The 11th NSPCS conference

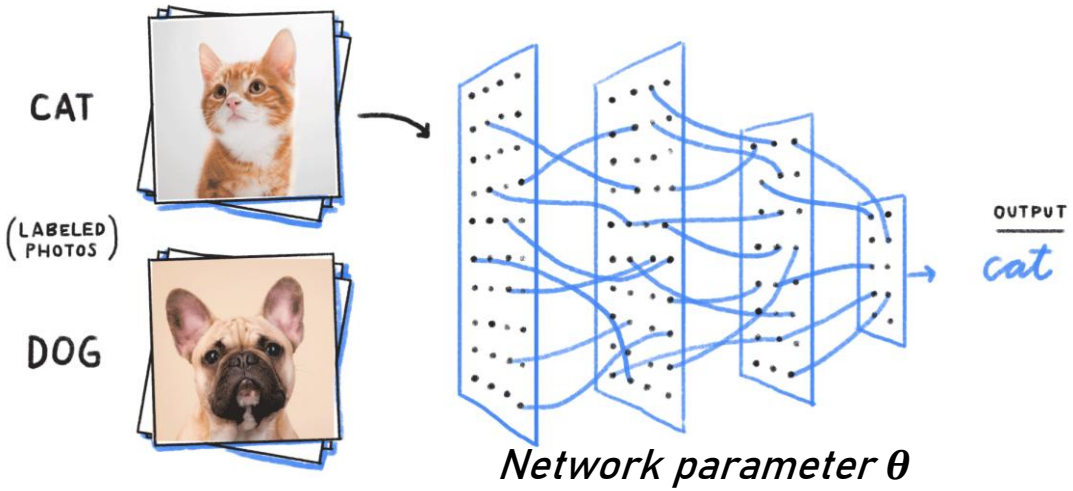
Mini-batch Gradient Descent without Replacement Drives Stable Convergence to Flat Minima



Presenter – Youngkyoung Bae

Department of Physics and Astronomy, Seoul National University, South Korea

| Why Does Deep Learning Work So Well? *Algorithms*



Network parameter θ

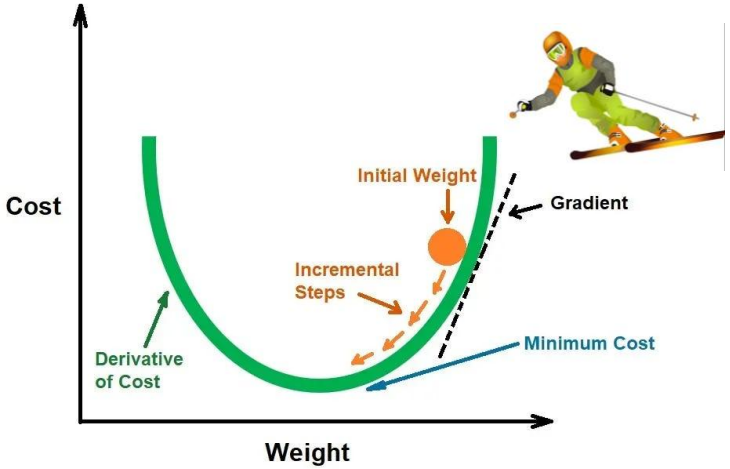
GOAL

Learn a neural network f_θ that maps input x to output y.

$$f_\theta(\text{🐱}) = `cat'$$

But, how to learn?

Minimize the loss (or cost) $\mathcal{L}(\theta) \equiv \sum_{(x,y)\in\mathcal{D}} \big(y - f_\theta(x)\big)^2$



Gradient Descent: Take a step in the opposite direction of $\nabla_\theta \mathcal{L}_\theta$

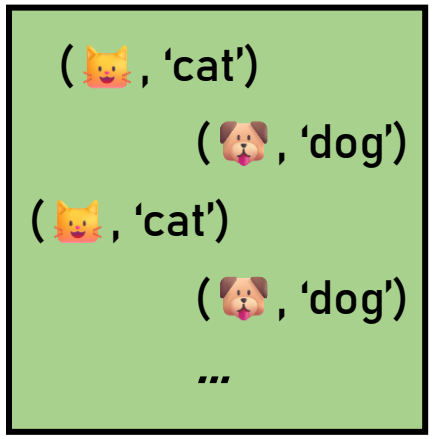
⇒ PROBLEM: When considering the large dataset,



Expensive computation & Memory…!

| Stochastic Gradient Descent (Mini-batch GD)

Dataset $\mathcal{D}$



Loss for dataset $\mathcal{D}$: $\mathcal{L}(\theta) \equiv \frac{1}{N}\sum_{(x,y)\in\mathcal{D}} \left(y - f_\theta(x)\right)^2$

Stochastic Gradient Descent (Mini-batch GD)

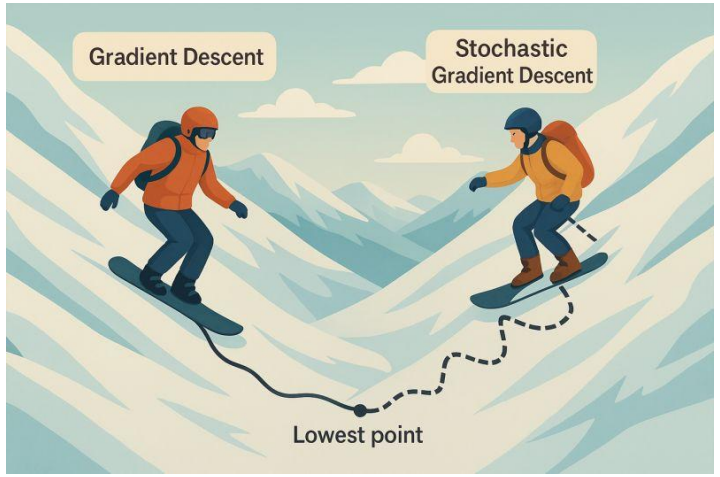
Dataset $\mathcal{D}$

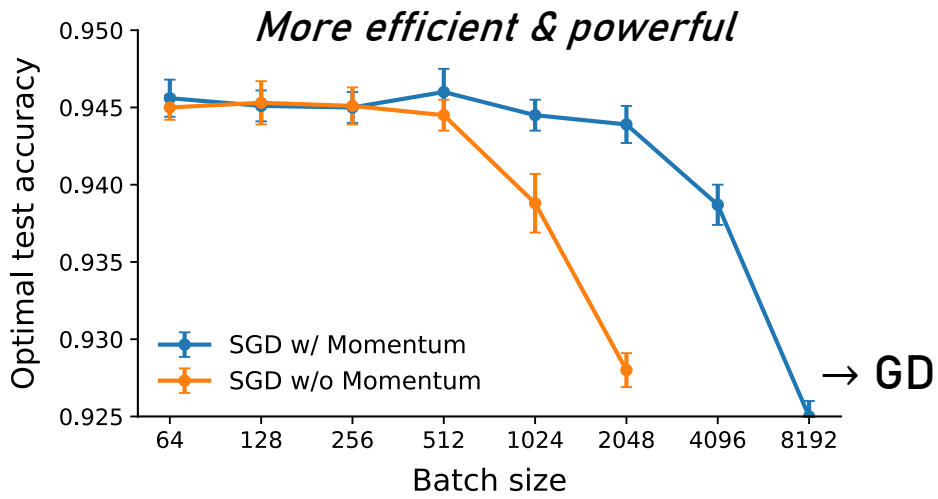
Batch B^1
Batch B^2
Batch B^3
Batch B^4
...

Loss for batch B^μ: $\mathcal{L}^\mu(\theta) \equiv \frac{1}{B}\sum_{(x,y)\in B^\mu} \left(y - f_\theta(x)\right)^2$,

where B^μ are typically sampled *w/o replacement* by randomly permuting the dataset.

⇒ *The randomness of SGD comes from batch sampling*

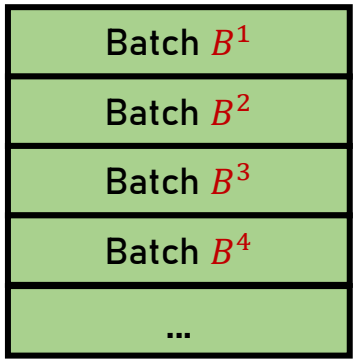




S. Smith, E. Elsen, and S. De, *ICML* (2020)

Stochastic Gradient Descent (Mini-batch GD)

Dataset $\mathcal{D}$



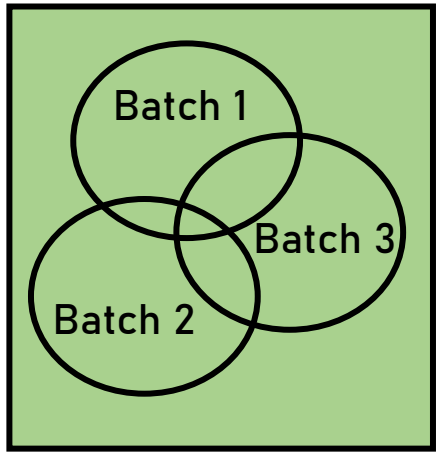
Loss for batch B^μ: $\mathcal{L}^\mu(\theta) \equiv \frac{1}{B}\sum_{(x,y)\in B^\mu} \left(y - f_\theta(x)\right)^2$,

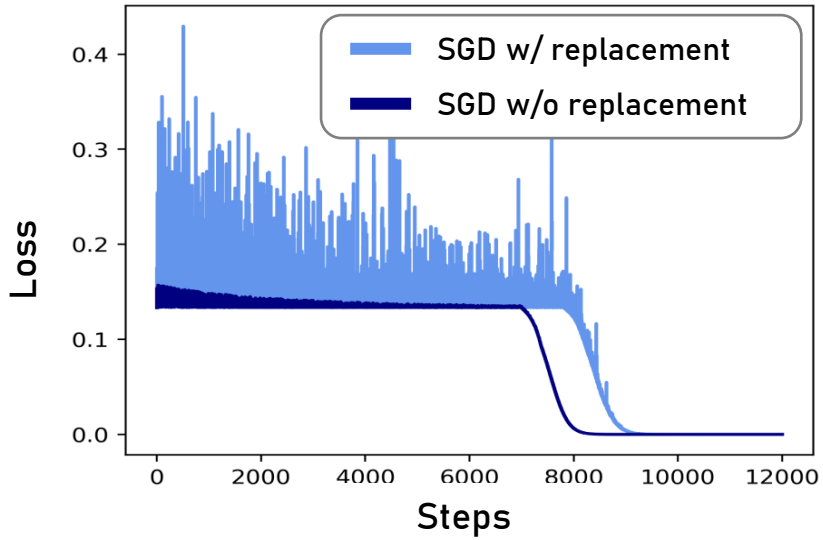
where B^μ are typically sampled *w/o replacement* by randomly permuting the dataset.

⇒ *The randomness of SGD comes from batch sampling*



w/ replacement

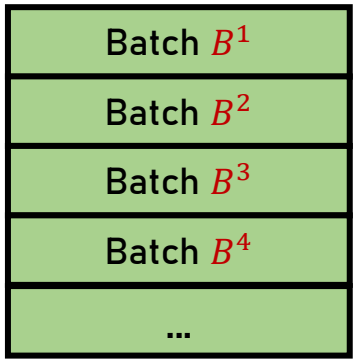




P. Beneventano, arXiv:2312.16143

Stochastic Gradient Descent (Mini-batch GD)

Dataset $\mathcal{D}$



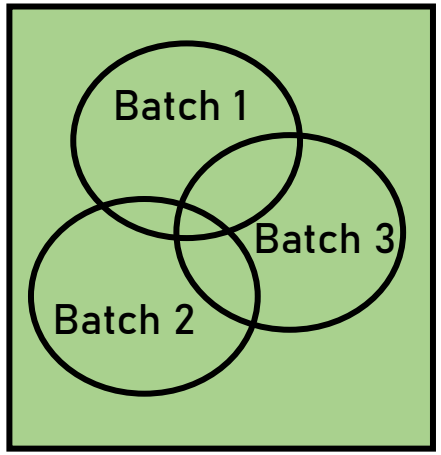
Loss for batch B^μ: $\mathcal{L}^\mu(\theta) \equiv \frac{1}{B}\sum_{(x,y)\in B^\mu} \left(y - f_\theta(x)\right)^2$,

where B^μ are typically sampled *w/o replacement* by randomly permuting the dataset.

⇒ *The randomness of SGD comes from batch sampling*



w/ replacement



Our Goals

1) Understand why SGD without replacement performs better.

2) Explore whether this benefit can be further enhanced.

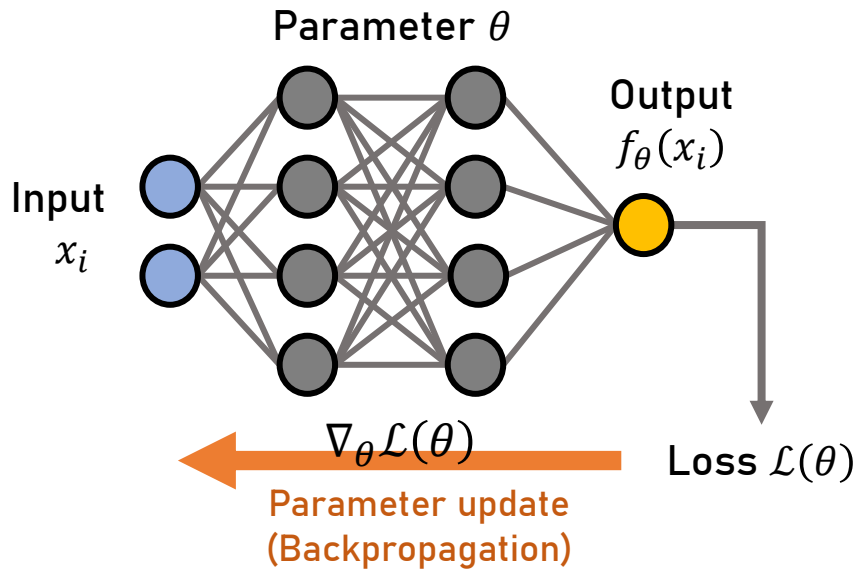
Steps

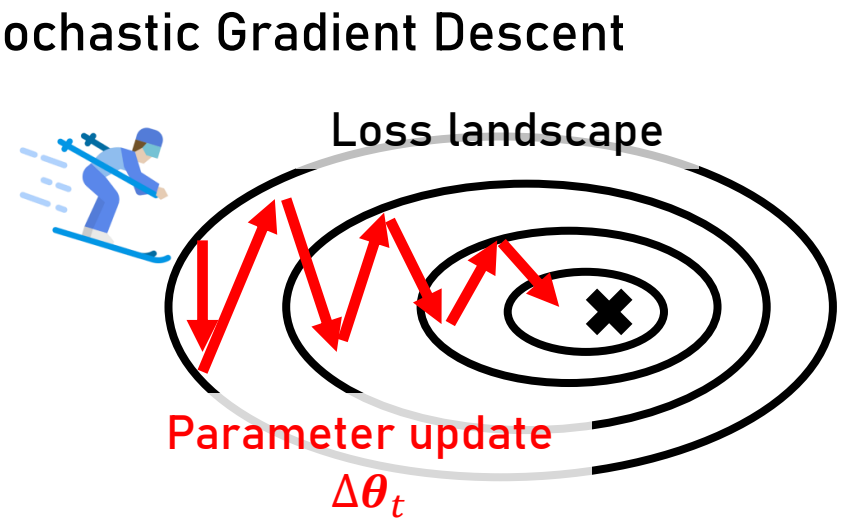
P. Beneventano, arXiv:2312.16143

Description

Stochastic Gradient Descent

For a given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, parameters θ are trained to minimize the loss l with a learning rate η





$$\textit{Stochastic process} \quad \Delta\theta_t = -\eta\nabla\frac{1}{B}\sum_{i=1}^{B} l(y_i, f_\theta(x_i)) = \underline{-\eta\nabla\mathcal{L}(\theta_t)} + \underline{\boldsymbol{\xi}_t\sqrt{\eta}}$$

Drift from dataset

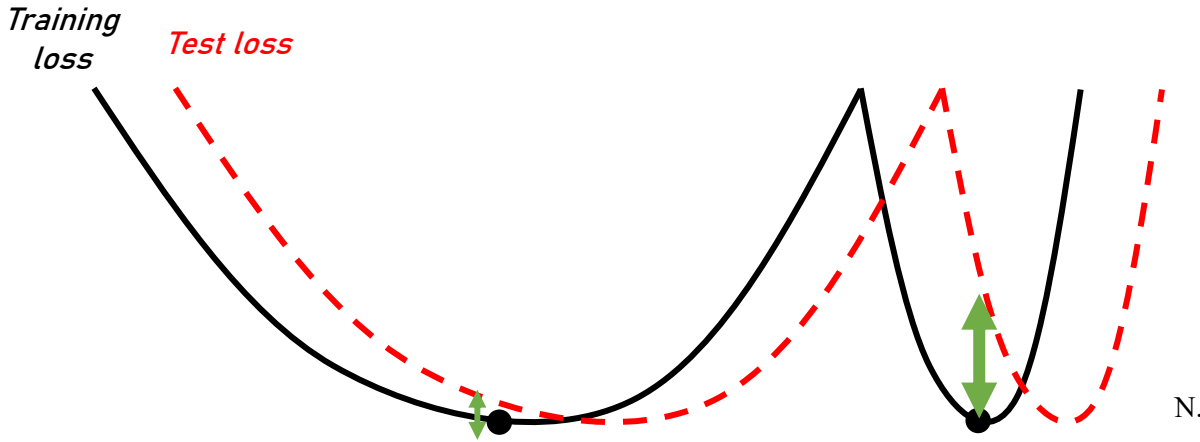
Random noise (i.e., $\langle\xi_t\rangle = 0$) *from batch-sampling*

Description

| Flat-minima Hypothesis

For analytical tractability, we consider *an overparameterized regime* (θ can rapidly enter a low-loss valley)





N. S. Keskar et al., ICLR (2017)

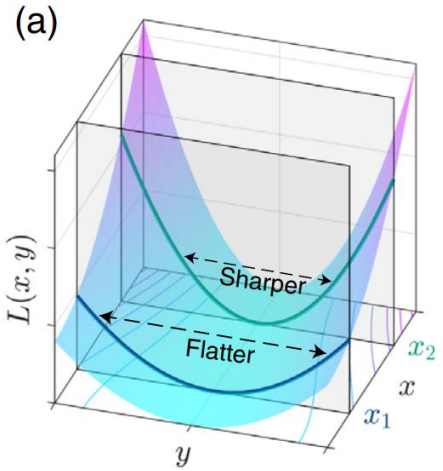
Among low training loss solutions, those in flatter regions tend to generalize better

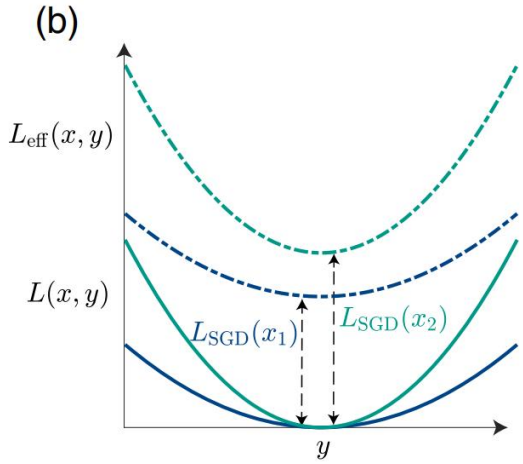
Description

| Flat-minima Hypothesis

For analytical tractability, we consider *an overparameterized regime* (θ can rapidly enter a low-loss valley)







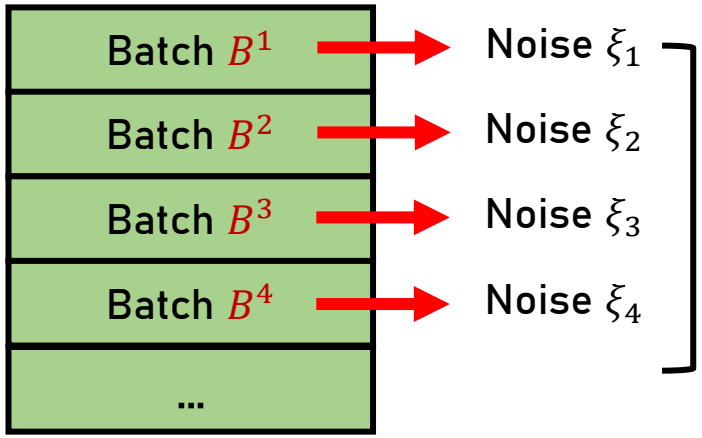
…Anisotropy of gradient noise drives the implicit preference of flat minima.

N. Yang et al., *Phys. Rev. Lett.***130**, 237101 (2023)

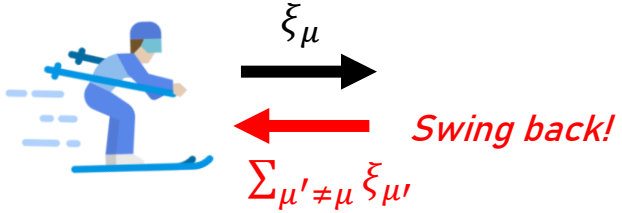
What effect does SGD without replacement have on the implicit preference for flat minima?

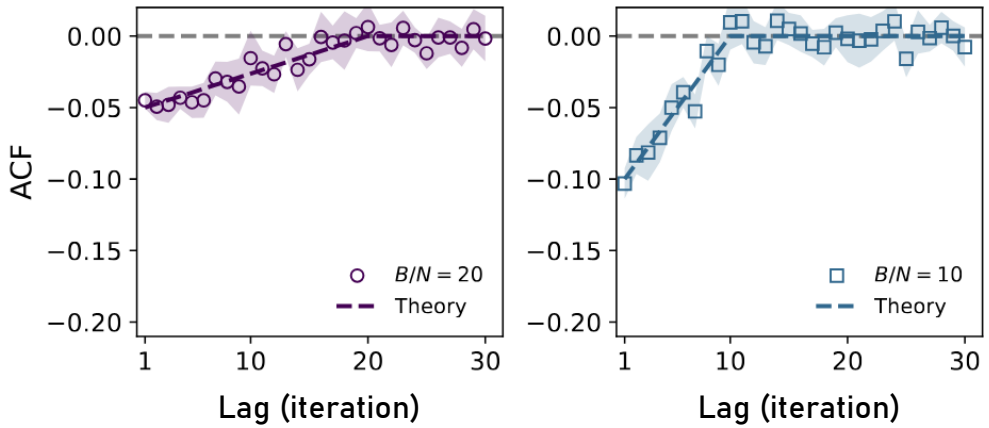
Negative Temporal Correlation of Gradient Noise

Dataset $\mathcal{D}$



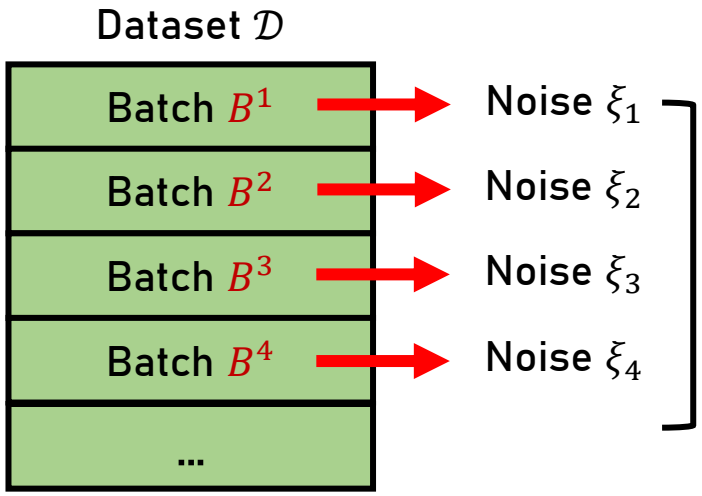
By definition, $\sum_\mu \xi_\mu = 0$ $\Rightarrow$ *Negative correlation between* ξ_μ





…The negative correlation varies depending on B/N

| Negative Temporal Correlation of Gradient Noise

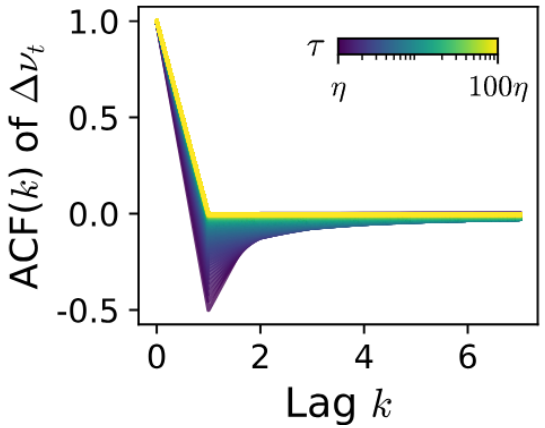


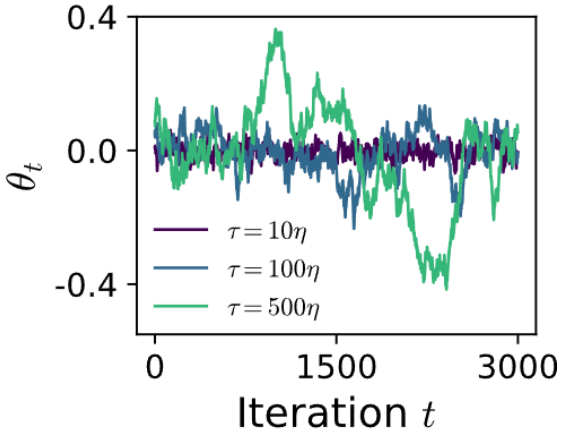
By definition, $\sum_\mu \xi_\mu = 0$ $\Rightarrow$ *Negative correlation between* ξ_μ

Let us model this negative temporal correlated noise!

$$\Delta\theta_t = -\eta\nabla\mathcal{L}_\mathcal{D}(\theta_t) + \Delta\nu_t \text{ , where } \Delta\nu_t = -(\eta/\tau)\nu_t + \sigma_A\sqrt{\eta}\xi_t$$

τ controls the temporal correlation ($\sim N/B$), σ_A controls the noise magnitude





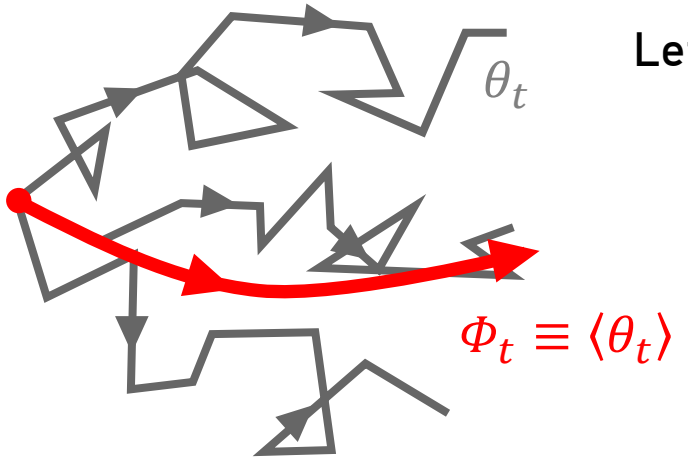
...As τ *increases, the lag-1 autocorrelation becomes less negative.*

$\Rightarrow$ *if* $\tau \to \infty$, *then* $\Delta\nu_t \to$ *No temporal correlation (w/ replacement)*

...As τ *increases, the fluctuation of* θ_t *increases (in free-space)*

$\Rightarrow$ *Stronger negative correlation (low* τ *) prevents* θ *from deviating too far in a random direction.*

| Implicit Regularization of SGD w/o Replacement



Let Φ_t be *a deterministic process* with $\Delta\Phi_t \coloneqq \langle\Delta\theta_t\rangle$: *The mean path of* θ_t

$\theta_t \equiv \Phi_t + \omega_t$, where ω_t is *a stochastic process* with $\langle\omega_t\rangle = 0$

How does the mean path Φ_t *evolve?*

$$\Delta\Phi_t \coloneqq \langle\Delta\theta_t\rangle = -\eta\langle\nabla\mathcal{L}(\theta_t)\rangle \coloneqq -\eta\nabla\tilde{\mathcal{L}}(\Phi_t)$$

$\nabla\mathcal{L}(\theta_t) \approx \nabla\mathcal{L}(\Phi_t) + H(\Phi_t)\omega_t + \frac{1}{2}\left[\left(\nabla H(\Phi_t)\right):\omega_t\omega_t^T\right] + \mathcal{O}(\omega_t^{3/2})$, where $H(\Phi) \coloneqq \nabla^2\mathcal{L}(\Phi)$ is the Hessian matrix.

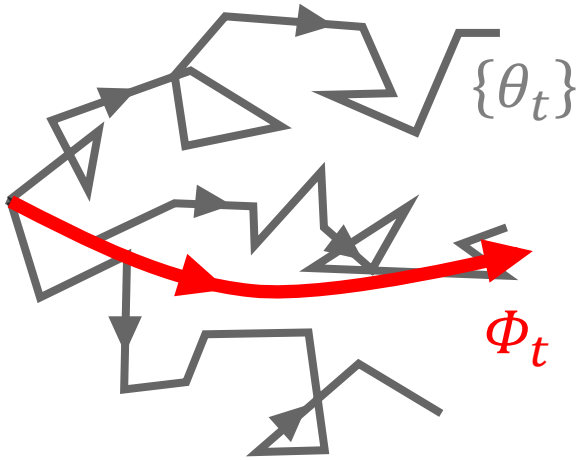
$\Rightarrow \quad \nabla\tilde{\mathcal{L}}(\Phi_t) \approx \nabla\mathcal{L}(\Phi_t) + \frac{1}{2}\left[\left(\boldsymbol{\nabla H(\Phi_t)}\right):\boldsymbol{\Sigma_\omega}\right]$ *Noise-induced term (Implicit Regularization)*

Using the expression $\nabla\tilde{\mathcal{L}}(\Phi_t)$ and the update rule of θ, obtain $\boldsymbol{\Delta\omega_t \coloneqq \Delta\theta_t - \Delta\Phi_t = -\eta H(\Phi_t)\omega_t + \Delta\nu_t}$ (OU process)

$\Rightarrow$ *We can obtain the parameter fluctuation* $\boldsymbol{\Sigma_\omega}$ *and the effective loss* $\boldsymbol{\tilde{\mathcal{L}}(\Phi_t)}$

| Implicit Regularization of SGD w/o Replacement

The effective loss $\tilde{\mathcal{L}}(\Phi_t) \approx \mathcal{L}(\Phi_t) + \sigma_A^2 \mathcal{R}_\tau(\Phi_t)$ where $\mathcal{R}_\tau = \frac{1}{4(1-\eta/2\tau)} \mathrm{Tr}\left[\ln|I_d + (\tau-\eta)H| - \ln\left|I_d - \frac{\eta}{2}H\right|\right]$



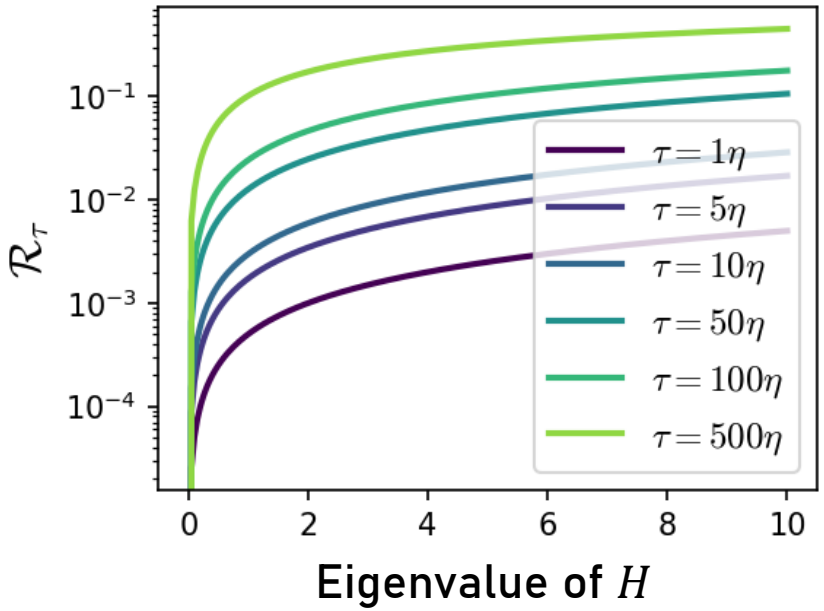
We can predict

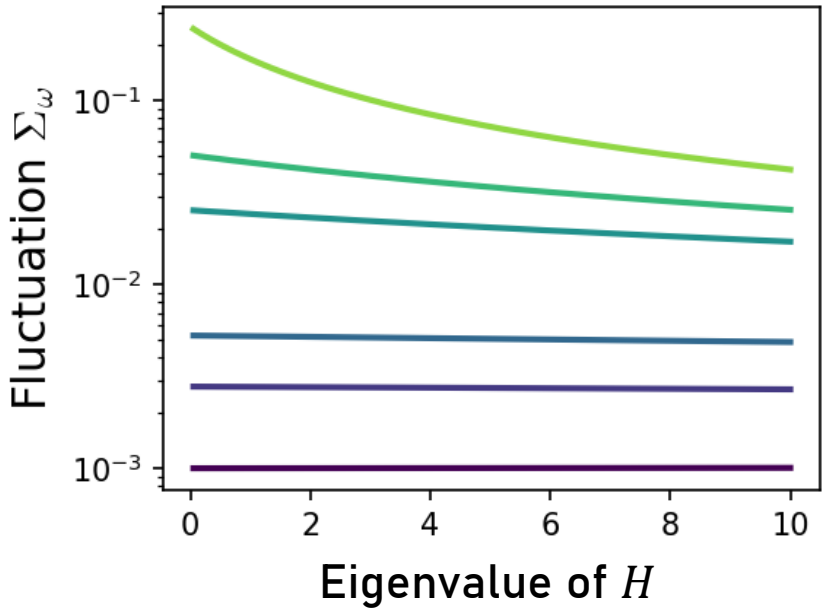
the mean path of SGD

$\Delta\Phi_t = -\eta\nabla\tilde{\mathcal{L}}_{\mathcal{D}}(\Phi_t)$

…As the Hessian eigenvalue increases, $\mathcal{R}_\tau$ increases ⇒ Penalize high curvature

…As τ increases, both the penalty $\mathcal{R}_\tau$ and the fluctuation Σ_ω increase ⇒ Trade-Off?

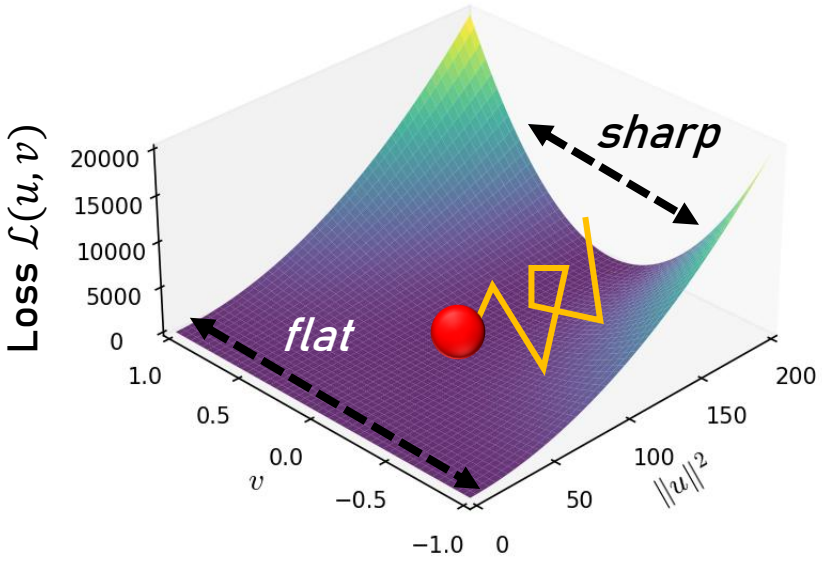


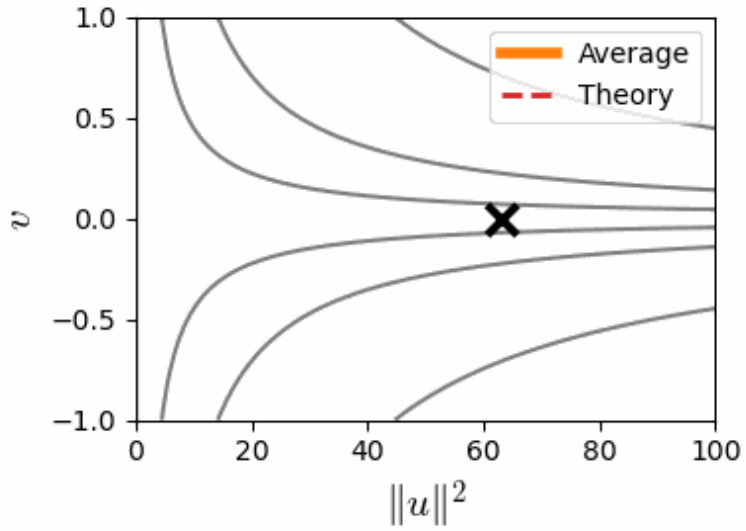


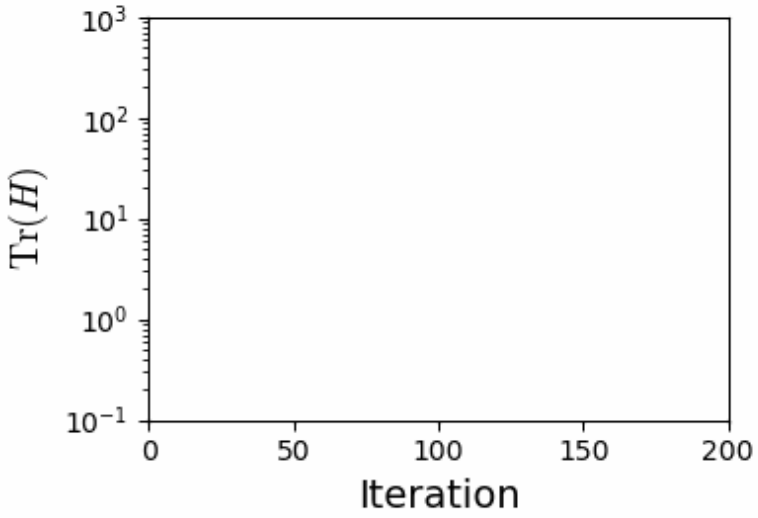
| Validation on Widening Valley

The *widening valley* is defined by the loss function $\mathcal{L}(u, v) = \frac{1}{2} v^2 \, ||u||^2$ with trainable parameters $v \in \mathbb{R}$ and $u \in \mathbb{R}^d$

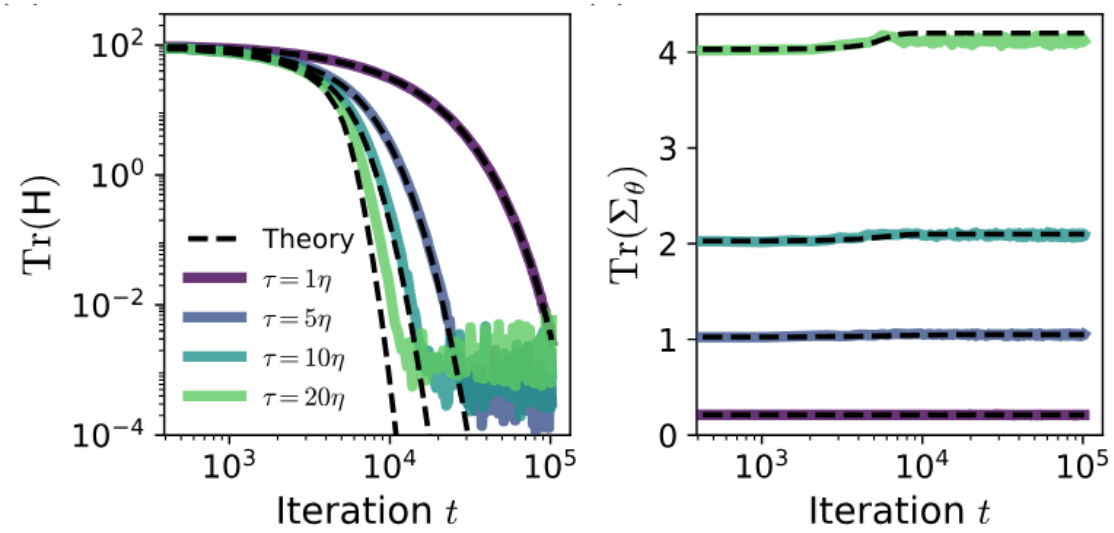
where the gradient $\nabla \mathcal{L}(u, v) = \begin{bmatrix} v^2 u \\ ||u||^2 v \end{bmatrix}$ and the trace of Hessian $\mathrm{Tr}(H) = dv^2 + ||u||^2$

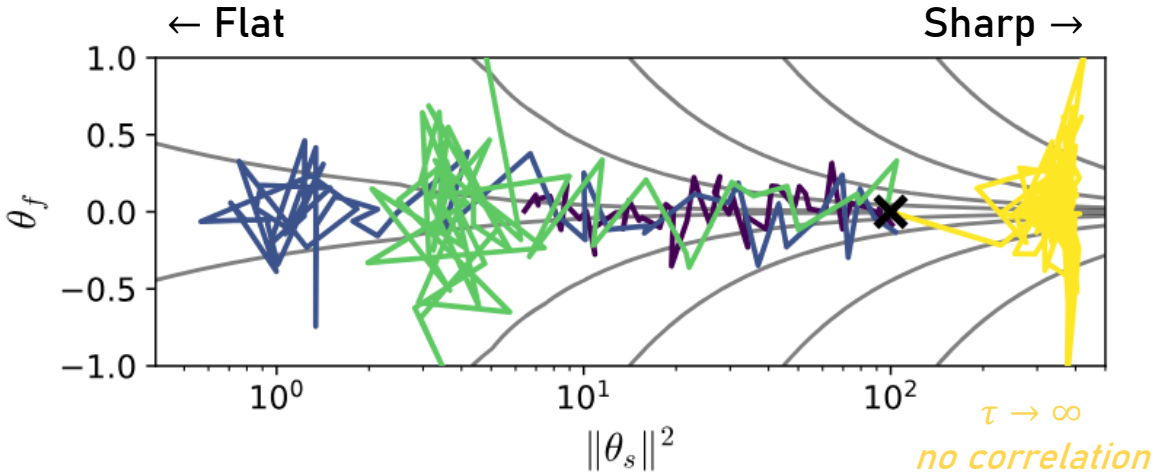






| Validation on Widening Valley

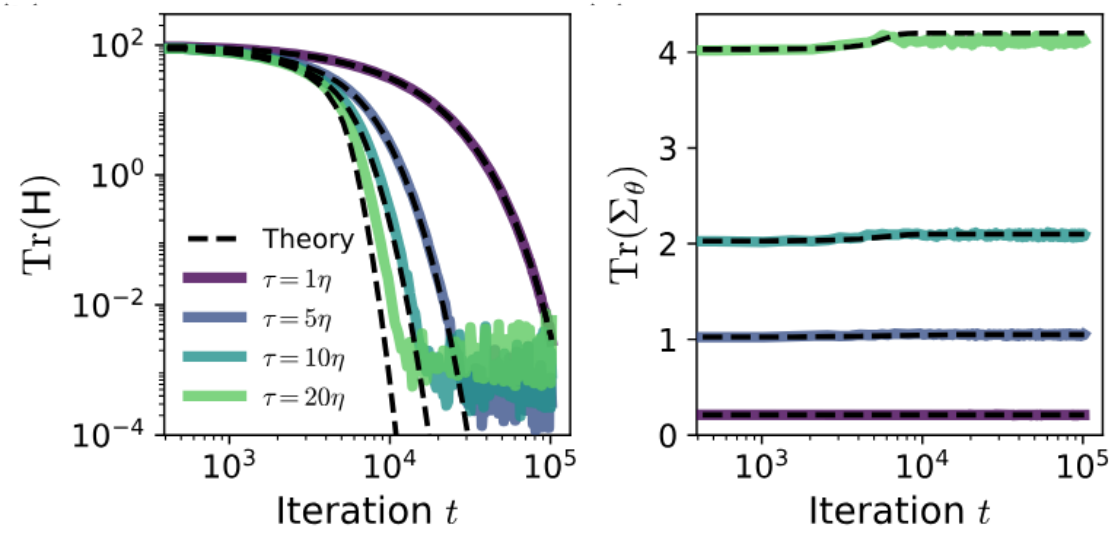


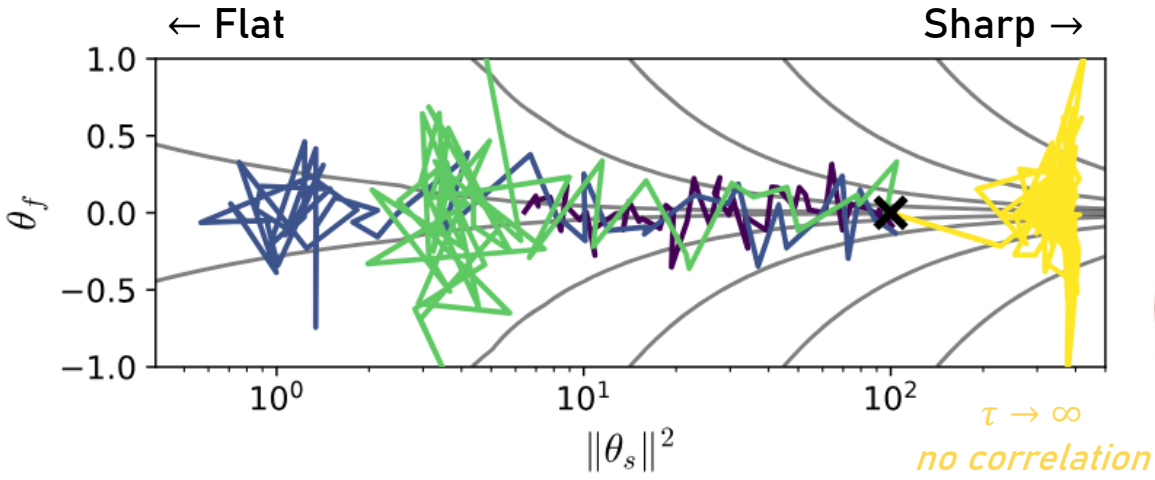


🌟 …Increasing τ speeds up the initial approach toward the flattest minima, but increased fluctuation lead to convergence farther away from them (Trade-off relation)

+ when $\tau \to \infty$ (no correlation), the parameter θ diverges

| Validation on Widening Valley

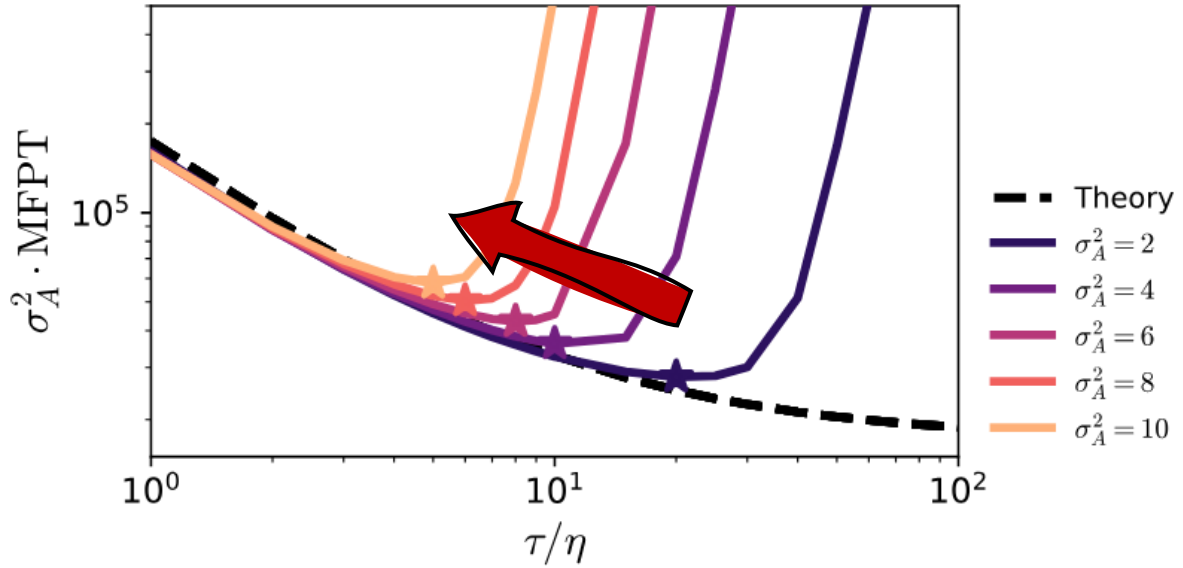




🌟…Increasing τ speeds up the initial approach toward the flattest minima, but increased fluctuation lead to convergence farther away from them (Trade-off relation)

+ when $\tau \to \infty$ (no correlation), the parameter θ diverges

Optimal τ from the trade-off: *Speed $\mathcal{R}_\tau$ vs. Stability Σ_θ*



| Empirical Results – Regression & Classification

Regression (Matrix Sensing)

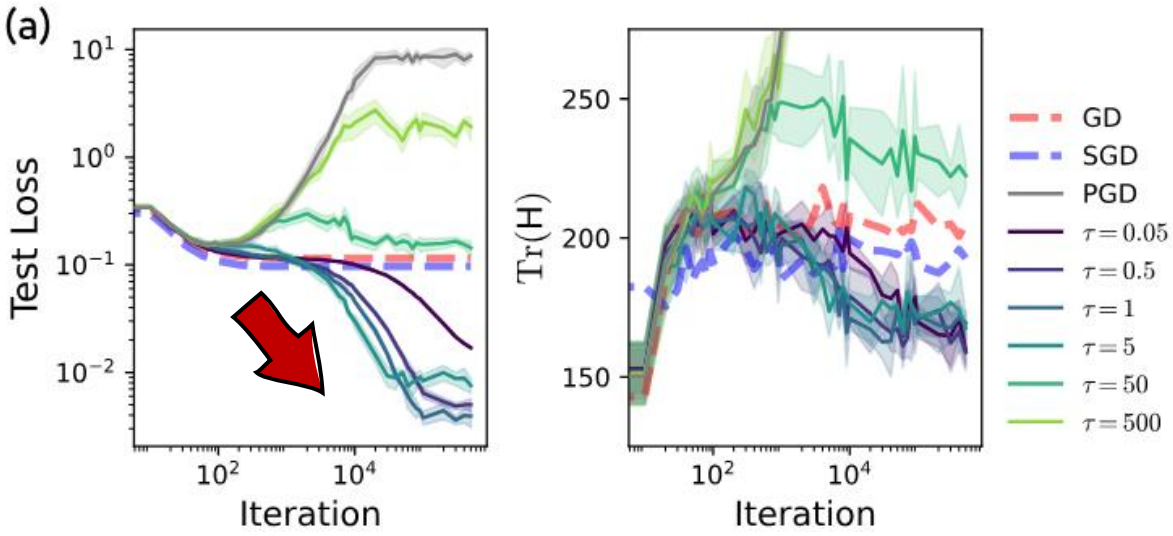
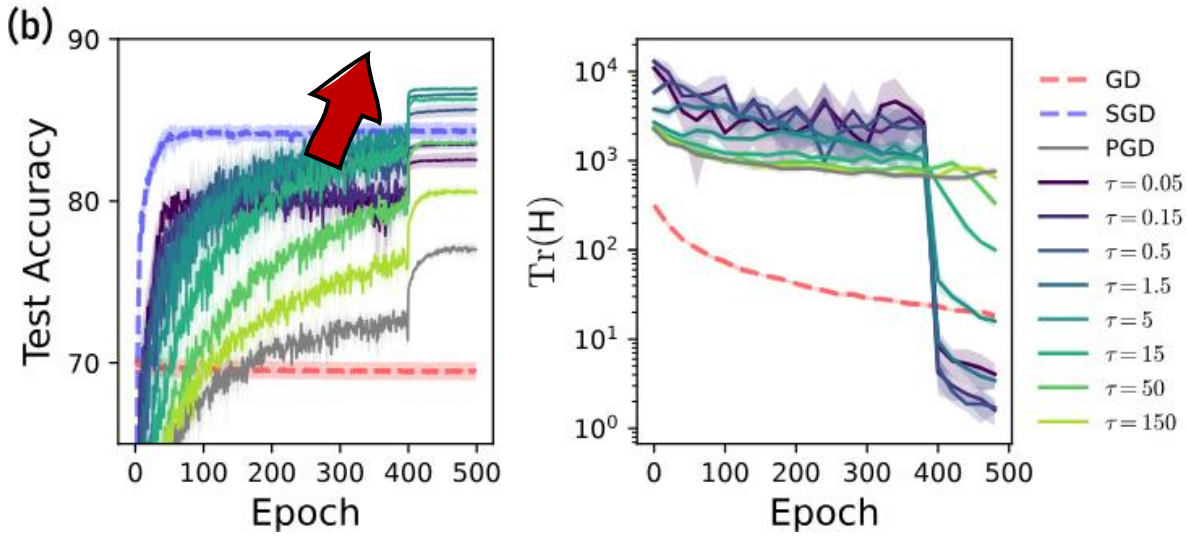


Image Classification (CIFAR-10)

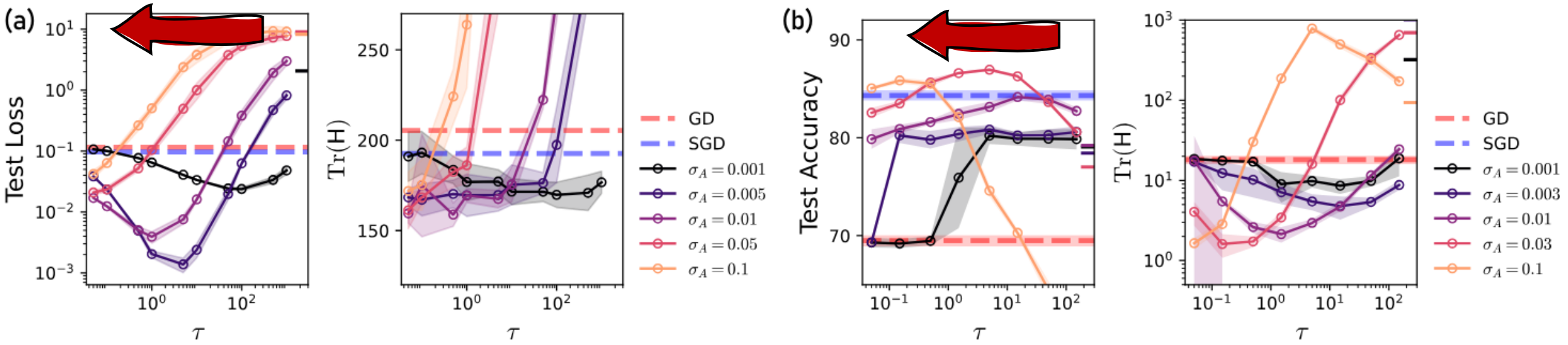


🌟 …Increasing τ speeds up the initial approach toward the flattest minima, but increased fluctuation lead to convergence farther away from them (Trade-off relation)

| Empirical Results – Regression & Classification

Regression (Matrix Sensing)

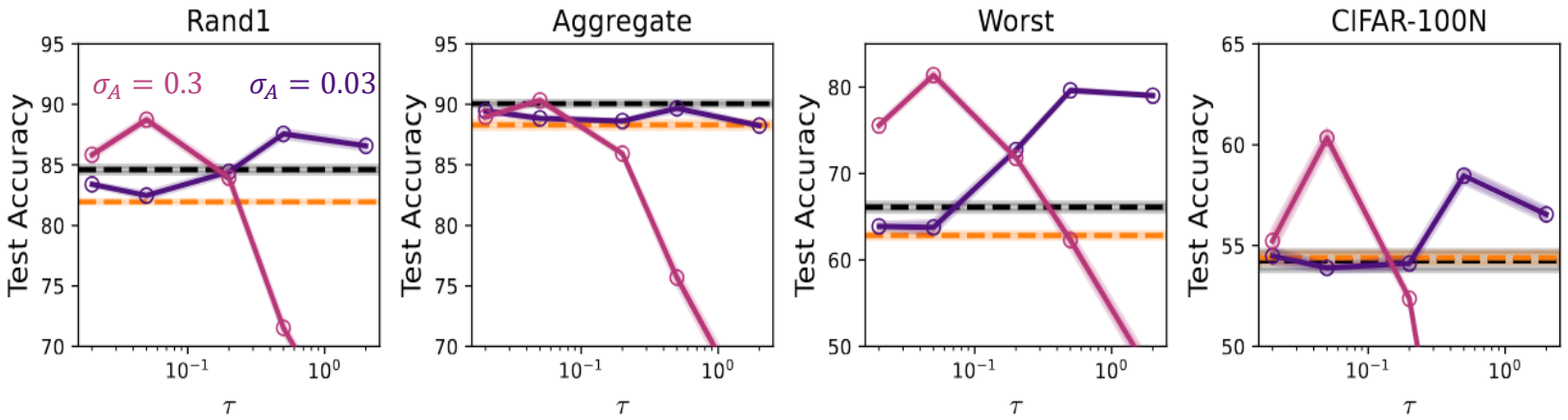
Image Classification (CIFAR-10)



🌟 …Increasing τ speeds up the initial approach toward the flattest minima, but increased fluctuation lead to convergence farther away from them (Trade-off relation)

🌟 …Optimal τ decreases as the fluctuation magnitude increases ($\sigma_A \uparrow$)

| Empirical Results in Real-world Datasets (Noisy-label dataset)



	CIFAR-10N			CIFAR-100N
	Rand1 (Nr $\sim$ 0.09)	Aggregate (Nr $\sim$ 0.17)	Worst(Nr $\sim$ 0.4)	Noisy(Nr $\sim$ 0.4)
SGD	84.51 ± 0.36	89.83 ± 0.33	66.38 ± 0.52	54.28 ± 0.40
+ (White Noise)	40.74 ± 1.35	43.15 ± 0.54	33.98 ± 1.06	31.66 ± 0.23
+ ($\tau = \eta$)	81.83 ± 0.21	88.30 ± 0.22	62.79 ± 0.40	54.29 ± 0.27
+ (Best τ)	$\mathbf{88.67 \pm 0.21}$	$\mathbf{90.34 \pm 0.16}$	$\mathbf{81.50 \pm 0.32}$	$\mathbf{58.38 \pm 0.32}$

🌟 *…We can enhance the beneficial mechanism of negative temporal correlation!*

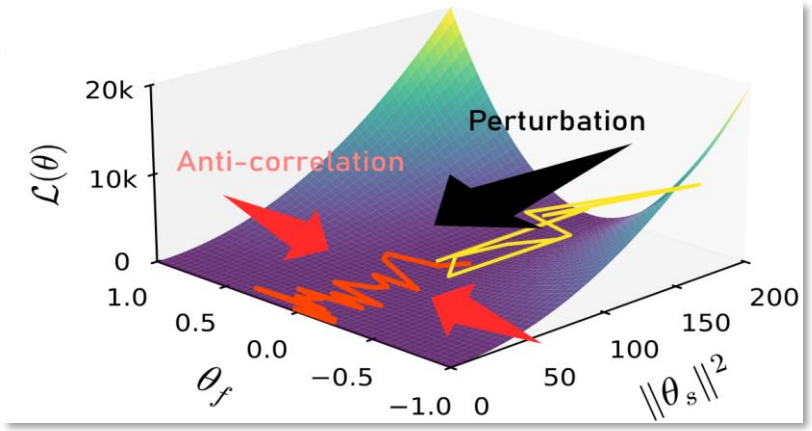
Conclusion

🌟…SGD without replacement induces negative temporal correlation in the gradient noise, which plays an important yet often overlooked role in generalization.

🌟…We analytically show that, although stronger negative correlation weakens the implicit regularization, it simultaneously reduces the parameter fluctuation, enabling *more stable convergence near flat minima*

(Trace-off between Speed vs. Stability)

🌟…We empirically demonstrate our theory on various tasks and show that the benefits of negative noise can be further enhanced in practical settings.





Take-home: Just use without-replacement batch sampling

– *but choose the batch size carefully…!*

Thank you for listening!

Presenter - Youngkyoung Bae ✉ tardis_95@snu.ac.kr



Minseok M. Kim



CANONEQal members



Prof. Yongjoo Baek

Empirical Results in Practice – Regression & Image Classification

Linear-regression model $y_i = u^T x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$

