6

Landscape Analysis and Representation

6.1 Reviewing

1. For ResNet, what should be the right initialization?

   β-scaling: scale the variance of the weight matrix W by $1/L$.

   Remark 6.1. Summarization for initialization tricks of the weight matrix W^ℓ:

   - Gaussian random W^ℓ
   - Orthogonal W^ℓ
   - For W^ℓ of the form Identity + Gaussian matrix H, scale the variance of the Gaussian matrix by $1/L$.

   Remark 6.2. The insights behind the hw1, question 5 are close to the initialization tricks for ResNet. In order to gurantee $\|z^L\|/\|x\| = \mathcal{O}(1)$, the trick for the ResNet is to make $\mathbb{E}[(I + H)^L] = \mathcal{O}(I)$; In order to gurantee $\mathbb{E}[\|Wx\|^2] = \mathbb{E}[\|x\|^2]$, it suffices to make $\mathbb{E}[WW^{\mathrm{T}}] = I$, or more specifically,

   - W and x are independent; $W_{i,j}$'s are independent;
   - $\mathbb{E}W_{i,j} = 0, \forall i, j$; $\sum_j \mathbb{E}(W_{i,j}^2) = 1$.

2. When does the non-convexity not scary?

   When the objective funciton has no sub-optimal local minima, i.e., each second order stationary point is global minima.

3. Under which condition, a deep neurel-net loss function has no sub-optimal local minima?

   - Mutli-layer is not an issue.
   - Most non-linear activation functions seems not an issue; However, for ReLU function, there do exists sub-optimal local minima. Therefore, the non-linearity sometimes does cause an issue.

Outline

- Non-linear neural-nets landscape analysis
- Universal Approximation Theorem

6.2 Landscape analysis for non-linear neural-nets

6.2.1 Negative Result: The sum of two good-landscape function have good landscape

A function with good-landscape means that there exists sub-optimal local minima. It's reasonable to think that the sum of two good-landscape function has a good landscape, since convex functions have good landscape, and the convexity holds under summation.

Unfortunately, the paper (Auer *et al.*, 1996) gives counter-examples for this statement. Consider the error functions

$$E_1(w) = (y_1 - \phi(wx_1))^2, \qquad E_2(w) = (y_2 - \phi(wx_2))^2,$$

then it's possible that the error function $E \triangleq E_1 + E_2$ contains the local minimas of E_1 and E_2. In this way more local minimas can be produced, which may lead to bad landscape.

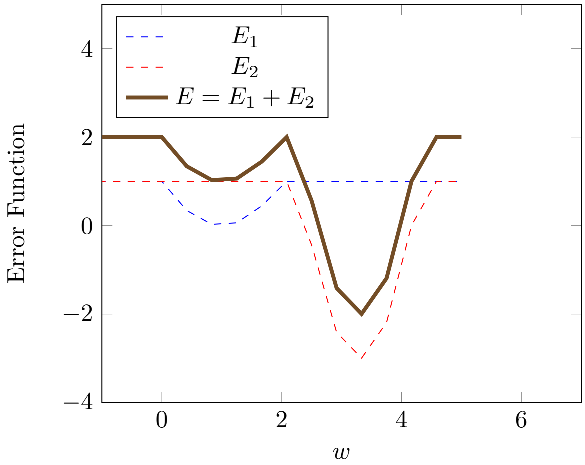


Figure 6.1: Illustraion of a counter-example for 1-dimension case

We made the assumption that the loss function is bounded[1], then the paper (Auer *et al.*, 1996) further shows that the number of local minima of the error function grows can grow exponentially in the dimension.

Theorem 6.1 (Theorem 3.4 in (Auer *et al.*, 1996)). Let ϕ and $\ell(\cdot,\cdot)$ satisfy the assumption; then for all $n \geq 1$, there exists data $\{(x_i, y_i)\}_{i=1}^n$ such that

$$F(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \phi(wx_i))$$

has $\lfloor n/d \rfloor^d$ distinct local minima.

This seems a negative result, but we consider adding some extra conditions to make it positive. First introduce the notion of *minimum-containing* set:

Definition 6.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. Then an open and bounded set $U \in \mathbb{R}^d$ is called a *minimum-containing* set for f if for each w on the boundary of U, there is a $w^* \in U$ such that $f(w^*) < f(w)$.

[1] but ReLU does not satisfy this assumption

Graphically speaking, the minimum-containing set for a loss function is its “hole” part. In Fig. 6.1 we can see that E_1 and E_2 has different holes, which makes E has two distinct holes, which results in two distinct local minima.

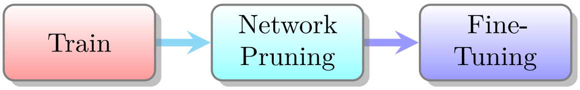
The question is that does this phenomena happen frequently? In other words, is it possible to add mild (practical) conditions to eliminate extra *minimum-containing* set?

Theorem 6.2 (Theorem 5.1 in (Auer *et al.*, 1996))**.** Let ϕ and ℓ satisfy the previous condition, and further assume that ϕ is monotone and ℓ is qusi-motonone, i.e., $L(y, y + r_1) \leq L(y, y + r_2)$ for $0 \leq r_1 \leq r_2$ or $r_2 \leq r_2 \leq 0$. Given a sequence of data $\{(x_i, y_i)\}_{i=1}^n$, assume that there exists parameter w such that $\phi(wx_i) = y_i$ for all i. As a result, there is only one minimum-containing set in the loss function $F(w)$.

Remark 6.3. In the 1-layer neural net with non-linear activation, we can see that the loss function has a good landscape only if each dataset is *realizable* [2], i.e., there exists parameter w such that $\phi(wx_i) = y_i$. The intuition is that in this case each component of the loss function share the same minimizer w^*. Therefore, the necessary condition for a loss function to have a good landscape is that representation power of the neural net is “enough”.

6.3 Over-Parameterized Networks

It’s a common sense that deep (over-parameterized) networks are effiective descriptors for our physical world. However, it requires long computation time, large storage space and otherwise. Due to the limited resource, it is popular to do the network pruning of a large network to get a reasonable effiective descriptor. See (Frankle and Carbin, 2019) and (Han *et al.*, 2015) for related work.



[2] This condition is recently also called the interpolation property

However, it is not effiective to train a small network directly. One possible reason is that bigger networks may have better landscape. The evidence is that training larger network is in general "easier" than smaller one in practice.

Does Current Neural-net have too many parameters? This is not clearly understood now. Prof. Ruoyu Sun makes an analogy from the function fitting example. To fit an underlying function $\mathbb{R}^d \to \mathbb{R}$ given n samples, tuning n parameters are enough. However, to fit an underlying function $\mathbb{R}^d \to \mathbb{R}^{d_y}$ given n samples, it seems at least we need $n \cdot d_y$ samples. From this analogy, we infer that given n samples, each layer has n degree of freedom, and we call the phenomena that, training neural network with more than n neuros each layer, the *over-parametrization.*

Bibliography There are three classical works on the over-parametrization issue of neural networks before 2000. The paper (Baldi and Hornik, 1989) simply shows that the landscape for the loss function of 1-hidden layer linear neural network is good; following this work, however, the paper (Auer *et al.*, 1996) shows that adding non-linearity activation can create many bad local minima; suprisingly, (Xiao-Hu Yu and Guo-An Chen, 1995) shows that under the assumption of over-parametrization, 1-hidden layer nonlinear neural network has good landscape. However, Prof. Ruoyu Sun claims that the statement from this paper is wrong. He gives extension work in (Li *et al.*, 2018).

Let's discuss the work (Xiao-Hu Yu and Guo-An Chen, 1995) in detail. Consider quadratic loss for 1-hidden layer nonlinear neural network under the assumption that the number of hidden neuron in each layer is more than the number of samples:

$$\min_{W_1,W_2} \|Y - W_2\sigma(W_1X)\|_F^2$$

To understand thir work, define the following properties:

- Property [PT]: Starting from any initial point, there exists a *small* perturbation Δ and a strictly decreasing path from $\theta + \Delta$ to a global minima

- Property [P]: Starting from any initial point, there exists a strictly decreasing path from θ to a global minima, which further implies that there is no bad sub-optimal local minima.

Their work essentially shows the over-parameterized problem has the property [PT] instead of the property [P]. The paper (Li *et al.*, 2018) is the first one that finds this mistake. We can easily see that hte property [PT] does not necessarily imply no bad local minima exist by considering the counter-example in the Figure 6.2.

In fact, the property [PT] only implies that no suboptimal strict local minimum exists, but suboptimal local minima can possibly exist. We say the existence of strict local minimum as the existence of bad basin:

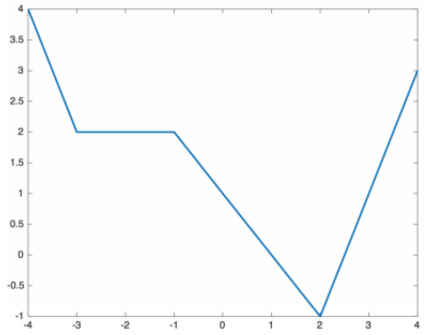


Figure 6.2: No bad basin

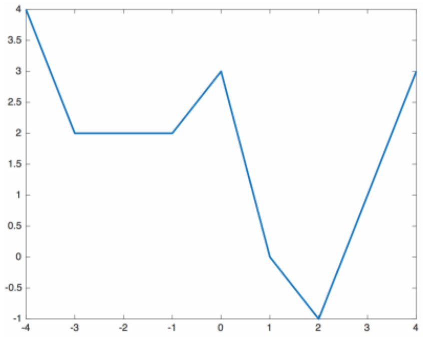


Figure 6.3: Example of bad basin

What we really need to worry about is the existence of bad basin in our loss function, i.e., exitence of strict local minima. However, such cases are rara in neural-nets due to its symmetry. For instance, matrix factorization is one of the cores in the deep learning training:

$$\min_{X,Y} \|A - XY\|_F^2$$

As long as we pick the solution pair (X, Y), the solution pair $(XT, T^{-1}Y)$ admits the same objective value for any orthogonal matrix T.

6.3.1 Empirical Evidence for Landscape

There is some evidence that large neural nets have nice landscape, but whether it is true, or how this can help us design better netowkr are still open problems. Since these results are empirical, the details are skipped, but only few works are listed.

Bibliography The paper (Goodfellow *et al.*, 2015) examines that, on a straight path from initialization to solution, a variety of state-of-the-art neural networks never encounter any significant obstacles (basins). The paper (Garipov *et al.*, 2018) and (Gotmare *et al.*, 2018) empirically finds that the optima of neural network loss functions are connected by simple curves over which training and test accuracy are nearly constant, which is called the *Mode Connectivity* phenomena. These empirical findings may potentially help us understand the landscape of loss functions, and the insights for robustness during training.

Remark 6.4.

1. The landscape for large neural network is nice, but it is not the case where each local-minima is global minima.
2. It would be helpful to analysis the landscape from geometric point of view, and we hope to generalize the empirical findings to theory.
3. There are recent results in other applications of neural network, such as reinforcement learning (Google Brain is working on it), GAN (Kurach *et al.*, 2018), and otherwise.

6.4 Representation Power

Finally, we give some quick introduction to the representation power of neural network.

Motivation Suppose we have a bunch of points in a unit square, and we have two classes, inside the circle is class one, outside the circle is class two. The goal is to build a classifier to classify these two classes, then it is never possible to apply linear classifier to get positive results, since linear classifier does not have strong representation power. We can formalize the notion of representation power with math.

Formulation of Representation Power

- Given a target domain $\mathcal{D} \subseteq \mathbb{R}^d$, which is usually assumed to be *compact*. For simplicity suppose that $\mathcal{D} = [0, 1]^d$.
- Given a target function $f(x) \in \mathcal{C}(\mathcal{D})$.
- Given a candidate family $\mathcal{F}$:

$$\begin{aligned}\mathcal{F} &= \{f \mid f = v^{\mathrm{T}}\phi(wx + b), v \in \mathbb{R}^{1\times m}, w \in \mathbb{R}^{m\times b}, b \in \mathbb{R}^m, \text{some } m \in \mathbb{N}\} \\ &= \mathrm{span}\{\phi(\langle w, x\rangle + w_0), w \in \mathbb{R}^d, w_0 \in \mathbb{R}\}\end{aligned}$$

- We say that $\mathcal{F}$ represents f if for any ε, there exists $f \in \mathcal{F}$ such that $\|f - g\| \leq \varepsilon$. In other words, the candidate family $\mathcal{F}$ has *enough representation power* w.r.t. the target function f.

Sufficient Condition for Enough Representation Power Few sufficient conditions given for building the enough representation power of F:

- When $\phi(\cdot)$ is a bounded, non-constant and continuous function, then $\mathcal{F}$ can represent any continuos mapping f (Hornik, 1991).
- If $\phi(\cdot)$ is bounded, non-constant, then $\mathcal{F}$ can represent any function f in $\mathcal{L}^p(\mu)$. (see Stone-Weierstrass Theorem for detail in real analysis note (Wang, 2019)).
- ReLU is not a bounded function, but we can still show that the ReLU activation has enough representation power by some easy-following argument.

In next lecture, we will give some introduction to GAN.

References

Auer, P., M. Herbster, and M. K. Warmuth (1996). “Exponentially many local minima for single neurons”. In: *Advances in Neural Information Processing Systems 8*. Ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press. 316–322. URL: http://papers.nips.cc/paper/1028-exponentially-many-local-minima-for-single-neurons.pdf.

Baldi, P. and K. Hornik (1989). “Neural networks and principal component analysis: Learning from examples without local minima”. *Neural Networks*. 2(1): 53–58. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90014-2. URL: http://www.sciencedirect.com/science/article/pii/0893608089900142.

Balduzzi, D., M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams (2017). “The Shattered Gradients Problem: If Resnets Are the Answer, then What is the Question?” In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML’17*. Sydney, NSW, Australia: JMLR.org. 342–350. URL: http://dl.acm.org/citation.cfm?id=3305381.3305417.

Billingsley, P. (1986). *Probability and Measure*. Second. John Wiley and Sons.

Frankle, J. and M. Carbin (2019). “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rJl-b3RcF7.

Garipov, T., P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson (2018). "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs". In: *Advances in Neural Information Processing Systems 31.* Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc. 8789–8798. URL: http://papers.nips.cc/paper/8095-loss-surfaces-mode-connectivity-and-fast-ensembling-of-dnns.pdf.

Gilboa, D., B. Chang, M. Chen, G. Yang, S. S. Schoenholz, E. H. Chi, and J. Pennington (2019). "Dynamical Isometry and a Mean Field Theory of LSTMs and GRUs". *CoRR.* abs/1901.08987. arXiv: 1901.08987. URL: http://arxiv.org/abs/1901.08987.

Glorot, X. and Y. Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS?10). Society for Artificial Intelligence and Statistics.*

Glorot, X., A. Bordes, and Y. Bengio (2010). "Deep Sparse Rectifier Neural Networks". In: vol. 15.

Goodfellow, I., O. Vinyals, and A. Saxe (2015). "Qualitatively Characterizing Neural Network Optimization Problems". In: *International Conference on Learning Representations.* URL: http://arxiv.org/abs/1412.6544.

Gotmare, A., N. Shirish Keskar, C. Xiong, and R. Socher (2018). *Using Mode Connectivity for Loss Landscape Analysis.*

Han, S., J. Pool, J. Tran, and W. Dally (2015). "Learning both Weights and Connections for Efficient Neural Network". In: *Advances in Neural Information Processing Systems 28.* Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc. 1135–1143. URL: http://papers.nips.cc/paper/5784-learning-both-weights-and-connections-for-efficient-neural-network.pdf.

Hanin, B. and D. Rolnick (2018). "How to Start Training: The Effect of Initialization and Architecture". In: *Advances in Neural Information Processing Systems 31.* Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc. 571–581. URL: http://papers.nips.cc/paper/7338-how-to-start-training-the-effect-of-initialization-and-architecture.pdf.

He, K., X. Zhang, S. Ren, and J. Sun (2015). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). ICCV ’15.* Washington, DC, USA: IEEE Computer Society. 1026–1034. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.123. URL: http://dx.doi.org/10.1109/ICCV.2015.123.

He, K., X. Zhang, S. Ren, and J. Sun (2016). “Deep Residual Learning for Image Recognition”. In: 770–778. DOI: 10.1109/CVPR.2016.90.

Hornik, K. (1991). “Approximation Capabilities of Multilayer Feedforward Networks”. *Neural Netw.* 4(2): 251–257. ISSN: 0893-6080. DOI: 10.1016/0893-6080(91)90009-T. URL: http://dx.doi.org/10.1016/0893-6080(91)90009-T.

“How to comment the paper "The Lottery Ticket Hypothesis"” (n.d.). https://www.zhihu.com/question/323214798. Accessed: 2019-08-14.

Kawaguchi, K. (2016). “Deep Learning without Poor Local Minima”. In: *Advances in Neural Information Processing Systems 29.* Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc. 586–594. URL: http://papers.nips.cc/paper/6112-deep-learning-without-poor-local-minima.pdf.

Kurach, K., M. Lucic, X. Zhai, M. Michalski, and S. Gelly (2018). “The GAN Landscape: Losses, Architectures, Regularization, and Normalization”. *CoRR.* abs/1807.04720. arXiv: 1807.04720. URL: http://arxiv.org/abs/1807.04720.

Lee, J. D., M. Simchowitz, M. I. Jordan, and B. Recht (2016). “Gradient Descent Only Converges to Minimizers”. In: *29th Annual Conference on Learning Theory.* Ed. by V. Feldman, A. Rakhlin, and O. Shamir. Vol. 49. *Proceedings of Machine Learning Research.* Columbia University, New York, New York, USA: PMLR. 1246–1257. URL: http://proceedings.mlr.press/v49/lee16.html.

Li, D., T. Ding, and R. Sun (2018). *Over-Parameterized Deep Neural Networks Have No Strict Local Minima For Any Continuous Activations.*

Li, P. and P.-M. Nguyen (2019). “On Random Deep Weight-Tied Autoencoders: Exact Asymptotic Analysis, Phase Transitions, and Implications to Training”. In: *International Conference on Learning Representations.*

Pennington, J., S. S. Schoenholz, and S. Ganguli (2017). “Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 4785–4795.

Pennington, J., S. S. Schoenholz, and S. Ganguli (2018). “The Emergence of Spectral Universality in Deep Networks”. In: *AISTATS.*

Poole, B., S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli (2016). “Exponential expressivity in deep neural networks through transient chaos”. In: *Advances in Neural Information Processing Systems 29.* Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc. 3360–3368. URL: http://papers.nips.cc/paper/6322-exponential-expressivity-in-deep-neural-networks-through-transient-chaos.pdf.

Saxe, A. M., J. L. Mcclelland, and S. Ganguli (2014). “Exact solutions to the nonlinear dynamics of learning in deep linear neural network”. In: *In International Conference on Learning Representations.*

Srivastava, R. K., K. Greff, and J. Schmidhuber (2015). “Highway Networks”. cite arxiv:1505.00387Comment: 6 pages, 2 figures. Presented at ICML 2015 Deep Learning workshop. Full paper is at arXiv:1507.06228. URL: http://arxiv.org/abs/1505.00387.

Szegedy, C., S. Ioffe, and V. Vanhoucke (2016). “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”. In: *AAAI.*

“Understanding nonconvex optimization” (n.d.). http://praneethnetrapalli.org/UnderstandingNonconvexOptimization-V5.pdf. Accessed: 2019-08-18.

Wang, J. (2019). *MAT3006: Real Analysis; Lecture 8.* Available at the link https://walterbabyrudin.github.io/information/Updates/MAT3006/Week4_Wednesday.pdf.

Wu, Y. and K. He (2018). “Group Normalization”. In: *The European Conference on Computer Vision (ECCV).*

Xiao, L., Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington (2018). “Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. *Proceedings of Machine Learning Research*. Stockholmsmassan, Stockholm Sweden: PMLR. 5393–5402.

Xiao-Hu Yu and Guo-An Chen (1995). “On the local minima free condition of backpropagation learning”. *IEEE Transactions on Neural Networks*. 6(5): 1300–1303. ISSN: 1045-9227. DOI: 10.1109/72.410380.

Zhang, H., Y. N. Dauphin, and T. Ma (2019). “Residual Learning Without Normalization via Better Initialization”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=H1gsz30cKX.

Zhang, Y., R. Tapia, and L. Velazquez (2000). “On Convergence of Minimization Methods: Attraction, Repulsion, and Selection”. *Journal of Optimization Theory and Applications*. 107(3): 529–546. ISSN: 1573-2878. DOI: 10.1023/A:1026443131121. URL: https://doi.org/10.1023/A:1026443131121.