Three Factors Influencing Minima in SGD

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, Amos Storkey

HEP-AI Journal Club, January 2018

The generalization puzzle

- Deep models are highly over-parameterized
  - ImageNet: 10^7 training samples vs. 10^8 parameters (VGG)
- Yet they often generalize well (do not completely overfit). Why?
  - Gradient descent?
  - Noise in gradient estimation of SGD?
  - Good priors built into the architecture? (‘inductive bias’)
  - Smaller-than-expected capacity?

Deep models have large capacity

- Zhang et al. (1611.03530) showed that typical models can typically memorize all training samples
    - Training labels are randomized
    - Deep models can still memorize the training samples (reach 100% training accuracy)
    - (But not generalize)
- Classical measures of capacity (Rademacher complexity) are probably not useful for explaining generalization

Generalization and flatness

- There is evidence that flat minima generalize better than sharp minima. Possible intuition:

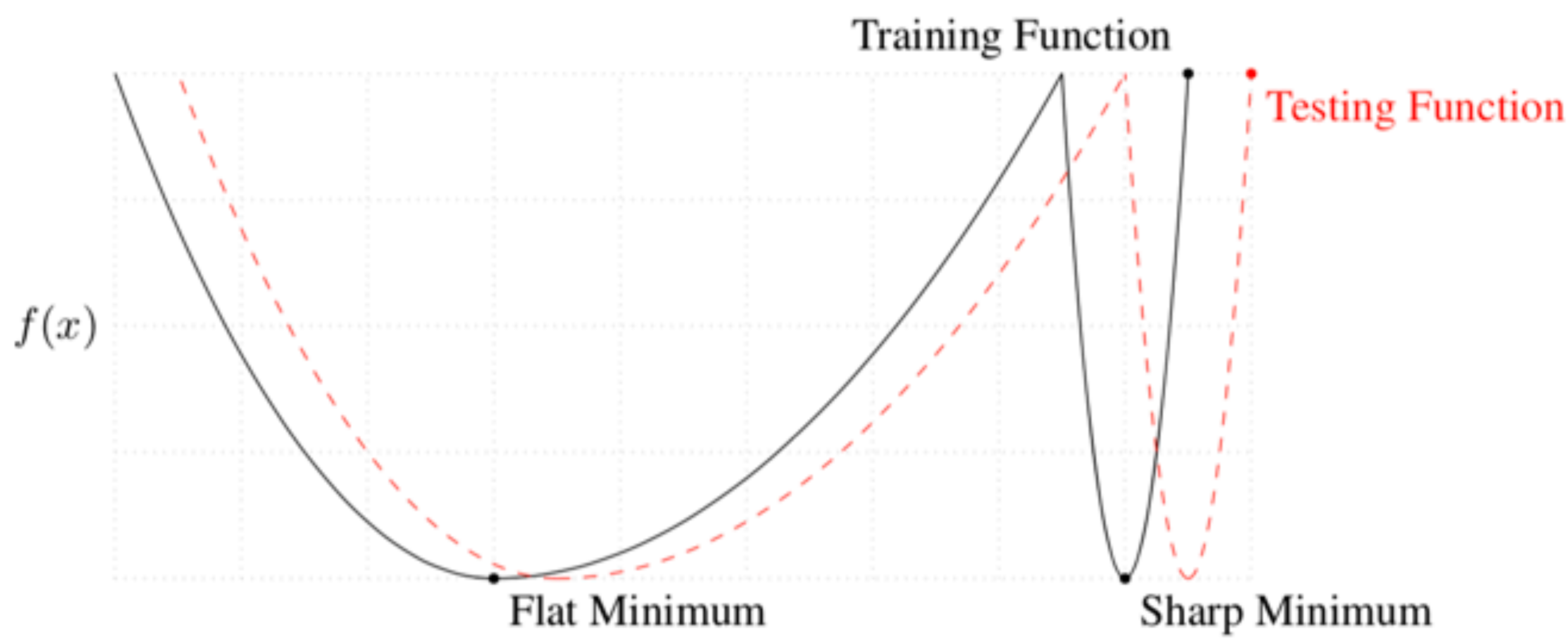


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

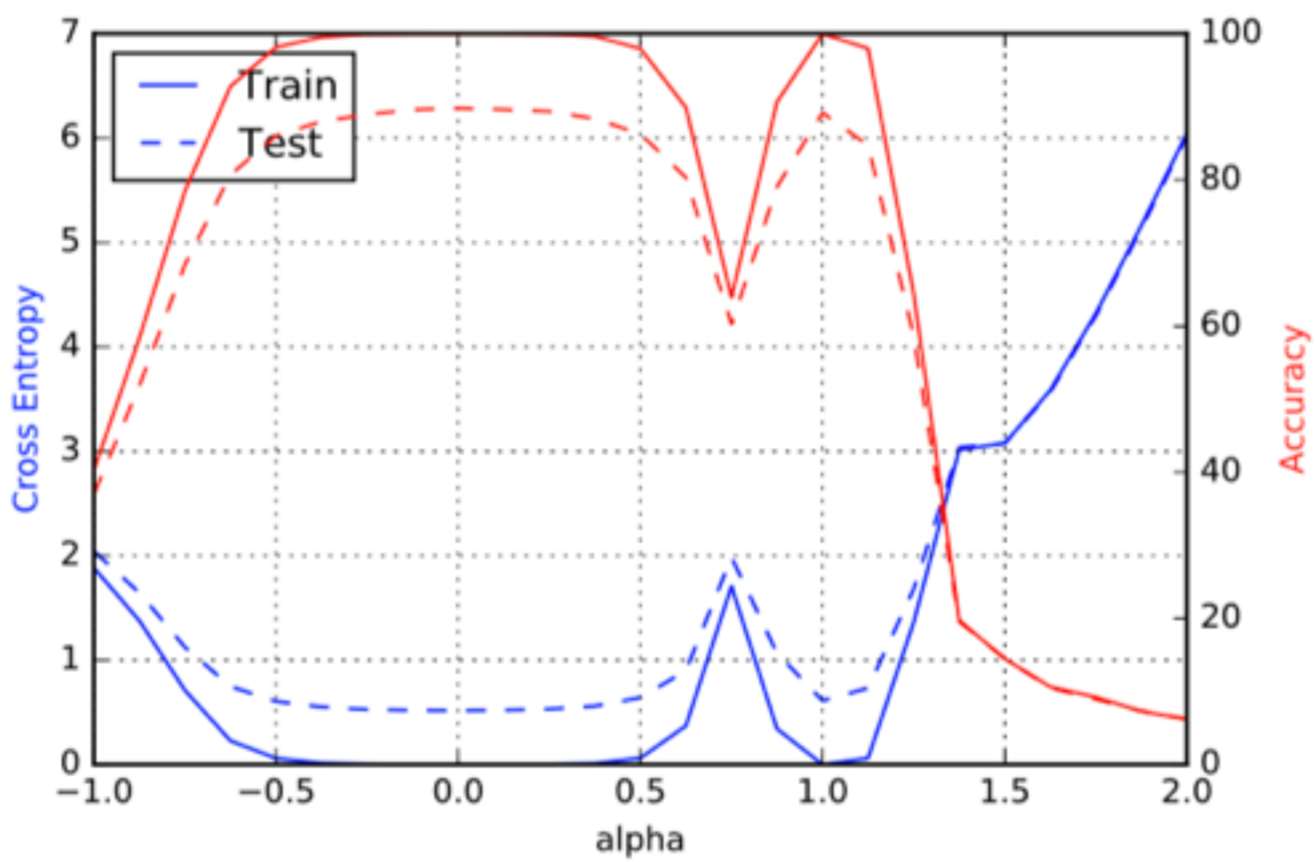
[1609.0836]

Stochastic gradient descent and flatness

- In SGD we estimate the gradient by sampling mini-batches from training set
- Introduces noise (variance) into the gradient
- Noisy gradients favor flat minima
- So it seems that noisy SGD improves generalization

Batch size and flatness

- Small batch <-> noisy gradient
- Interpolate loss between small/large-batch minima



(d) C_2

$$L(\alpha) = L(\alpha\theta_{\text{large-batch}} + (1-\alpha)\theta_{\text{small-batch}})$$

[1609.0836]

Three Factors Influencing Minima in SGD

- Take continuum limit of SGD
- Compute equilibrium distribution of learned weights
- Show that SGD favors deeper, wider minima
- Higher noise makes probabilities of deep / shallow minima closer

$$\frac{p_A}{p_B} = \sqrt{\frac{\det \mathbf{H}_B}{\det \mathbf{H}_A}} \exp\left(\frac{2}{n\sigma^2}\left(L_B - L_A\right)\right) \qquad n \equiv \frac{\eta}{S} = \frac{\textbf{learning rate}}{\textbf{batch size}}$$

Three factors: learning rate, batch size, gradient variance

Three Factors Influencing Minima in SGD

- Continuum limit of SGD gives Langevin equation

$$\frac{d\boldsymbol{\theta}}{dt} = -\eta \boldsymbol{g}(\boldsymbol{\theta}) + \frac{\eta}{\sqrt{S}} \mathbf{B}(\boldsymbol{\theta}) \mathbf{f}(t)$$

- Describes stochastic dynamics of a single training run

- Fokker-Planck equation describes evolution of distribution

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \left[\eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot [\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t)] \right].$$

- Equilibrium solution is Boltzmann distribution

$$P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2L(\boldsymbol{\theta})}{n\sigma^2} \right) \qquad n \equiv \frac{\eta}{S} = \frac{\textbf{learning rate}}{\textbf{batch size}}$$

SGD dynamics in continuum limit

$$L^{(S)}(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}} l(\boldsymbol{\theta}, \boldsymbol{x}_n) \ , \qquad \mathbf{g}^{(S)}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} L^{(S)}(\boldsymbol{\theta}) \ .$$

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \boldsymbol{g}^{(S)}(\boldsymbol{\theta}) \ .$$

Central limit theorem (large N, large batch size)

$$\mathbf{g}^{(S)}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) + \frac{1}{\sqrt{S}} \Delta \mathbf{g}(\boldsymbol{\theta}), \text{ where } \Delta \mathbf{g}(\boldsymbol{\theta}) \sim N(0, \mathbf{C}(\boldsymbol{\theta})) \ . \qquad \mathbf{C}(\boldsymbol{\theta}) = \mathbf{B}(\boldsymbol{\theta}) \mathbf{B}^\top(\boldsymbol{\theta})$$

Continuum limit is Langevin equation

$$\frac{d\boldsymbol{\theta}}{dt} = -\eta \boldsymbol{g}(\boldsymbol{\theta}) + \frac{\eta}{\sqrt{S}} \mathbf{B}(\boldsymbol{\theta}) \mathbf{f}(t)$$

learning rate should be different though…

Noise term correlations

$$\langle f(t) \rangle = 0 \qquad \langle f(t) f(t') \rangle = \delta(t - t')$$

Fokker-Planck equation

Continuum limit is Langevin equation

$$\frac{d\boldsymbol{\theta}}{dt} = -\eta \boldsymbol{g}(\boldsymbol{\theta}) + \frac{\eta}{\sqrt{S}}\mathbf{B}(\boldsymbol{\theta})\mathbf{f}(t) \qquad \mathbf{C}(\boldsymbol{\theta}) = \mathbf{B}(\boldsymbol{\theta})\mathbf{B}^\top(\boldsymbol{\theta})$$

Noise term correlations

$$\langle f(t) \rangle = 0 \qquad \langle f(t) f(t') \rangle = \delta(t - t')$$

$$\frac{\partial P(\boldsymbol{\theta}, t)}{\partial t} = \nabla_{\boldsymbol{\theta}} \cdot \left[\eta \boldsymbol{g}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t) + \frac{\eta^2}{2S} \nabla_{\boldsymbol{\theta}} \cdot [\mathbf{C}(\boldsymbol{\theta}) P(\boldsymbol{\theta}, t)] \right].$$

Fokker-Planck equation (1d)

$$P(\theta, t_2) = \int \mathcal{D}f_{t_1 \to t_2}\, p(f) \int d\theta_0 P(\theta_0, t_1)\, \delta\left(\theta - \theta(t_2; \theta(t_1), f)\right)$$

$$p(f) \sim \exp\left(-\frac{1}{2\sigma^2}\int dt f(t)^2\right)$$

Langevin

$$P(\theta, t+dt) = \int \mathcal{D}f_{t \to t+dt}\, p(f) \int d\theta_0 P(\theta_0, t)\, \delta\left(\theta - \theta_0 + g(\theta)dt - \int_t^{t+dt} f(t')dt'\right)$$

Expand the delta function, use noise correlations $\langle f(t) f(t') \rangle = \sigma^2 \delta(t - t')$

$$\frac{\partial}{\partial t} P(\theta, t) = \frac{\partial}{\partial \theta}\left(P(\theta, t) g(\theta)\right) + \frac{1}{2}\frac{\partial^2}{\partial \theta^2}\left(\sigma^2 P(\theta, t)\right)$$

Drift term **Diffusion term**

Equilibrium distribution

$$\partial_t P + \nabla \cdot J = 0\,, \quad -J = \eta g P + \frac{\eta^2}{2S} \nabla \cdot (CP)$$

Stationary solutions

$$\partial_t P = 0$$

Assume constant isotropic noise: $\mathbf{C}(\boldsymbol{\theta}) = \sigma^2 \mathbf{I}$.

Stationary solutions only depend on $\dfrac{\sigma^2 \eta}{S}$

Equilibrium solutions (detailed balance): $J = 0$

Equilibrium distribution

Equilibrium solution is Boltzmann

$$J = 0 \quad \Rightarrow \quad P(\boldsymbol{\theta}) = P_0 \exp\left(-\frac{2L(\boldsymbol{\theta})}{n\sigma^2}\right)$$

noise coefficient: $n \equiv \frac{\eta}{S}$ = $\frac{\textbf{learning rate}}{\textbf{batch size}}$

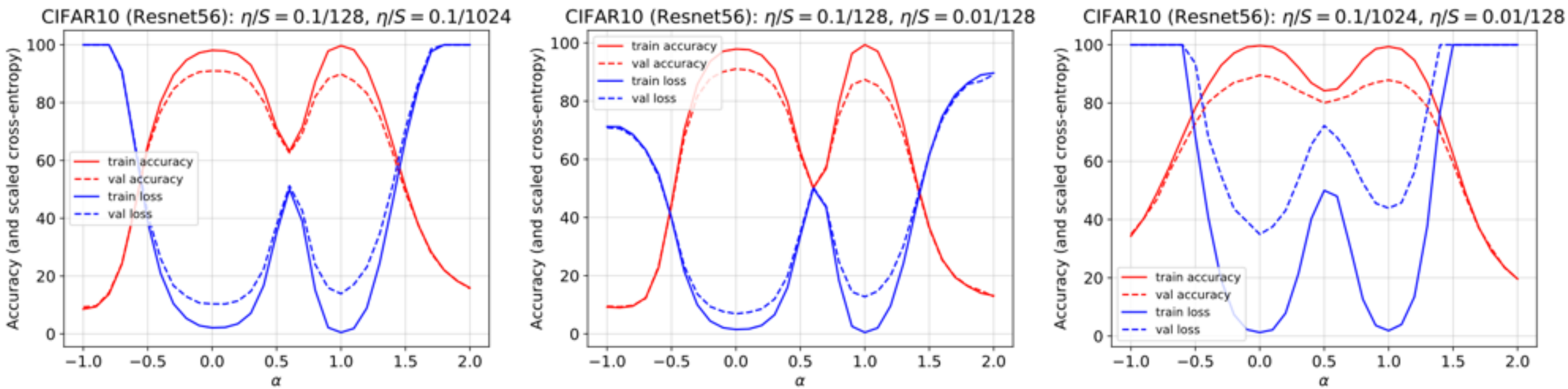
Flat vs. sharp minima

$$L(\boldsymbol{\theta}) \approx L_A + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_A)^\top \mathbf{H}_A(\boldsymbol{\theta} - \boldsymbol{\theta}_A).$$

$$\begin{aligned}
p_A &\approx P_0 \int_{R_A} \exp\left(-\frac{2S}{\eta\sigma^2} L(\boldsymbol{\theta})\right) \\
&\approx P_0 \int_{R_A} \exp\left(-\frac{2S}{\eta\sigma^2}\left[L_A + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_A)^\top \mathbf{H}_A(\boldsymbol{\theta} - \boldsymbol{\theta}_A)\right]\right) \\
&\approx \tilde{P}_0 \exp\left(-\frac{2SL_A}{\eta\sigma^2}\right)\sqrt{\frac{1}{\det \mathbf{H}_A}}
\end{aligned}$$

$$\frac{p_A}{p_B} = \sqrt{\frac{\det \mathbf{H}_B}{\det \mathbf{H}_A}} \exp\left(\frac{2}{n\sigma^2}\left(L_B - L_A\right)\right)$$

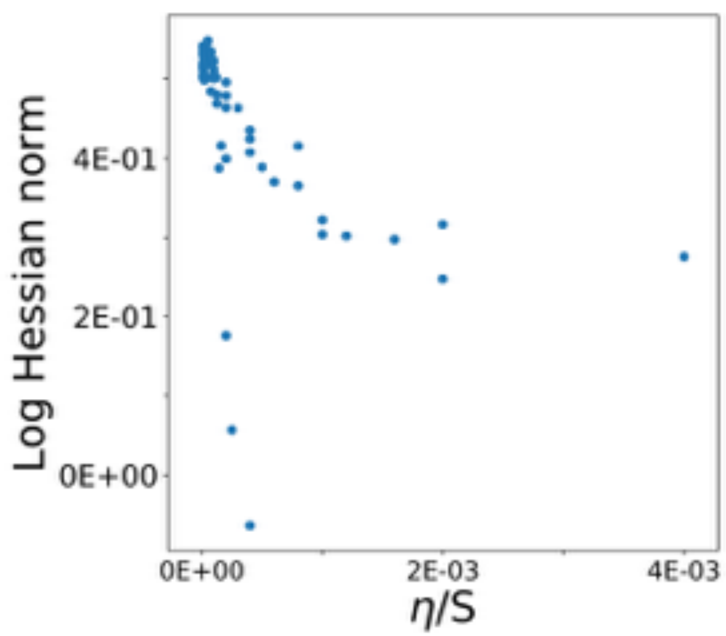
Experimental evidence



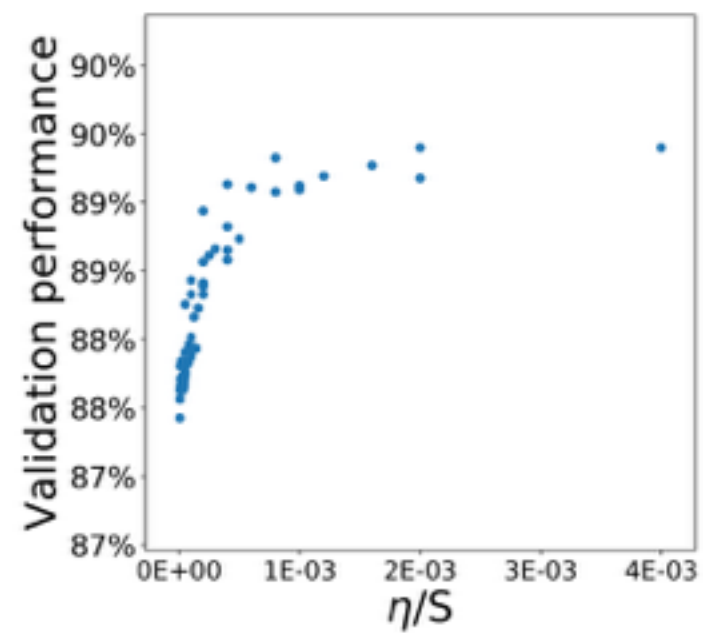
(a) left $\frac{\eta=0.1}{S=128}$, right $\frac{\eta=0.1}{S=1024}$ (b) left $\frac{\eta=0.1}{S=128}$, right $\frac{\eta=0.01}{S=128}$ (c) left $\frac{\eta=0.1}{S=1024}$, right $\frac{\eta=0.01}{S=128}$

$$\tilde{p}_A = \frac{1}{\sqrt{\det \mathbf{H}_A}} \exp\left(-\frac{2L_A}{n\sigma^2}\right) \qquad n = \frac{\eta}{S}$$

Experimental evidence



(a) Correlation of $\frac{\eta}{S}$ with logarithm of norm of Hessian.



(b) Correlation of $\frac{\eta}{S}$ with validation accuracy.

Each experiment is run for 200 epochs; most models reach approximately 100% accuracy on train set. As n grows, we observe that the norm of the Hessian at the minima also decreases, suggesting that higher $\frac{\eta}{S}$ pushes the optimization towards flatter minima. This agrees with Theorem 2, Eq. (3), that higher $\frac{\eta}{S}$ favors flatter over sharper minima.

$$\tilde{p}_A = \frac{1}{\sqrt{\det \mathbf{H}_A}} \exp\left(-\frac{2L_A}{n\sigma^2}\right) \qquad n = \frac{\eta}{S}$$

Are they assuming sharper minima have smaller loss? Why?

Questions

- Momentum? (they comment on it but don’t have conclusions)
- Natural gradient?
- More realistic covariance matrices for the noise?
- Why equilibrium and not just stationary solutions?