# A Dynamical Stability Perspective for Understanding Neural Network Training

**Lei Wu**
Center of Machine Learning Research
School of Mathematical Sciences, PKU

February 14, 2025



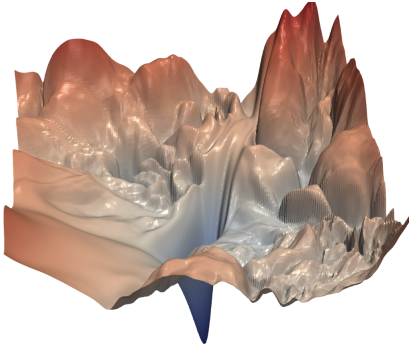

Joint work with Weinan E (PKU), Chao Ma (Stanford), Weijie Su (UPenn), Mingze Wang (PKU).

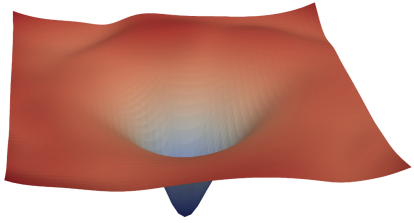
# Outline


1. Motivation
2. Inspiring phenomena in NN training
3. The connection between stability and generalization
   - Why does SGD tend to find flat minima?
   - Why do flat minima generalize well?
4. The connection between stability and optimization

# Understanding Neural Network Training is Challenging

- The landscape is high-dimensional, non-convex, degenerate, and consists of enormous local minima and saddles.
- However, in practice, the training **algorithm often explores only a small, well-behaved region of the landscape**.





- Understanding this inductive bias of training algorithm is important.
  - **Generalization**.
  - **Optimization**.

# Phenomenon 1: SGD generalizes unreasonably well

- SGD can converge to solutions that generalize well without needing any explicit regularizations.

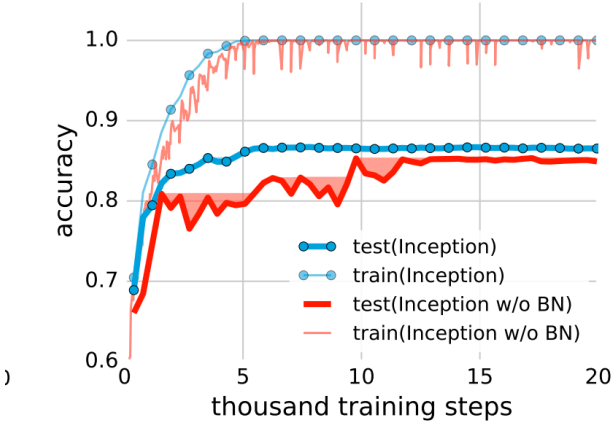



Figure 1: Classifying CIFAR-10 with Inception networks (Taken from [Chiyuan Zhang, et al, ICLR2017])

# Phenomenon 2: SGD generalizes better than GD

| Experiment | Mini-batching | Epochs | Steps | Modifications | Val. Accuracy % |
|---|---|---|---|---|---|
| Baseline SGD | ✓ | 300 | 117,000 | - | 95.70(±0.11) |
| Baseline FB | ✗ | 300 | 300 | - | 75.42(±0.13) |
| FB train longer | ✗ | 3000 | 3000 | - | 87.36(±1.23) |
| FB clipped | ✗ | 3000 | 3000 | clip | 93.85(±0.10) |
| FB regularized | ✗ | 3000 | 3000 | clip+reg | 95.36(±0.07) |
| FB strong reg. | ✗ | 3000 | 3000 | clip+reg+bs32 | 95.67(±0.08) |
| FB in practice | ✗ | 3000 | 3000 | clip+reg+bs32+shuffle | 95.91(±0.14) |

Table 2: Summary of validation accuracies in percent on the CIFAR-10 validation dataset for each of the experiments with data augmentations considered in Section 3. All validation accuracies are averaged over 5 runs.

Figure 2: Taken from (Geiping et al., 2021)

- Full-batch GD already generalizes well.
- SGD generalizes much better than GD, implying noise is crucial for generalization.

$$\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \xi_t).$$

# Phenomenon 3: Flat minima hypothesis

- Flat minima generalize better (Hochreiter and Schmidhuber, 1995);
- SGD finds flatter minima than GD (Keskar et al., 2016).

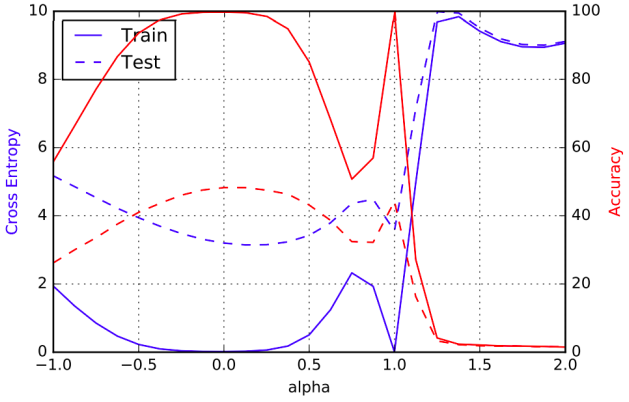



Figure 3: The landscape along $\{\theta(\alpha) := (1-\alpha)\theta_{\text{sgd}} + \alpha\theta_{\text{gd}} : \alpha \in [-1, 2]\}$ (taken from Keskar et al. (2016)).

# Phenomena 4: Stability can distinguish SGD with GD

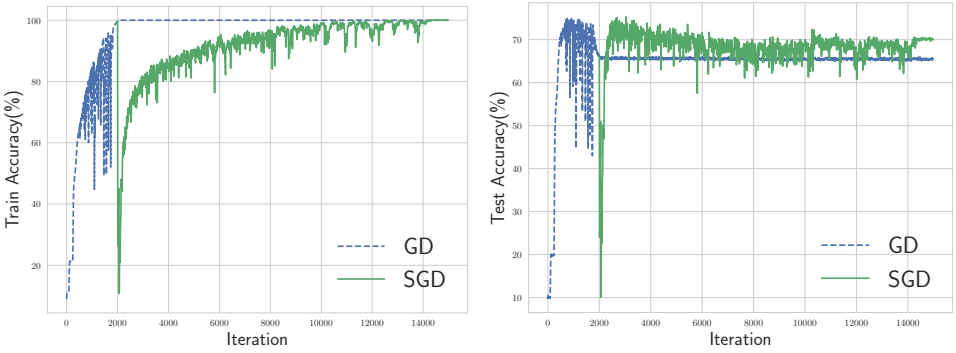



Figure 4: Fast escape phenomenon in fitting corrupted FashionMNIST (**Wu**-Ma-E, NeurIPS 2018)

**Observation:**

- This escape phenomenon indicates that GD solutions are **dynamically unstable** for SGD.

# Phenomena 4: Stability can distinguish SGD with GD

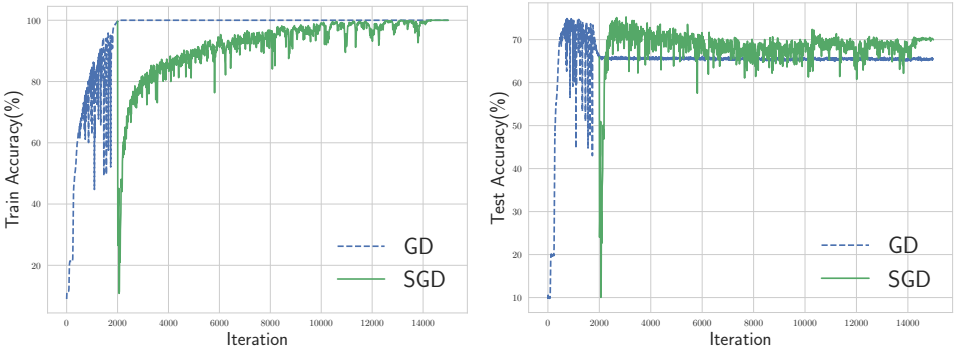



Figure 4: Fast escape phenomenon in fitting corrupted FashionMNIST (**Wu**-Ma-E, NeurIPS 2018)

**Observation:**

- This escape phenomenon indicates that GD solutions are **dynamically unstable** for SGD.
- The escape is *unreasonably* fast.

# Why does SGD tend to find flat minima?

# The stability of gradient descent (GD)

GD updates as $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$, for which all $A, B, C, D$ are fixed points.

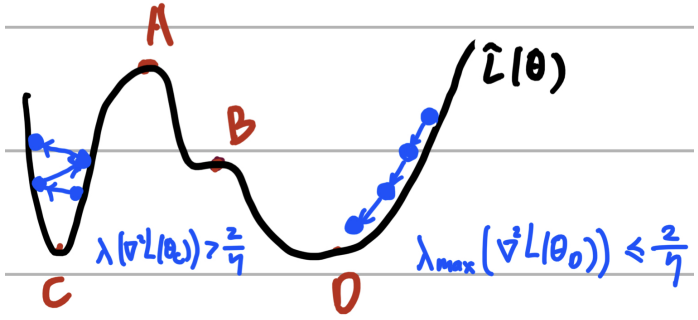



Figure 5: GD with a large LR only selects $D$. The minimum C is stable for GF but not for GD with a relatively large LR.

- When $\eta \approx 0$, GD becomes $\dot{\theta}_t = -\nabla L(\theta_t)$. Then, A,B are unstable.
- Increasing $\eta$ makes C unstable.
- Further increasing $\eta$, all fixed points are unstable.

# Linear stability analysis

- **Local Linearization of GD:** Let $H(\theta) = \nabla^2 L(\theta)$. Then, linearising GD around $\theta^*$ gives

$$\begin{aligned}\theta_{t+1} - \theta^* &= \theta_t - \theta^* - \eta(\nabla L(\theta_t) - \nabla L(\theta^*)) \\ &= (I - \eta H(\theta^*))^t(\theta_0 - \theta^*).\end{aligned}$$

# Linear stability analysis

- **Local Linearization of GD:** Let $H(\theta) = \nabla^2 L(\theta)$. Then, linearising GD around $\theta^*$ gives

$$\begin{aligned}\theta_{t+1} - \theta^* &= \theta_t - \theta^* - \eta(\nabla L(\theta_t) - \nabla L(\theta^*)) \\ &= (I - \eta H(\theta^*))^t(\theta_0 - \theta^*).\end{aligned}$$

- **Stability condition:** The minima that is linearly stable for GD must satisfy

$$\eta \leqslant \frac{2}{\lambda_{\max}(H(\theta^*))} \quad \implies \quad \underbrace{\lambda_{\max}(H(\theta^*))}_{\text{sharpness}} \leqslant \frac{2}{\eta}.$$

# Linear stability analysis

- **Local Linearization of GD:** Let $H(\theta) = \nabla^2 L(\theta)$. Then, linearising GD around $\theta^*$ gives

$$\begin{aligned}\theta_{t+1} - \theta^* &= \theta_t - \theta^* - \eta(\nabla L(\theta_t) - \nabla L(\theta^*)) \\ &= (I - \eta H(\theta^*))^t(\theta_0 - \theta^*).\end{aligned}$$

- **Stability condition:** The minima that is linearly stable for GD must satisfy

$$\eta \leqslant \frac{2}{\lambda_{\max}(H(\theta^*))} \quad \implies \quad \underbrace{\lambda_{\max}(H(\theta^*))}_{\text{sharpness}} \leqslant \frac{2}{\eta}.$$

# Linear stability analysis

- **Local Linearization of GD:** Let $H(\theta) = \nabla^2 L(\theta)$. Then, linearising GD around $\theta^*$ gives

$$\begin{aligned}\theta_{t+1} - \theta^* &= \theta_t - \theta^* - \eta(\nabla L(\theta_t) - \nabla L(\theta^*)) \\ &= (I - \eta H(\theta^*))^t(\theta_0 - \theta^*).\end{aligned}$$

- **Stability condition:** The minima that is linearly stable for GD must satisfy

$$\eta \leqslant \frac{2}{\lambda_{\max}(H(\theta^*))} \quad \Longrightarrow \quad \underbrace{\lambda_{\max}(H(\theta^*))}_{\text{sharpness}} \leqslant \frac{2}{\eta}.$$

**Implications:**

- **Stability controls the spectral norm of Hessian matrix of minima that GD selects**.
- GD escapes from that minimum exponentially fast:

$$|\theta_t - \theta^*| \geqslant |1 - \eta\lambda_1(H(\theta^*)|^t$$

# The edge of stability (EoS) phenomenon

In practice, GD often occurs on the **edge of stability** (EoS) in training neural nets.

Table 1: Sharpness of the solutions found by GD with different learning rates. Each experiment is repeated for $5$ times with independent random initialization.

| $\eta$ | 0.01 | 0.05 | 0.1 | 0.5 | 1 |
|---|---|---|---|---|---|
| FashionMNIST | $53.5 \pm 4.3$ | $39.3 \pm 0.5$ | $19.6 \pm 0.15$ | $3.9 \pm 0.0$ | $1.9 \pm 0.0$ |
| CIFAR10 | $198.9 \pm 0.6$ | $39.8 \pm 0.2$ | $19.8 \pm 0.1$ | $3.6 \pm 0.4$ | - |
| The edge of stability: $2/\eta$ | 200 | 40 | 20 | 4 | 2 |

See follow-up works (Jastrzebski et al., ICLR 2020; Cohen et al., ICLR 2021) on this striking phenomenon.

# Linear stability of SGD

- The local dynamical behavior of SGD around $\theta^*$ can be characterized by linearizing the dynamics [1]:

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \frac{\eta}{B} \sum_{i \in I_t} \nabla^2 L_i(\theta^*)(\tilde{\theta}_t - \theta^*), \tag{1}$$

[1] Assuming the validity of local quadratic approximation for all $L_i$.

# Linear stability of SGD

- The local dynamical behavior of SGD around $\theta^*$ can be characterized by linearizing the dynamics [1]:

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \frac{\eta}{B} \sum_{i \in I_t} \nabla^2 L_i(\theta^*)(\tilde{\theta}_t - \theta^*), \tag{1}$$

- We can rewrite the SGD as follows

$$\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \xi_t),$$

where $(\xi_t)_{t \geqslant 1}$ are any noises satisfying $\mathbb{E}[\xi_t] = 0, \mathbb{E}[\xi_t \xi_t^T] = \frac{1}{B} \Sigma(\theta_t)$, where

$$\mathbb{E}[\xi_t \xi_t^T] = \Sigma(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla L_i(\theta) \nabla L_i(\theta)^T - \nabla L(\theta) L(\theta)^T$$

[1] Assuming the validity of local quadratic approximation for all $L_i$.

# Linear stability of SGD

- The local dynamical behavior of SGD around $\theta^*$ can be characterized by linearizing the dynamics [1]:

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \frac{\eta}{B}\sum_{i\in I_t}\nabla^2 L_i(\theta^*)(\tilde{\theta}_t - \theta^*), \tag{1}$$

- We can rewrite the SGD as follows

$$\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \xi_t),$$

where $(\xi_t)_{t\geqslant 1}$ are any noises satisfying $\mathbb{E}[\xi_t] = 0, \mathbb{E}[\xi_t\xi_t^T] = \frac{1}{B}\Sigma(\theta_t)$, where

$$\mathbb{E}[\xi_t\xi_t^T] = \Sigma(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla L_i(\theta)\nabla L_i(\theta)^T - \nabla L(\theta)L(\theta)^T$$

- We shall use $H = \nabla^2 L(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}\nabla L_i(\theta^*)$.

[1] Assuming the validity of local quadratic approximation for all $L_i$.

# Examing stability via one-step update

**Lemma 1**

*We have*

$$\mathbb{E}[L(\theta_{t+1})] = \mathbb{E}[\underbrace{r(\theta_t)L(\theta_t)}_{\textbf{GD}} + \underbrace{\frac{\eta^2}{2B}\, v(\theta_t)}_{\textbf{Noise}}], \tag{2}$$

# Examing stability via one-step update

**Lemma 1**

*We have*

$$\mathbb{E}[L(\theta_{t+1})] = \mathbb{E}[\underbrace{r(\theta_t)L(\theta_t)}_{\textbf{GD}} + \underbrace{\frac{\eta^2}{2B}\, v(\theta_t)}_{\textbf{Noise}}], \tag{2}$$

*where*

$$\nu(\theta) = \mathit{Tr}(H\Sigma(\theta)).$$

- SGD is stable only if $\mathbb{E}[L(\theta_{t+1})] \leqslant \mathbb{E}[L(\theta_t)]$.
- The **alignment** between noise covariance and local Hessian is crucial.

# Hessian and noise structure

- Consider the regression with square loss: $L_i(\theta) = \frac{1}{2}(f_\theta(x_i) - y_i)^2$.
- To distinguish different global minima, we define

$$H(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_i;\theta)\nabla f(x_i;\theta)^T + \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\theta) - y_i)\nabla^2 f(x_i;\theta)$$

$$G(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_i;\theta)\nabla f(x_i;\theta)^T,$$

when $\theta^*$ is a global minimum

$$G(\theta^*) = H(\theta^*).$$

# The alignment property of SGD noise: An intuitive derivation

- Recall

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(f(x_i;\theta) - y_i)^2 = \frac{1}{2n}\sum_{i=1}^{n} e_i^2.$$

- Decoupling approximation:

$$\begin{aligned}
\Sigma(\theta) &= \frac{1}{n}\sum_i e_i \nabla f(x_i;\theta) e_i \nabla f(x_i;\theta)^\top - \nabla L(\theta)\nabla L(\theta)^\top \\
&\approx \frac{1}{n}\sum_i e_i^2 \nabla f(x_i;\theta)\nabla f(x_i;\theta)^\top \\
&\approx \left(\frac{1}{n}\sum_i e_i^2\right)\frac{1}{n}\sum_i \nabla f(x_i;\theta)\nabla f(x_i;\theta)^\top \\
&\approx 2L(\theta)G(\theta), \qquad (3)
\end{aligned}$$

# The alignment property of SGD noise: An intuitive approximation

The decoupling approximation is crude but informative

$$\Sigma(\theta) \approx 2 \underbrace{L(\theta)}_{\text{Magnitude}} \underbrace{G(\theta)}_{\text{Direction}} .$$

We observe that

- **Magnitude**: The noise magnitude is proportional to the loss value.

# The alignment property of SGD noise: An intuitive approximation

The decoupling approximation is crude but informative

$$\Sigma(\theta) \approx 2 \underbrace{L(\theta)}_{\text{Magnitude}} \underbrace{G(\theta)}_{\text{Direction}} .$$

We observe that

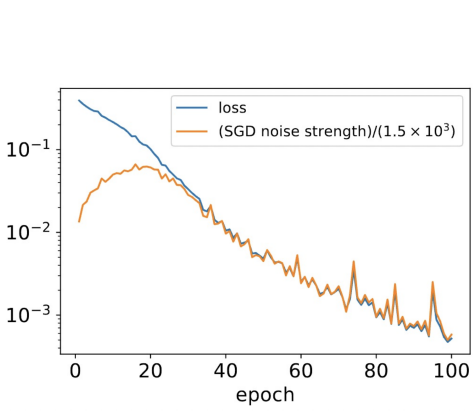
- **Magnitude**: The noise magnitude is proportional to the loss value.
- **Directions**: that $\Sigma(\theta)$ aligns with $G(\theta)$ suggests

  *noise concentrates on sharp directions of local landscape*.

  Specifically, the noise magnitude along $v$=loss $\times$ curvature.

$$\underbrace{\mathbb{E}[(v^T\xi)^2]}_{\text{noise energy}} = v^\top \Sigma(\theta) v = 2 \underbrace{L(\theta)}_{\text{loss value}} \underbrace{v^\top H(\theta) v}_{\text{curvature}} .$$

# Some experiments



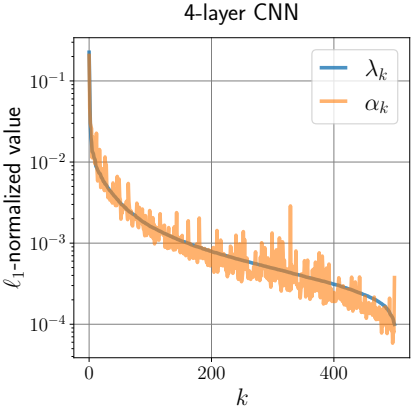




Figure 6: Fitting CIFAR-10 with a small CNN. **(Left)** The total noise energy is proportional to the loss value (taken from Mori et al. (2021)); **(Right)** The component of noise energy along one direction is proportional to the curvature.

# Towards a quantitative measurement of alignment

We define

$$\mu(\theta) = \frac{\nu(\theta)}{\bar{\nu}(\theta)} = \frac{\mathrm{Tr}(\Sigma(\theta)G(\theta))}{2L(\theta)\|G(\theta)\|_F^2}.$$

Where $\bar{\nu}(\theta)$ denotes the noise contribution when the decoupled approximation.

**Key Observation (Wu-Wang-Su, NeurIPS 2022; Wang-Wu, arXiv:2310.00692)**

There exist $C_1$ and $C_2$ that are independent of model size such that

$$C_1 \leqslant \mu(\theta) \leqslant C_2.$$

- This alignment can be rigorously proved for linear nets, random feature models, and two-layer nonlinear nets.
- It can be also validated on practical model training.

# Experiments on small nets

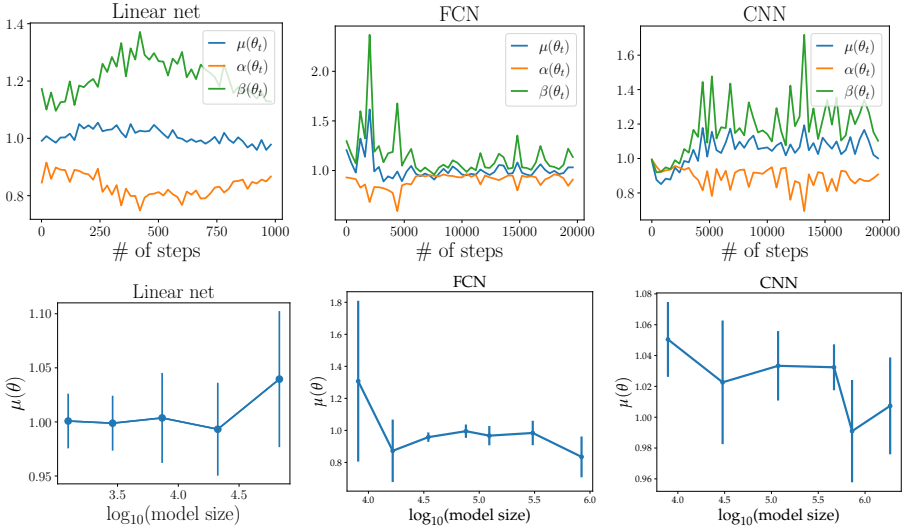



Figure 7: (Up) The strength of various alignments during the SGD training. (Bottom) How the alignment strength changes with increasing degrees of over-parameterization.

# Larger nets

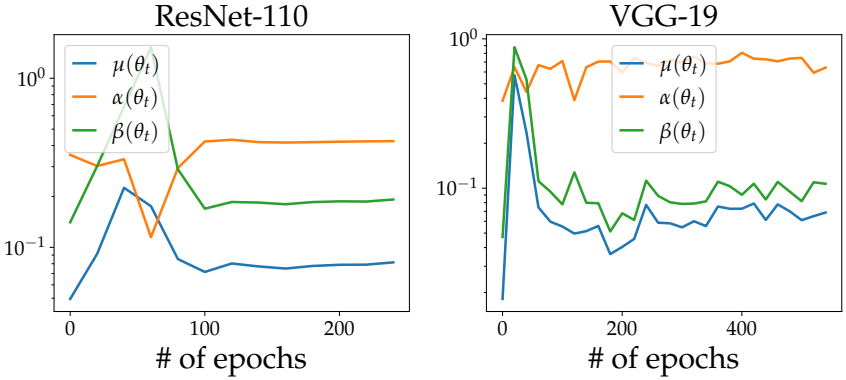



Figure 8: Fitting CIFAR-10.

# A stability condition via noise alignment

Assuming $\mu(\theta) \geqslant \mu_0$, we have

$$\mathbb{E}[L(\theta_{t+1})] = \mathbb{E}[r(\theta_t)L(\theta_t) + \frac{\eta^2}{2B}\nu(\theta_t)] \geqslant \frac{\eta^2}{2B}\mu(\theta)\bar{\nu}(\theta)$$
$$\geqslant \frac{\mu_0\eta^2\|H\|_F^2}{B}\mathbb{E}[L(\theta_t)]$$

# A stability condition via noise alignment

Assuming $\mu(\theta) \geqslant \mu_0$, we have

$$\mathbb{E}[L(\theta_{t+1})] = \mathbb{E}[r(\theta_t)L(\theta_t) + \frac{\eta^2}{2B}\nu(\theta_t)] \geqslant \frac{\eta^2}{2B}\mu(\theta)\bar{\nu}(\theta)$$
$$\geqslant \frac{\mu_0\eta^2\|H\|_F^2}{B}\mathbb{E}[L(\theta_t)]$$

**Theorem 2**

*Let $\theta^*$ be a global minimum that is stable. If $\mu(\theta) \geqslant \mu_0$, then*

$$\|H(\theta^*)\|_F \leqslant \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}.$$

# Implication: A size-independent sharpness control

$$\|H(\theta^*)\|_F \leqslant \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}.$$

- **The upper bound of sharpness is independent of the sample and parameter size, regardless the level of overparameterization.**

# Implication: A size-independent sharpness control

$$\|H(\theta^*)\|_F \leqslant \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}.$$

- **The upper bound of sharpness is independent of the sample and parameter size, regardless the level of overparameterization**.
- Different optimizers induce different sharpness control:

$$\underbrace{\sum_{j=1}^{m}\lambda_j^2(H(\theta^*))}_{\text{SGD}} \quad \text{vs} \quad \underbrace{\lambda_1(H(\theta^*))}_{\text{GD}}.$$

# Tightness of our bounds for practical models

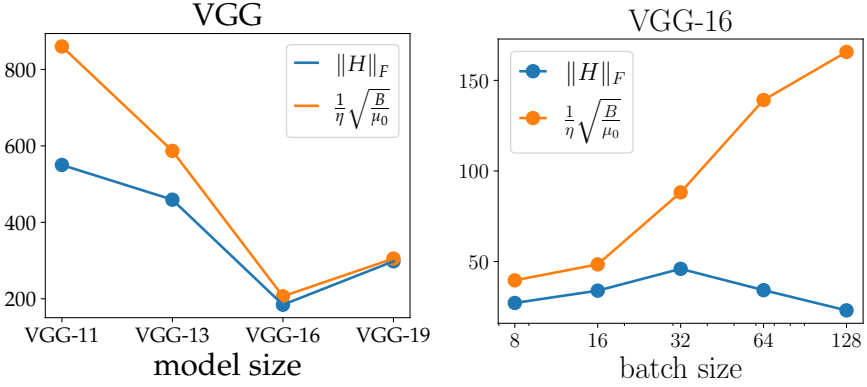



Figure 9: (Left) The flatness and upper bound for various VGG nets. (Right) How the tightness of upper bound changes with decreasing batch size.

# What happens if the stability condition is violated?

## Theorem 3 (Escape from sharp minima)

*If* $\|H(\theta^*)\|_F > \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}$, *then we have*

$$\mathbb{E}[\hat{L}(\theta_t)] \geqslant \gamma_0^t \mathbb{E}[\hat{L}(\theta_0)]$$

*where* $\gamma_0 = \frac{\eta^2 \mu_0}{B}\|H(\theta^*)\|_F^2 > 1$.

- The noise-driven escape is exponentially fast. This is due to that the **noise magnitude is proportional to the loss value**.
- The directional alignment ensures the rate is independent of $d$.

# What happens if the stability condition is violated?

## Theorem 3 (Escape from sharp minima)

*If* $\|H(\theta^*)\|_F > \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}$, *then we have*

$$\mathbb{E}[\hat{L}(\theta_t)] \geqslant \gamma_0^t \mathbb{E}[\hat{L}(\theta_0)]$$

*where* $\gamma_0 = \frac{\eta^2 \mu_0}{B}\|H(\theta^*)\|_F^2 > 1$.

- The noise-driven escape is exponentially fast. This is due to that the **noise magnitude is proportional to the loss value**.
- The directional alignment ensures the rate is independent of $d$.

If the noise is uniformly bounded, the escape is at most linear in time.

## Lemma 4

*Under the setting of Lemma 1, assume* $\eta \leqslant 2/\lambda_1(H)$ *and* $\mathbb{E}[H\Sigma(\theta)] \leqslant 2\sigma^2$. *Then* $\mathbb{E}[\hat{L}(\theta_t)] \leqslant \mathbb{E}[\hat{L}(\theta_0)] + \eta^2\sigma^2 t$.

# A zoom-in of the escape process

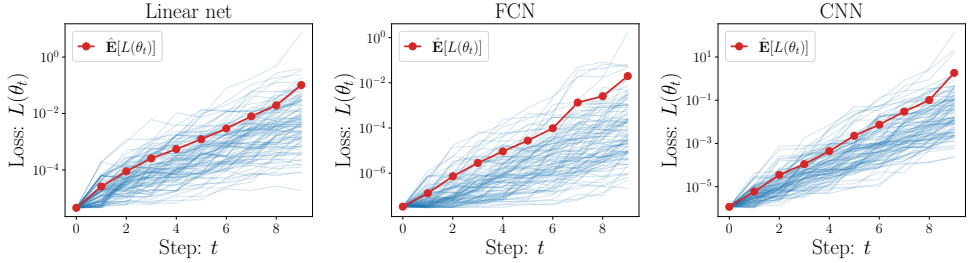



Figure 10: **The exponentially fast escape from sharp minima.** The blue curves are 200 trajectories of SGD; the red curve corresponds to the average. The sharp minimum is found by GD. When GD nearly converge, we switch to SGD with the same learning rate. This choice ensures that the minimum is stable for GD, and thus the escape is purely driven by SGD noise.

# Why do flat minima generalize well?

# Flatness vs generalization

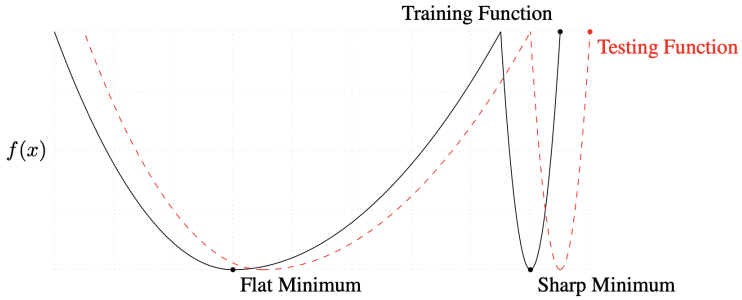



Figure 11: Taken from https://openreview.net/pdf?id=H1oyRlYgg. This illustration is very popular but **misleading**.

# Flatness vs generalization

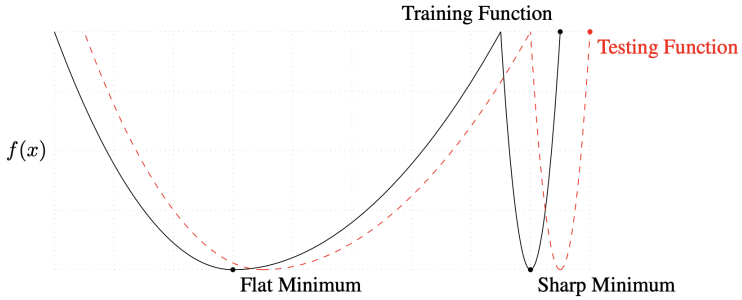



Figure 11: Taken from https://openreview.net/pdf?id=H1oyRlYgg. This illustration is very popular but **misleading**.

- Sharp minima can generalize well (Dinh et al., ICML 2017):

$$\sum_j a_j \mathrm{ReLU}(w_j^\top x) = \sum_j \frac{a_j}{\lambda} \mathrm{ReLU}(\lambda w_j^\top x) \text{ for any } \lambda > 0.$$

Thus, **flatness can be only sufficient condition** at the most.

# Flatness vs generalization

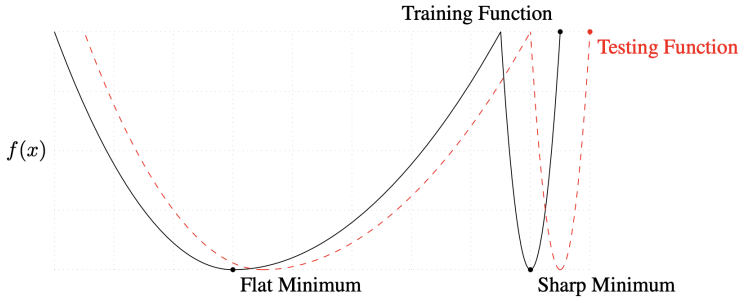



Figure 11: Taken from https://openreview.net/pdf?id=H1oyRlYgg. This illustration is very popular but **misleading**.

- Sharp minima can generalize well (Dinh et al., ICML 2017):

$$\sum_j a_j \mathrm{ReLU}(w_j^\top x) = \sum_j \frac{a_j}{\lambda} \mathrm{ReLU}(\lambda w_j^\top x) \text{ for any } \lambda > 0.$$

  Thus, **flatness can be only sufficient condition** at the most.
- There exist models such that flat minima generalize worse (Wu et al., NeurIPS 2018; Zhiyuan Li et al.2023).

# Why do flat minima generalize?

Consider two-layer neural nets: $f(x;\theta) = \sum_{j=1}^{m} a_j \sigma(w_j^T x)$ where $\theta = \{(a_j, w_j)\}_{j=1}^{m}, \sigma(z) = \max(0, z)$. Assume $x \sim \rho = \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$.

# Why do flat minima generalize?

Consider two-layer neural nets: $f(x;\theta) = \sum_{j=1}^m a_j \sigma(w_j^T x)$ where $\theta = \{(a_j, w_j)\}_{j=1}^m$, $\sigma(z) = \max(0, z)$. Assume $x \sim \rho = \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$.

## Proposition 5 (sharpness=weight decay; Wu-Su, ICML 2023)

*Given any $\delta \in (0,1)$, assume $n \gtrsim \max(d, \log(1/\delta))$. Then w.p. $1-\delta$ over the random sampling of $x_1, \ldots, x_n$, we have for any $\theta$ that*

$$\|G(\theta)\|_2 \sim \sum_{j=1}^m (\|w_j\|^2 + a_j^2)$$

$$\|G(\theta)\|_F \sim \sum_{j=1}^m (\|w_j\|^2 + \sqrt{d} a_j^2)$$

$$Tr(G(\theta)) \sim \sum_{j=1}^m (\|w_j\|^2 + d a_j^2)$$

Let $\|\theta\|_{2,\gamma} = \sum_{j=1}^m (\|w_j\|^2 + \gamma a_j^2)$ be the weighted square $\ell_2$ norm.

# Diagnoal linear networks

Consider diagonal linear net

$$f(x;\theta) = \langle a \odot b, x \rangle = \sum_{j=1}^{d} a_j b_j x_j.$$

## Theorem 6 (Wu-Su, ICML 2023)

*Let $\alpha = a \odot a + b \odot b$. For any $\delta \in (0,1)$, if $n \gtrsim \log(d/\delta)$, we have*

$$\|G(\theta)\|_2 \sim \|\alpha\|_\infty \tag{4}$$
$$\|G(\theta)\|_F \sim \|\alpha\|_2 \tag{5}$$
$$Tr(G(\theta)) \sim \|\alpha\|_1. \tag{6}$$

Note that

$$\|\alpha\|_p = \sqrt[p]{\sum_{j=1}^{d} (a_j^2 + b_j^2)^p} \geqslant 2 \sqrt[p]{\sum_{j}^{d} |a_j b_j|^p} \geqslant 2\|a \odot b\|_p.$$

Then, the stability of SGD can effectively control the model capacity, while that of GD cannot.

# Takeways

- Stability can guarantee that SGD find only flat minima.
- The anisotropic structure of SGD noise is crucial for ensuring the size independence of sharpness.
- Flat minima provably generalize well for two-layer neural nets and (diagonal) linear nets. This is due to that sharpness is equivalent to certain parameter norms for these models.

# Understanding the connection between stability and optimization.

# Revisiting the EoS phenomenon

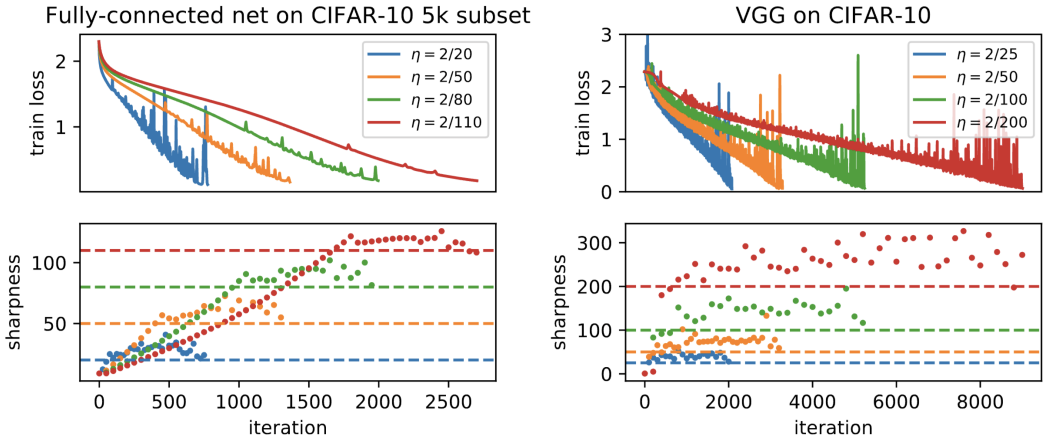


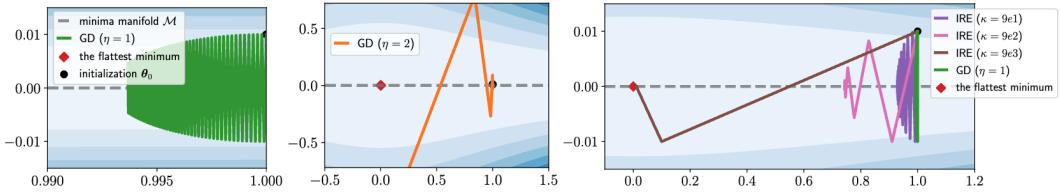

Figure 12: The edge of stability persists throughout the middle and late phases of training (Cohen et al., ICLR 2021).

# How does stability impact the sharpness reduction

Consider to use GD to solve following problem

$$\min_{\boldsymbol{u},\boldsymbol{v}} \frac{1}{2}\boldsymbol{v}^\top H(\boldsymbol{u})\boldsymbol{v},$$

where $\boldsymbol{v} \in \mathbb{R}^m, \boldsymbol{u} \in \mathbb{R}^{p-m}$. Let $H(\boldsymbol{u}) = \mathrm{diag}(\lambda_1(\boldsymbol{u}), \ldots, \lambda_m(\boldsymbol{u}))$.




(a) GD with $\eta = 1$. (b) GD with $\eta = 2$. (c) IRE with various $\kappa$'s v.s. GD with $\eta = 1$.

- **Flat direction:** $\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \frac{\eta}{2}\sum_{i=1}^m v_{t,i}^2 \nabla\lambda_i(\boldsymbol{u}_t)$. flat direction $\boldsymbol{u}_t$'s dynamics monotonically reduces the sharpness $\boldsymbol{\lambda}(\boldsymbol{u})$ as long as $\boldsymbol{v}_t$ is nonzero;
- **Sharp direction:** $\boldsymbol{v}_{t+1} = (\mathbf{1} - \eta\boldsymbol{\lambda}(\boldsymbol{u}_t)) \odot \boldsymbol{v}_t$. sharp direction $\boldsymbol{v}_t$'s dynamics determines the training stability.

# Accelerate sharpness reduction via decoupling stable and unstable direction

**Algorithm 1: Practical IRE** (A practical framework of implementing IRE)

**Input:** $\boldsymbol{\theta}_0$, $T$, $K$, learning rate $\{\eta_t\}_t$, warm-up time $T_{\mathrm{w}}$, IRE hyperparams: $\kappa \geqslant 0$ and $\gamma \in (0.5, 1)$;

**for** $t = T_{\mathrm{w}}, \cdots, T-1$ **do**

  Compute the original update direction $\boldsymbol{g}_t$ according to the base optimizer;

  **if** $(t - T_{\mathrm{w}}) \bmod K = 0$ **then**

    Estimate the diagonal Hessian $\boldsymbol{h}_t \in \mathbb{R}^p$ using Eq. (6);

    Update the mask $\boldsymbol{n}_t \in \mathbb{R}^p$ using Eq. (5);

  **else**

    $\boldsymbol{n}_t = \boldsymbol{n}_{t-1}$

  $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\big(\boldsymbol{g}_t + \kappa \boldsymbol{n}_t \odot \boldsymbol{g}_t\big)$;

**Output:** $\boldsymbol{\theta}_T$.

# Numerical results

**IRE can consistently improve generalization.**

Table 2: ViT-T/S on CIFAR-100.

| | ViT-T | ViT-S |
|---|---|---|
| AdamW | 63.90 | 65.43 |
| AdmIRE | **67.05** (**+3.15**) | **68.39** (**+2.96**) |
| SAM | 64.25 | 66.93 |
| SAM-IRE | **67.33** (**+3.08**) | **70.47** (**+3.54**) |

Table 3: ViT-S on ImageNet.

| | Top-1 | Top-5 |
|---|---|---|
| AdamW | 78.7 | 94.0 |
| AdmIRE ($\kappa = 2, \gamma = 0.6$) | **79.0** (**+0.3**) | **94.3** (**+0.3**) |
| AdmIRE ($\kappa = 2, \gamma = 0.8$) | **79.1** (**+0.4**) | **94.2** (**+0.2**) |

# A principle of block-wise sharpness for transformers

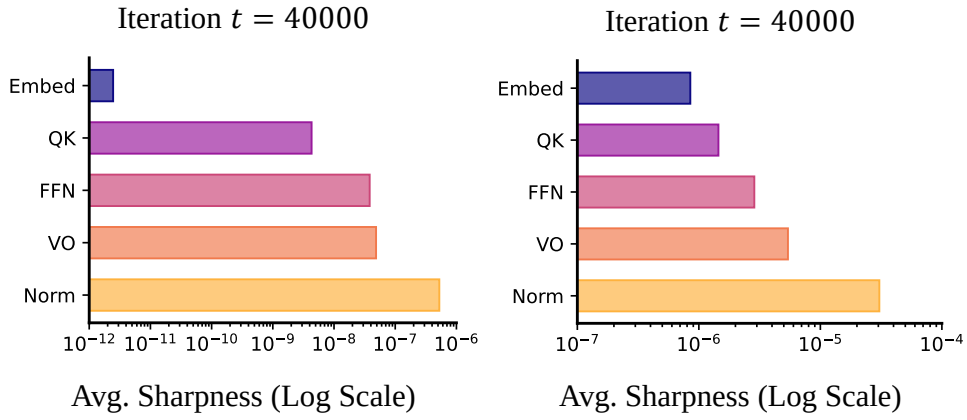
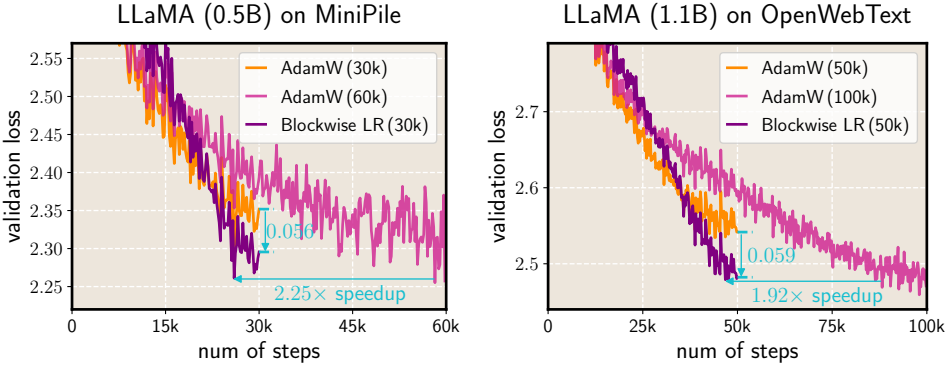



Figure 13: (Left) GPT-2(small) for openwebtext; (right) LLaMA (0.25B) for openwebtext.

# Acceleration via increasing LR along stable direction



# Summary

- The **stability imposes capacity control during the training**, which partially explains the implicit regularization of SGD.
- For speed up convergence, we can decouple the **stable** and **unstable** directions, and then increase the learning rates along stable directions.

# Thank You!


- **Lei Wu**, Chao Ma, and Weinan E, *How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective*. NeurIPS 2018.
- **Lei Wu**, Mingze Wang, Weijie J. Su, *The alignment property of SGD noise and how it helps select flat minima: A stability analysis*. NeurIPS 2022.
- **Lei Wu**, Weijie J. Su, *The implicit regularization of dynamical stability in stochastic gradient descent*. ICML 2023.
- Mingze Wang, **Lei Wu**, *The noise geometry of stochastic gradient descent*. Submitted.
- Mingze Wang, Haotian He, Jinbo Wang, Zilin Wang, Guanhua Huang, Feiyu Xiong, Zhiyu Li, **Lei Wu**, *Improving generalization and convergence by enhancing implicit regularization*, NeurIPS 2024.
- Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Weinan E, **Lei Wu**, *The Sharpness Disparity Principle in Transformers for Accelerating Language Model Pre-Training*, submitted.