The 10th Nonequilibrium Statistical Physics of Complex Systems

Stochastic Restarting to Overcome Overfitting in Neural Networks with Noisy Labels

***Youngkyoung Bae*[1], Yeongwoo Song[2], and Hawoong Jeong[3]**

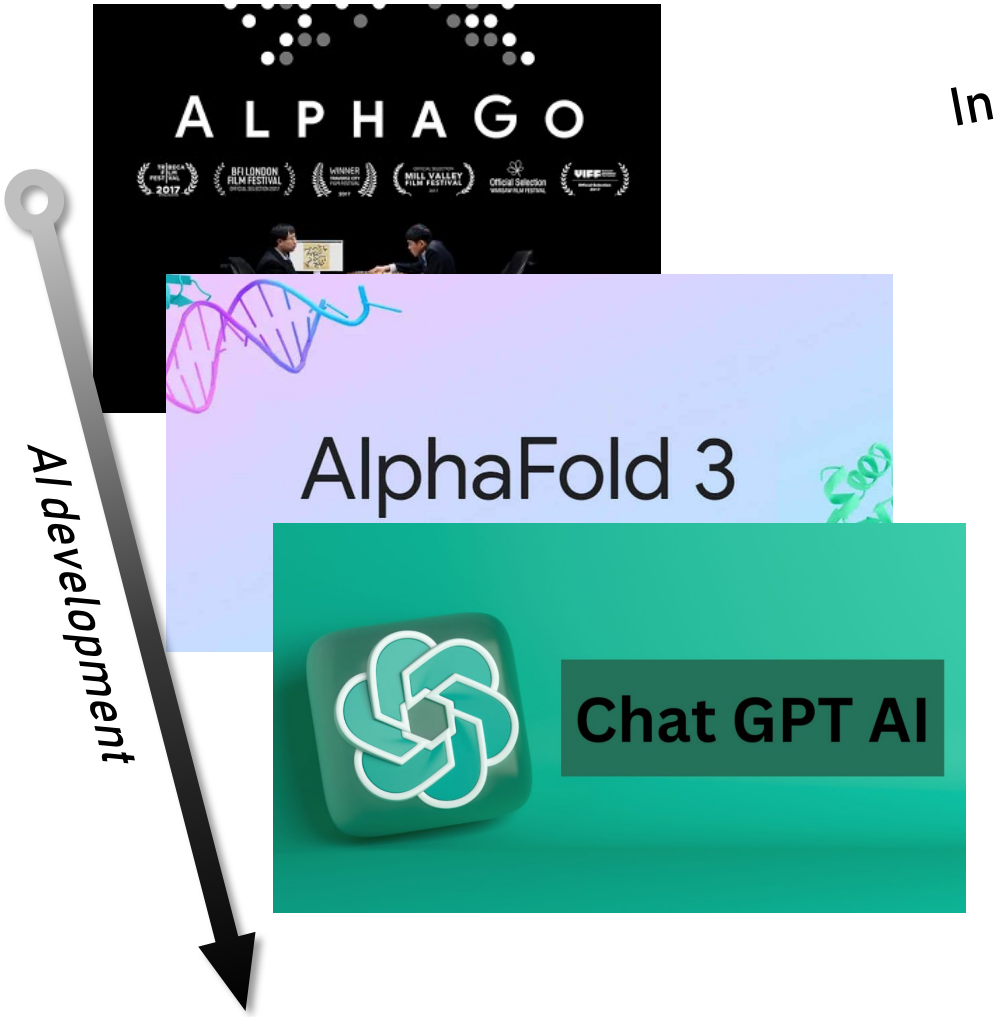
[1]Department of Physics and Astronomy, Seoul National University, Korea

[2]Department of Physics, Korea Advanced Institute of Science and Technology, Korea

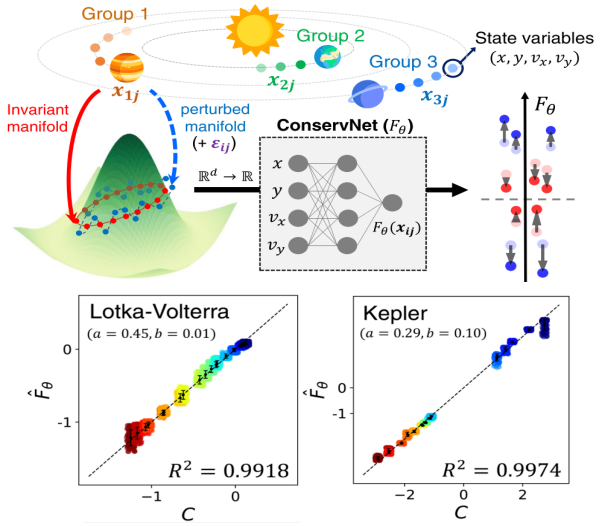
[3]Center for Complex Systems , Korea Advanced Institute of Science and Technology, Korea

Introduction

| Success of deep learning

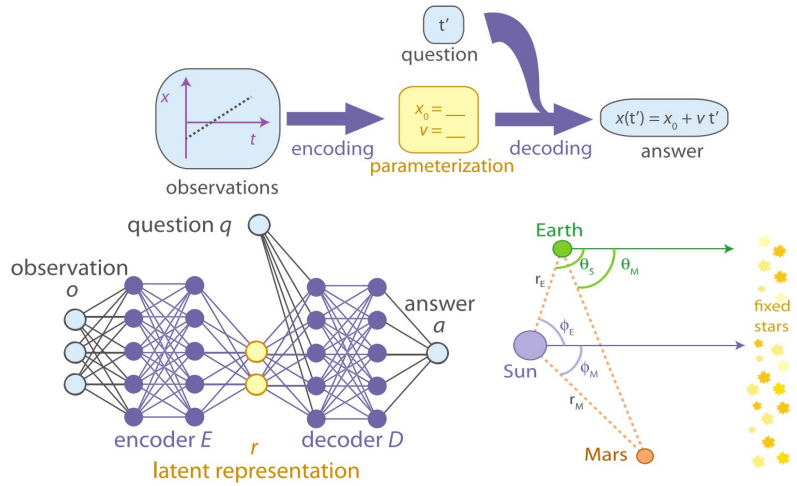


In the field of physics



Discovering invariants

S. Ha et al., Phys. Rev. Res. **3**, L042035 (2021)



Discovering Physical Concepts

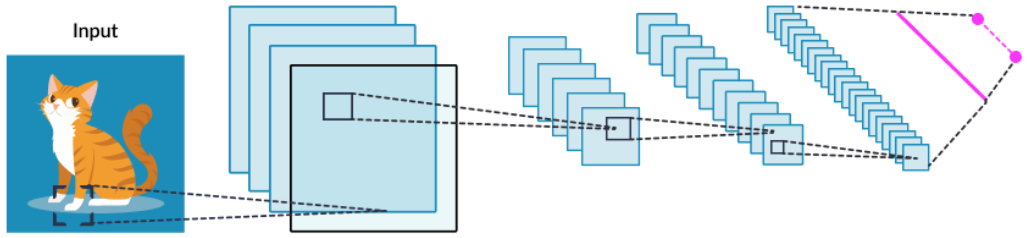
R. Iten et al, Phys. Rev. Lett. **124**, 010508 (2019)

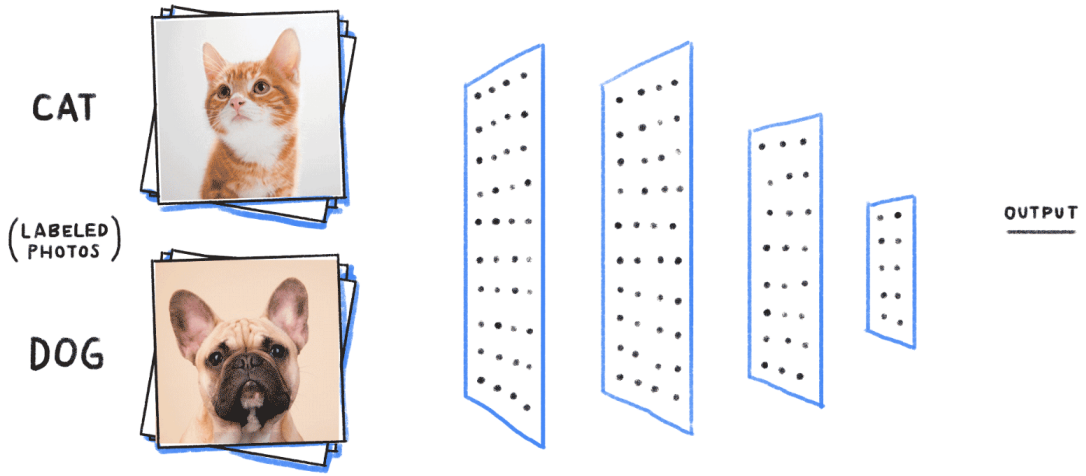
Introduction

| Success of deep learning

Ex] *Image classification*

Convolutional Neural Network (CNN)

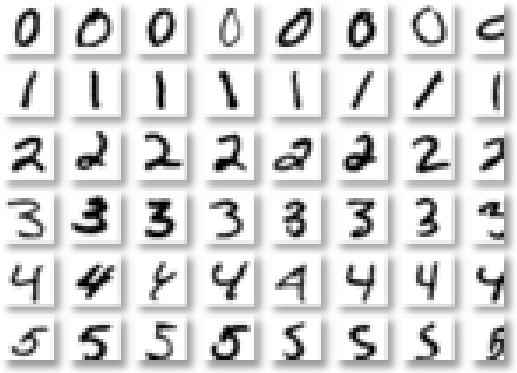




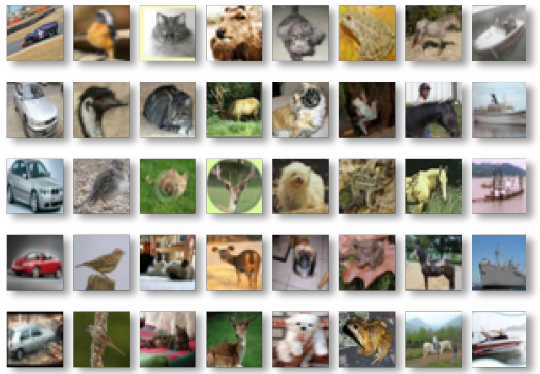
Objective function (e.g., *Cross entropy*)

$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\theta}) = -\sum_i \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \log(\boldsymbol{y}_i)$$

$\boldsymbol{x}_i$: input data (image), $\boldsymbol{y}_i$: label, $\boldsymbol{f}_{\boldsymbol{\theta}}$: neural network



MNIST ($\approx$ 99.9%)



CIFAR-100 ($\approx$ 96.1%)

Introduction

| Success of deep learning *with noisy labels…?*

Chihuahua or ***Muffin***?



Noisy labels (≡ unreliable labels) are induced by

- human mistakes
- the lack of domain expertise (e.g., medical data)
- incomplete labeling techniques

…

➡ ***Degrade the generalization performance***

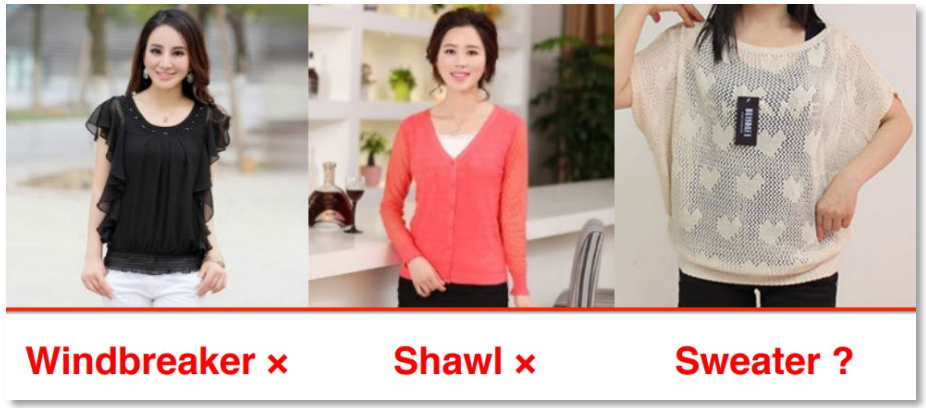
Introduction

| Success of deep learning *with noisy labels…?*

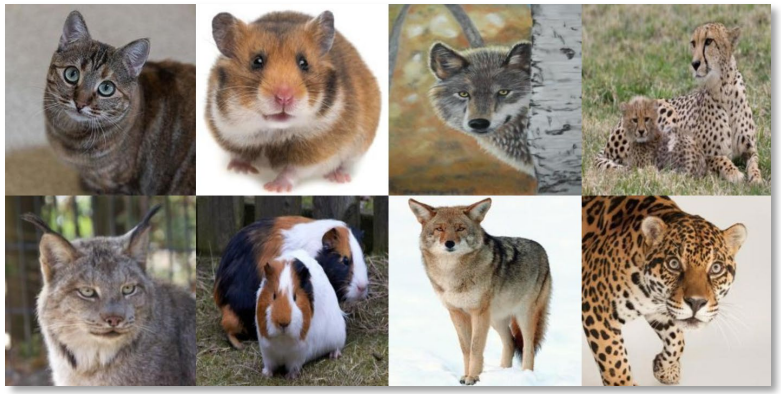
Chihuahua or Muffin



Clothing-1M dataset, noise rate ≈ 38.5%



Animal-10N dataset, noise rate ≈ 8.0%



Introduction

| Success of deep learning *with noisy labels…?*

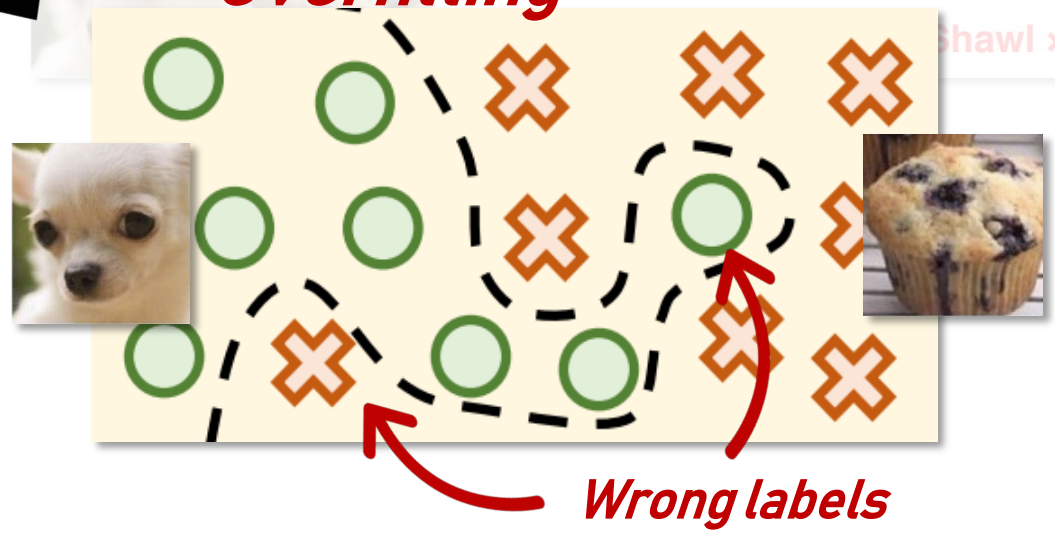
Chihuahua or Muffin

Clothing-1M dataset, noise rate ≈ 38.5%

Animal-10N dataset, noise rate ≈ 8.0%

The ratio of wrong labels** in real-world datasets is to range **from 8.0% to 38.5%

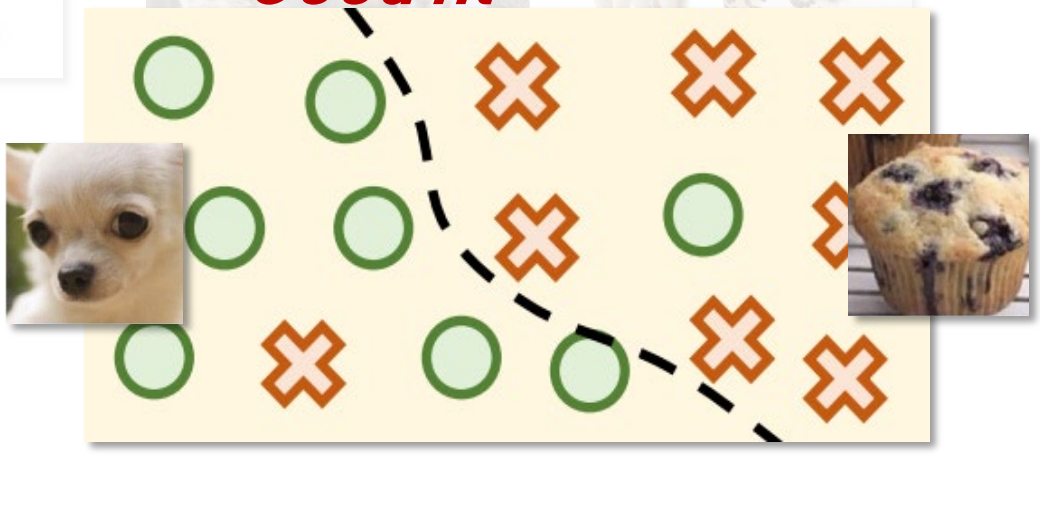
Overfitting



Wrong labels

How??

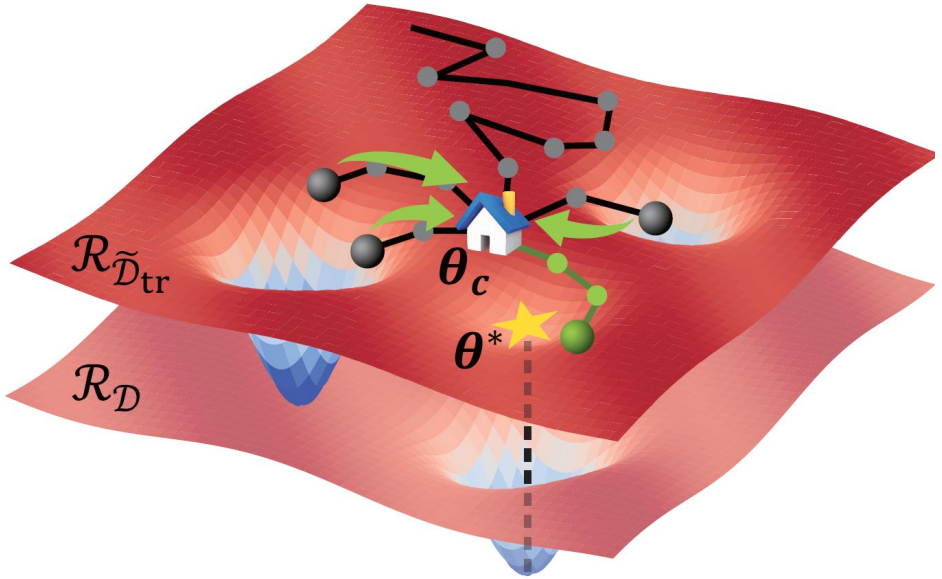
Good fit

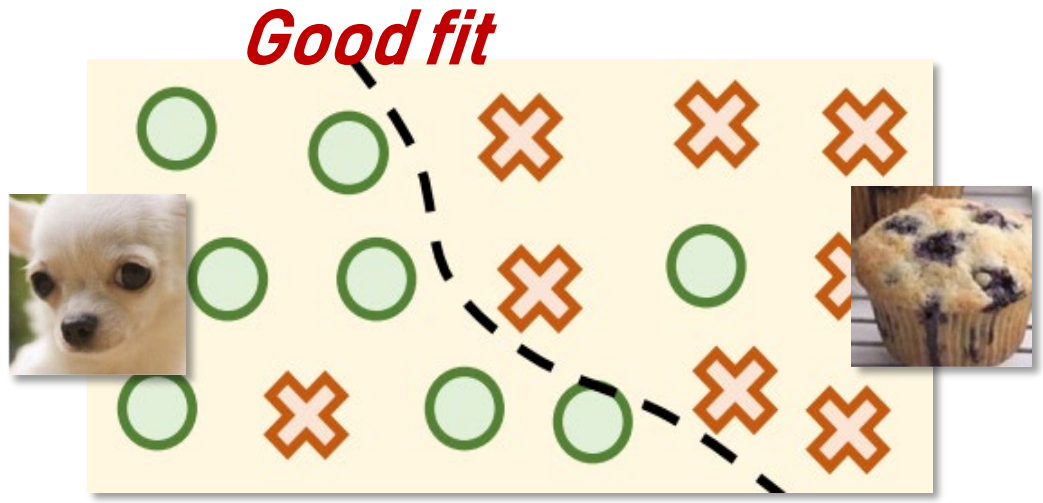


Introduction

| Stochastic restarting to overcome the overfitting issue

*Can we overcome **the overfitting issue** through the method of **statistical physics**?*

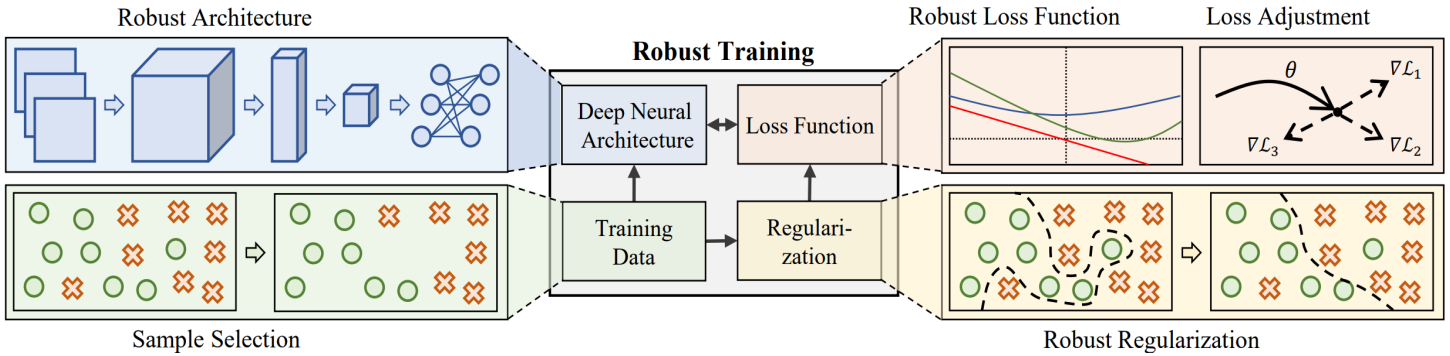




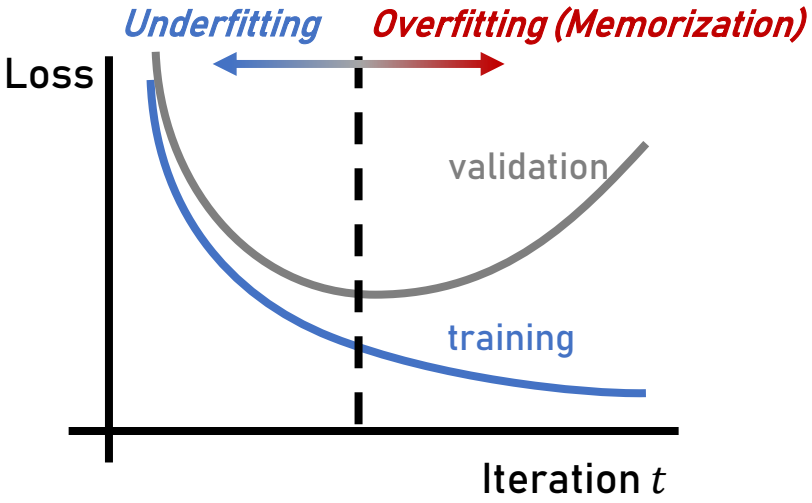
Stochastic restarting* + *Deep learning

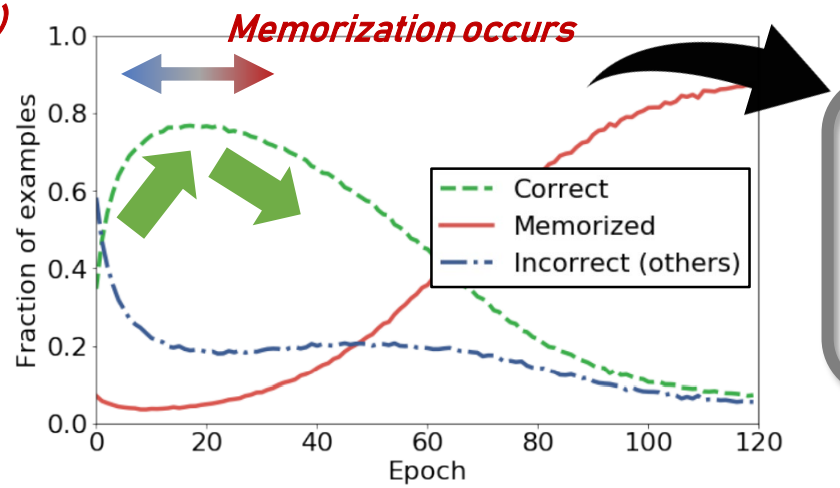
Previous works

Deep learning with noisy labels



H. Song et al., IEEE Trans. Neur. Netw. Learn. Sys., 34, pp.8135-8153 2023





S. Liu et al., NeurIPS, 33, pp. 20331–20342, 2020.

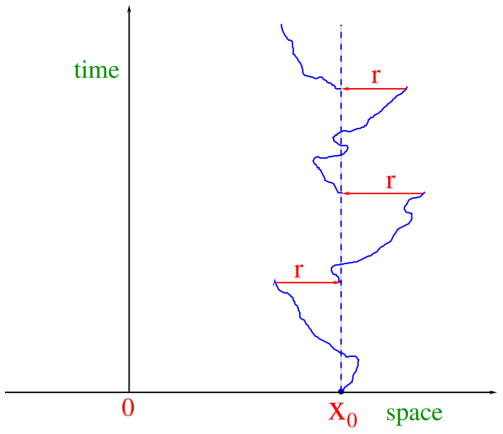
Firstly learn the general pattern

then gradually memorize

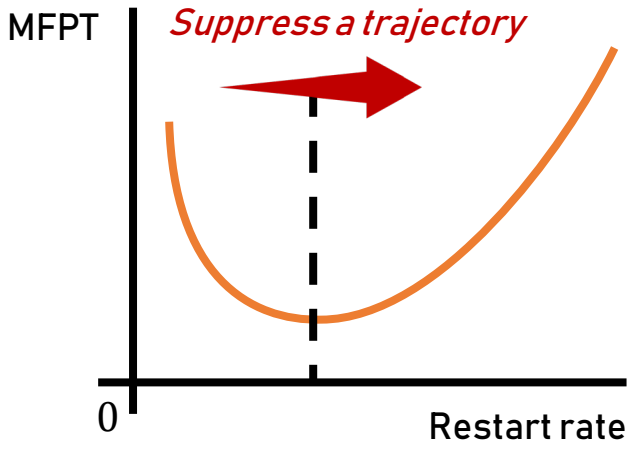
the specific pattern

Previous works

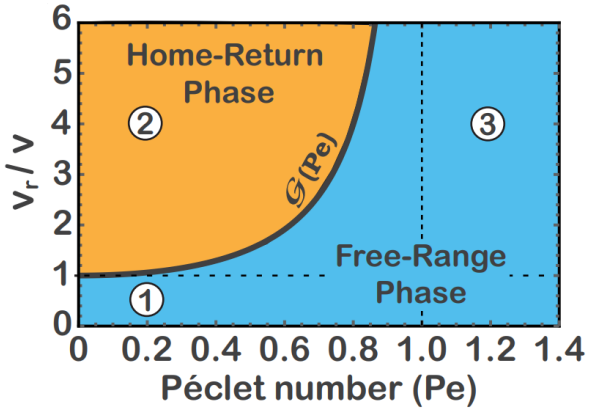
| Stochastic restarting in search process

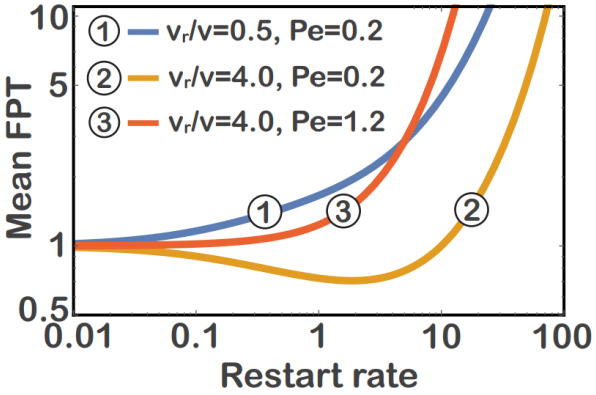


M. R. Evans et al., Phys. Rev. Lett., 106, 160601, 2011



Peclet number $Pe \equiv Lv/2D$





A. Pal et al., Phys. Rev. Research, 2, 043174, 2020

*When **the stochasticity** is sufficiently larger than **the drift toward a target**, restarting can be beneficial for random searches*

Description

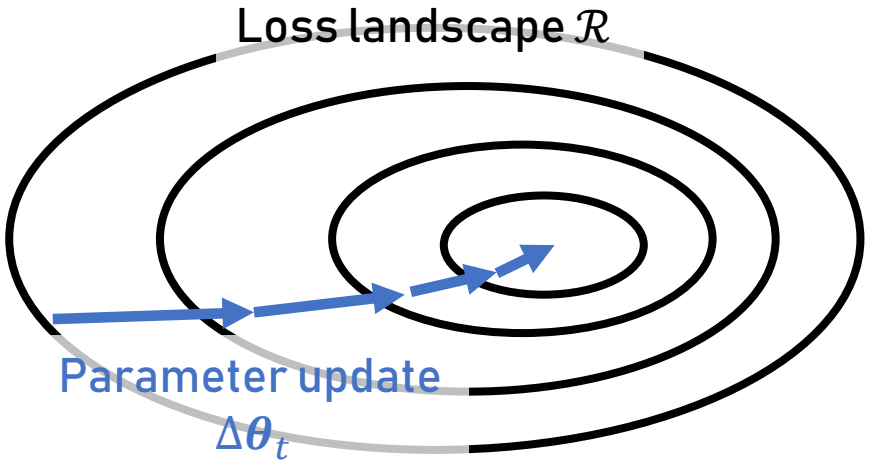
| SGD dynamics → Langevin dynamics

Notation & Definition

Dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i$ is the input data (e.g., image) and $\boldsymbol{y}_i$ is the label (e.g., cat or dog),

Loss function $\mathcal{L}_i(\boldsymbol{\theta}) \equiv \mathcal{L}(x_i, y_i; \boldsymbol{\theta})$ for a sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, Risk $\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}) \equiv \langle \mathcal{L}(\boldsymbol{\theta}) \rangle_{\mathcal{D}}$

Goal of DNNs – ***Minimizing the risk $\mathcal{R}$ for general dataset***



(Full-batch) Gradient Descent

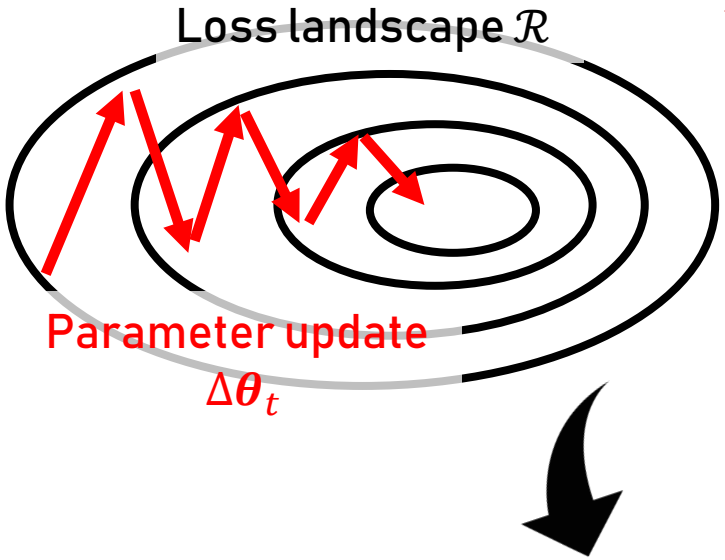
the learning rate η

$$\Delta\boldsymbol{\theta}_t = -\frac{\eta}{N} \sum_{(x_i, y_i) \in \mathcal{D}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_i(\boldsymbol{\theta}_t) = -\eta \boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{R}(\boldsymbol{\theta}_t)$$

similar with $\Delta\boldsymbol{x} = -\boldsymbol{\nabla}_{\boldsymbol{x}} V(\boldsymbol{x}) \Delta t$

Description

| SGD dynamics → Langevin dynamics



Stochastic Gradient Descent (SGD)

$$\Delta\boldsymbol{\theta}_t = -\frac{\eta}{B}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{B}_t}\boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_i(\boldsymbol{\theta}_t)$$

$$= \underbrace{-\frac{\eta}{N}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{D}}\boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_i(\boldsymbol{\theta}_t)}_{\equiv -\boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}_t)\eta} + \underbrace{\left(\frac{\eta}{N}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{D}}\boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_i(\boldsymbol{\theta}_t) - \frac{\eta}{B}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{B}_t}\boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_i(\boldsymbol{\theta}_t)\right)}_{\equiv \xi_t\sqrt{\eta} \text{ with } \langle\xi_t\rangle_{\mathcal{D}} = 0 \text{ and } \langle\xi_t^2\rangle \propto \eta/(2B)}$$

Dataset $\mathcal{D}$

Random sample → ***(Mini) batch*** $\mathcal{B}_t$

The SGD dynamics *can be described by the (overdamped) Langevin equation*

$$\Delta\boldsymbol{\theta}_t = -\boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}_t)\eta + \boldsymbol{\xi}_t\sqrt{\eta}$$

(Overdamped) Langevin eq. $\Delta x_t = -\nabla_\theta V(x_t)\Delta t + \xi_t\sqrt{\Delta t}$

Result

| SGD dynamics on noisy dataset $\widetilde{D}$

For noisy dataset $\widetilde{D} = [\mathcal{D}^c\ (= \text{correct}), \mathcal{D}^w (= \text{wrong})]$ with noise ratio τ,

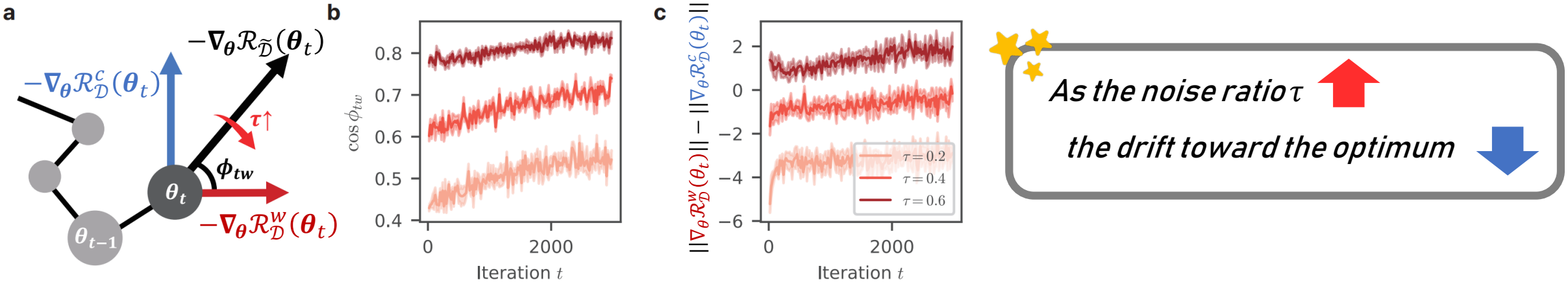
$$\Delta\boldsymbol{\theta}_t = -\nabla_{\boldsymbol{\theta}}\mathcal{R}_{\widetilde{\mathcal{D}}}(\boldsymbol{\theta}_t)\eta + \boldsymbol{\xi}_t\sqrt{\eta}$$

$$= -\Big[\underbrace{\nabla_{\boldsymbol{\theta}}\mathcal{R}^c_{\widetilde{\mathcal{D}}}(\boldsymbol{\theta}_t)}_{\text{Correct}} + \underbrace{\nabla_{\boldsymbol{\theta}}\mathcal{R}^w_{\widetilde{\mathcal{D}}}(\boldsymbol{\theta}_t)}_{\text{Wrong}}\Big]\eta + \boldsymbol{\xi}_t\sqrt{\eta}$$

where $\nabla_{\boldsymbol{\theta}}\mathcal{R}^c_{\widetilde{\mathcal{D}}} = (1-\tau)\nabla_{\boldsymbol{\theta}}\mathcal{R}_{\mathcal{D}^c}$ and $\nabla_{\boldsymbol{\theta}}\mathcal{R}^w_{\widetilde{\mathcal{D}}} = \tau\nabla_{\boldsymbol{\theta}}\mathcal{R}_{\mathcal{D}^w}$

Due to the high-dimensionality of $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}}\mathcal{R}^c_{\mathcal{D}}(\boldsymbol{\theta}_t) \perp \nabla_{\boldsymbol{\theta}}\mathcal{R}^c_{\mathcal{D}}(\boldsymbol{\theta}_t)$$

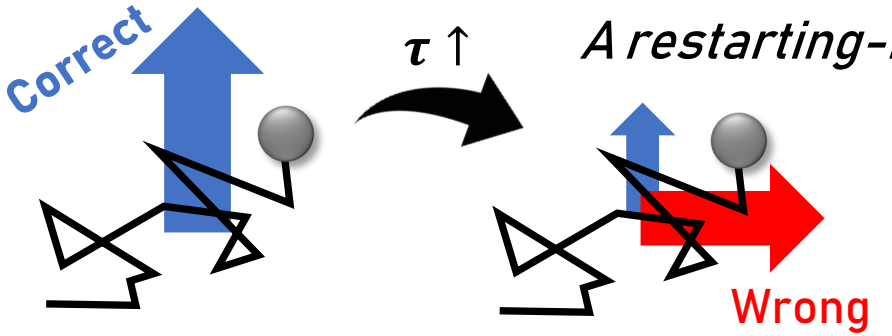


As the noise ratio τ ↑

the drift toward the optimum ↓

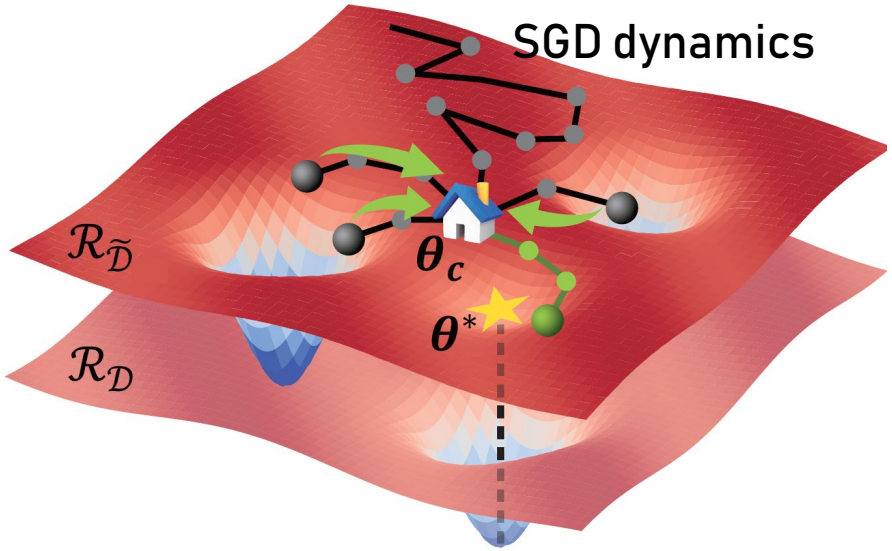
Result

| SGD dynamics on noisy dataset $\widetilde{D}$



A restarting-beneficial criterion…?

$$\boldsymbol{\theta}_{t+1} = \begin{cases} \boldsymbol{\theta}_c, & \text{with probability } \boldsymbol{r}, \\ \boldsymbol{\theta}_t - \boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{R}_{\widetilde{\mathcal{D}}}(\boldsymbol{\theta}_t)\eta + \boldsymbol{\xi}_t\sqrt{\eta}\,, & \text{otherwise.} \end{cases}$$



Main Questions

- Where should we restart (i.e., the restart point $\boldsymbol{\theta}_c$)?

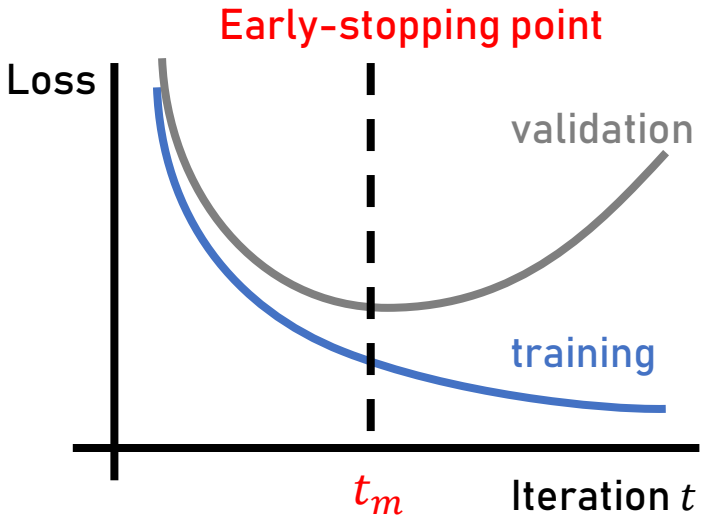
- Which factors affect the advantageous of restarting?

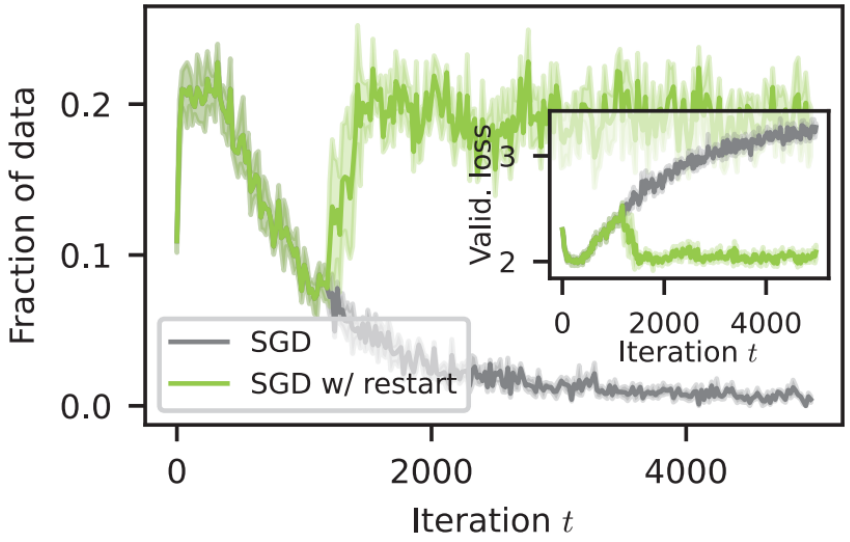
Result

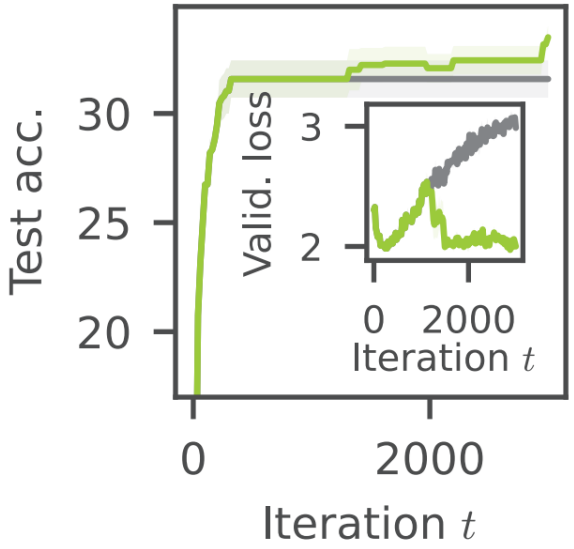
| Where should we restart?

The early-stopping is simple and powerful method → ***Let θ_c be the early-stopping point !!***

Expt. Setup] ciFAIR-10 dataset, noise ratio $\tau = 0.4$







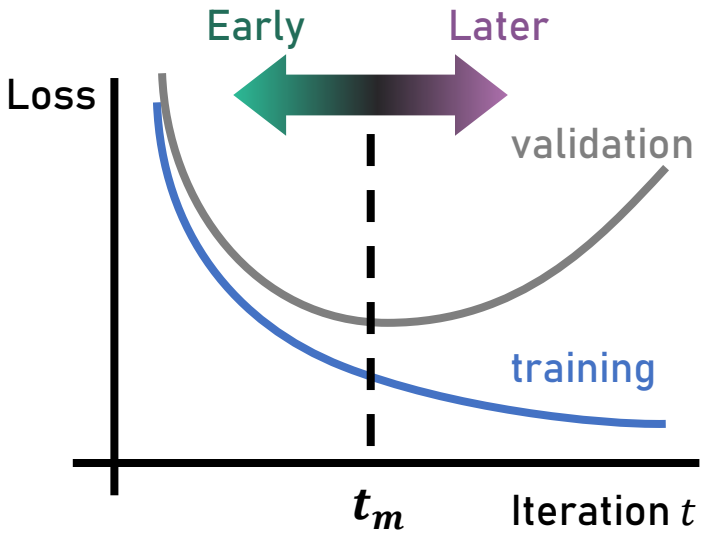
Stochastic restarting significantly improves the generalization performance on dataset with noisy labels…!

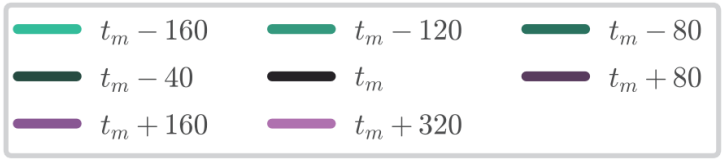
Result

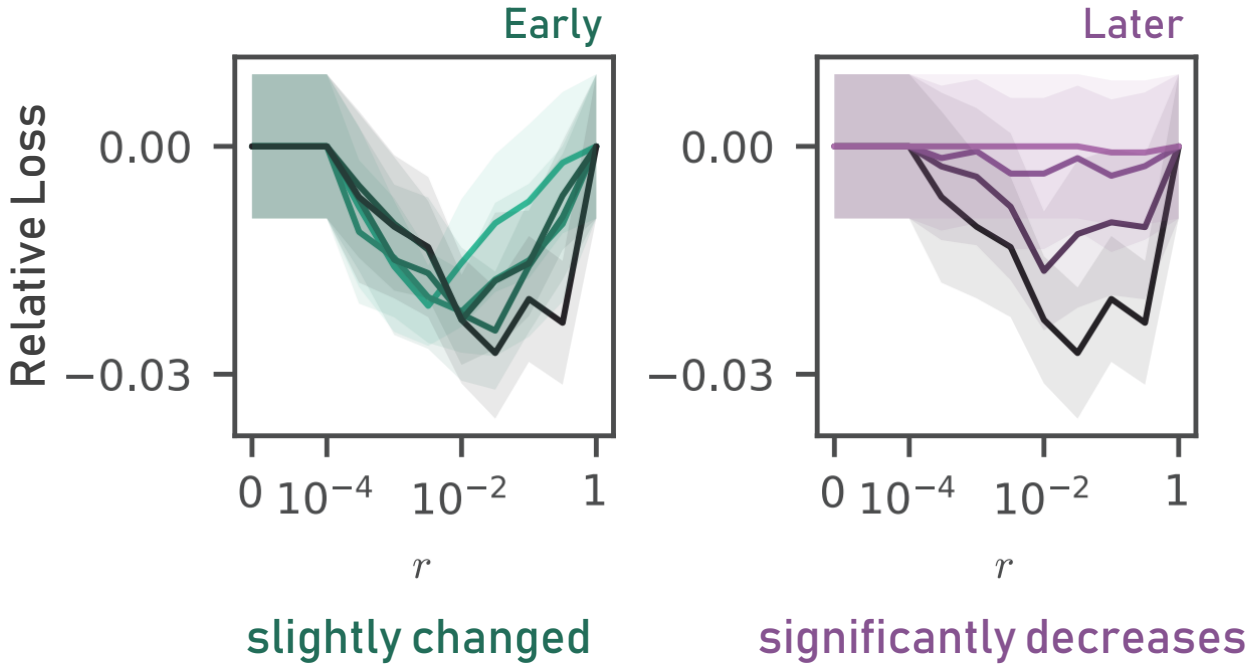
| Where should we restart?

The early-stopping is simple and powerful method → ***Let θ_c be the early-stopping point !!***

Expt. Setup] ciFAIR-10 dataset, noise rate $\tau = 0.4$







slightly changed

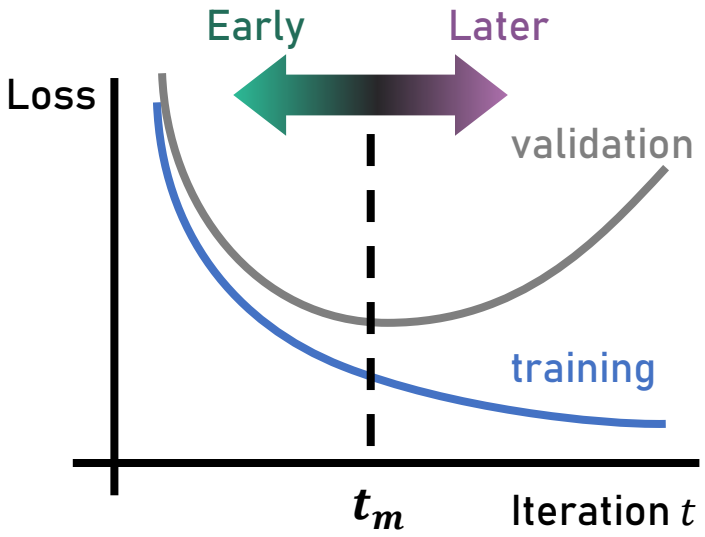
significantly decreases

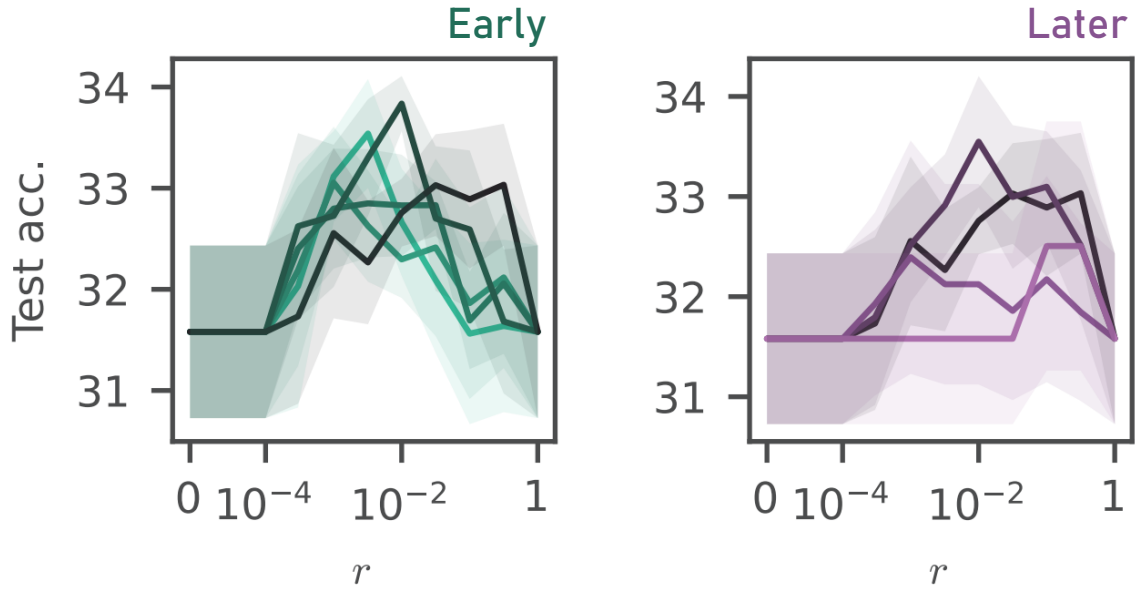
Result

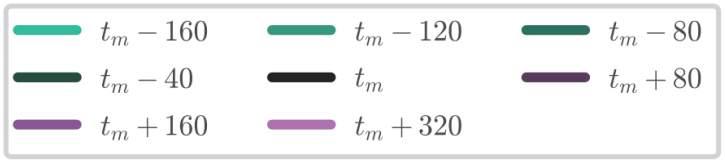
| Where should we restart?

The early-stopping is simple and powerful method → ***Let θ_c be the early-stopping point !!***

Expt. Setup] ciFAIR-10 dataset, noise rate $\tau = 0.4$





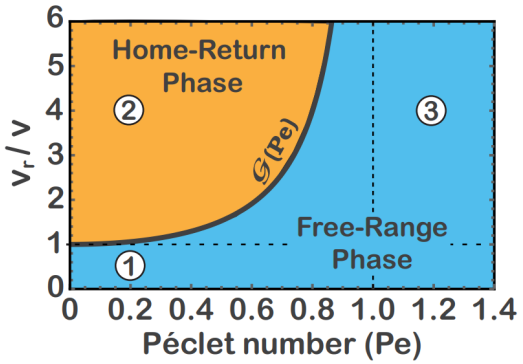


The restart point θ_c should be chosen in the early-learning stage!

Result

| Which factors affect the advantageous of restarting?

⚠ To improve the performance, we now adaptively move θ_c to the newly updated early-stopping point.

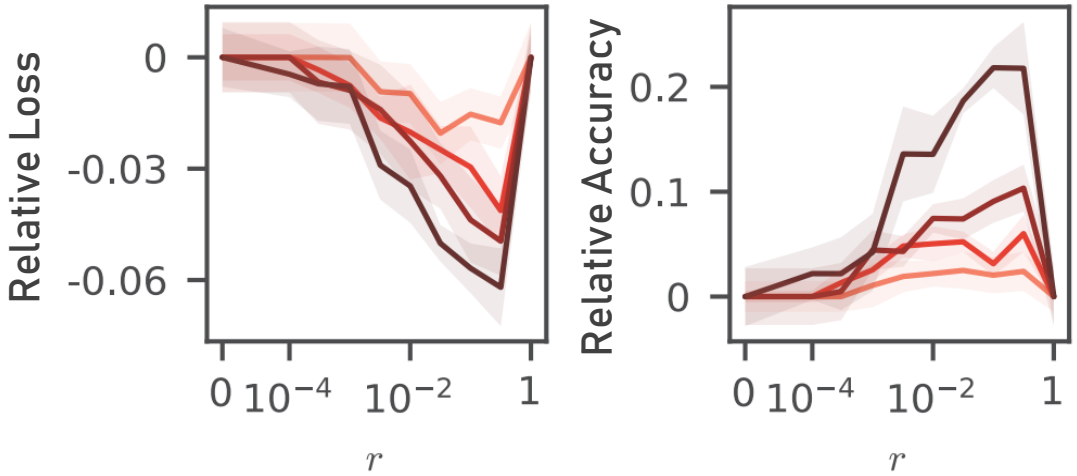


A. Pal et al., Phys. Rev. Research, 2, 043174, 2020

Peclet number $Pe \propto \frac{\text{(Drift toward a target)}}{\text{(Diffusion)}}$

SGD dynamics $\Delta\theta_t = -\left[\nabla_{\boldsymbol{\theta}} \mathcal{R}^{c}_{\tilde{\mathcal{D}}}(\boldsymbol{\theta}_t) + \nabla_{\boldsymbol{\theta}} \mathcal{R}^{w}_{\tilde{\mathcal{D}}}(\boldsymbol{\theta}_t)\right]\eta + \boldsymbol{\xi}_t\sqrt{\eta}$

(1) Noise rate τ (2) Batch size B



As the noise ratio τ ↑

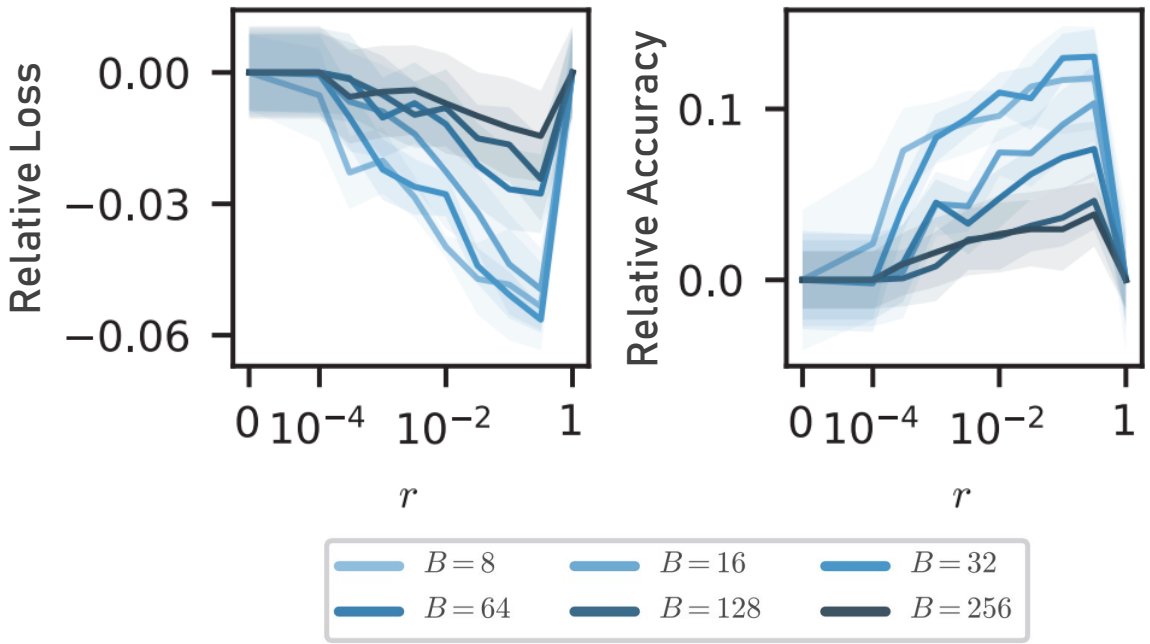
the improvement of performance ↑

Result

| Which factors affect the advantageous of restarting?

Peclet number $Pe \propto \frac{(\text{Drift toward a target})}{(\text{Diffusion})}$

(1) Noise rate τ (2) Batch size B ※ $(\text{Diffusion}) \propto 1/B$



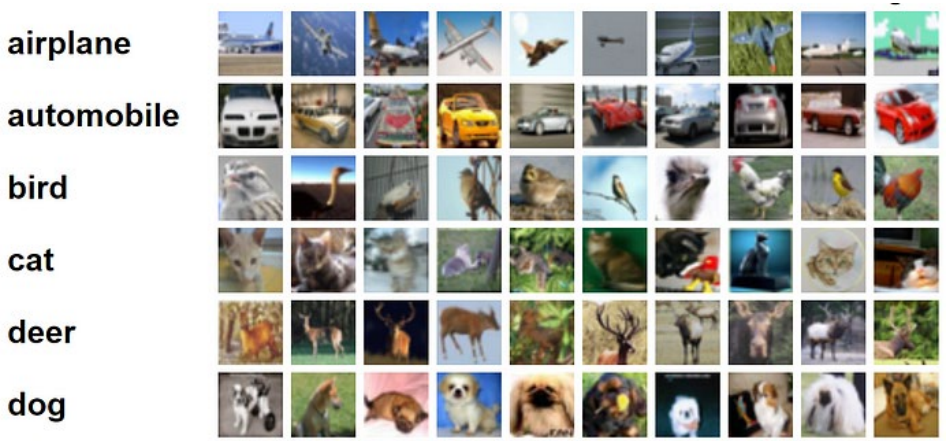
As the batch size B ↑

the improvement of performance ↓

Result

| Improved results on benchmark datasets & methods

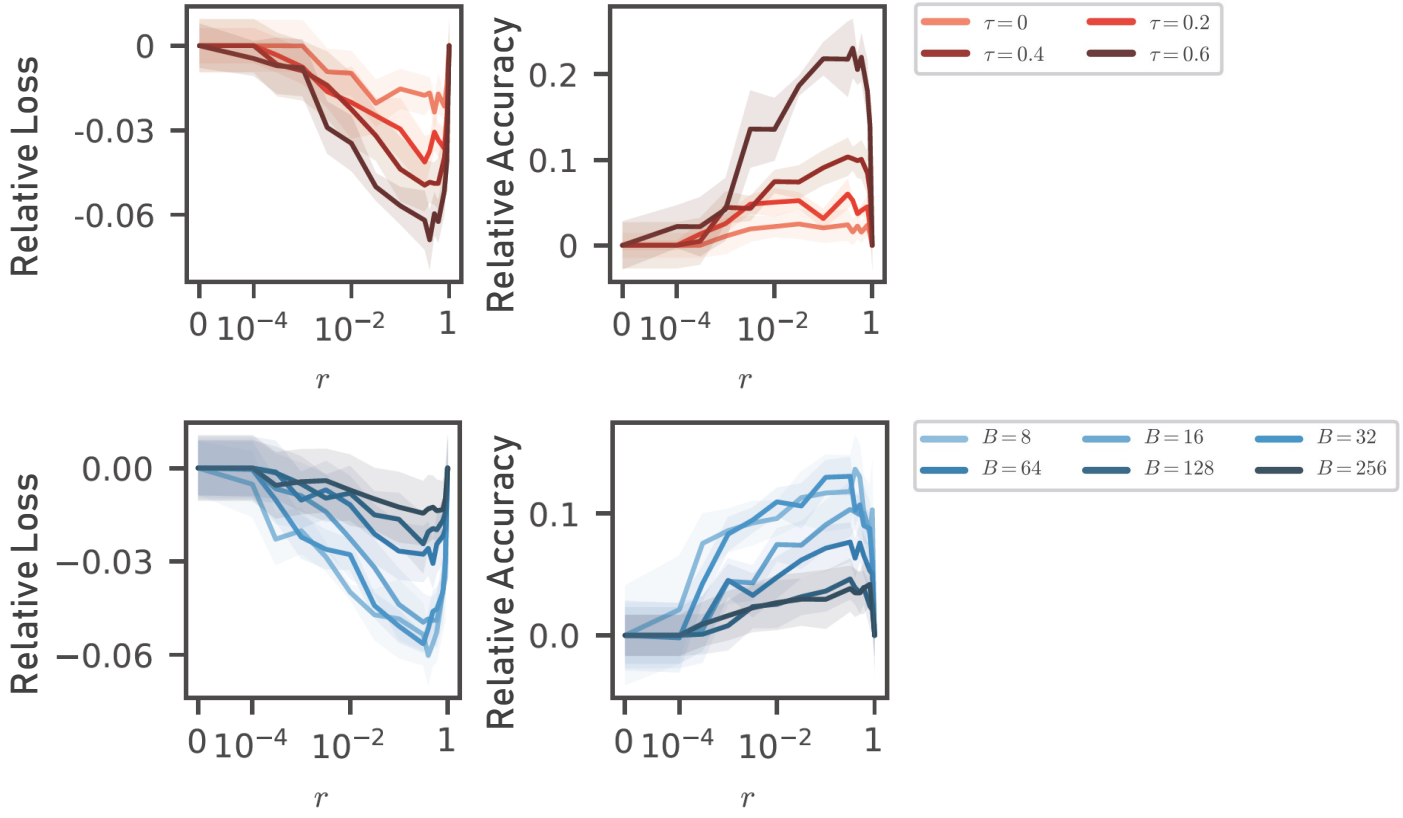
Expt. Setup] Dataset *CIFAR-10* and *CIFAR-100*



...

CIFAR-10: 50k, ***10*** classes

CIFAR-100: 50k, ***100*** classes



Result

| Improved results on benchmark datasets & methods

Expt. Setup] Dataset - *CIFAR-10* and *CIFAR-100*, *restart probability* $r = 0.001$

Dataset	*Method*	Noise rate 0.2		Noise rate 0.4		Noise rate 0.6	
		No	Restart	No	Restart	No	Restart
CIFAR-10	CE	84.8 ± 0.4	$\mathbf{90.0 \pm 0.5}^{***}$	80.8 ± 0.7	$\mathbf{86.3 \pm 0.7}^{***}$	72.9 ± 0.7	$\mathbf{79.5 \pm 0.9}^{***}$
	PartRestart	—	$\mathbf{88.5 \pm 0.4}^{***}$	—	$\mathbf{83.7 \pm 0.9}^{***}$	—	$\mathbf{75.1 \pm 0.4}^{***}$
	MAE	91.0 ± 0.2	91.0 ± 0.2	85.6 ± 3.3	$\mathbf{85.9 \pm 3.3}$	67.1 ± 6.8	$\mathbf{67.3 \pm 6.7}$
	GCE	90.6 ± 0.1	$\mathbf{91.0 \pm 0.2}^{*}$	85.3 ± 0.4	$\mathbf{87.1 \pm 0.3}^{***}$	76.2 ± 0.5	$\mathbf{79.1 \pm 0.8}^{***}$
	SCE	91.3 ± 0.3	$\mathbf{91.4 \pm 0.1}$	86.6 ± 0.2	$\mathbf{87.9 \pm 0.2}^{***}$	80.1 ± 0.5	$\mathbf{81.6 \pm 0.6}^{**}$
	ELR	$\mathbf{91.5 \pm 0.2}$	91.4 ± 0.2	87.8 ± 0.4	87.8 ± 0.4	81.1 ± 0.3	81.1 ± 0.3
CIFAR-100	CE	62.7 ± 5.6	$\mathbf{64.5 \pm 1.5}$	45.6 ± 2.3	$\mathbf{56.9 \pm 3.0}^{***}$	32.9 ± 1.6	$\mathbf{44.1 \pm 3.1}^{***}$
	PartRestart	—	$\mathbf{64.0 \pm 0.9}$	—	$\mathbf{55.5 \pm 2.1}^{***}$	—	$\mathbf{43.5 \pm 1.5}^{***}$
	MAE	19.9 ± 2.8	19.9 ± 2.8	$\mathbf{11.0 \pm 3.8}$	10.8 ± 3.7	6.7 ± 1.2	$\mathbf{6.8 \pm 1.3}$
	GCE	68.3 ± 0.4	$\mathbf{69.3 \pm 0.2}^{**}$	61.2 ± 0.6	$\mathbf{63.2 \pm 0.3}^{***}$	50.1 ± 0.6	$\mathbf{53.1 \pm 0.8}^{***}$
	SCE	54.5 ± 1.2	$\mathbf{62.0 \pm 1.0}^{***}$	44.4 ± 1.4	$\mathbf{52.8 \pm 0.7}^{***}$	29.3 ± 2.5	$\mathbf{36.3 \pm 1.8}^{***}$
	ELR	66.4 ± 0.4	$\mathbf{67.2 \pm 1.2}$	57.4 ± 1.0	$\mathbf{60.9 \pm 2.0}^{**}$	47.2 ± 1.4	$\mathbf{48.0 \pm 2.6}$

Improved performance!

Consistently improve the generalization performance & Flexibility of restarting strategy…!

Conclusion

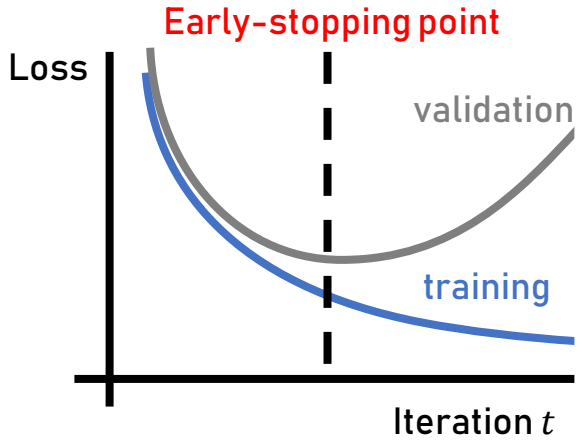
| Summary & Discussion

- We empirically demonstrate that ***stochastic restarting can mitigate the overfitting issue*** in deep learning on noisy dataset.

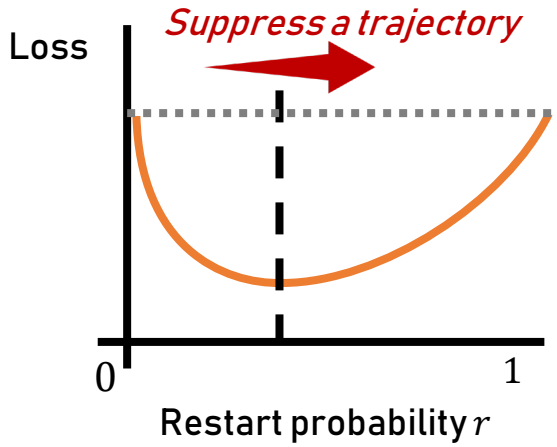
- We theoretically identified ***factors that influence the effectiveness of restarting*** and empirically validated our method as well as the impact of these factors.

Limitations

(i) *Double descent*



(ii) *Theoretical analysis & Optimal r…?*



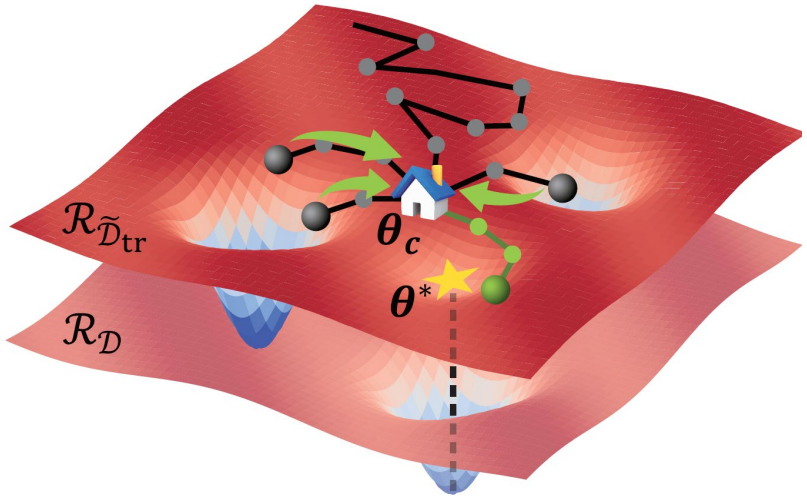
Can we find the optimal relation

e.g., $\sigma(T)/\langle T \rangle = 1$ …?

SCAN ME

Thank you for listening!

Presenter - Youngkyoung Bae ✉☐ *dudrud3085@gmail.com*



Thank our coauthors and advisors!!



Hawoong Jeong
KAIST, Korea



Yeongwoo Song
KAIST, Korea



Yongjoo Baek
SNU, Korea