
Visualizing, Rethinking, and Mining the Loss Landscape of Deep Neural Networks

Xin-Chun Li^{1,2}, Lan Li^{1,2}, De-Chuan Zhan^{1,2}

¹ School of Artificial Intelligence, Nanjing University, Nanjing, China

² National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
 {l1xc, l1l}@lamda.nju.edu.cn, zhandc@nju.edu.cn

Abstract

The loss landscape of deep neural networks (DNNs) is commonly considered complex and wildly fluctuated. However, an interesting observation is that the loss surfaces plotted along Gaussian noise directions are almost v-basin ones with the perturbed model lying on the basin. This motivates us to rethink whether the 1D or 2D subspace could cover more complex local geometry structures, and how to mine the corresponding perturbation directions. This paper systematically and gradually categorizes the 1D curves from simple to complex, including v-basin, v-side, w-basin, w-peak, and vvv-basin curves. Notably, the latter two types are already hard to obtain via the intuitive construction of specific perturbation directions, and we need to propose proper mining algorithms to plot the corresponding 1D curves. Combining these 1D directions, various types of 2D surfaces are visualized such as the saddle surfaces and the bottom of a bottle of wine that are only shown by demo functions in previous works. Finally, we propose theoretical insights from the lens of the Hessian matrix to explain the observed several interesting phenomena.

1 Introduction

It is commonly recognized that deep neural networks (DNNs) are difficult to train without the proposal of proper architectures [34, 30], proper initialization [23, 24, 29], or effective optimization algorithms [53, 69, 40]. A guess is that the loss landscape of DNNs is too complex to search for a qualified solution [23, 19, 52, 44]. Capturing a global view of the high-dimensional loss landscape of DNNs is still a mystery to the community [43, 18, 16, 7, 59, 51, 32], but projecting it into the 1D or 2D subspace is a common way to visualize the local geometry of DNNs [25, 33, 44, 27, 17, 67].

Previous works either show 1D or 2D surfaces that are nearly smooth and are not as complex as expected [25, 68, 33, 21], or they only illustrate the complex DNN loss landscape through simple demo functions [9, 14, 12]. For example, the linear interpolation between the initialization and the converged solution shows monotonic decreasing losses, i.e., the Monotonic Linear Interpolation (MLI) phenomenon [25, 20, 47, 61, 64]; the linear interpolation between two independent solutions will commonly encounter one and only one loss barrier, and they are amazingly connected by simple Bezier or quadratic curves, i.e., the Linear Mode Connectivity (LMC) phenomenon [25, 21, 58, 1]. The existence of MLI and LMC makes us rethink the loss landscape of DNNs: *could the 1D curves or 2D surfaces display more complex patterns? If so, could we search for definite ways to mine and visualize them in an explainable manner?*

This paper systematically and gradually visualizes the loss surface of mainstream DNNs in 1D or 2D subspace. We first find an interesting observation that perturbing DNN parameters by Gaussian noise directions commonly leads to monotonic increasing curves on both two sides. These types of curves are categorized as v-basin curves which look like the shape of “v” and the perturbed model lies on the basin. Setting the perturbation direction to the negative gradient or the direction

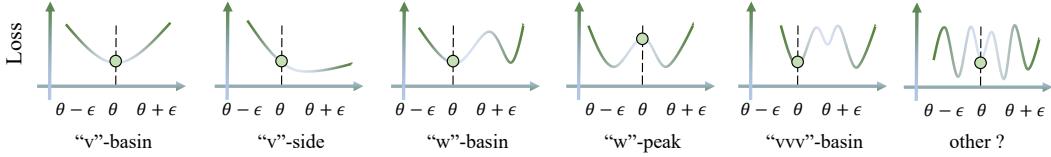


Figure 1: The illustration of the categorized 1D curves. For each categorized type, we provide definite perturbation directions or mining algorithms to plot them in an explainable way. An open question is whether we can mine more perturbation directions and plot complex 1D curves correspondingly.

to subsequent checkpoints could display v-side curves. The w-basin curves are inspired by the loss barrier phenomenon in LMC [25, 60, 13, 21], where we take the direction to an independently trained checkpoint as the perturbation direction. Plotting w-peak curves should dive into the Hessian eigenvalues and eigenvectors of the DNN [9, 54, 55, 26, 2, 36, 22, 57]. However, the most negative eigenvalue of the Hessian is absolutely small, which makes the w-peak curves own small curvatures and only show a short loss decreasing trend on both sides. We propose an algorithm to mine obvious w-peak curves and analyze the relationship to the Hessian eigenvector directions. There are no intuitive perturbation directions to plot more complex curves such as vvv-basin ones, and we also provide an algorithm to mine them. The illustration can be found in Fig. 1.

Combining the above 1D directions, we could plot various types of 2D surfaces such as saddle surfaces and a gutter structure like the bottom of a bottle. Notably, these surfaces are seldom plotted in previous works, which are instead illustrated by simple demo functions such as in [9]. Aside from empirical visualization, this paper additionally provides theoretical insights to explain the phenomenon of monotonic loss increasing when perturbed by Gaussian noise. Furthermore, this explanation is bridged to the MLI phenomenon, which leads to a novel insight for explaining MLI.

To conclude, this paper has several advantages and contributions as follows: (1) systematically visualizing and mining the embedded 1D loss curves of DNNs by category; (2) plotting several types of 2D surfaces which are only illustrated by demo functions in previous works; (3) defining as GMI the monotonic loss increasing phenomenon when perturbed by Gaussian noise and providing theoretical insights from the lens of Hessian for GMI and MLI; (4) making it easier to understand and interpret the loss surfaces of DNNs in a step-by-step manner; (5) leaving open problems and interesting guesses about the low-dimensional loss landscape visualization of DNNs.

2 Empirical Visualization of Loss Landscape

This section introduces basic notations and experimental settings, then visualizes and mines 1D curves by category, and finally presents various types of 2D surfaces.

2.1 Basic Notations and Experimental Settings

Given a model θ , we could perturb it along a single direction ϵ or two orthogonal directions ϵ_1 and ϵ_2 . The former shows the 1D curve embedded in the global landscape, while the latter plots the 2D surface centered around θ . We uniformly name ϵ , ϵ_1 , and ϵ_2 perturbation directions or noise directions, and name θ the perturbed model. ϵ has the same shape as θ , and is usually element-wisely sampled for every trainable parameter. By default, the 1D curves are plotted in the range of $\lambda \in [-1, 1]$ with the perturbation equation as $\theta + \lambda\epsilon$. We use the equation of $\theta + \lambda_1\epsilon_1 + \lambda_2\epsilon_2$ to plot the 2D surfaces with $\lambda_1 \in [-1, 1]$ and $\lambda_2 \in [-1, 1]$. Sometimes, we set the range of λ , λ_1 , or λ_2 as $[-s, s]$ to show a micro or macro landscape view. The norm of the perturbation direction may also be scaled to the same as the perturbed model by the equation $\epsilon \leftarrow \|\theta\| * \epsilon / \|\epsilon\|$.

We train an MLP with two layers on CIFAR-10 (C10) [41], and train ResNet32/110 (RN32/110) [30] on CIFAR-100 (C100) [41]. We also finetune the pre-trained MobileNet-V2 (MV2) [56] on CUB [62]. The pre-trained ResNeXt101 (RNX101) [66] could be directly downloaded and verified on ImageNet [10]. Pre-trained models are from PyTorch¹. We do not consider perturbing the running

¹<https://pytorch.org/>

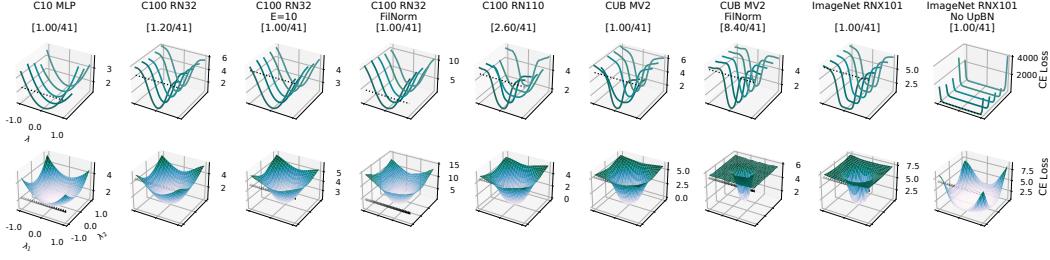


Figure 2: The 1D v-basin curves and 2D surfaces along Gaussian noise directions. The plots are amazingly smoother than we previously thought under various conditions.

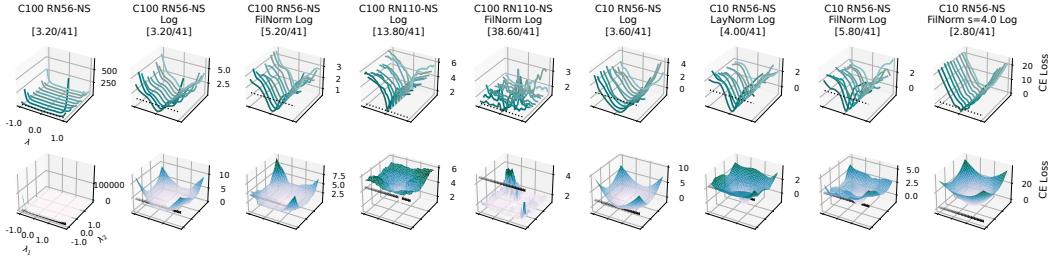


Figure 3: The 1D v-basin curves and 2D surfaces plotted around models with no skip connection along Gaussian noise directions. “Log” denotes that the y-axis is in the log scale.

statistics in the BN layers [34], and we will forward the interpolated model one pass on the dataset to re-calculate them. More experimental details can be found in Appendix A.

2.2 Visualization of v-basin Curves

A common way to view the local landscape of DNNs is setting ϵ , ϵ_1 , and ϵ_2 as random Gaussian directions. The 1D curves and 2D surfaces are displayed in Fig. 2. The horizontal dotted black lines represent the random prediction loss threshold, i.e., the loss of $\log(C)$ with C being the number of classes. Various conditions are considered, which include: (1) the number of training epochs (E), where we show the results of RN32 on C100 when $E = 10$, and the others are $E = 200$ by default; (2) the filter normalization (“FilNorm”) method proposed by [44], which normalizes each filter in the noise direction to have the same norm of corresponding filters in the perturbed model; (3) the processing of BN statistics, and “No UpBN” denotes that the BN running states are not re-calculated. Five independent 1D curves are visualized in the same plot. The average number of stationary points in the five curves is reported in the “[J]”. We plot 41 discrete points for each 1D curve, and most of them only have 1 stationary point on average (i.e., the perturbed model itself when $\lambda = 0.0$). One exception is the 7th plot that utilizes FilNorm on CUB, while the additional stationary points almost exist in line segments that surpass the random prediction loss. We denote these types of 1D curves as v-basin curves with the perturbed model lying on the basin.

The previous work [44] plots the 2D surface of ResNet without skip connections (RN56/110-NS), which is extremely chaotic. We carefully checked their code and found several interesting phenomena. First, perturbing RN-NS could easily lead to exploded losses (e.g., a loss value of 10^5 on C100 using RN56-NS), which makes the landscape look like a horizontal hyperplane (i.e., the first column in Fig. 3). As a trick, they plot the losses in a log scale to enhance the distinctness between points. Second, RN110-NS is hard to optimize, which only performs slightly better than random guess (i.e., a test accuracy of 7.08% on C100 and 10.09% on C10). Perturbing models that are not well-trained is less meaningful, and the loss landscape around them is chaotic (e.g., the 5th column in Fig. 3). Third, RN56-NS could achieve much better performances on C100 (57.06%) and C10 (87.01%), and the chaotic loss landscape almost exists above the random guess curves (i.e., above the dotted black lines). Additionally, with a macro view scale (i.e., $s = 4.0$), the loss landscape of RN56-NS on C10

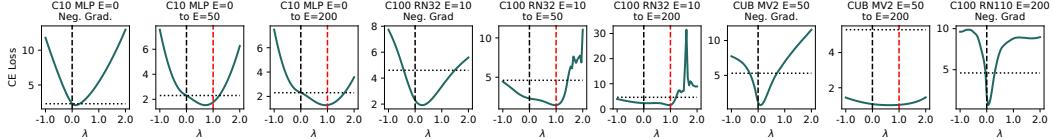


Figure 4: The 1D v-side curves plotted around different checkpoints. The vertical black/red dotted line shows the position of the perturbed/subsequent checkpoint.

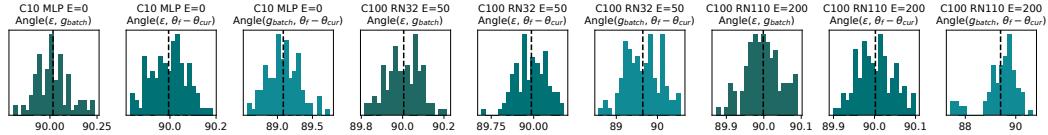


Figure 5: The angles between randomly sampled Gaussian directions and the descent directions. g_{batch} denotes the negative of batch gradient, and $\theta_{\text{cur}}/\theta_f$ denotes the current/final checkpoint.

becomes smooth again (i.e., the last column in Fig. 3). With the same number of scatter points, a macro view means a larger step between adjacent scatters, which may skip the fluctuating segments. The theoretical insights will be provided in Sect. 3.

Overall, chaotic surfaces seem to show up only on poorly trained DNNs or in regions worse than random predictions, which are of little significance to study. For well-trained and mainstream models, the loss surfaces along Gaussian perturbations are extremely smoother than we previously thought. In other words, it seems difficult to plot 1D curves other than the v-basin ones by Gaussian perturbation alone. We denote the phenomenon that Gaussian perturbation leads to double-side monotonic increasing losses as Gaussian Monotonic Increasing (GMI). A theoretical parallel will be bridged between GMI and MLI in Sect. 3. Some experimental details are explained in Appendix A.3.

2.3 Visualization of v-side Curves

Considering the first-order and second-order approximation (f.o.a and s.o.a) of the loss $\mathcal{L}(\theta + \lambda\epsilon)$:

$$\mathcal{L}_{\text{f.o.a}} \approx \mathcal{L}(\theta) + \lambda\epsilon^T g_\theta, \quad \mathcal{L}_{\text{s.o.a}} \approx \mathcal{L}(\theta) + \lambda\epsilon^T g_\theta + \frac{1}{2}\lambda^2\epsilon^T H_\theta\epsilon, \quad (1)$$

where $g_\theta = \nabla_\theta \mathcal{L}$ and $H_\theta = \nabla \nabla^T \mathcal{L}$ denote the gradient and Hessian matrix calculated at the point of θ . The first-order approximation implies that the negative gradient is the sharpest descent direction, i.e., $\epsilon = -g_\theta$. However, the gradient norm of converged models is nearly zero, only showing negligible descent losses. Hence, we consider the intermediate checkpoints and take the negative gradient as one type of descent direction. The direction to the subsequent checkpoint (e.g., $\epsilon = \theta_{E=200} - \theta_{E=50}$) is also a descent direction for $\theta_{E=50}$, which is inherently an accumulation of multiple gradient steps. Fig. 4 shows the plotted results, where we take MLP with $E = 0$ on C10, RN32 with $E = 10$ on C100, MV2 with $E = 50$ on CUB, and RN110 with $E = 200$ on C100. The curves are almost v-side ones with the perturbed model lying on the valley side (i.e., the vertical black dotted line when $\lambda = 0.0$). RN32 on C100 again shows chaotic segments whose loss surpasses that of a random guess.

According to the first-order approximation, the loss will descend in a specific range if we could sample a Gaussian vector satisfying $\epsilon^T g_\theta < 0$. Easily, the mean and variance of $\epsilon^T g_\theta$ is $\mathbb{E}_\epsilon[\epsilon^T g_\theta] = 0$ and $\mathbb{V}_\epsilon[\epsilon^T g_\theta] = \|g_\theta\|_2^2$. The probability of sampling a same vector as g_θ is $\exp(-\frac{1}{2}\|g_\theta\|_2^2)/(2\pi)^{d/2}$, and d is the total number of trainable parameters. Take the MLP on C10 as an example, it has about $d=395K$ parameters, and $\|g_\theta\|_2^2$ is about 119.7 under $E = 0$, and the logarithmic probability of sampling g_θ is about -3.6×10^5 . That is, we could hardly sample a Gaussian direction having a larger overlap with the gradient direction in the high-dimensional space. We sample 100 groups of Gaussian directions and calculate their angles with the negative gradient direction and the direction to the subsequent checkpoint. Fig. 5 shows the distribution of angles. The gradient directions are calculated on random data batches. The angle range between the Gaussian direction and descent

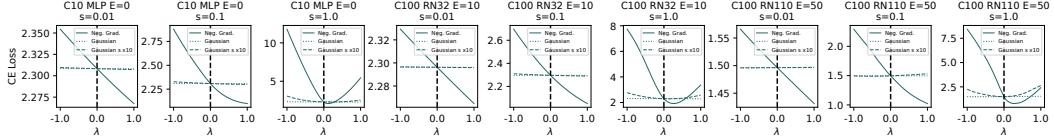


Figure 6: The descent curves of the negative gradient and the most overlapped Gaussian noise.

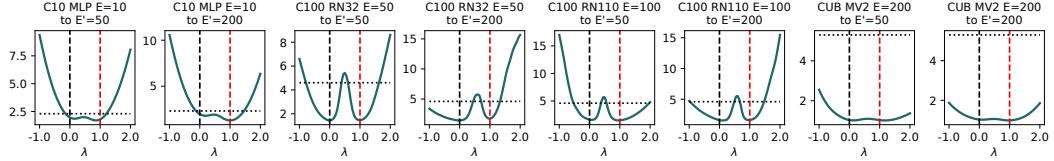


Figure 7: The 1D w-basin curves inspired by loss barrier between independently trained checkpoints. The vertical black/red dotted lines show the positions of two independent checkpoints.

directions is about [89.75, 90.25]. For comparison, we also calculate the angle between the negative gradient and the direction of the converged model, which are almost in the range of [88.0, 90.0]. This shows that the one-step batch gradient may slightly imply the global converged direction, while it is nearly impossible to sample a Gaussian noise that has an acceptable overlap with descent directions, e.g., a noise with an angle smaller than 89.5.

We select the Gaussian noise most overlapped with the negative gradient direction from the 100 groups, and plot the descent curves in Fig. 6. We set s in $\{0.01, 0.1, 1.0\}$ and scale the norm of the Gaussian direction to that of the negative gradient. In each plot, the x-axis with Gaussian noise is also scaled by $10\times$ for better comparison. Even with a $10\times$ scale, the loss decreasing brought by the Gaussian direction is negligible. This verifies the GMI phenomenon shown in Fig. 2 from a micro perspective that the Gaussian noise could hardly lead to v-side curves.

2.4 Visualization of w-basin/w-peak Curves

Previous researches about the mode connectivity of DNNs provide an intuitive way to plot w-basin curves [25, 13, 5]. Commonly, the linear interpolation between two independent converged models encounters one and only one barrier. That is, the loss $\mathcal{L}((1-\lambda)\theta_1 + \lambda\theta_2)$ within $\lambda \in [0, 1]$ looks like a hill. In other words, given a model θ_1 , we could plot the 1D curve by the equation of $\theta_1 + \lambda(\theta_2 - \theta_1)$. If we take $\lambda \in [-1, 2]$, then we could obtain w-basin curves as shown in Fig. 7. Notably, we extend the phenomenon of loss barrier to different checkpoints. For example, the vertical black dotted line denotes the position of the MLP checkpoint with $E = 10$, and the red dotted line denotes another independent checkpoint with $E' = 50$ (i.e., the 1st plot in Fig. 7). All plots present perfectly smooth w-basin curves, showing no chaotic segments.

Then, we would like to search for a direction that could make the loss double-side decrease, i.e., w-peak curves. If we assume g_θ is zero for a converged moel θ , then the second-order approximation in Eq. 1 will become $\mathcal{L}_{\text{s.o.a}} - \mathcal{L}(\theta) \approx \frac{1}{2}\lambda^2\epsilon^T H_\theta\epsilon$. Setting ϵ to the eigenvectors of the Hessian matrix corresponding to the negative eigenvalues (abbreviated as N.E.) could make $\mathcal{L}_{\text{s.o.a}}$ smaller than $\mathcal{L}(\theta)$ within a definite small range of λ . We use the `sparse.linalg.eigsh` in the Scipy² package to calculate the most 5 positive or negative eigenpairs [2]. Specifically, `eigsh` accepts as input the `LinearOperator` constructed by the “Jacobian Vector Product” function that returns $H_\theta v$ for a given vector v . The detail and demo code can be found in Appendix B.1. Fig. 8 plots the 1D curves along the positive eigenvectors (P.E.) and N.E. ones. We show results of MLP on C10, RN32 on C100, and MV2 on CUB, and each pair considers the checkpoint of $E \in \{10, 50, 200\}$. The 1D curves along P.E. directions almost consistently show v-basin curves, while the cases for N.E. ones are complex. The initial checkpoints of simple models (e.g., MLP on C10 and RN32 on C100) display perfect w-peak

²<https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.linalg.eigsh.html>

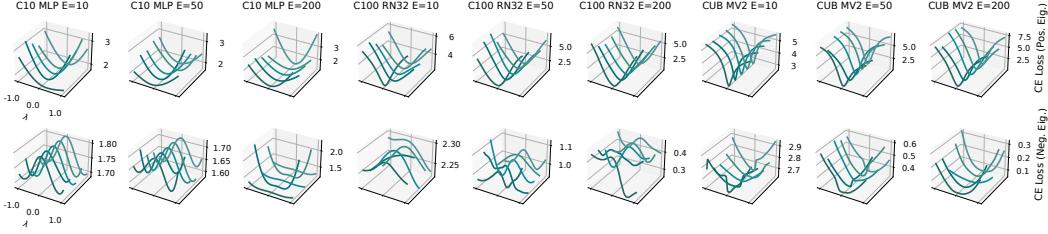


Figure 8: The 1D curves plotted along the eigenvector directions of the Hessian matrix. The second row shows the w-peak curves along N.E. directions. Top 5 directions of P.E. and N.E. are plotted.

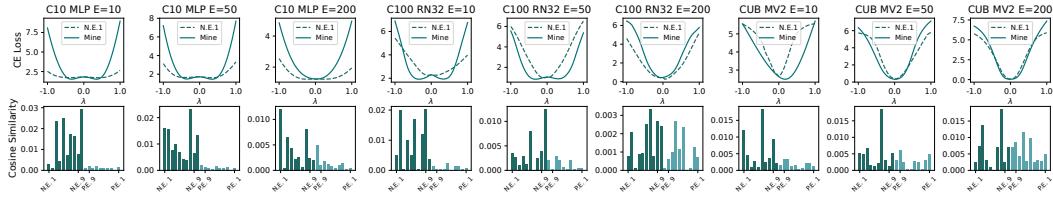


Figure 9: The mined w-peak curves via Algo. 1 and the cosine similarity with P.E./N.E. directions.

curves when $E = 10$ or $E = 50$. MV2 on CUB does not display obvious w-peak curves. Additionally, with the model becoming converged, the “height of the peak” decreases.

We also propose an algorithm to mine the w-peak curves without calculating the negative eigenvalues and corresponding eigenvectors of the Hessian matrix. Our optimization formulation is:

$$\arg \min_{\epsilon} \mathbb{E}_{\lambda \in [-1, 1]} [\mathcal{L}(\theta + \lambda \epsilon)]. \quad (2)$$

To solve this optimization problem, we first initialize the elements in ϵ as zero. For each data batch, we sample λ from $[-\alpha, \alpha]$ and obtain the interpolated model $\hat{\theta} = \theta + \lambda \epsilon$. The loss and gradient are calculated on $\hat{\theta}$ and we update ϵ by $\epsilon - \eta \lambda \nabla_{\hat{\theta}} \mathcal{L}$. η is the learning rate and α will gradually increase from 0 to 1 during the whole optimization process. The pseudo-code is listed in Algo. 1. With the optimized ϵ , we plot the 1D curves of $\theta + \lambda \epsilon$ in the first row of Fig. 9. Similarly, it is hard to obtain w-peak curves for MV2 fine-tuned on CUB. For MLP on C10 and RN32 on C100, the w-peak curves are obvious when E is smaller. When $E = 200$, it is hard to obtain double-side loss decreasing curves because the converged model already reaches a quite low-loss area. We also study the relationship between the mined direction and the eigenvectors of H_θ . The cosine similarities are calculated and reported in the second row of Fig. 9. The first 10 bars show the absolute cosine similarity with “N.E.1” to “N.E.10” while the following 10 bars show that with “P.E.10” to “P.E.1”. The mined direction is not just one of the Hessian eigenvectors but seems to be a weighted combination of all eigenvectors with the weights of N.E. slightly larger than that of P.E. ones.

Algorithm 1 Mine w-peak Curves

```

1: for each epoch  $e = 1, 2, \dots, E$  do
2:   Set  $\alpha = e/E$ 
3:   for each batch do
4:     Sample  $\lambda \in [-\alpha, \alpha]$  uniformly
5:     Interpolate model  $\hat{\theta} = \theta + \lambda \epsilon$ 
6:      $\epsilon \leftarrow \epsilon - \eta \lambda \nabla_{\hat{\theta}} \mathcal{L}$ 
7:   end for
8: end for

```

Algorithm 2 Mine vvv-basin Curves

```

1: for each epoch  $e = 1, 2, \dots, E$  do
2:   for each batch do
3:     Sample  $\lambda \in [0.5 - \alpha, 0.5 + \alpha]$ 
4:     Interpolate  $\hat{\phi} = (1 - \lambda)\theta + \lambda\phi$ 
5:      $\phi \leftarrow \phi - \eta (\nabla_{\phi} \mathcal{L} + \gamma \lambda \nabla_{\hat{\phi}} \mathcal{L})$ 
6:   end for
7: end for
8:  $\epsilon = \phi - \theta$ 

```

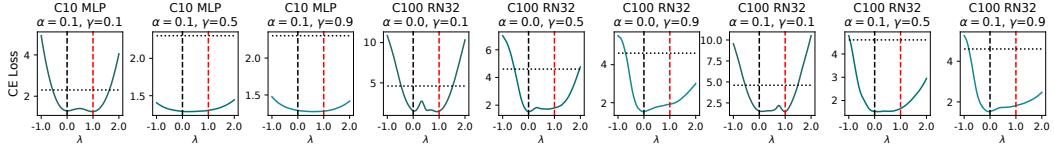


Figure 10: The mined vvv-basin curves via Algo. 2 under various hyperparameters of γ and α .

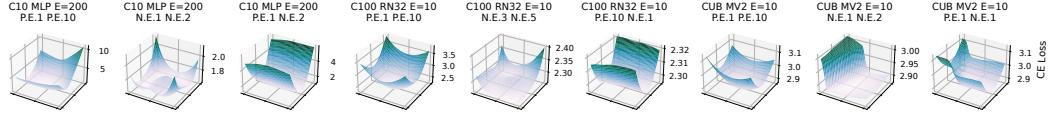


Figure 11: The 2D surfaces plotted by combining the eigenvector directions. P.E./N.E. denotes eigenvectors corresponding to the positive/negative eigenvalues.

2.5 Visualization of vvv-basin Curves

We have not yet found an intuitive perturbation direction to plot vvv-basin curves, but we provide a possible optimization problem to mine possible directions:

$$\arg \min_{\phi} \mathbb{E}_{\lambda \in [0.5-\alpha, 0.5+\alpha]} [\mathcal{L}(\phi) + \gamma \mathcal{L}((1-\lambda)\theta + \lambda\phi)], \quad (3)$$

where θ is a converged model, and γ is the regularization coefficient, and α controls the interpolation range. This formula optimizes the loss of ϕ and simultaneously decreases the loss around the middle interpolation with θ . The optimization process is similar to Eq. 2 and the pseudo-code is in Algo. 2. When the optimization finishes, the direction of $\epsilon = \phi - \theta$ is utilized to plot 1D curves around θ . We set $\alpha \in \{0.0, 0.1\}$ and $\gamma \in \{0.1, 0.5, 0.9\}$, respectively. The mined curves are shown in Fig. 10, where the vertical black/red dotted lines mark the position of θ/ϕ . MLP on C10 is a simple DNN and shows no vvv-basin patterns. A proper hyperparameter group on C100 with RN32 may present vvv-basin curves (e.g., the 4th plot with $\alpha = 0.0$ and $\gamma = 0.1$). Other plots show patterns looking like the transitional shape from “w” to “vvv”. Overall, vvv-basin curves have already become not so intuitive to mine and visualize.

2.6 Visualization of 2D Surfaces by Combining 1D Directions

First, the Hessian eigenvector directions could be combined to plot the 2D surfaces. Eigenvectors are orthogonal to each other and we do not need to make them orthogonal again. The plots are in Fig. 11. Combining two directions of P.E. leads to a surface looking like the 2D surfaces of Gaussian perturbation shown in Fig. 2. Combining directions of N.E. shows a surface looking like the bottom of a bottle of wine, which is only illustrated by demo functions in previous works [9]. An exceptional case is the 8th plot, which is resulted from the smaller negative curvature of MV2 on CUB shown in Fig. 8. The combination of P.E. and N.E. leads to saddle surfaces, and the curvature along the N.E. direction is smaller because the absolute value of the corresponding eigenvalue is small (Fig. 13). Then, the 1D directions that lead to v-basin, v-side, w-basin, w-peak, and vvv-basin curves could be combined to display more complex 2D surfaces. The possible 2D surfaces are shown in Fig. 12. The summary of the mined perturbation directions and their abbreviations are in Appendix A.4.

3 Theoretical Insights from the Lens of the Hessian

This section provides some initial theoretical explanations for the observed interesting phenomena from the lens of the Hessian [22, 2, 55]. Specifically, we utilize the PyHessian [68]³ tool to calculate the approximated eigenvalue density of the Hessian matrix. The details and demo code are in Appendix B.2. The logarithmic probability density of Hessian eigenvalues is shown in the first row

³<https://github.com/amirgholami/PyHessian>

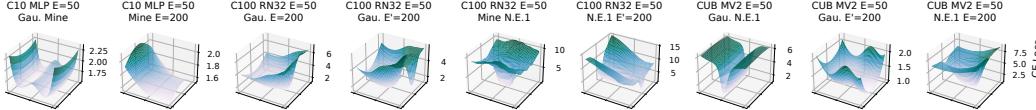


Figure 12: The 2D surfaces plotted by combining the mined 1D perturbation directions. The second line of the title shows the abbreviations of the utilized two directions.

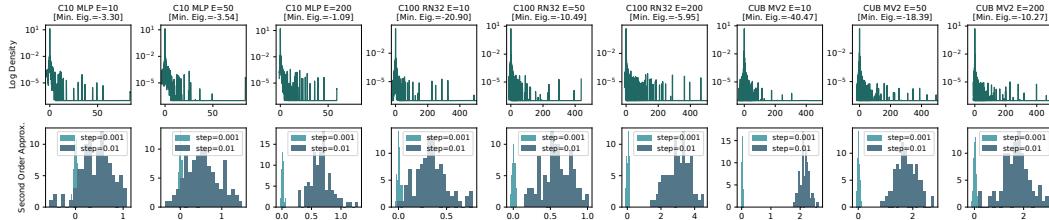


Figure 13: The eigenvalue density of the Hessian matrix (first row) and the distribution of loss change when perturbed by Gaussian noises with $\lambda \in \{0.001, 0.01\}$ (second row).

of Fig. 13. The Hessian density presents a shape with a bulk of values around zero and several large positive outliers [55, 26]. The smallest eigenvalue is reported in “[]”, and its absolute value becomes smaller when E increases. This means that the singularity of the Hessian matrix decreases and it is harder to mine w-peak curves when the model goes to a more convex area (Fig. 8 and Fig. 9).

Then, we rethink the interpolation formula in MLI [25, 20, 47, 64] and the finding of GMI in our paper. The former plots the losses of $(1 - \alpha)\theta_0 + \alpha\theta_f$ with $\alpha \in [0, 1]$, where θ_0 and θ_f denote the initial and converged model, respectively. We could re-formulate this equation as $\theta_f + \lambda\epsilon$ with $\lambda = \alpha - 1 \in [-1, 0]$, and $\epsilon = \theta_f - \theta_0$. Hence, MLI and GMI differ only in the perturbation direction, which implies that they may have the same explanation. Considering the second-order approximation in Eq. 1, the loss change when perturbed by the Gaussian noise is $\delta\mathcal{L} = \lambda\epsilon^T g_\theta + \frac{1}{2}\lambda^2\epsilon^T H_\theta\epsilon$. The mean and variance of this term is:

$$\mathbb{E}_\epsilon [\delta\mathcal{L}] = \frac{1}{2}\lambda^2\sigma^2\text{tr}(H_\theta), \quad \mathbb{V}_\epsilon [\delta\mathcal{L}] = \lambda\|g_\theta\|_2^2 + \frac{1}{2}\lambda^2\sigma^4\text{tr}(H_\theta H_\delta), \quad (4)$$

where we assume ϵ is element-wisely sampled from $\mathcal{N}(0, \sigma^2)$, and $\text{tr}(\cdot)$ denotes the trace of a matrix. The trace of the Hessian during the training process is almost always positive as presented in previous studies [55, 26, 19]. This could also be observed from the Hessian density as shown in Fig. 13. Hence, the average of loss change perturbed by a random Gaussian vector is about $\frac{1}{2}\lambda^2\sigma^2\text{tr}(H_\theta)$, which is commonly positive. The variance is hard to calculate because of $\text{tr}(H_\delta H_\delta)$. Instead, we simulate the $\delta\mathcal{L}$ by sampling 100 groups of ϵ and plot its distribution in the second row of Fig. 13. The details can be found in Appendix B.3. We set $\sigma = 1.0$ and plot distributions under $\lambda \in \{0.001, 0.01\}$. The loss change is almost centered around 0.0 when $\lambda = 0.001$, while it becomes almost positive when $\lambda = 0.01$. When plotting the 1D curves, we could only sample a limited number of points, and the steps among each other are much greater than 0.001. For example, if we plot 41 points in the range of $[-1.0, 1.0]$, then the step is 0.05. This step size may commonly lead to loss increment, which explains the GMI phenomenon. Due to formula similarity, MLI could also be explained by this well.

Finally, we take 11 discrete points to plot the GMI and MLI curves as in Fig. 14. For each discrete point, we plot the second-order approximation curve under a neighbor range of $[-0.05, 0.05]$. The MLI curves look similar to the GMI curves, which verifies that the MLI phenomenon could be explained by perturbing the final model with $\epsilon = \theta_f - \theta_0$. Second, for most discrete points, the s.o.a. shows a decreasing trend when the x-axis becomes larger. It is worth noting that the s.o.a. does not work well when $\lambda = -1$, where the interpolated model is close to the initialization or random noise. The Hessian eigenvalue distribution of the initialization has a slightly large proportion of negative values (Fig. 13), which also implies that the chaotic loss surface in Fig. 2 and Fig. 4 is more likely to appear on the model worse than the random guess.

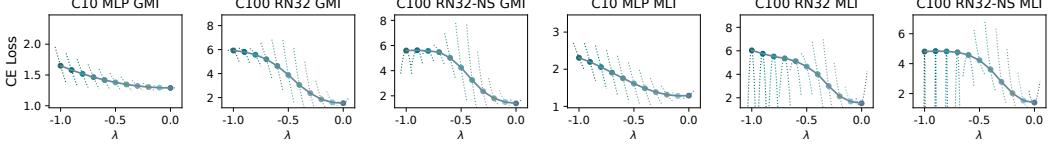


Figure 14: The second-order approximation for the interpolation of $\theta_f + \lambda\epsilon$ with $\lambda \in [-1, 0]$. The first three plots use ϵ as Gaussian noise (i.e., GMI), while the last ones set $\epsilon = \theta_f - \theta_0$ (i.e., MLI).

4 Related Works

Low-Dimensional Visualization of Loss Landscape. [25] plots 1D curves between the initialized and converged model, and between two independently converged models, presenting the interesting monotonic linear interpolation (MLI) phenomenon [20, 47, 64] and the barrier in linear mode connectivity (LMC) [60, 5]. Amazingly, the linear interpolation of two independent solutions only crosses one loss peak, and they could be connected by a slightly complex curve with low losses [13, 21]. [33] empirically shows the surface between solutions found by different optimizers. [44] provides visualization of loss landscape with or without skip connections. [27] plots 1D curves and 2D surfaces for fine-tuned BERT [11]. Some works also show the guesses about the loss surface of DNNs by plotting demo functions [9, 14, 19]. For example, [9] plots some saddle surfaces and gutter structures by simple 2D functions such as $z = (x^2 + y^2 - 1)^2$.

Global View of Loss Landscape. The saddle points may challenge the optimization process of high-dimensional non-convex DNNs [9]. [8] analyzes the global landscape of multi-layer DNNs by replacing the ReLU activation with some assumptions. [38] proves that deep linear networks have no poor local minima, and [52] searches for a special category of DNNs with no bad minima. [19] proposes a notion of Goldilocks zone to show the effectiveness of proper initialization methods. [18] proposes a toy loss landscape model named n-wedges to present surprising and counter-intuitive properties of DNNs in a more explainable way. Exploring the Hessian matrix of DNNs also explains some interesting properties of DNNs [55, 57]. For example, [54] points out that the Hessian eigenvalue distribution composes the bulk part and the positive outliers, and the number of the latter may be the number of classes. This makes the gradient during optimization lie in a small tiny subspace [26].

Applications of Studying Loss Landscape. Exploring the loss landscape could provide some helpful insights for practical applications. The valley flatness or sharpness around a converged model may reflect the generalization performance [31, 39, 49, 37], while later works debate against their relation [12, 3]. Motivated by the flat minima, some advanced optimization methods are proposed [70, 42, 15, 6]. Mitigating the loss barrier between two independent models is studied for better model fusion [4, 58, 1, 14], which has also been applied to federated learning [48, 45, 63]. The asymmetric valley [28] explains the success of stochastic weight averaging [35]. Fusing multiple model soups fine-tuned from the same pre-trained model could lead to a better loss area [50, 65].

5 Limitations and Future Works

This paper does not strictly prove how complex the 1-D loss curves of DNNs can be, and it is still an open problem to answer. Additionally, the mined vvv-basin curves are not as perfect as we expected. If an arbitrarily complex 1D curve can be mined, then the global loss surface of the DNNs is more complex. But if we can only mine finite complex 1D curves, then the global loss surface may be either finite complex or may still be complicated by high-dimensional combinations. This is also an open problem to verify. Mining and visualizing more complex types of local geometry structures for various types of DNNs are future works.

6 Conclusion

We systematically mine and plot several types of 1D curves embedded in the global landscape of DNNs. Various types of 2D surfaces are further mined from the basis of 1D perturbation directions. Theoretical analysis from the view of Hessian properties explains the GMI and MLI phenomenon.

References

- [1] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.
- [2] Guillaume Alain, Nicolas Le Roux, and Pierre-Antoine Manzagol. Negative eigenvalues of the hessian in deep neural networks. In *6th International Conference on Learning Representations, Workshop Track*, 2018.
- [3] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning*, pages 840–902, 2023.
- [4] Stephen C. Ashmore and Michael S. Gashler. A method for finding similarity between multi-layer perceptrons by forward bipartite alignment. In *International Joint Conference on Neural Networks*, pages 1–7, 2015.
- [5] Gregory W. Benton, Wesley J. Maddox, Sanae Lotfi, and Andrew Gordon Wilson. Loss surface simplexes for mode connecting volumes and fast ensembling. In *Proceedings of the 38th International Conference on Machine Learning*, pages 769–779, 2021.
- [6] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *The 5th International Conference on Learning Representations*, 2017.
- [7] Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023.
- [8] Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.
- [9] Yann N. Dauphin, Razvan Pascanu, Çaglar Gülcehre, KyungHyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems 27*, pages 2933–2941, 2014.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.
- [12] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1019–1028, 2017.
- [13] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1308–1317, 2018.
- [14] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *The Tenth International Conference on Learning Representations*, 2022.
- [15] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *The 9th International Conference on Learning Representations*, 2021.
- [16] Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *CoRR*, abs/1910.05929, 2019.
- [17] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *CoRR*, abs/1912.02757, 2019.
- [18] Stanislav Fort and Stanislaw Jastrzebski. Large scale structure of neural network loss landscapes. In *Advances in Neural Information Processing Systems 32*, pages 6706–6714, 2019.

- [19] Stanislav Fort and Adam Scherlis. The goldilocks zone: Towards better understanding of neural network loss landscapes. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 3574–3581, 2019.
- [20] Jonathan Frankle. Revisiting "qualitatively characterizing neural network optimization problems". *CoRR*, abs/2012.06898, 2020.
- [21] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems 31*, pages 8803–8812, 2018.
- [22] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2232–2241, 2019.
- [23] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.
- [24] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.
- [25] Ian J. Goodfellow and Oriol Vinyals. Qualitatively characterizing neural network optimization problems. In *3rd International Conference on Learning Representations*, 2015.
- [26] Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *CoRR*, abs/1812.04754, 2018.
- [27] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4141–4150, 2019.
- [28] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems 32*, pages 2549–2560, 2019.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [31] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.
- [32] W. Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K. Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. In *"I Can't Believe It's Not Better!" at NeurIPS Workshops*, pages 87–97, 2020.
- [33] Daniel Jiwoong Im, Michael Tao, and Kristin Branson. An empirical analysis of the optimization of deep network loss surfaces. *CoRR*, abs/1612.04010, 2016.
- [34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, 2015.
- [35] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 876–885, 2018.
- [36] Stanislaw Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos J. Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *7th International Conference on Learning Representations*, 2019.
- [37] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *The 8th International Conference on Learning Representations*, 2020.
- [38] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, pages 586–594, 2016.
- [39] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *The 5th International Conference on Learning Representations*, 2017.
- [40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

- [41] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2012.
- [42] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5905–5914, 2021.
- [43] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *6th International Conference on Learning Representations*, 2018.
- [44] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems 31*, pages 6391–6401, 2018.
- [45] Xin-Chun Li, Yi-Chu Xu, Shaoming Song, Bingshuai Li, Yinchuan Li, Yunfeng Shao, and De-Chuan Zhan. Federated learning with position-aware neurons. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10082–10091, 2022.
- [46] Lin Lin, Yousef Saad, and Chao Yang. Approximating spectral densities of large matrices. *SIAM Review*, 58(1):34–65, 2016.
- [47] James Lucas, Juhan Bae, Michael R. Zhang, Stanislav Fort, Richard S. Zemel, and Roger B. Grosse. Analyzing monotonic linear interpolation in neural network loss landscapes. *CoRR*, abs/2104.11044, 2021.
- [48] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.
- [49] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, pages 5947–5956, 2017.
- [50] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems 33*, 2020.
- [51] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2603–2612, 2017.
- [52] Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. In *7th International Conference on Learning Representations*, 2019.
- [53] Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.
- [54] Levent Sagun, Léon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *CoRR*, abs/1611.07476, 2016.
- [55] Levent Sagun, Utku Evci, V. Ugur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *CoRR*, abs/1706.04454, 2017.
- [56] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [57] Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N. Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 9481–9488, 2021.
- [58] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems 33*, 2020.
- [59] Ruoyu Sun, Dawei Li, Shiyu Liang, Tian Ding, and Rayadurgam Srikant. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020.
- [60] N. Joseph Tattro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. Optimizing mode connectivity via neuron alignment. In *Advances in Neural Information Processing Systems 33*, 2020.
- [61] Tiffany J. Vlaar and Jonathan Frankle. What can linear interpolation of neural network loss landscapes tell us? In *International Conference on Machine Learning*, pages 22325–22341, 2022.
- [62] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. (CNS-TR-2011-001), 2011.

- [63] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris S. Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *8th International Conference on Learning Representations*, 2020.
- [64] Xiang Wang, Annie N. Wang, Mo Zhou, and Rong Ge. Plateau in monotonic linear interpolation - A "biased" view of loss landscape for deep networks. In *The Eleventh International Conference on Learning Representations*, 2023.
- [65] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998, 2022.
- [66] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5987–5995, 2017.
- [67] Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E. Gonzalez, Kannan Ramchandran, and Michael W. Mahoney. Taxonomizing local versus global structure in neural network loss landscapes. In *Advances in Neural Information Processing Systems 34*, pages 18722–18733, 2021.
- [68] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *IEEE International Conference on Big Data*, pages 581–590, 2020.
- [69] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.
- [70] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning*, pages 26982–26992, 2022.

A Experimental Details

This section reports the details about the DNNs and datasets utilized in this paper. Then, the training details and figure details are presented.

A.1 Datasets, DNNs, and Training Details

An MLP with two layers is investigated on CIFAR-10 [41]. We flatten the images in CIFAR-10 to vectors with the size of $32 \times 32 \times 3 = 3072$ as inputs, and the hidden size of the MLP is 128. Then, a classifier layer is used to output 10 classes. ResNet [30] with 32, 56, and 110 layers are trained on CIFAR-10 and CIFAR-100 [41], and the versions with no-skip connections are also trained. A pre-trained MobileNet-V2 [56] is fine-tuned on CUB [62]. These DNNs are trained for $E = 200$ epochs, and the intermediate checkpoints such as $E = 10, 50, 100$ are also saved for experimental studies. The learning rate is 0.1 for MLP and ResNet, and a smaller one is 0.01 for pre-trained MobileNet-V2. The batch size is 128 and the weight decay is 5×10^{-4} . Additionally, the pre-trained ResNeXt101 (RNX101) [66] could be directly downloaded from PyTorch and verified on ImageNet [10].

A.2 Processing of BN Layers

Some utilized networks contain the BN layers [34], which contain two types of parameters. The first type contains “BN.weight” and “BN.bias”, which are trainable during the loss backward pass. The other ones contain running statistics such as “BN.running_mean” and “BN.running_var”, which calculate the mean and variance of hidden representations channel by channel in the forward pass. When perturbing θ by the noise direction ϵ , the running statistics are not considered. Hence, the interpolated model $\theta + \lambda\epsilon$ may have inconsistent running statistics, which should be updated by additionally taking a forward pass of the data to re-calculate them. The processing of this manner is referred to as “UpBN” by default. If we do not additionally update the running statistics and directly use the ones in θ for $\theta + \lambda\epsilon$, we name this manner as “No UpBN”.

A.3 Figure Details and More Plots

We then respectively present some details when plotting the figures of Fig. 2 and Fig. 3. For each Gaussian noise direction, we scale its norm to the same as the perturbed model by $\epsilon \leftarrow \|\theta\| * \epsilon / \|\epsilon\|$. The range of λ is $[-1, 1]$ by default, i.e., $s = 1.0$. For each 1D curve, we calculate the number of stationary points. The stationary point means that its y -axis value is both larger or smaller than that of its left and right point, i.e., satisfying $(y_t - y_{t+1}) * (y_t - y_{t-1}) > 0$. We do not consider the two endpoints in the 1D curve. Perfect v-basin curves have 1 stationary point, and perfect w-basin/w-peak curves have 3 stationary points. The number of stationary points that one 1D curve has could reflect its smoothness. As empirically pointed out by this paper, it is hard to find and plot 1D loss curves of DNNs that have more than 5 stationary points.

Various conditions are considered to verify the Gaussian Monotonic Increasing (GMI) phenomenon, including the number of training epochs ($E \in \{0, 10, 50, 100, 200\}$), the normalization ways of the Gaussian noise (“Norm”, “LayNorm” and “FilNorm”), the processing of BN statistics (“UpBN” and “No UpBN”), and the view scale ($s \in \{0.1, 1.0, 4.0\}$).

“Norm” means that we may normalize the perturbation direction to have the same norm of the perturbed model, i.e., $\epsilon \leftarrow \|\theta\| \frac{\epsilon}{\|\epsilon\|}$. This is utilized in [33]. The filter normalization (“FilNorm”) is proposed by [44], which normalizes each filter in the noise direction to have the same norm of corresponding filters in the perturbed model. We also utilize the “LayNorm” which normalizes each layer in the noise direction to have the same norm of corresponding layers in the perturbed model. “UpBN” means that we re-calculate the running statistics in the BN layer (i.e., the running mean and variance) by taking an additional forward pass. “No UpBN” means that we keep the BN statistics as the ones in the perturbed model. The view scale means that we plot 1D curves or 2D surfaces in the range of $[-s, s]$ instead of $[-1, 1]$.

With these conditions, we additionally propose several groups of plots to verify the GMI phenomenon. Fig. 15 shows the conditions under different checkpoints of MV2 on CUB. Fig. 16 shows the conditions under various view scales of $s \in \{0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$ on CUB with MV2.

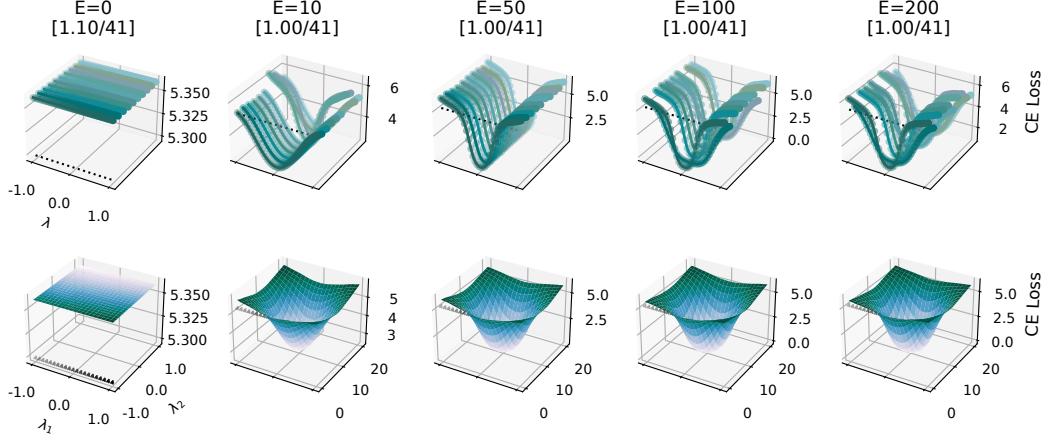


Figure 15: The GMI phenomenon under various checkpoints of MV2 on CUB.

Fig. 17 shows the conditions under various view scales and various normalization ways on C100 with RN32. All plots show obvious GMI phenomenon.

A.4 A Summary of Mined 1D Curves and Corresponding Perturbation Directions

In this paper, we gradually visualize and mine several types of 1D curves by category. The following lists the summary of these types of curves.

- **v-basin curves:** amazingly, random Gaussian directions could almost lead to v-basin curves (Fig. 2), and it is hard to generate other types of curves such as v-side ones (Fig. 6). The Gaussian perturbation is abbreviated as “Gau.” in Fig. 12. Aside from the Gaussian direction, the eigenvectors corresponding to the positive eigenvalues of the Hessian matrix also display v-basin curves (Fig. 8), which are abbreviated as “P.E.x” in Fig. 12. “x” ranges from 1 to 10.
- **v-side curves:** the negative gradient and the direction to the subsequent checkpoints are intuitive descent directions (Fig. 4). The negative gradient is abbreviated as “Neg. Grad.”. The direction to subsequent checkpoints is abbreviated as “ $E = 200$ ” (or other checkpoints) in Fig. 12.
- **w-basin curves:** the direction to an independent checkpoint is an intuitive direction for plotting w-basin curves (Fig. 7), which is abbreviated as “ $E' = 200$ ” (or other checkpoints) in Fig. 12.
- **w-peak curves:** the eigenvectors corresponding to the negative eigenvalues of the Hessian matrix lead to the w-peak curves (Fig. 8), which are abbreviated as “N.E.x” in Fig. 12. “x” ranges from 1 to 10. The w-peak curves could also be plotted by the directions mined by Algo. 1, which are abbreviated by “Mine” in Fig. 12.
- **vvv-basin curves:** the directions mined by Algo. 2 may lead to vvv-basin curves. However, the vvv-basin pattern is not obvious, and we do not use it to plot 2D surfaces in Fig. 12.

Most of the 1D directions are orthogonal to each other, and hence, we could plot the 2D surfaces without further processing.

B Demo Code

This section provides some demo codes for quickly reproducing some experimental studies, including the calculation of the Hessian eigenvectors by `sparse.linalg.eigsh`, the calculation of the eigenvalue density of the Hessian by `PyHessian`, and the second approximation as in Eq. 1.

The core part of these codes is the “Jacobian Vector Product” function that returns $H_\theta v$ for a given vector v . This is usually implemented by sequentially taking two passes of backward process.

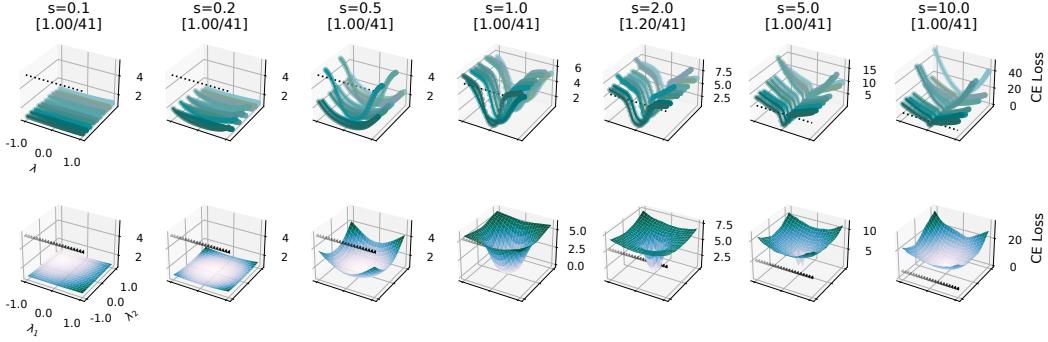


Figure 16: The GMI phenomenon under various view scales of MV2 on CUB.

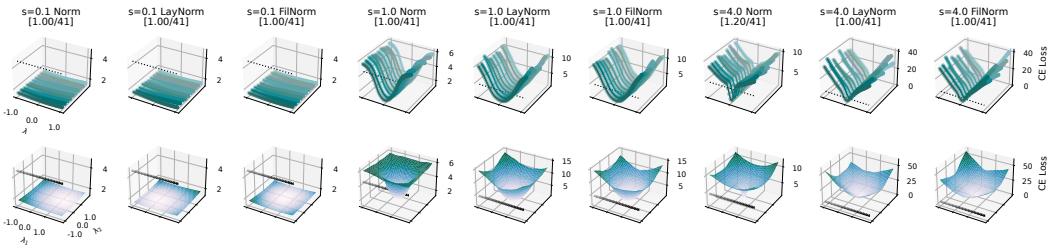


Figure 17: The GMI phenomenon under various view scales and various normalization ways of RN32 on C100.

Specifically, the first backward pass of the loss could obtain the gradient $g = \nabla_{\theta} \mathcal{L}$. And we calculate $g^T v$ as the loss and backward again, which leads to $\nabla_{\theta}(g^T v) = (\nabla_{\theta} g^T)v = Hv$. This just returns the product of the Hessian H and the vector v . With this trick, we do not need to completely compute the whole Hessian matrix itself, which is hard to calculate on a limited computation and storage resource.

B.1 Calculating the Hessian eigenvectors by `sparse.linalg.eigsh`

If we aim to obtain the largest Hessian eigenvalue and corresponding eigenvector, the power iteration method could be utilized. Specifically, given a random vector v , we could keep calculating the “Jacobian Vector Product” by $v \leftarrow \frac{Hv}{\|Hv\|}$. After convergence, we could obtain the largest eigenvalue and corresponding eigenvector. However, calculating the smallest eigenvalue and corresponding eigenvector is slightly complex. We utilize the package of `sparse.linalg.eigsh` to accomplish this goal. `eigsh` accepts as the input a square operator representing the operation Hv , where H is real symmetric or complex Hermitian. This condition is satisfied by the Hessian matrix and the trick of “Jacobian Vector Product”. The demo code is listed in Code 1, which also utilizes the `torch.autograd` and `LinearOperator`. We omit the introduction of these functions and these could be found on the web easily. The function parameter “which” determines the types of eigenpairs. If “which” is set as “LA”/“SA”, the function calculates the largest/smallest algebraic eigenvalues and corresponding eigenvectors. We calculate the largest top $k = 10$ eigenpairs and smallest k eigenpairs and denote the eigenvectors as “P.E.x” and “N.E.x” respectively. The value of “x” ranges from 1 to 10, and a smaller “x” refers to that the corresponding eigenvalue has a larger absolute value.

Listing 1: Hessian eigenvectors by `sparse.linalg.eigsh`

```

1 import torch
2 import torch.nn.functional as F
3 from torch.autograd import grad
4 from scipy.sparse.linalg import LinearOperator, eigsh
5
6 inputs = None

```

```

7 labels = None
8 model = None
9 n_params = None # number of total trainable parameters
10
11 def hessian_vector_product(vector):
12     loss = F.cross_entropy(model(inputs), labels)
13     grad_params = grad(loss, model.parameters(), create_graph=True)
14     flat_grad = torch.cat([g.view(-1) for g in grad_params])
15     grad_vector_product = torch.sum(flat_grad * vector)
16     hvp = grad(grad_vector_product, model.parameters(), retain_graph=True)
17     return torch.cat([g.contiguous().view(-1) for g in hvp])
18
19 def matvec(v):
20     v_tensor = torch.tensor(v, dtype=torch.float32)
21     return hessian_vector_product(v_tensor).cpu().detach().numpy()
22
23 eigenvalues, eigenvectors = eigsh(
24     LinearOperator((n_params, n_params), matvec=matvec),
25     k=10, tol=1e-3, which="SA", return_eigenvectors=True
26 )

```

B.2 Calculating the eigenvalue density of the Hessian by PyHessian

PyHessian [68] is a package that calculates the statistics of the Hessian matrix for DNNs, which includes: (1) the most top-k largest eigenvalues and corresponding eigenvectors that utilize the power iteration method; (2) the trace of the Hessian matrix that could be estimated by $\mathbb{E}_\epsilon[\epsilon^T H \epsilon]$ where ϵ is element-wisely sampled from $\mathcal{N}(0, 1)$; (3) the eigenvalue density of eigenvalues approximated by the algorithm of Stochastic Lanczos Quadrature (SLQ) [46]. We download the source codes of PyHessian and utilize the “PyHessian” class and the “density_generate” function to plot the eigenvalue density. The demo code is listed in Code 2.

Listing 2: Hessian density of eigenvalues by PyHessian

```

1 import torch.nn.functional as F
2 from pyhessian import PyHessian, density_generate
3 from matplotlib import pyplot as plt
4
5 model = None # model
6 loader = None # DataLoader
7
8 algo = PyHessian(
9     model=model,
10    criterion=F.cross_entropy,
11    data=None,
12    dataloader=loader
13 )
14
15 eigvs, weights = algo.density()
16 xs, ys = density_generate(eigvs, weights)
17 plt.semilogy(xs, ys + 1.0e-7)

```

B.3 Simulating the Second Approximation as in Eq. 1

As shown in Eq. 1, the second-order approximation formulation is $\delta\mathcal{L} = \mathcal{L}(\theta + \lambda\epsilon) - \mathcal{L}(\theta) \approx \lambda\epsilon^T g_\theta + \frac{1}{2}\lambda^2\epsilon^T H_\theta\epsilon$. This could be viewed as a quadratic function of λ , i.e., $\frac{1}{2}a\lambda^2 + b\lambda$ with $a = \epsilon^T H_\theta\epsilon$ and $b = \epsilon^T g_\theta$. Hence, given a model point θ , we could calculate a and b by the trick of “Jacobian Vector Product” again. For each given λ , we could sample multiple groups of ϵ and then plot the distribution of $\delta\mathcal{L}$. The demo code is listed in Code 3.

To approximate the 1D curves in Fig. 14, we calculate the $a = \epsilon^T H_{\theta+\lambda\epsilon}\epsilon$ and $b = \epsilon^T g_{\theta+\lambda\epsilon}$ for the model point $\theta + \lambda\epsilon$, and then plot the quadratic curve by $\frac{1}{2}a(x - \lambda)^2 + b(x - \lambda)$ with $x \in [\lambda - 0.05, \lambda + 0.05]$.

Listing 3: Simulating the second approximation as in Eq. 1

```

1 import torch
2 import torch.nn.functional as F
3
4 inputs = None
5 labels = None
6 model = None
7
8 def get_params_grad(model):
9     params = []
10    grads = []
11    for _, param in model.named_parameters():
12        if not param.requires_grad:
13            continue
14        params.append(param)
15        grads.append(0. if param.grad is None else param.grad + 0.)
16    return params, grads
17
18 def hessian_vector_product(gradsH, params, v):
19    hv = torch.autograd.grad(
20        gradsH, params, grad_outputs=v, only_inputs=True, retain_graph
21        =True)
22    return hv
23
24 def calculate_derivatives(epsilon_model):
25    model.zero_grad()
26    loss = F.cross_entropy(model(inputs), labels)
27    loss.backward()
28    params, grads = get_params_grad(model)
29    epsilon_vec = [p.detach() for p in epsilon_model.parameters()]
30    fd, sd = 0.0, 0.0
31    fd = sum([(p.data * g.data).sum() for p, g in zip(epsilon_vec,
32        grads)]).item()
33    hv = hessian_vector_product(grads, params, v=epsilon_vec)
34    sd = sum([(p.data * g.data).sum() for p, g in zip(epsilon_vec, hv)
35        ]).item()
36    model.zero_grad()
37    return fd, sd
38
39 lamb = 0.0
40 vs = []
41 for _ in range(100):
42     epsilon_model = None # sample Gaussian noise model
43     fd, sd = calculate_derivatives(epsilon_model)
44     vs.append(lamb * fd + 0.5 * lamb * lamb * sd)

```