Double Descent, flat minima, and SGD

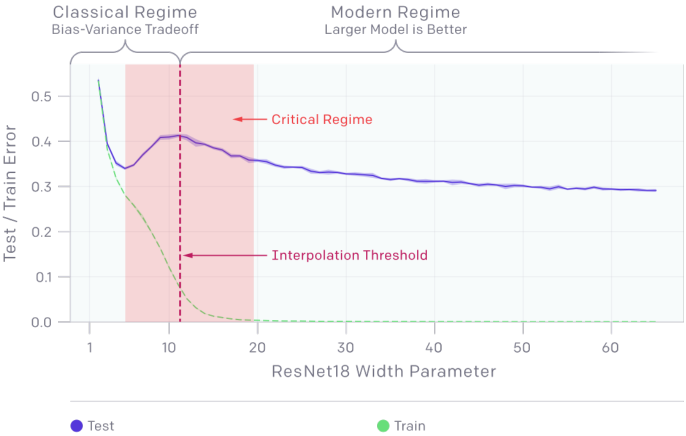
Maxim Kodryan

Samsung-HSE Laboratory
National Research University Higher School of Economics

November 27, 2020

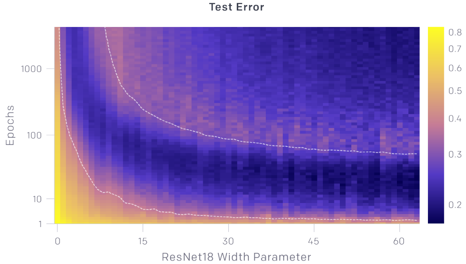
The Double Descent (DD) phenomenon [1]

Model-wise DD



A vast range of studies tackle the *model-wise* DD both empirically and theoretically [1–7]. But what about the *epoch-wise* DD?..

Epoch-wise DD [8, 9]



The "flat minima" intuition [10]

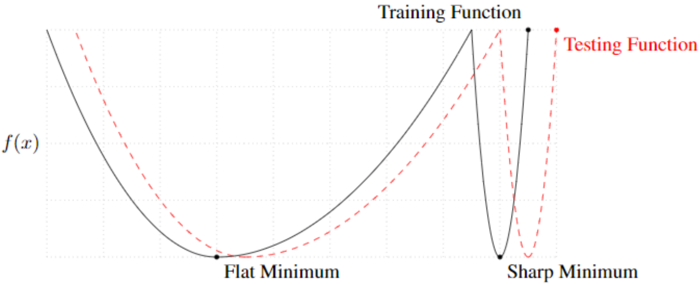
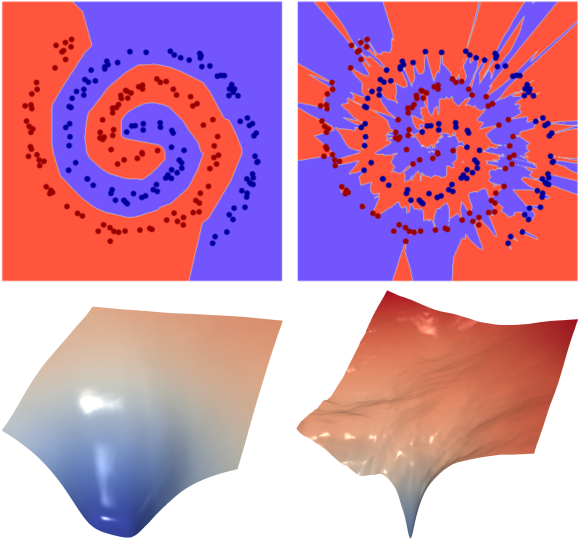


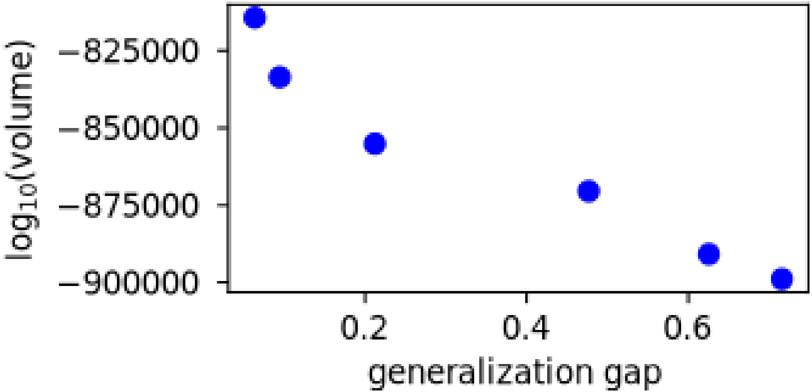
Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

There exist a whole bunch of "flatness" definitions (with critique) [10–18], but the intuition is simple: *the "wider" the minimum the better it generalizes*.

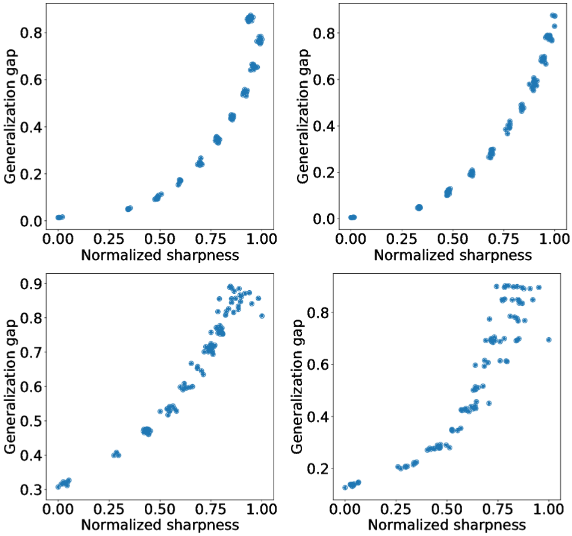
Flat minima visualization [14]



Volume of the minimum vs. generalization gap [14]



Normalized sharpness vs. generalization gap [16]



Fisher Information Matrix (FIM)

- Suppose we have a discriminative model $p_w(y \mid x)$ parameterized by w and a data distribution $Q(x)$
- The *Fisher Information Matrix (FIM)* is defined as

$$F \coloneqq \mathbb{E}_{x\sim Q(x)}\mathbb{E}_{y\sim p_w(y|x)}\left[\nabla_w \log p_w(y \mid x)\nabla_w \log p_w(y \mid x)^T\right] = \\ = -\mathbb{E}_{x\sim Q(x)}\mathbb{E}_{y\sim p_w(y|x)}\left[\nabla_w^2 \log p_w(y \mid x)\right]$$

FIM properties

- FIM is positive semidefinite: $F \succeq 0$
- Let $w' = w + \delta w$, then

$$\mathbb{E}_{x \sim Q(x)} \operatorname{KL}\left(p_{w'}(y \mid x) \| p_w(y \mid x)\right) = \delta w^T F \delta w + o\left(\delta w^2\right)$$

- FIM is a semidefinite approximation of the loss Hessian [19]
- FIM trace is easy to estimate and measures the *average model robustness to small parameters perturbations* [20]:

$$\operatorname{tr}(F) = \mathbb{E}_{x \sim Q(x)} \mathbb{E}_{y \sim p_w(y \mid x)} \left[\left\|\nabla_w \log p_w(y \mid x)\right\|^2\right]$$

FIM, loss Hessian, and gradient noise [21]

- $\mathbf{C}$ — (uncentered) covariance matrix of the gradients
- $\mathbf{H}$ — Hessian of the loss
- $\mathbf{F}$ — FIM

$$\mathbf{C} \propto \mathbf{F} \approx \mathbf{H}$$

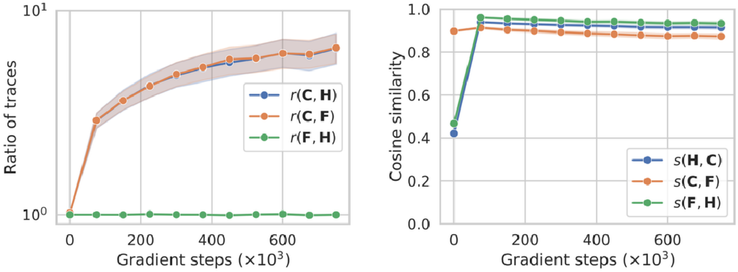
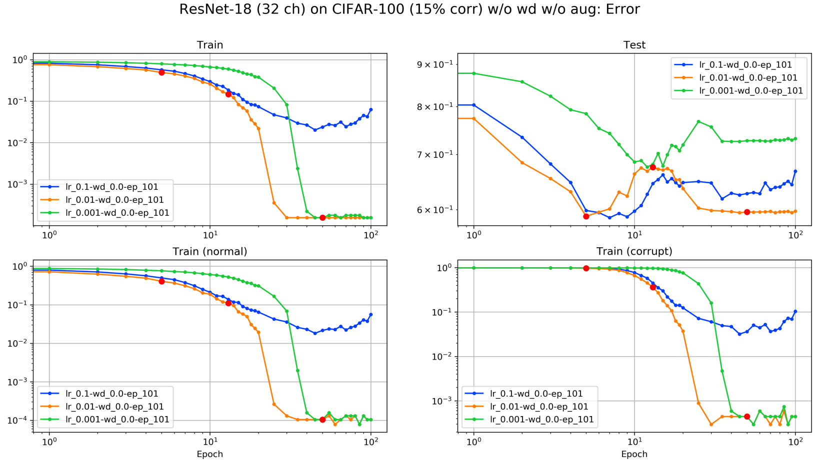


Figure 3: Scale and angle similarities between information matrices.

FIM is a good proxy of both loss curvature and gradient noise.

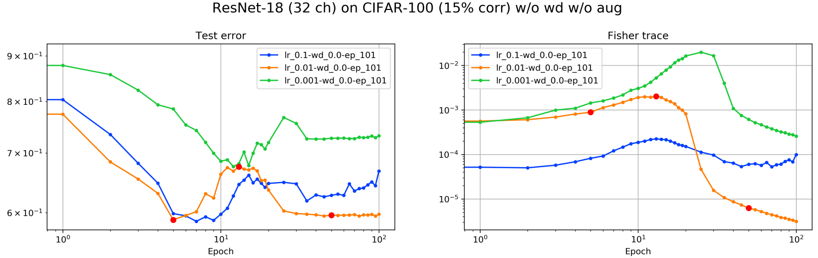
Epoch-wise DD and generalization vs. memorization



Epoch-wise DD = generalization + memorization + consolidation

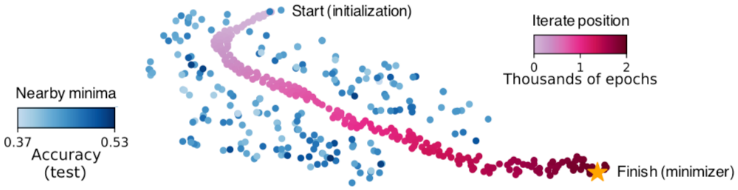
1. At first, model learns simple useful features and *generalizes* on normal examples [22–24] — test error decreases. This can be partially explained by *clustering of gradients* [25, 26].
2. Then it starts *memorizing* noise examples [22, 27] — test error increases.
3. Finally, network *consolidates* [9, 20]: removes redundancy, enters flat regions, improves generalization — test error decreases again.

Epoch-wise DD and FIM



FIM sheds light on model dynamics after the test error peak: *the model enjoys the second test risk descent exactly when it traverses from the firstly found sharp unstable regions to flat well-generalizing minima.*

Minefields in loss landscape [14]



It seems that *most* minima are *bad* [14, 28]!
What helps neural networks avoid them?

NNs avoid bad minima due to:

- Small volume of bad optima [14]
- Architecture tricks: surprisingly, it's mostly Batch Norm, not Skip Connections [29]
- *SGD noise* induced by small batch size [10, 30–33], large LR [28, 30–34], gradient covarience structure [30, 33, 35], implicit regularization [36, 37]...

Implicit Gradient Regularization (IGR) [36] sketch

- GD updates $\theta_{i+1} = \theta_i - h\nabla L(\theta_i)$ are the explicit Euler approximation of the following ODE: $\dot{\theta}(t) = -\nabla L(\theta(t))$
- Consider Taylor expansion of the exact solution: $\theta(h) = \theta_0 - h\nabla L(\theta_0) + \frac{h^2}{2}\nabla^2 L(\theta_0)\nabla L(\theta_0) + O(h^3)$
- Then one-step difference is $\|\theta_1 - \theta(h)\| = O(h^2)$
- Consider modified loss $\tilde{L}(\theta) = L(\theta) + \frac{h}{4}\|\nabla L(\theta)\|^2$
- Then one-step difference between GD and modified dynamics is $\left\|\theta_1 - \tilde{\theta}(h)\right\| = O(h^3)$, where $\dot{\tilde{\theta}}(t) = -\nabla \tilde{L}\left(\tilde{\theta}(t)\right)$
- This implies that modified loss $\tilde{L}$, which encourages the discovery of flatter optima, better mimics the regularization effect of discreteness of GD steps!

Implicit Stochastic Gradient Regularization (ISGR) [37]

- Generalization of IGR for the SGD case
- Let the loss be $L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta)$
- Then ISGR loss is

$$\tilde{L}_{SGD}(\theta) = L(\theta) + \frac{h}{4m}\sum_{k=0}^{m-1}\left\|\nabla \hat{L}_k(\theta)\right\|^2 =$$

$$= L(\theta) + \frac{h}{4}\left\|\nabla L(\theta)\right\|^2 + \frac{h}{4m}\sum_{k=0}^{m-1}\left\|\nabla \hat{L}_k(\theta) - \nabla L(\theta)\right\|^2,$$

  where m is #mini-batches, $\hat{L}_k$ is the k-th mini-batch loss
- This confirms that SGD selects not only wide, but also *uniform* optima, i.e., satisfying each mini-batch [38]!

Takeaways

- Epoch-wise DD is important and interesting, yet not well-studied phenomenon
- Another spectacular fact is the connection between optimum flatness and its ability to generalize
- Linking them together via loss geometry and information theory (e.g., FIM) can be a promising direction to put further our understanding of DNNs optimization and generalization
- The implicit noise of SGD explicitly helps neural networks to converge into wide and "uniform" optima

References I

[1] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[2] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[3] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[4] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. *arXiv preprint arXiv:2003.01054*, 2020.

References II

[5] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. *arXiv preprint arXiv:2002.08404*, 2020.

[6] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

[7] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. *arXiv preprint arXiv:2002.11328*, 2020.

[8] Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. *arXiv preprint arXiv:2007.10099*, 2020.

References III

[9] Xiao Zhang and Dongrui Wu. Rethink the connections among generalization, memorization and the spectral bias of dnns. *arXiv preprint arXiv:2004.13954*, 2020.

[10] 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. *arXiv preprint arXiv:1609.04836*, 2016.

[11] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[12] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

References IV

[13] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems*, pages 2553–2564, 2019.

[14] W Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. *arXiv preprint arXiv:1906.03291*, 2019.

[15] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 888–896. PMLR, 2019.

References V

[16] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. *arXiv preprint arXiv:1901.04653*, 2019.

[17] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. In *Advances in Neural Information Processing Systems*, pages 4949–4959, 2018.

[18] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

References VI

[20] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018.

[21] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020.

[22] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*, 2017.

References VII

[23] Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[24] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, pages 3496–3506, 2019.

[25] Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*, 2019.

References VIII

[26] Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

[27] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020.

[28] Nikhil Iyer, V Thejas, Nipun Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *arXiv preprint arXiv:2003.03977*, 2020.

[29] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. *arXiv preprint arXiv:1901.10159*, 2019.

References IX

[30] Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[31] Stanislaw Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018.

[32] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.

[33] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

References X

[34] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11674–11685, 2019.

[35] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. *arXiv preprint arXiv:1906.07405*, 2019.

[36] David GT Barrett and Benoit Dherin. Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*, 2020.

[37] Anonymous. On the origin of implicit regularization in stochastic gradient descent. In *Submitted to International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=rq_Qr0c1Hyo`. under review.

References XI

[38] Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018.