi K. Depth creates no bad local minima[J]. arXiv preprint arXiv:1702.08580, 2017.
[28] Zhou Y, Liang Y. Critical points of neural networks: Analytical forms and landscape properties[J]. arXiv preprint arXiv:1710.11205, 2017.
[29] Yun C, Sra S, Jadbabaie A. Global optimality conditions for deep neural networks[J]. arXiv preprint arXiv:1707.02444, 2017.
[30] Hardt M, Ma T. Identity matters in deep learning[J]. arXiv preprint arXiv:1611.04231, 2016.
[31] Nguyen Q, Hein M. The loss surface of deep and wide neural networks[C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70: JMLR. org, 2017: 2603-2612.
[32] Nguyen Q, Hein M. Optimization landscape and expressivity of deep cnns[C]// International conference on machine learning: PMLR, 2018: 3730-3739.
[33] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[J]. arXiv preprint arXiv:1806.07572, 2018.
[34] Lee J, Xiao L, Schoenholz S, et al. Wide neural networks of any depth evolve as linear models under gradient descent[J]. Advances in Neural Information Processing Systems, 2019, 32: 8572-8583.
[35] Chen Z, Cao Y, Zou D, et al. How much over-parameterization is sufficient to learn deep relu networks?[J]. arXiv preprint arXiv:1911.12360, 2019.
[36] Cao Y, Gu Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32: 10836-10846.
[37] Sankararaman K A, De S, Xu Z, et al. The impact of neural network over-

parameterization on gradient confusion and stochastic gradient descent[C]// International Conference on Machine Learning: PMLR, 2020: 8469-8479.
[38] Arora S, Cohen N, Golowich N, et al. A convergence analysis of gradient descent for deep linear neural networks[J]. arXiv preprint arXiv:1810.02281, 2018.
[39] Du S, Lee J, Li H, et al. Gradient descent finds global minima of deep neural networks[C]// International Conference on Machine Learning: PMLR, 2019: 1675-1685.
[40] Zou D, Cao Y, Zhou D, et al. Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks. arXiv e-prints, art[J]. arXiv preprint arXiv:1811.08888, 2018.
[41] Allen-Zhu Z, Li Y, Song Z. On the Convergence Rate of Training Recurrent Neural Networks[J]. Advances in Neural Information Processing Systems, 2019, 32: 6676-6688.
[42] Zou D, Gu Q. An improved analysis of training over-parameterized deep neural networks[C]// Advances in Neural Information Processing Systems, 2019: 2055-2064.
[43] Daniely A. SGD learns the conjugate kernel class of the network[J]. arXiv preprint arXiv:1702.08503, 2017.
[44] Li H, Xu Z, Taylor G, et al. Visualizing the loss landscape of neural nets[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 6391-6401.
[45] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.
[46] Buja A, Swayne D F, Littman M L, et al. Data visualization with multidimensional scaling[J]. Journal of Computational and Graphical Statistics, 2008, 17(2): 444-472.
[47] Liao Q, Poggio T. Theory II: Landscape of the empirical risk in deep learning[J]. arXiv preprint arXiv:1703.09833, 2017.
[48] Moon K R, Van Dijk D, Wang Z, et al. Visualizing structure and transitions in high-dimensional biological data[J]. Nature biotechnology, 2019, 37(12): 1482-1492.
[49] Horoi S, Huang J, Wolf G, et al. Visualizing high-dimensional trajectories on the loss-landscape of ANNs[J]. arXiv preprint arXiv:2102.00485, 2021.
[50] Pascanu R, Dauphin Y N, Ganguli S, et al. On the saddle point problem for non-convex optimization[J]. arXiv preprint arXiv:1405.4604, 2014.

Note: Figure translations are in progress. See original paper for figures.