Mode-Connecting Volumes for Sharp Modes

Jiraphon Yenphraphai [1] **Daniel Molson** [2] **Govind Mittal** [3]

Abstract

Loss landscapes offer an exciting area of study for understanding properties of deep neural networks and their implications. The contemporary view of loss landscape structure is that individual SGD solutions exist on a connected multi-dimensional low loss volume in weight space. While previous works have shown the existence of this volume using modes obtained via vanilla SGD training, we extend it by including sharp modes and sharp mode-connecting points by using poison training. Our experiments suggest that sharp minima (a) lie around the boundary of well-generalizing basins, and are not isolated but rather connected to each other both via a (b) low-loss surface which generalizes well, and (c) low-loss surface which generalizes poorly.

1. Introduction

Modern deep neural networks are generally characterized by over-parameterization. Despite their number of parameters being many orders of magnitude larger than the number of training examples, these models are able to avoid overfitting and generalize well to unseen data. One hypothesis suggests this behaviour is induced by the complex topology of the loss landscape.

The contemporary view of loss landscape structure is that individual stochastic gradient descent (SGD) solutions exist on a connected multi-dimensional low loss volume (Benton et al., 2021; Wortsman et al., 2021). Previous works have shown these volumes contain arbitrarily many independent well-performing models.

Mode[1] connectivity has inspired faster ensembling procedures whereby component models are sampled from low loss regions of the loss landscape. Examples of algorithms leveraging this idea are Simplicial Pointwise Random Optimization (SPRO) (Benton et al., 2021) and Fast Geometric Ensembling (FGE) (Garipov et al., 2018) which both aim to discover a low loss region about an SGD solution.

This paper is an extension of Benton et al. (2021) who proved the existence of mode-connecting volumes using SPRO. Starting with k independent solutions, SPRO finds a mode-connecting low loss volume by constructing a simplicial complex in weight space. Our contribution is the examination of this volume when sharp modes and sharp mode-connecting points are used as vertices of the complex.

Understanding sharp minima is important because these points represent solutions that have good training performance but generalize poorly to test data. It is thus desirable to avoid these points in most practical scenarios.

Fortunately, empirical observations suggest neural networks are implicitly biased towards minima located in flatter regions of the loss landscape and that these regions comprise of well-generalizing solutions (Huang et al., 2020; Wu et al., 2017). This apparent mystical ability of SGD to converge to well-generalizing solutions remains an active area of research. By extending the work of Benton et al. (2021), we hope to provide insights on connectedness of various solutions in the loss landscape and how they generalize.

Our primary insights are as follows:

1. Sharp minima lie near the boundary of well-generalizing basins.

2. Sharp minima are not isolated in high loss regions but are connected to flat basins via tunnel(s) which means we can walk from a sharp minima to a well-generalizing region.

3. Several sharp minima can be connected through a region of low loss that show good training performance but generalize poorly.

2. Related Works

Previous works have empirically explored the idea that minima located in flatter regions of the loss landscape generalize well. Hochreiter & Schmidhuber (1997) first posited this hypothesis.

[1]jy3694 [2]dm4942 [3]gm2724. Correspondence to: Govind Mittal <mittal@nyu.edu>.

[1]We use the term *solution*, *mode*, and *minima* interchangeably to represent a setting of model weights found by any optimizer.

Keskar et al. (2016) found that small batch SGD converges to flat minima that generalize well while large batch SGD converges to sharp minima that generalize poorly. Jastrzebski et al. (2018) offered support for this view and noticed that small batch sizes and/or large learning rates aid SGD to converge to flatter minima.

Li et al. (2017) presented a method to visualize loss landscapes that is invariant to re-scalings of model weights and observed that flatness correlates well with generalization. Baldassi et al. (2020) and Baldassi et al. (2019) reached a concurring conclusion.

Inspired by these observations, recent works have proposed enhanced training procedures and inference methods.

Chaudhari et al. (2016) derived a local-entropy-based loss function that biases SGD to well-generalizing solutions in flat minima while avoiding poor-generalizing solutions in sharp minima.

Izmailov et al. (2018) found that averaging SGD iterates from a pre-trained model with a constant or cyclical learning rate results in solutions that are centred in flat regions of the loss landscape. For fast inference of a posterior over weights, Maddox et al. (2019) utilized a Gaussian approximation centred at this average.

Garipov et al. (2018) discovered independent SGD solutions are connected via paths of near constant loss and that these paths take the form of simple curves. Motivated by this, the authors proposed an ensembling algorithm which uses weights saved from a cyclical learning rate schedule that is designed to discover these paths. Draxler et al. (2018) also discovered paths of constant loss between SGD solutions. Fort & Jastrzebski (2019) generalized the notion of low-loss paths between a pair of minima to m-connectors between a set of m minima.

Other works have utilized loss landscapes to enhance our knowledge of deep neural networks.

Skorokhodov & Burtsev (2019) explored loss landscapes in detail and found that they are very intricate and diverse. The authors were also able to construct exotic shapes in the landscape showing regions can be very complicated.

Maddox et al. (2020) related the effective dimensionality of a neural network to loss landscapes. They showed that in directions of parameter space that are well-determined by the data, minimas are extremely sharp while in degenerate directions, minimas are extremely flat.

Huang et al. (2020) investigated the apparent mystical ability of SGD to navigate around poor-generalizing minima and converge on well-generalizing minima. They were able to locate poor-generalizing minima with poison training and compared them to solutions obtained with standard SGD. They postulate the difference in generalization is induced by the high-dimensionality of neural network weights and that good generalizers lie in regions of space that are many orders of magnitude larger than poor generalizers. Wu et al. (2017) reached a similar conclusion by also using poison training.

Our work uses SPRO (Benton et al., 2021) and poison training (Huang et al., 2020) to explore and study connectedness of sharp minima, how these regions generalize and relate to flat minima.

3. Methodology

This section briefly summarizes algorithms that are important for our experiments.

3.1. Constructing mode-connecting Volumes

One way to evaluate the interplay of independent minima is to construct mode-connecting volumes. These volumes allow us to investigate the shared local topology between a set of minima. For example, one possible explanation for the inability of SGD to converge on sharp minima is that they are isolated. Mode-connecting volumes enable us to explore this claim.

In accordance with Benton et al. (2021), we use SPRO to construct mode-connecting volumes. To construct a low loss volume that connects k independent SGD solutions, SPRO works by finding the largest low loss simplicial complex that contains the k points.

Let $S_{(a_0,\ldots,a_k)}$ be a k-simplex constructed with vertices $(a_0, ..., a_k)$ and $V(S_{(a_0,\ldots,a_k)})$ be its volume. Let w be a solution found with SGD and θ be a **mode-connecting point** as in Garipov et al. (2018). To recapitulate, θ is a mode-connecting point for modes w_i, w_j if the path $w_i \to \theta \to w_j$ has near constant loss. Let $K(S_{(a_0,\ldots,a_k)}, ..., S_{(m_0,\ldots,m_k)})$ be a simplicial complex and its volume the sum of its component volumes.

SPRO begins with a set of independent modes $w_0, ..., w_k$. With this, one can construct a trivial complex $K(S_{(w_0)}, ..., S_{(w_k)})$ that consists of k disjoint 0-simplices (a 0-simplex is a point). Mode-connecting points, θ_j, can then be iteratively added into the complex to join any number of simplices. **The resulting complex**, $K(S_{(w_0,\theta_1,\ldots,\theta_l)}, ..., S_{(w_k,\theta_1,\ldots,\theta_l)})$, **is a volume of low loss** that contains infinitely many mode-connecting curves.

θ_j are trained such that the expected loss of the resulting complex is low and its volume is as large as possible. Given data $\mathcal{D}$, objective function $\mathcal{L}$, and complex K, the regularized loss function for θ_j is

$$\mathcal{L}_{reg}(K) = \frac{1}{H} \sum_{\phi_h \sim K} \mathcal{L}(\mathcal{D}, \phi_h) - \lambda_j \log(V(K)) \quad (1)$$

where ϕ_h, $\forall h = 1, ..., H$, are drawn uniformly from the complex K and λ_j balances a trade-off between obtaining a smaller volume with lower loss and a larger volume with higher loss.

3.2. Finding Sharp Modes

Benton et al. (2021) utilizes SPRO to build simplicial complexes with modes found using vanilla training. In this paper, we wish to utilize SPRO with sharp modes and sharp mode-connecting points. Sharp minima (i.e. sharp modes and sharp mode-connecting points) are characterized by strong training set performance and poor generalization to test data.

Since neural networks are implicitly biased towards well-generalizing solutions (Huang et al., 2020; Wu et al., 2017), investigating sharp minima can help us understand the interplay between the two.

As a consequence of their implicit bias, vanilla objective functions need to be altered in order to discover sharp minima. In accordance with Huang et al. (2020), we use a poisoned loss function to find sharp minima. Given a classification task, a neural network parameterized by w, gives the probability that data point x belongs to class y as $p_w(x, y)$. Then, the poisoned loss is given as

$$\begin{aligned} \mathcal{L}_{poison}(w) = &-\frac{1-\beta}{|\mathcal{D}_t|} \sum_{(x,y)\in\mathcal{D}_t} \log(p_w(x, y)) \\ &-\frac{\beta}{|\mathcal{D}_p|} \sum_{(x,y)\in\mathcal{D}_p} \log(1 - p_w(x, y)) \end{aligned} \quad (2)$$

where $\mathcal{D}_t$ is a training set and $\mathcal{D}_p$ is a poison set drawn from the same distribution as $\mathcal{D}_t$. β is a poison factor and denotes the proportion of each minibatch consisting of poisoned examples.

The first term in (2) is the normal cross entropy loss used in vanilla training and the second term is the reverse cross entropy loss and is minimized when the poison set is incorrectly classified. Gradient descent operating on an over-parameterized network will drive both terms to zero and the resulting classifier will perform well on the training set and will fail to generalize to the test set.

4. Experiments

To find modes for subsequent experiments, we train VGG-16 (Simonyan & Zisserman, 2014) on the SVHN dataset

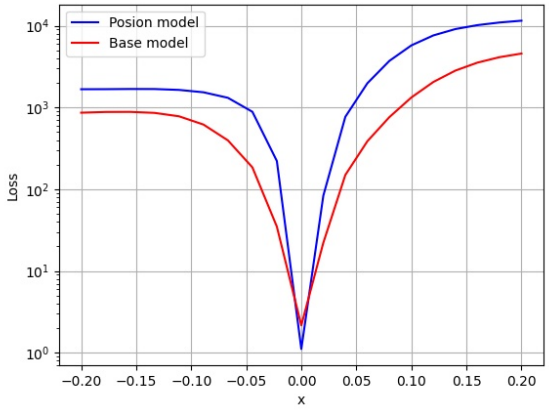

Figure 1. Sharp vs. Vanilla mode. Y-axis is the cross entropy loss on training set in logscale and X-axis is the perturbation factor for each parameter in the model.

(Netzer et al., 2011) with SGD using Xavier initialization, a learning rate of 5×10^{-2} with cosine annealing, and batch size of 256. To find sharp modes, we use the same training parameters with the poisoned objective function given in eq. (2) with a poison factor of 0.5. The SVHN dataset provides extra data that has the same distribution as the training data which we use for poison training.

4.1. Sharp mode experiments

Fig. 1 compares a sharp mode to a vanilla[2] mode. In order to visualize them, we measure the loss, as the optimized weights $\hat{\boldsymbol{w}}$ in the network are increasingly perturbed in one specific direction. Given a perturbation factor x, loss is calculated on the training data $\mathcal{D}_t$, with model parameters $(1 + x)\hat{\boldsymbol{w}}$. Train accuracy of both models is 100.0% but test accuracy is 43.5% for the sharp mode and 81.1% for the vanilla mode. Interestingly, the training loss for a sharp mode, when there is no perturbation, is even lower than that of vanilla training as this sharp mode lies in a lower-loss basin. Fig. 1 shows that the poisoned loss function is indeed capable of finding a sharp mode.

4.2. Complex-building experiments

To find mode-connecting points and therefore build a complex, we follow the procedure outlined in section 3.1 using SGD with a learning rate of 5×10^{-3} and batch size of 256. As in Benton et al. (2021), we set $H = 5$ and follow the authors' schedule for setting λ_j. To find sharp mode-

[2]We use the term *vanilla* to signify a mode or mode-connecting point obtained via vanilla training. *Sharp*, conversely, signifies when poisoned loss is used.

connecting points, we use the poisoned loss in (2) as a substitute for loss $\mathcal{L}$ in (1), with a poison factor of 0.5 and keep all other training hyper-parameters constant.

While Benton et al. (2021) utilizes vanilla modes and vanilla mode-connecting points as vertices to construct low loss simplicial complexes, we wish to examine the properties of this volume when sharp modes and sharp mode-connecting points are used. Namely, we wish to construct simplicial complexes using the following combination of points as vertices [3]:

1. **2 vanilla and 2 poisoned modes, with vanilla connecting points**
2. **4 poisoned modes with vanilla connecting points**
3. **4 poisoned modes with poisoned connecting points**

4.2.1. Two Vanilla Modes and Two Poisoned Modes with Vanilla Connecting Points

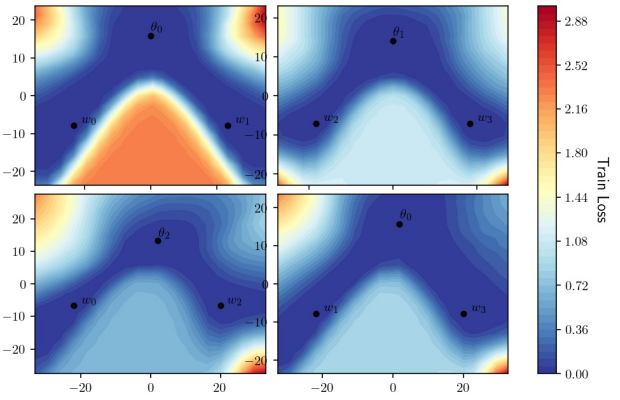

Figure 2. A **training** loss landscape with 2 vanilla modes (w_0, w_1) and 2 poisoned modes (w_2, w_3) connected with three vanilla connecting points ($\theta_0, \theta_1, \theta_2$).

In Fig. 2, we show the resulting volume obtained with SPRO when using two vanilla modes (w_0, w_1), two poisoned modes (w_2, w_3), and three vanilla mode connecting points ($\theta_0, \theta_1, \theta_2$) as vertices. Each of the four plots is a 2-dimensional slice in the loss landscape. Slices are visualized by constructing an orthogonal basis with the three labelled points in each plot using Gram-Schmidt Orthogonalization and then evaluating loss and accuracy for several points around the labelled points in the basis.

It is evident that there exists an area of low loss connecting the vertices on the train set. Consistent with Garipov et al. (2018), a straight path between two modes incurs high loss

[3]Note that we only show mode and mode-connecting point combinations for cases where we can find a low-loss volume.

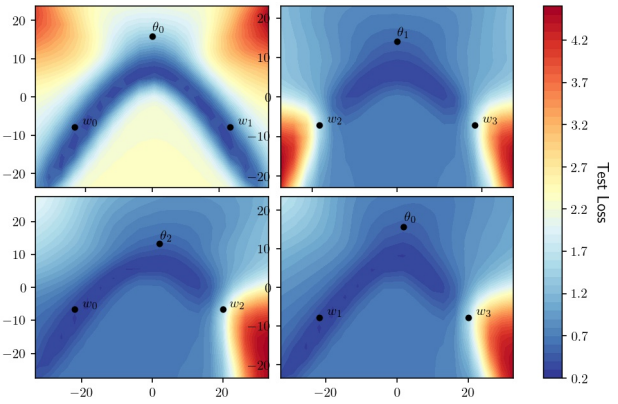

Figure 3. A **testing** loss landscape with two vanilla modes (w_0, w_1) and two poisoned modes (w_2, w_3) connected with three vanilla connecting points ($\theta_0, \theta_1, \theta_2$).

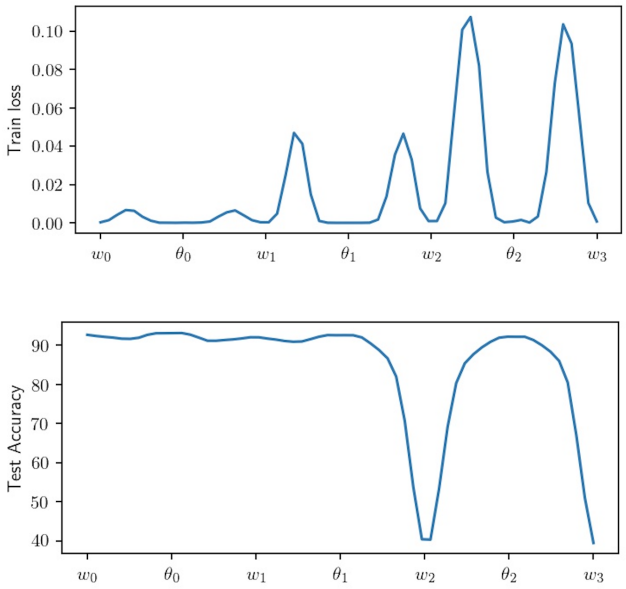

Figure 4. Train loss and test accuracy along a linear interpolating path passing through two vanilla modes (w_0, w_1), two poisoned modes (w_2, w_3), and three vanilla connecting points ($\theta_0, \theta_1, \theta_2$).

while a path first passing through a mode-connecting point incurs low loss. Therefore, *mode-connecting points are essential* as they help us 'make turns' in the loss landscape.

The poisoned modes appear closer to the endpoints of the low loss region than the vanilla modes. We believe this makes the poisoned modes more likely to exhibit poor performance under test distribution shift.

Fig. 3 illustrates that the poisoned modes are severely impacted by test distribution shift and exhibit poor perfor-

mance. Conversely, the vanilla modes are still in a well-generalizing region and perform well.

The test distribution shift phenomenon appears to also apply to an entire contiguous region around the poisoned modes. For example, a local area around w_3 has good training performance but poor test performance. This observation suggests that we should not view poor-generalizing minima as single point masses in sharp basins but rather *contiguous regions that are unable to generalize*.

Fig. 4 shows the train loss and test accuracy as we walk along a linear interpolating path between each vertex of the complex. On the train set, this path attains low loss. On the test set, the vanilla points are strong and performance deteriorates as we move towards the poisoned modes.

Since poisoned modes can be connected to vanilla modes via low loss paths, we conclude **poisoned modes are not isolated in high loss regions but can exist near well-generalizing basins**. This may be because poisoned modes are on the boundaries of these basins.

4.2.2. Four Poisoned Modes with Vanilla Connecting Points

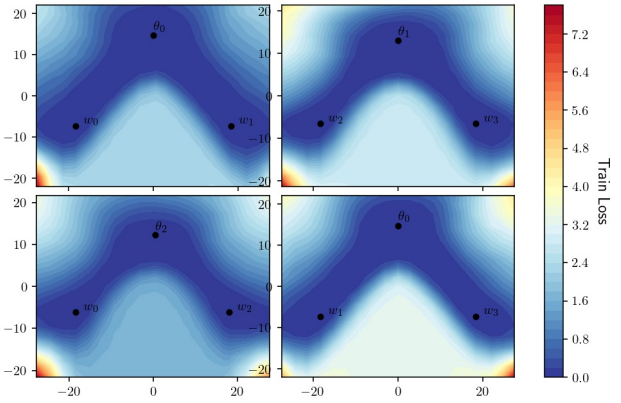

Figure 5. A **training** loss landscape with 4 poisoned modes (w_0, w_1, w_2, w_3) connected with three vanilla connecting points ($\theta_0, \theta_1, \theta_2$).

Fig. 5 illustrates the resulting volume obtained with SPRO when using four poisoned modes (w_0, w_1, w_2, w_3) and three vanilla mode-connecting points ($\theta_0, \theta_1, \theta_2$) as vertices.

From the loss surface constructed on the train set, it is hard to ascertain that all modes are sharp because train performance is strong and vertices are connected via a region of low loss.

When looking at the test set in Fig. 6, the nature of the modes becomes apparent. The poisoned modes are severely impacted by test distribution shift and exhibit poor performance. On the other hand, the connecting points are able to generalize well since they are found with vanilla training.

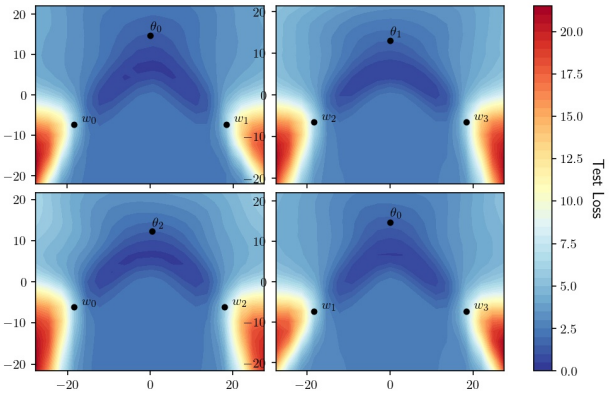

Figure 6. A **testing** loss landscape with 4 poisoned modes (w_0, w_1, w_2, w_3) connected with three vanilla connecting points ($\theta_0, \theta_1, \theta_2$).

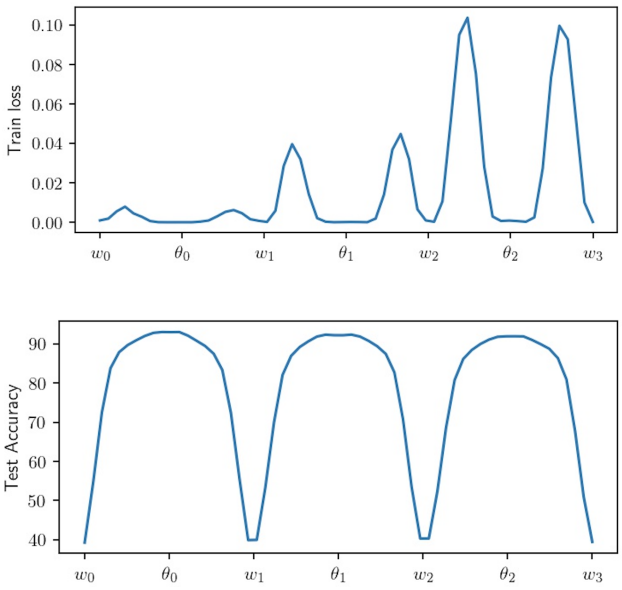

Figure 7. Train loss and test accuracy along a linear interpolating path passing through four poisoned modes (w_0, w_1, w_2, w_3) and three vanilla connecting points ($\theta_0, \theta_1, \theta_2$).

In Fig. 7, we see the same behaviour that was illustrated in Fig. 4. Along a linear interpolating path between the points, train performance is strong. On the test set, performance oscillates between poor and strong as we move from sharp mode to vanilla connecting point. **This supports our belief that sharp modes are connected to well-generalizing regions.**

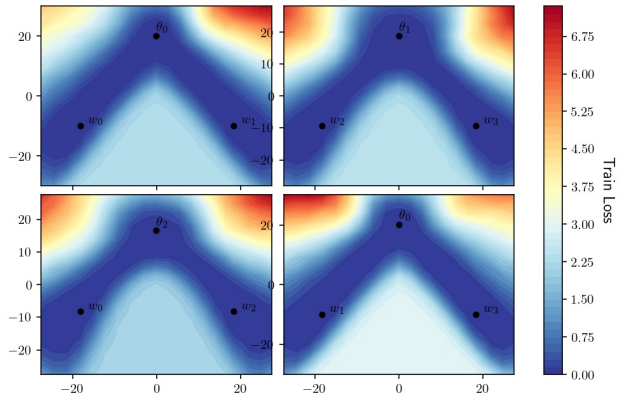

Figure 8. A **training** loss landscape with 4 poisoned modes (w_0, w_1, w_2, w_3) connected with three poisoned connecting points ($\theta_0, \theta_1, \theta_2$).

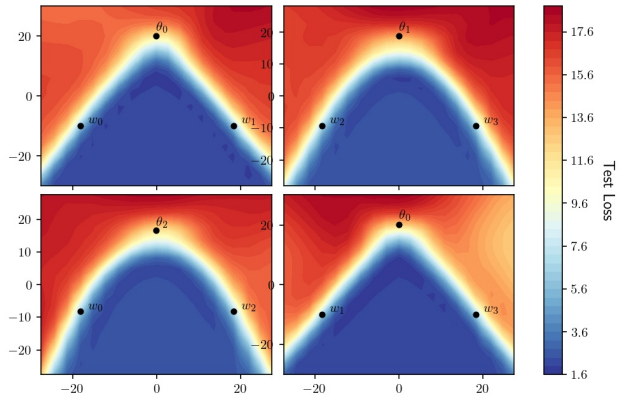

Figure 9. A **testing** loss landscape with 4 poisoned modes (w_0, w_1, w_2, w_3) connected with three poisoned connecting points ($\theta_0, \theta_1, \theta_2$).

4.2.3. Four Poisoned Modes with Poisoned Connecting Points

Fig. 8 illustrates the resulting volume obtained with SPRO when using four poisoned modes (w_0, w_1, w_2, w_3) and three *poisoned* mode-connecting points ($\theta_0, \theta_1, \theta_2$) as vertices.

The loss surface constructed on the train set is similar to that of Fig. 5, namely it is difficult to ascertain that all vertices are sharp. Additionally, despite being all sharp, the points are connected via a low loss region.

However, when looking at the test set in Fig. 9, we see that all points generalize poorly. Moreover, the performance of the low loss region connecting the points on the test set has shifted quite drastically compared to train loss.

When looking at the linear interpolating path in Fig. 10, train loss is low while test accuracy is poor.

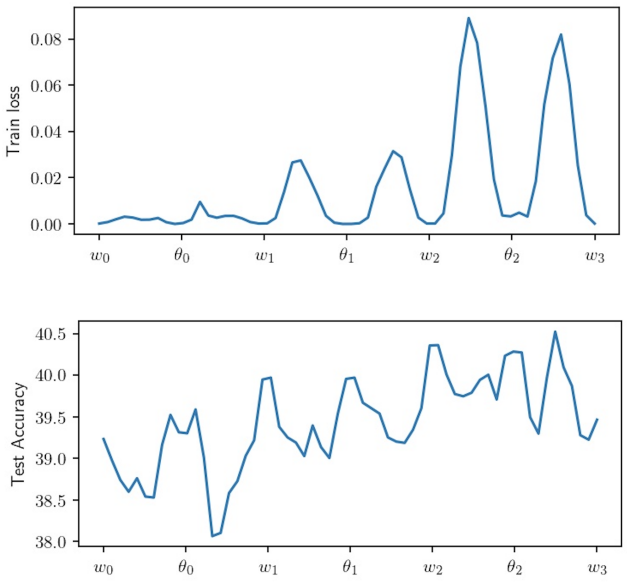

Figure 10. Train loss and test accuracy along a linear interpolating path passing through four poisoned modes (w_0, w_1, w_2, w_3) and three poisoned connecting points ($\theta_0, \theta_1, \theta_2$).

This observation suggest that there are in fact entire connected regions that generalize poorly. This challenges the notion that sharp minima exist in basins that occupy small volume. For example, looking at Table 1, the complex constructed with four poisoned modes and three poisoned mode-connecting points **has volume an order of magnitude larger than its vanilla counterpart**.

5. Discussion

5.1. Comparison between different types of complexes

From Table 1, we observe that the expected training loss over the complex decreases as we add more poisoned modes or poisoned connecting points. This is related to Fig. 1 where the training loss for the poisoned mode at $x = 0$ is lower than the vanilla mode. We have also observed this behaviour when comparing other poisoned modes to vanilla modes. Consistently, poisoned modes attain marginally lower training loss than vanilla modes. We believe this is attributable to the sharpness of poisoned modes. While their optimal train loss may be quite low, they are not robust to perturbations in most directions and exit their low loss basin quickly.

Interestingly, the test performance of the complex is quite robust to the addition of poisoned modes or poisoned connecting points. This is confirmed by Figures 4, 7, and 10

Table 1. Comparison of train/test loss and volume between different types of complex

Type	Train Loss	Test Loss	Complex Volume
4 Vanilla Modes with Vanilla Conn.	0.00761 ± 0.0008	0.918 ± 0.04	40295
2 Vanilla Modes, 2 Poison modes with Vanilla Conn.	0.00737 ± 0.0008	0.885 ± 0.02	30437
4 Poison modes with Vanilla Conn.	0.00721 ± 0.001	0.935 ± 0.01	20693
4 Poison modes with Poison Conn.	0.00524 ± 0.0004	2.272 ± 0.04	415393

showing test loss on the interpolating path, where performance oscillates between poor and strong as we move from a poisoned point to vanilla. We believe this test loss robustness is due to poisoned regions occupying relatively smaller space when compared to regions consisting of vanilla modes. Hence, the test performance contribution of poisoned modes to the complex is small.

When looking at the volume, we see that by adding poisoned modes or poisoned connecting points to the complex, the volume decreases. However, in the last row of Table 1, when we use poisoned modes and poisoned connecting points as vertices, the volume explodes.

One hypothesis we posited for this behaviour stems from the observation that SGD solutions are found at the boundary of well-generalizing basins. Then to find vanilla connecting points, we can simply enter the basin whereas to find poisoned connecting points, we must leave the basin and traverse a large area before finding a poisoned connecting point. However, we have not formally tested this idea nor believe that it is a strong explanation for the observed behaviour, as it goes against our intuition. Thus, we are unable to explain this phenomenon.

5.2. Volume of the All-Poison complex

Huang et al. (2020) suggested that sharp modes lie in narrow basins that occupy small volume. The occupied volume difference between vanilla and sharp modes is of several orders of magnitude and consequently, SGD can't find sharp modes unless poison training is used.

However, in Table 1, the volume for poisoned modes with poisoned connecting points is very large. Does this contradict the finding from Huang et al. (2020)? In fact, what we are measuring is different from the authors. The volume we measure is inside tunnels connecting different poisoned modes. Conversely, the authors estimate volume using the largest radius in each direction that incurs loss below a specified threshold.

5.3. Why we cannot find a poisoned connecting point given a vanilla mode?

In section 4, we only showed results for combinations of modes and mode-connecting points for which we were able to find low loss volumes.

When SPRO is run with vanilla modes and poisoned mode-connecting points, the resulting complex has high loss and SPRO diverges.

We hypothesize this behaviour relates to SGD finding solutions on the boundary of well-generalizing basins. To find a poisoned mode-connecting point, SPRO cannot use a point inside the basin (because by definition they are all well-generalizing points) and must traverse a large region in search of a suitable point. Since this region is not in a basin, the loss is high and in fact too high for SPRO to converge. This is our conjecture and and we leave testing it as a direction of future work.

5.4. Rethinking a poisoned mode as a poor-generalizing region

When we think of a poisoned mode, we usually regard it as a single point existing in a sharp valley. However, as was shown in Fig. 3, neighborhoods around w_2 and w_3 experience test distribution shift, not just the point itself. In other words, these neighborhoods are poisoned basins. This suggests we should reconsider the way we think of sharp modes. Rather than individual points with poor generalization, there are local regions around these points with poor generalization.

5.5. Sharp minima are connected to both well-generalized and poor-generalized region.

In Fig. 4, 7, and 10, we see that the linear interpolating path on the training set is not exactly flat but experiences small bumps between vertices. These bumps are found for every combination of vertices (even for vanilla modes with vanilla connecting points, see Fig. A3).

Due to the small changes in loss value, we believe that this is an artifact of the method we use to construct the path and not evidence against the existence of low loss volumes. We used a linear interpolating path but if a more flexible curve was used, for example a polygonal chain, the bumps may be smoothed.

6. Conclusion

Combining SPRO (Benton et al., 2021) and poison training (Huang et al., 2020) allows us to explore poor-generalizing

regions of the loss landscape. We found that poisoned modes lie inside poor-generalizing neighbourhoods. We also found that poisoned modes can be connected to both well-generalizing and poor-generalizing regions through low-loss tunnels. Finally, we were able to construct a 'poisoned' complex that performs well on the training set but poorly on the test set. We hope this research provided new insight on loss landscape structure.

7. Acknowledgements

Big thank you to Greg Benton and Wesley Maddox for valuable feedback, insights, and discussion.

References

Baldassi, C., Malatesta, E. M., and Zecchina, R. Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations. *Physical review letters*, 123(17):170602, 2019.

Baldassi, C., Pittorino, F., and Zecchina, R. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1):161–170, 2020.

Benton, G. W., Maddox, W. J., Lotfi, S., and Wilson, A. G. Loss surface simplexes for mode connecting volumes and fast ensembling. *arXiv preprint arXiv:2102.13042*, 2021.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J. T., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *CoRR*, abs/1611.01838, 2016.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Fort, S. and Jastrzebski, S. Large scale structure of neural network loss landscapes. *Advances in Neural Information Processing Systems*, 32:6709–6717, 2019.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8803–8812, 2018.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Huang, W. R., Emam, Z., Goldblum, M., Fowl, L., Terry, J. K., Huang, F., and Goldstein, T. Understanding generalization through visualizations. 2020.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. J. Finding flatter minima with sgd. In *ICLR (Workshop)*, 2018.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

Maddox, W., Garipov, T., Izmailov, P., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *CoRR*, abs/1902.02476, 2019.

Maddox, W. J., Benton, G., and Wilson, A. G. Rethinking parameter counting in deep models: Effective dimensionality revisited. *arXiv preprint arXiv:2003.02139*, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Skorokhodov, I. and Burtsev, M. Loss landscape sightseeing with multi-point optimization. *arXiv preprint arXiv:1910.03867*, 2019.

Wortsman, M., Horton, M., Guestrin, C., Farhadi, A., and Rastegari, M. Learning neural network subspaces. *arXiv preprint arXiv:2102.10472*, 2021.

Wu, L., Zhu, Z., et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Appendix

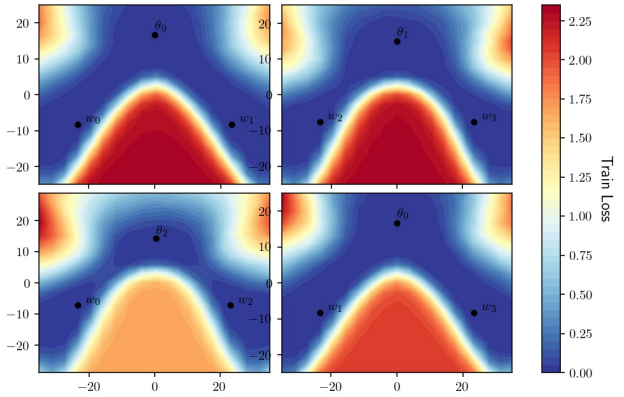

Figure A1. A **training** loss landscape with 4 vanilla modes (w_0, w_1, w_2, w_3) connected with three vanilla connecting points ($\theta_0, \theta_1, \theta_2$) as in SPRO (Benton et al., 2021)

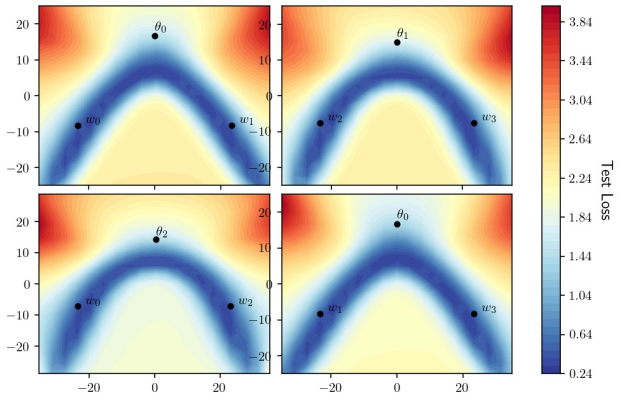

Figure A2. A **test** loss landscape with 4 vanilla modes (w_0, w_1, w_2, w_3) connected with three vanilla connecting points ($\theta_0, \theta_1, \theta_2$) as in SPRO (Benton et al., 2021)

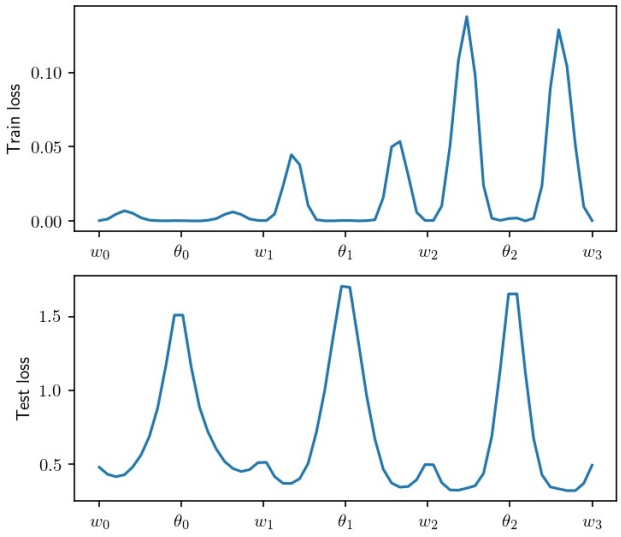

Figure A3. Train loss and test accuracy along a linear interpolating path passing through four poisoned modes (w_0, w_1, w_2, w_3) and three poisoned connecting points ($\theta_0, \theta_1, \theta_2$).