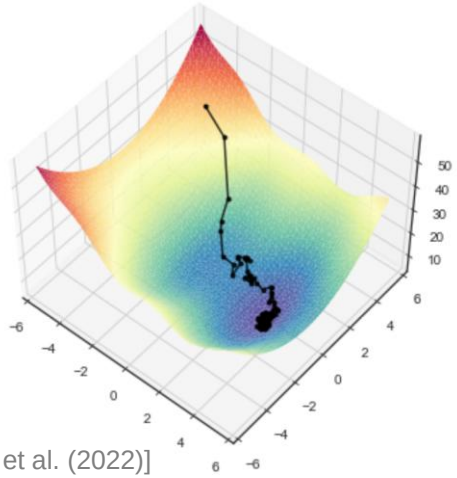
UNIVERSITÄT
LEIPZIG

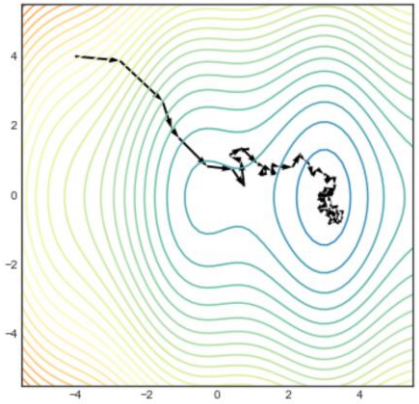
Anti-Correlated Noise in Epoch-Based Stochastic Gradient Descent and its Implications

Prague, 10.04.2025

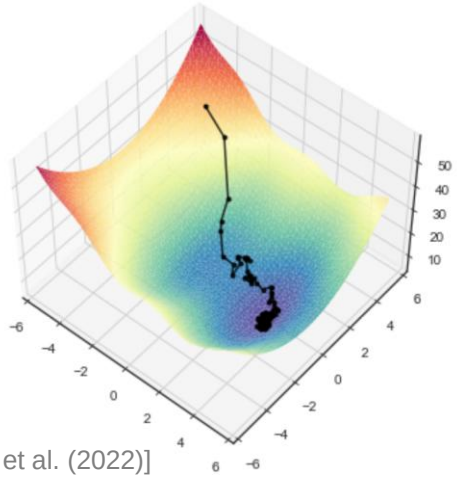
Marcel Kühn and Bernd Rosenow – Institute for Theoretical Physics



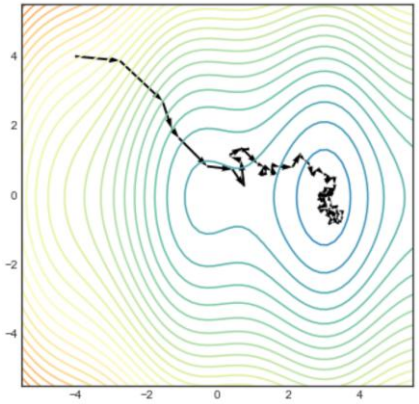
[Berner et al. (2022)]



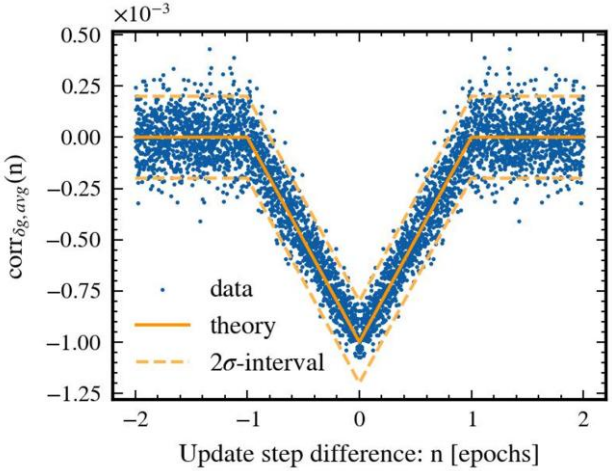
Stochastic Gradient Descent



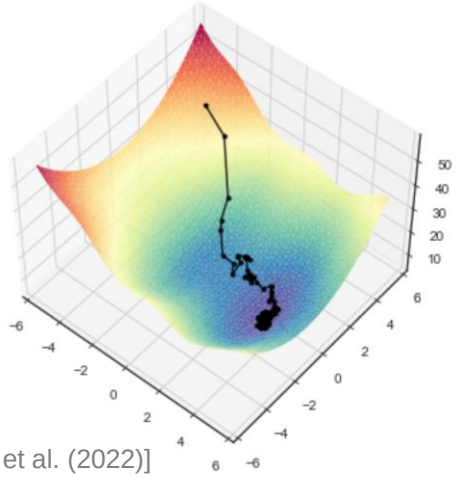
[Berner et al. (2022)]



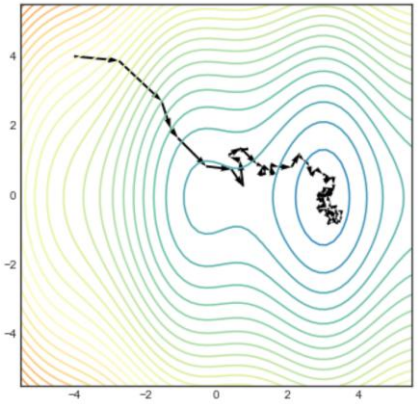
Stochastic Gradient Descent



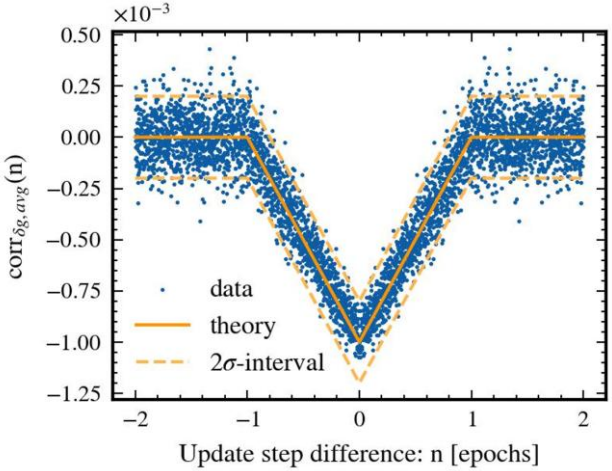
Anti-correlated Noise



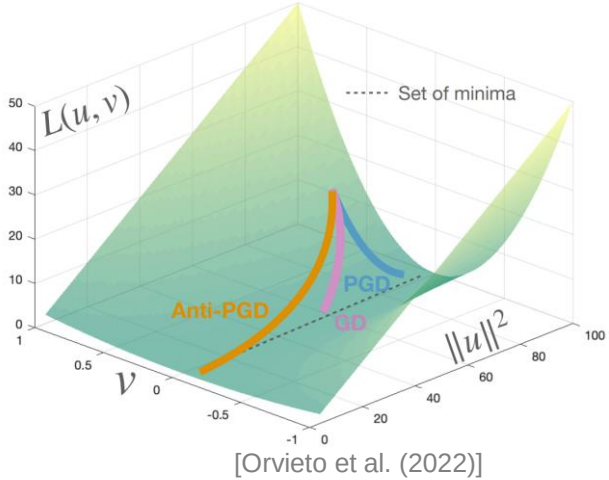
[Berner et al. (2022)]



Stochastic Gradient Descent



Anti-correlated Noise



[Orvieto et al. (2022)]

Implications for Training-Dynamics

Supervised Learning with Datasets

- learning from examples

$\{x_n, y_n \mid n=1, \ldots, N\}$ ... dataset

x ... example

y ... label

Supervised Learning with Datasets

- learning from examples

$$\{x_n, y_n \mid n=1, \ldots, N\} \quad \ldots \text{ dataset}$$
$$x \quad \ldots \text{ example}$$
$$y \quad \ldots \text{ label}$$

Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A:	No. [They were in the finals once, and lost.]

[Clark et al. (2019)]

FIG 4.
BoolQ Dataset

Supervised Learning with Datasets

- learning from examples

$$\{x_n, y_n \mid n=1, \ldots, N\} \quad \ldots \text{ dataset}$$
$$x \quad \ldots \text{ example}$$
$$y \quad \ldots \text{ label}$$

Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A:	No. [They were in the finals once, and lost.]

FIG 4.
BoolQ Dataset

[Clark et al. (2019)]

[Krizhevsky and Hinton (2009)]

FIG 5. CIFAR10 Dataset

Supervised Learning with Datasets

- learning from examples

$$\{x_n, y_n \mid n=1, \ldots, N\} \quad \ldots \text{ dataset}$$
$$x \quad \ldots \text{ example}$$
$$y \quad \ldots \text{ label}$$

Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A:	No. [They were in the finals once, and lost.]

[Clark et al. (2019)]

FIG 4.
BoolQ Dataset

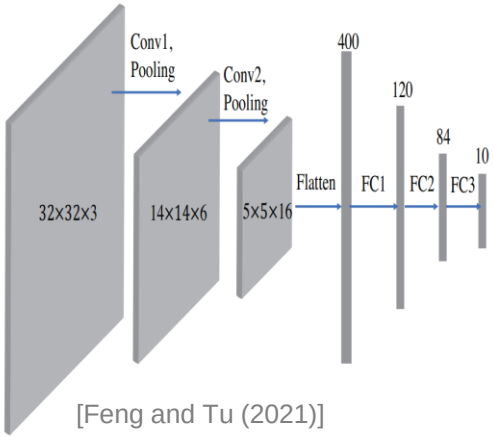
[Krizhevsky and Hinton (2009)]

FIG 5. CIFAR10 Dataset

Neural networks

$f(\boldsymbol{\theta}, x)$ ... network

$\boldsymbol{\theta} \in \mathbb{R}^d$ ... network weights



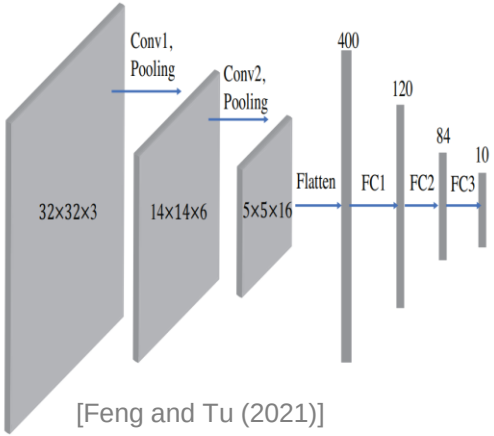
[Feng and Tu (2021)]

FIG 6.
LeNet

Neural networks

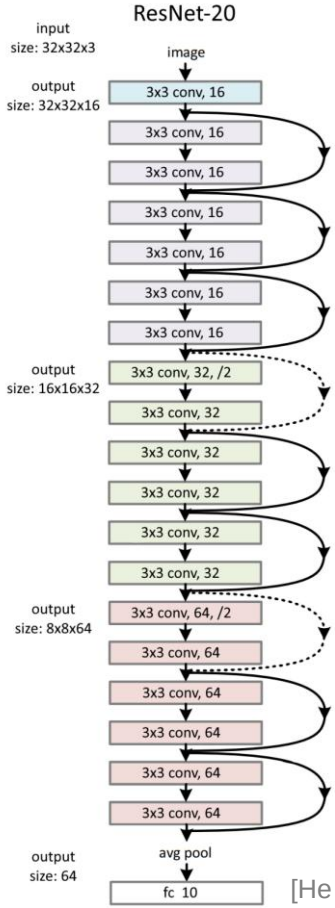
$f(\boldsymbol{\theta}, x)$ ... network

$\boldsymbol{\theta} \in \mathbb{R}^d$ ... network weights



[Feng and Tu (2021)]

FIG 6.
LeNet



[He et al. (2016)]

FIG 7.
ResNet

Training: Loss Landscape

$$\boldsymbol{\theta} \in \mathbb{R}^d \quad \ldots \text{ network weights}$$

$$l_n(\boldsymbol{\theta}) \coloneqq l(f(\boldsymbol{\theta}, x_n), y_n) \quad \ldots \text{ example loss}$$

$$L(\boldsymbol{\theta}) \coloneqq \frac{1}{N} \sum_{n=1}^{N} l_n(\boldsymbol{\theta}) \quad \ldots \text{ total loss}$$

- Quantify mismatch with loss

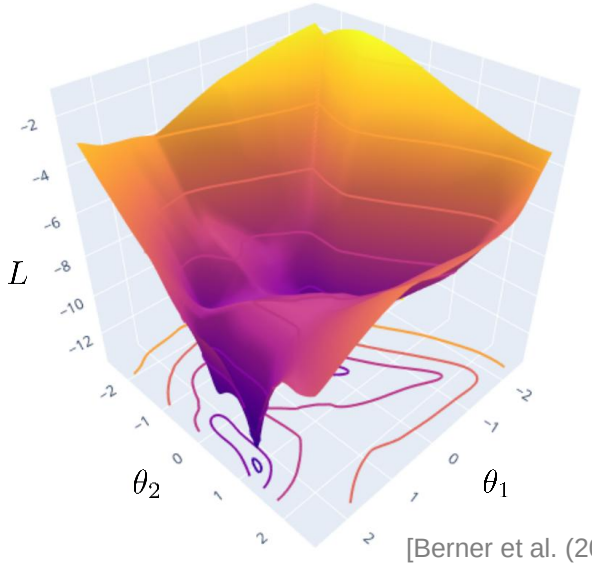
Training: Loss Landscape

$$\boldsymbol{\theta} \in \mathbb{R}^d \quad \ldots \text{ network weights}$$

$$l_n(\boldsymbol{\theta}) \coloneqq l(f(\boldsymbol{\theta}, x_n), y_n) \quad \ldots \text{ example loss}$$

$$L(\boldsymbol{\theta}) \coloneqq \frac{1}{N} \sum_{n=1}^{N} l_n(\boldsymbol{\theta}) \quad \ldots \text{ total loss}$$

- Quantify mismatch with loss



[Berner et al. (2022)]

FIG 8. Loss Landscape

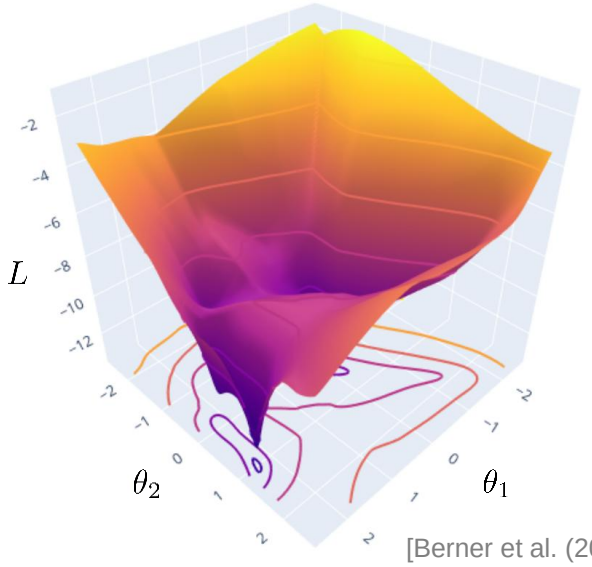
Training: Loss Landscape

$$\boldsymbol{\theta} \in \mathbb{R}^d \quad \ldots \text{ network weights}$$

$$l_n(\boldsymbol{\theta}) := l(f(\boldsymbol{\theta}, x_n), y_n) \quad \ldots \text{ example loss}$$

$$L(\boldsymbol{\theta}) := \frac{1}{N} \sum_{n=1}^{N} l_n(\boldsymbol{\theta}) \quad \ldots \text{ total loss}$$

- Quantify mismatch with loss
- Aim: minimize loss



[Berner et al. (2022)]

FIG 8. Loss Landscape

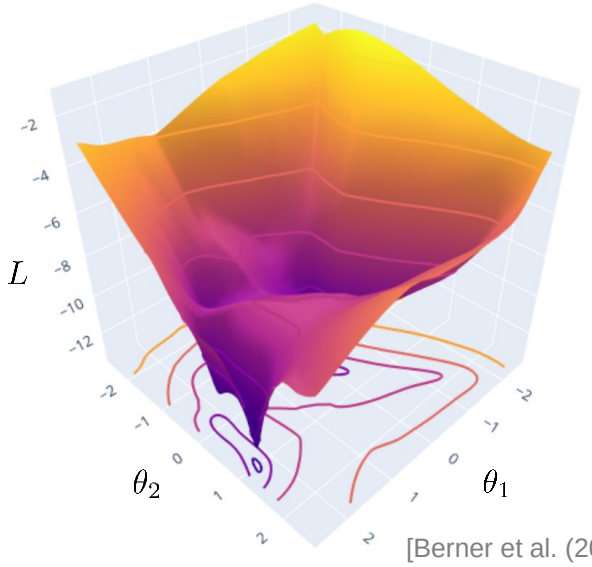
Training: Loss Landscape

$$\boldsymbol{\theta} \in \mathbb{R}^d \quad \ldots \text{ network weights}$$

$$l_n(\boldsymbol{\theta}) := l(f(\boldsymbol{\theta}, x_n), y_n) \quad \ldots \text{ example loss}$$

$$L(\boldsymbol{\theta}) := \frac{1}{N} \sum_{n=1}^{N} l_n(\boldsymbol{\theta}) \quad \ldots \text{ total loss}$$

- Quantify mismatch with loss
- Aim: minimize loss
- How? Gradient Descent!

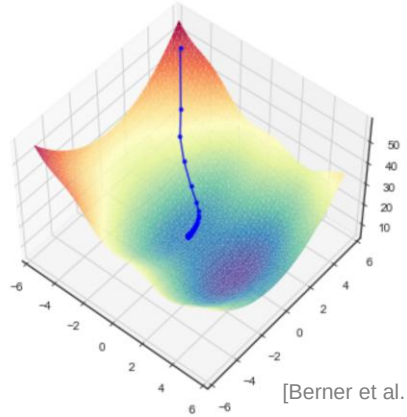


[Berner et al. (2022)]

FIG 8. Loss Landscape

Gradient Descent

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{\nabla} L(\boldsymbol{\theta}_k)$$



[Berner et al. (2022)]

FIG 9. Gradient Descent Trajectory

Gradient Descent

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla L(\boldsymbol{\theta}_k)$$

Stochastic Gradient Descent (SGD)

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \eta \mathbf{g}_k(\boldsymbol{\theta}_{k-1})$$

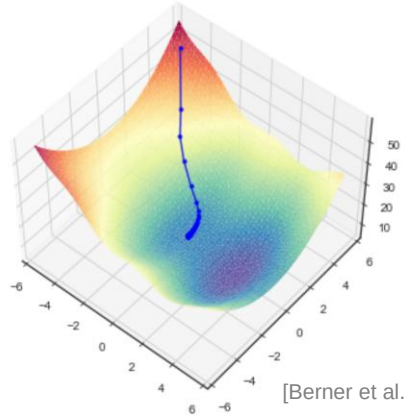
k ... update step index

η ... learning rate

S ... batch size

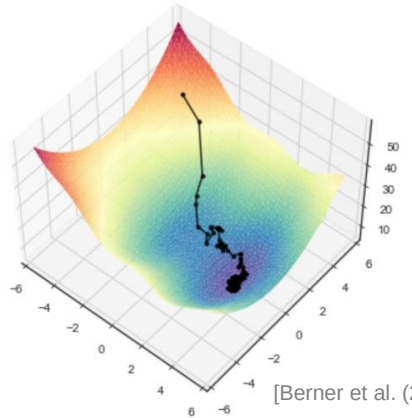
$\mathcal{B}_k = \{n_1, ..., n_S\}$ ... $S \ll N$ randomly selected examples

$(n_j \in \{1, ..., N\})$



[Berner et al. (2022)]

FIG 9. Gradient Descent Trajectory



[Berner et al. (2022)]

FIG 10. Stochastic Gradient Descent Trajectory

Gradient Descent

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla L(\boldsymbol{\theta}_k)$$

Stochastic Gradient Descent with Momentum (SGD)

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \qquad (1a)$$

$$\mathbf{v}_k = -\eta \mathbf{g}_k(\boldsymbol{\theta}_{k-1}) + \beta \mathbf{v}_{k-1} \qquad (1b)$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \mathbf{v}_k \qquad (1c)$$

k ... update step index

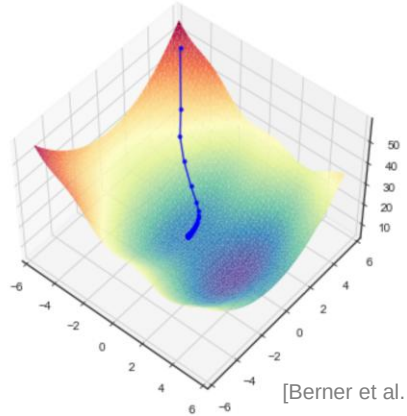
η ... learning rate

S ... batch size

β ... momentum parameter

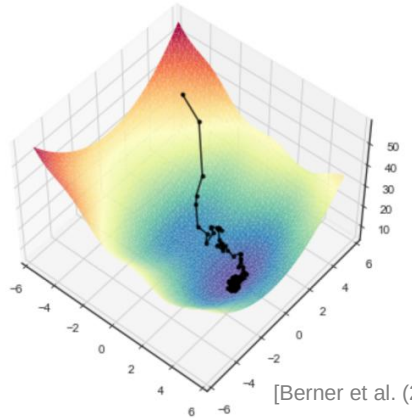
$\mathcal{B}_k = \{n_1, ..., n_S\}$ ... $S \ll N$ randomly selected examples

$(n_j \in \{1, ..., N\})$



[Berner et al. (2022)]

FIG 9. Gradient Descent Trajectory



[Berner et al. (2022)]

FIG 10. Stochastic Gradient Descent Trajectory

Stochastic Gradient Descent

- How to choose examples?

Stochastic Gradient Descent

- How to choose examples?
  - With or Without replacement (in epochs)

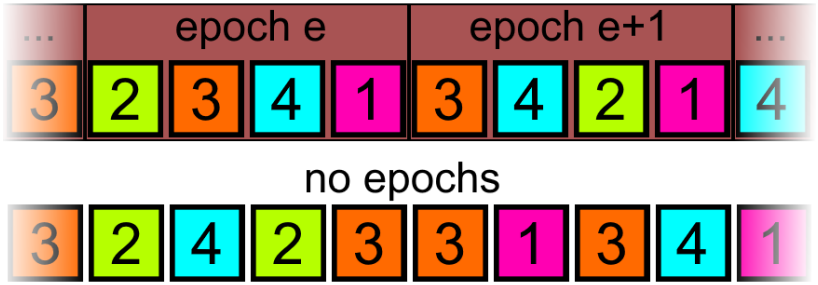


FIG 11. Epoch-based example selection

Stochastic Gradient Descent

- How to choose examples?
  - With or Without replacement (in epochs)

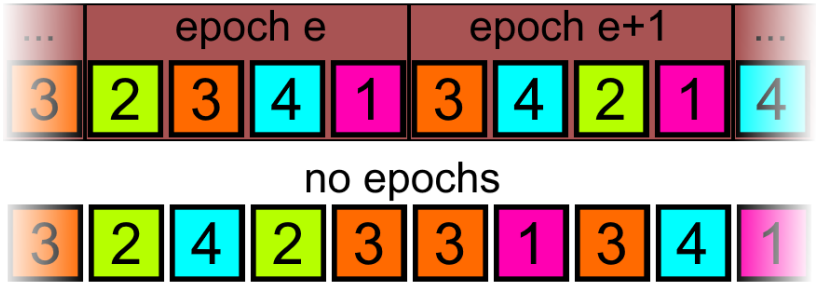


FIG 11. Epoch-based example selection

Stochastic Gradient Descent

- How to choose examples?
  - With or Without replacement (in epochs)
- Reason for anti-correlations

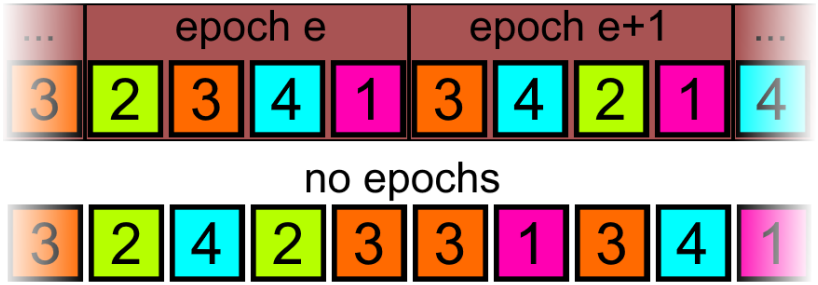


FIG 11. Epoch-based example selection

Anti-Correlations of the Noise

- gradient noise:

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \qquad \text{(1a)}$$

$$\boldsymbol{\delta}\mathbf{g}_k \coloneqq \mathbf{g}_k(\boldsymbol{\theta}) - \boldsymbol{\nabla} L(\boldsymbol{\theta})$$

$$\mathbf{C} = \operatorname*{cov}_k(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_k)$$

Anti-Correlations of the Noise

- gradient noise:

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \quad \text{(1a)}$$

$$\boldsymbol{\delta}\mathbf{g}_k \coloneqq \mathbf{g}_k(\boldsymbol{\theta}) - \boldsymbol{\nabla} L(\boldsymbol{\theta})$$

$$\mathbf{C} = \operatorname*{cov}_k(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_k)$$

- noise of one epoch has zero mean $\rightarrow$ anti-correlations:

$$\frac{1}{M} \sum_{k \in \text{epoch } e} \mathbf{g}_k(\boldsymbol{\theta}) = \boldsymbol{\nabla} L(\boldsymbol{\theta}) \quad M \coloneqq \frac{N}{S} \ldots \text{batches per epoch}$$

$$\Rightarrow \sum_{k \in \text{epoch } e} \boldsymbol{\delta}\mathbf{g}_k = 0$$

Anti-Correlations of the Noise

- gradient noise:

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \quad \text{(1a)}$$

$$\boldsymbol{\delta}\mathbf{g}_k \coloneqq \mathbf{g}_k(\boldsymbol{\theta}) - \boldsymbol{\nabla} L(\boldsymbol{\theta})$$

$$\mathbf{C} = \underset{k}{\mathrm{cov}}(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_k)$$

- noise of one epoch has zero mean $\rightarrow$ anti-correlations:

$$\frac{1}{M} \sum_{k \in \text{epoch } e} \mathbf{g}_k(\boldsymbol{\theta}) = \boldsymbol{\nabla} L(\boldsymbol{\theta}) \quad M \coloneqq \frac{N}{S} \ldots \text{batches per epoch}$$

$$\Rightarrow \sum_{k \in \text{epoch } e} \boldsymbol{\delta}\mathbf{g}_k = 0 \qquad \Rightarrow - \sum_{j \in e/\{k\}} \boldsymbol{\delta}\mathbf{g}_j = \boldsymbol{\delta}\mathbf{g}_k$$

Anti-Correlations of the Noise

- gradient noise:

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \qquad (1a)$$

$$\boldsymbol{\delta}\mathbf{g}_k := \mathbf{g}_k(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta})$$

$$\mathbf{C} = \underset{k}{\mathrm{cov}}(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_k)$$

- noise of one epoch has zero mean $\rightarrow$ anti-correlations:

$$\frac{1}{M} \sum_{k \in \text{epoch } e} \mathbf{g}_k(\boldsymbol{\theta}) = \nabla L(\boldsymbol{\theta}) \qquad M := \frac{N}{S} \ldots \text{batches per epoch}$$

$$\Rightarrow \sum_{k \in \text{epoch } e} \boldsymbol{\delta}\mathbf{g}_k = 0 \qquad \Rightarrow - \sum_{j \in e/\{k\}} \boldsymbol{\delta}\mathbf{g}_j = \boldsymbol{\delta}\mathbf{g}_k$$

- average over chance for two batches from same epoch

$$\underset{k}{\mathrm{cov}}(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_{k+h}) = \mathbf{C} \cdot \left(\delta_{h,0} - \mathbf{1}_{\{1,\ldots,M\}}(|h|) \frac{M - |h|}{M(M-1)} \right) \qquad (2)$$

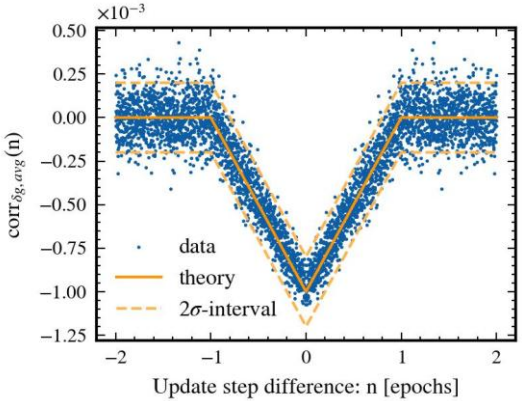


FIG 12. Noise autocorrelation: Examples selected without replacement

Anti-Correlations of the Noise

- gradient noise:

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S}\sum_{n\in\mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \qquad (1a)$$

$$\boldsymbol{\delta}\mathbf{g}_k \coloneqq \mathbf{g}_k(\boldsymbol{\theta}) - \boldsymbol{\nabla} L(\boldsymbol{\theta})$$

$$\mathbf{C} = \underset{k}{\mathrm{cov}}(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_k)$$

- noise of one epoch has zero mean $\rightarrow$ anti-correlations:

$$\frac{1}{M}\sum_{k\,\in\,\text{epoch}\,e} \mathbf{g}_k(\boldsymbol{\theta}) = \boldsymbol{\nabla} L(\boldsymbol{\theta}) \qquad M \coloneqq \frac{N}{S} \ldots \text{batches per epoch}$$

$$\Rightarrow \sum_{k\,\in\,\text{epoch}\,e} \boldsymbol{\delta}\mathbf{g}_k = 0 \qquad \Rightarrow -\sum_{j\,\in\,e/\{k\}} \boldsymbol{\delta}\mathbf{g}_j = \boldsymbol{\delta}\mathbf{g}_k$$

- average over chance for two batches from same epoch

$$\underset{k}{\mathrm{cov}}(\boldsymbol{\delta}\mathbf{g}_k, \boldsymbol{\delta}\mathbf{g}_{k+h}) = \mathbf{C}\cdot\left(\delta_{h,0} - \mathbf{1}_{\{1,\ldots,M\}}(|h|)\frac{M-|h|}{M(M-1)}\right) \qquad (2)$$

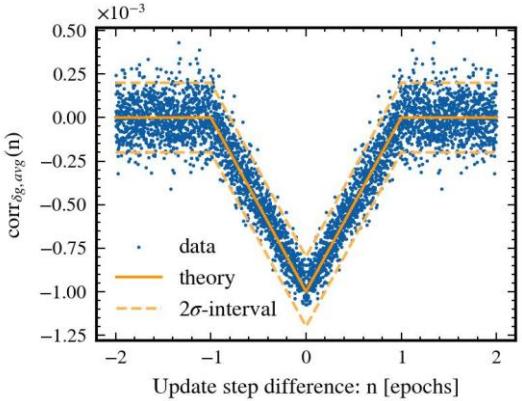


FIG 12. Noise autocorrelation: Examples selected without replacement

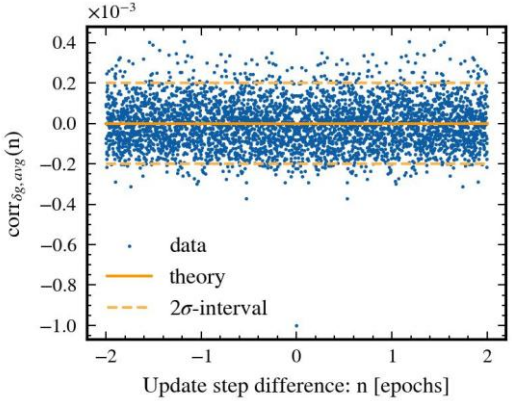


FIG 13. Noise autocorrelation: Examples selected with replacement

Anti-Correlations of the Noise

- calculations done for static weights
  - numerics also show anti-correlations at start of training

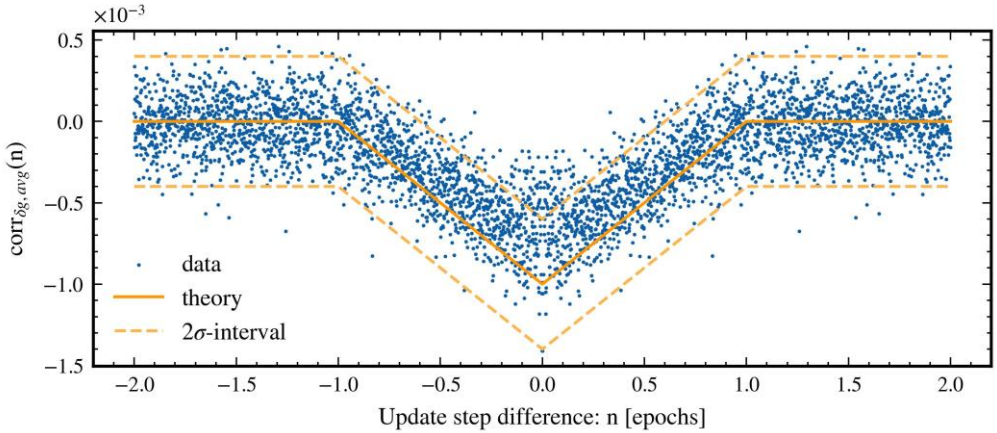


FIG 14. Noise autocorrelation: Beginning of training

Questions about Anti-correlations?

Questions about Anti-correlations?

- **Implications for Weight Fluctuations**

Weight Fluctuations at a Minimum

- late phase of training:

$$L(\boldsymbol{\theta}) \approx \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta} \qquad (3)$$

$\mathbf{H}$ . . . Hessian matrix

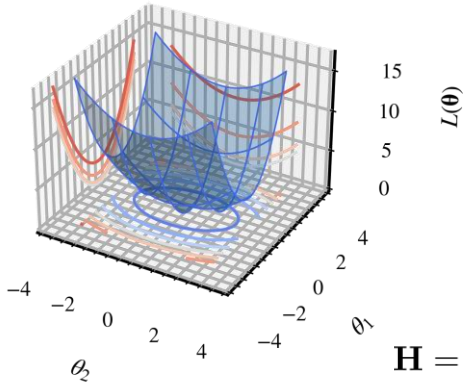


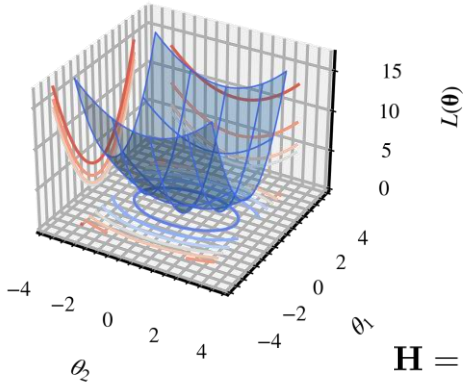
FIG 15. Toy loss and Hessian

Weight Fluctuations at a Minimum

- late phase of training:

$$L(\boldsymbol{\theta}) \approx \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta} \qquad (3)$$

$\mathbf{H}$ ... Hessian matrix



$$\mathbf{H} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

FIG 15. Toy loss and Hessian

- analyze weight covariance
  - resulting from stochasticity of SGD
  - measure for parameter exploration
  - insight about escape from minima

$$\boldsymbol{\Sigma} = \underset{k}{\mathrm{cov}}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k) \quad \text{... weight fluctuations} \qquad (4)$$

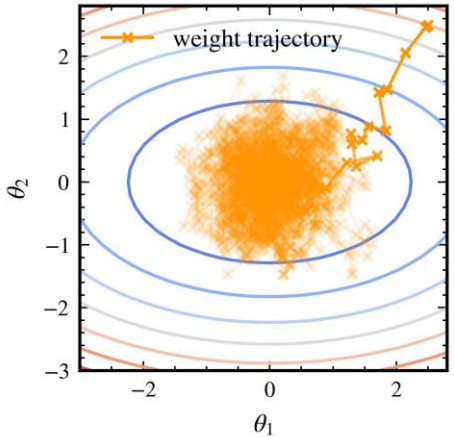


FIG 16. Weight trajectory in toy loss

Weight Fluctuations at a Minimum

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \tag{1a}$$

$$\mathbf{v}_k = -\eta \mathbf{g}_k(\boldsymbol{\theta}_{k-1}) + \beta \mathbf{v}_{k-1} \tag{1b}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \mathbf{v}_k \tag{1c}$$

- default case in physics: thermal noise

$$\boldsymbol{\Sigma} \propto \mathbf{H}^{-1} \ \ldots \text{ (see Einstein Relation)}$$

$$\text{for } \mathbf{C} := \operatorname*{cov}_k(\mathbf{g}_k, \mathbf{g}_k) \propto \mathbb{1} \ \ldots \text{ isotropic noise}$$

Weight Fluctuations at a Minimum

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \tag{1a}$$

$$\mathbf{v}_k = -\eta \mathbf{g}_k(\boldsymbol{\theta}_{k-1}) + \beta \mathbf{v}_{k-1} \tag{1b}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \mathbf{v}_k \tag{1c}$$

- default case in physics: thermal noise

$$\boldsymbol{\Sigma} \propto \mathbf{H}^{-1} \ \ldots \text{ (see Einstein Relation)}$$

$$\text{for } \mathbf{C} := \operatorname*{cov}_k(\mathbf{g}_k, \mathbf{g}_k) \propto \mathbb{1} \ \ldots \text{ isotropic noise}$$

- in neural networks: non-isotropic noise $\rightarrow$ previously in literature:
  (see Jastrzębski et al. (2018), Liu et al. (2021))

$$\mathbf{C} \propto \mathbf{H} \ \ldots \text{ Hessian noise approximation}$$

$$\Rightarrow \boldsymbol{\Sigma} \propto \mathbb{1} \ \ldots \text{ isotropic weight fluctuations}$$

Weight Fluctuations at a Minimum

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \tag{1a}$$

$$\mathbf{v}_k = -\eta \mathbf{g}_k(\boldsymbol{\theta}_{k-1}) + \beta \mathbf{v}_{k-1} \tag{1b}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \mathbf{v}_k \tag{1c}$$

- default case in physics: thermal noise

$$\boldsymbol{\Sigma} \propto \mathbf{H}^{-1} \ \ldots \text{ (see Einstein Relation)}$$

$$\text{for } \mathbf{C} := \operatorname*{cov}_k(\mathbf{g}_k, \mathbf{g}_k) \propto \mathbb{1} \ \ldots \text{ isotropic noise}$$

- in neural networks: non-isotropic noise $\rightarrow$ previously in literature:
  (see Jastrzębski et al. (2018), Liu et al. (2021))

$$\mathbf{C} \propto \mathbf{H} \ \ldots \text{ Hessian noise approximation}$$

$$\Rightarrow \boldsymbol{\Sigma} \propto \mathbb{1} \ \ldots \text{ isotropic weight fluctuations}$$

- recent empirical results contradict, more likely suggest
  (see Feng and Tu (2022))

$$\boldsymbol{\Sigma} \propto \mathbf{H} \ \ldots \text{ anisotropic weight fluctuations}$$

Weight Fluctuations at a Minimum

$$\mathbf{g}_k(\boldsymbol{\theta}) = \frac{1}{S} \sum_{n \in \mathcal{B}_k} \nabla l_n(\boldsymbol{\theta}) \tag{1a}$$

$$\mathbf{v}_k = -\eta \mathbf{g}_k(\boldsymbol{\theta}_{k-1}) + \beta \mathbf{v}_{k-1} \tag{1b}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \mathbf{v}_k \tag{1c}$$

- default case in physics: thermal noise

$$\boldsymbol{\Sigma} \propto \mathbf{H}^{-1} \ \ldots \ \text{(see Einstein Relation)}$$

$$\text{for } \mathbf{C} := \operatorname*{cov}_k(\mathbf{g}_k, \mathbf{g}_k) \propto \mathbb{1} \ \ldots \ \text{isotropic noise}$$

- in neural networks: non-isotropic noise $\rightarrow$ previously in literature:
  (see Jastrzębski et al. (2018), Liu et al. (2021))

$$\mathbf{C} \propto \mathbf{H} \ \ldots \ \text{Hessian noise approximation}$$

$$\Rightarrow \boldsymbol{\Sigma} \propto \mathbb{1} \ \ldots \ \text{isotropic weight fluctuations}$$

- recent empirical results contradict, more likely suggest
  (see Feng and Tu (2022))

$$\boldsymbol{\Sigma} \propto \mathbf{H} \ \ldots \ \text{anisotropic weight fluctuations}$$

- solution: noise anti-correlations

Weight Fluctuations at a Minimum

- weights covariance $\boldsymbol{\Sigma} := \mathrm{cov}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k\right)$, similarly velocity covariance $\boldsymbol{\Sigma}_{\mathbf{v}} := \mathrm{cov}\left(\mathbf{v}_k, \mathbf{v}_k\right)$ $(\mathbf{v}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1})$

Weight Fluctuations at a Minimum

- weights covariance $\boldsymbol{\Sigma} \coloneqq \mathrm{cov}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k\right)$, similarly velocity covariance $\boldsymbol{\Sigma}_{\mathbf{v}} \coloneqq \mathrm{cov}\left(\mathbf{v}_k, \mathbf{v}_k\right)$ $(\mathbf{v}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1})$

- **Assumption 1: Quadratic Approximation**
  $L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta}$ with Hessian $\mathbf{H}$

Weight Fluctuations at a Minimum

- weights covariance $\mathbf{\Sigma} := \mathrm{cov}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k\right)$, similarly velocity covariance $\mathbf{\Sigma}_\mathbf{v} := \mathrm{cov}\left(\mathbf{v}_k, \mathbf{v}_k\right)$ $(\mathbf{v}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1})$

- **Assumption 1: Quadratic Approximation**
  $L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta}$ with Hessian $\mathbf{H}$

- **Assumption 2: Anti-correlated Noise**
  Previously calculated SGD noise autocorrelation holds even for non static weights.

Weight Fluctuations at a Minimum

- weights covariance $\boldsymbol{\Sigma} \coloneqq \mathrm{cov}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k\right)$, similarly velocity covariance $\boldsymbol{\Sigma}_{\mathbf{v}} \coloneqq \mathrm{cov}\left(\mathbf{v}_k, \mathbf{v}_k\right)$ $(\mathbf{v}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1})$

- **Assumption 1: Quadratic Approximation**
  $L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta}$ with Hessian $\mathbf{H}$

- **Assumption 2: Anti-correlated Noise**
  Previously calculated SGD noise autocorrelation holds even for non static weights.

- **Assumption 3: Hessian and noise covariance commute**
  Covariance of noise commutes with Hessian, $[\mathbf{C}, \mathbf{H}] = 0$

Weight Fluctuations at a Minimum

- weights covariance $\boldsymbol{\Sigma} \coloneqq \mathrm{cov}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k\right)$, similarly velocity covariance $\boldsymbol{\Sigma}_{\mathbf{v}} \coloneqq \mathrm{cov}\left(\mathbf{v}_k, \mathbf{v}_k\right)$ $(\mathbf{v}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1})$

- **Assumption 1: Quadratic Approximation**
  $L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta}$ with Hessian $\mathbf{H}$

- **Assumption 2: Anti-correlated Noise**
  Previously calculated SGD noise autocorrelation holds even for non static weights.

- **Assumption 3: Hessian and noise covariance commute**
  Covariance of noise commutes with Hessian, $[\mathbf{C}, \mathbf{H}] = 0$

- with given assumptions: matrices commute, denote eigenvalues for common eigenvectors with $i = 1, ..., d$ as: λ_i for $\mathbf{H}$, $\sigma^2_{\delta g,i}$ for $\mathbf{C}$, $\sigma^2_{\theta,i}$ for $\boldsymbol{\Sigma}$, $\sigma^2_{v,i}$ for $\boldsymbol{\Sigma}_{\mathbf{v}}$

Weight Fluctuations at a Minimum - Derivation

- SGD with momentum in one dimension (with assumptions):

$$g_k = \frac{\partial}{\partial \theta}\left[\frac{1}{2}\theta_{k-1}\lambda\theta_{k-1}\right] + \delta g_k\,, \quad v_k = -\eta g_k + \beta v_{k-1}\,, \quad \theta_k = \theta_{k-1} + v_k \tag{5}$$

$$\theta_k = (1-\eta\lambda)\theta_{k-1} + \beta(\theta_{k-1} - \theta_{k-2}) - \eta\delta g_k \tag{6}$$

Weight Fluctuations at a Minimum - Derivation

- SGD with momentum in one dimension (with assumptions):

$$g_k = \frac{\partial}{\partial \theta}\left[\frac{1}{2}\theta_{k-1}\lambda\theta_{k-1}\right] + \delta g_k\,, \quad v_k = -\eta g_k + \beta v_{k-1}\,, \quad \theta_k = \theta_{k-1} + v_k \tag{5}$$

$$\theta_k = (1-\eta\lambda)\theta_{k-1} + \beta(\theta_{k-1} - \theta_{k-2}) - \eta\delta g_k \tag{6}$$

- in matrix form:

$$\mathbf{x}_k = \mathbf{D}\mathbf{x}_{k-1} - \eta\delta g_k \mathbf{e}_1\,, \quad \mathbf{x}_k := \begin{pmatrix}\theta_k \\ \theta_{k-1}\end{pmatrix}\,, \quad \mathbf{e}_1 := \begin{pmatrix}1 \\ 0\end{pmatrix} \tag{7}$$

$$\mathbf{D} := \begin{pmatrix}1+\beta-\eta\lambda & -\beta \\ 1 & 0\end{pmatrix} \tag{8}$$

Weight Fluctuations at a Minimum - Derivation

- SGD with momentum in one dimension (with assumptions):

$$g_k = \frac{\partial}{\partial \theta}\left[\frac{1}{2}\theta_{k-1}\lambda\theta_{k-1}\right] + \delta g_k\,, \quad v_k = -\eta g_k + \beta v_{k-1}\,, \quad \theta_k = \theta_{k-1} + v_k \tag{5}$$

$$\theta_k = (1-\eta\lambda)\theta_{k-1} + \beta(\theta_{k-1} - \theta_{k-2}) - \eta\delta g_k \tag{6}$$

- in matrix form:

$$\mathbf{x}_k = \mathbf{D}\mathbf{x}_{k-1} - \eta\delta g_k \mathbf{e}_1\,, \quad \mathbf{x}_k := \begin{pmatrix}\theta_k \\ \theta_{k-1}\end{pmatrix}\,, \quad \mathbf{e}_1 := \begin{pmatrix}1 \\ 0\end{pmatrix} \tag{7}$$

$$\mathbf{D} := \begin{pmatrix}1+\beta-\eta\lambda & -\beta \\ 1 & 0\end{pmatrix} \tag{8}$$

- respective covariance matrix:

$$\left\langle \mathbf{x}_k \mathbf{x}_k^\top \right\rangle = \begin{pmatrix}\sigma_\theta^2 & \langle\theta_k\theta_{k-1}\rangle \\ \langle\theta_k\theta_{k-1}\rangle & \sigma_\theta^2\end{pmatrix} \tag{9}$$

Weight Fluctuations at a Minimum - Derivation

$$\mathbf{x}_k = \mathbf{D}\mathbf{x}_{k-1} - \eta\delta g_k \mathbf{e}_1\,, \quad \mathbf{x}_k \coloneqq \begin{pmatrix} \theta_k \\ \theta_{k-1} \end{pmatrix} \tag{7}$$

$$\left\langle \mathbf{x}_k \mathbf{x}_k^\top \right\rangle = \mathbf{D} \left\langle \mathbf{x}_{k-1} \mathbf{x}_{k-1}^\top \right\rangle \mathbf{D}^\top + \eta^2 \left\langle \delta g_k \delta g_k \right\rangle \mathbf{e}_1 \mathbf{e}_1^\top - \eta \left(\mathbf{D} \left\langle \mathbf{x}_{k-1} \delta g_k \right\rangle \mathbf{e}_1^\top + \left(\mathbf{D} \left\langle \mathbf{x}_{k-1} \delta g_k \right\rangle \mathbf{e}_1^\top \right)^\top \right) \tag{10}$$

Weight Fluctuations at a Minimum - Derivation

$$\mathbf{x}_k = \mathbf{D}\mathbf{x}_{k-1} - \eta\delta g_k \mathbf{e}_1\,, \quad \mathbf{x}_k \coloneqq \begin{pmatrix}\theta_k \\ \theta_{k-1}\end{pmatrix} \tag{7}$$

$$\left\langle \mathbf{x}_k \mathbf{x}_k^\top \right\rangle = \mathbf{D}\left\langle \mathbf{x}_{k-1}\mathbf{x}_{k-1}^\top \right\rangle \mathbf{D}^\top + \eta^2 \left\langle \delta g_k \delta g_k \right\rangle \mathbf{e}_1\mathbf{e}_1^\top - \eta\left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top + \left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top\right)^\top\right) \tag{10}$$

- notice $v_k = \theta_k - \theta_{k-1} \rightarrow \sigma_v^2 = 2\sigma_\theta^2 - 2\langle\theta_k\theta_{k-1}\rangle$, and apply on $\mathbf{e}_1$:

$$\begin{pmatrix}\sigma_\theta^2 \\ \sigma_v^2\end{pmatrix} = \mathbf{F}\left[\eta^2\sigma_{\delta g}^2 \mathbf{e}_1\mathbf{e}_1^\top - \eta\left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top + \left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top\right)^\top\right)\right]\mathbf{e}_1 \tag{11}$$

$$\mathbf{F} = \frac{1}{(1-\beta)\big(2(1+\beta)-\eta\lambda\big)}\begin{pmatrix}\frac{1+\beta}{\eta\lambda} & \frac{2\beta(\eta\lambda-1-\beta)}{\eta\lambda} \\ 2 & 2(\eta\lambda-2)\end{pmatrix} \tag{12}$$

Weight Fluctuations at a Minimum - Derivation

$$\mathbf{x}_k = \mathbf{D}\mathbf{x}_{k-1} - \eta\delta g_k \mathbf{e}_1\,, \quad \mathbf{x}_k \coloneqq \begin{pmatrix}\theta_k \\ \theta_{k-1}\end{pmatrix} \tag{7}$$

$$\left\langle \mathbf{x}_k \mathbf{x}_k^\top \right\rangle = \mathbf{D}\left\langle \mathbf{x}_{k-1}\mathbf{x}_{k-1}^\top \right\rangle \mathbf{D}^\top + \eta^2 \left\langle \delta g_k \delta g_k \right\rangle \mathbf{e}_1\mathbf{e}_1^\top - \eta\left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top + \left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top\right)^\top\right) \tag{10}$$

- notice $v_k = \theta_k - \theta_{k-1} \to \sigma_v^2 = 2\sigma_\theta^2 - 2\langle\theta_k\theta_{k-1}\rangle$, and apply on $\mathbf{e}_1$:

$$\begin{pmatrix}\sigma_\theta^2 \\ \sigma_v^2\end{pmatrix} = \mathbf{F}\left[\eta^2\sigma_{\delta g}^2 \mathbf{e}_1\mathbf{e}_1^\top - \eta\left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top + \left(\mathbf{D}\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle \mathbf{e}_1^\top\right)^\top\right)\right]\mathbf{e}_1 \tag{11}$$

$$\mathbf{F} = \frac{1}{(1-\beta)\big(2(1+\beta)-\eta\lambda\big)}\begin{pmatrix}\frac{1+\beta}{\eta\lambda} & \frac{2\beta(\eta\lambda-1-\beta)}{\eta\lambda} \\ 2 & 2(\eta\lambda-2)\end{pmatrix} \tag{12}$$

- iterate SGD equation, use anti-correlation assumption:

$$\left\langle \mathbf{x}_{k-1}\delta g_k \right\rangle = \eta\sigma_{\delta g}^2 \frac{\mathbf{D}^M + (\mathbf{1}-\mathbf{D})M - \mathbf{1}}{(\mathbf{1}-\mathbf{D})^2 M(M-1)}\mathbf{e}_1 \tag{13}$$

Weight Fluctuations at a Minimum - Solution

- eigenvalues with $i = 1, ..., d$ as: λ_i for $\mathbf{H}$, $\sigma^2_{\delta g,i}$ for $\mathbf{C}$, $\sigma^2_{\theta,i}$ for $\mathbf{\Sigma}$, $\sigma^2_{v,i}$ for $\mathbf{\Sigma_v}$

- solution: $\left(\mathbf{e}_1 = (1 \;\; 0)^\top\right)$

$$\begin{pmatrix}\sigma^2_{\theta,i}\\ \sigma^2_{v,i}\end{pmatrix} = \eta^2\sigma^2_{\delta g,i}\mathbf{F_i}\left[\mathbf{e}_1 - \left(\mathbf{E_i} + \mathbf{E_i}^\top\right)\mathbf{e}_1\right] \tag{14}$$

$$\mathbf{F_i} = \frac{1}{(1-\beta)\big(2(1+\beta)-\eta\lambda_i\big)}\begin{pmatrix}\frac{1+\beta}{\eta\lambda_i} & \frac{2\beta(\eta\lambda_i-1-\beta)}{\eta\lambda_i}\\ 2 & 2(\eta\lambda_i-2)\end{pmatrix} \tag{15}$$

$$\mathbf{E_i} := \mathbf{D_i}\frac{\mathbf{D_i}^M + (\mathbf{1}-\mathbf{D_i})M - \mathbf{1}}{(\mathbf{1}-\mathbf{D_i})^2 M(M-1)}\mathbf{e}_1\mathbf{e}_1^\top \tag{16}$$

$$\mathbf{D_i} := \begin{pmatrix}1+\beta-\eta\lambda_i & -\beta\\ 1 & 0\end{pmatrix} \tag{17}$$

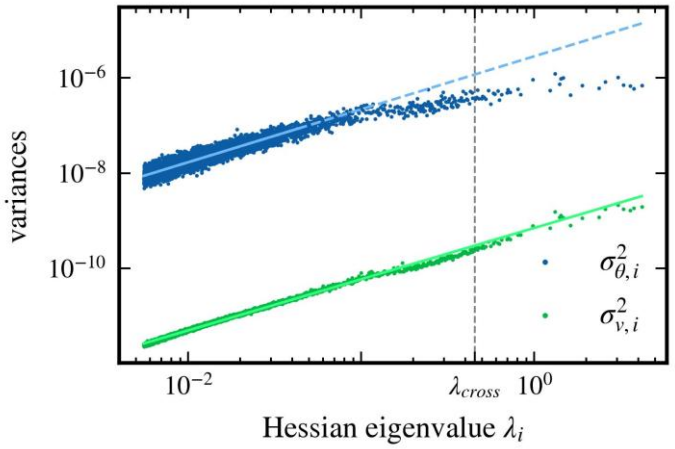


FIG 17. Empirical weight variance and velocity variance against Hessian eigenvalue

Velocity Correlation Time

- more comprehensive understanding with velocity-correlation-time τ_i $\begin{pmatrix} \theta_{k,i} := \boldsymbol{\theta}_k \cdot \mathbf{p_i} \\ v_{k,i} := \mathbf{v}_k \cdot \mathbf{p_i} \end{pmatrix}$ for common eigenvector $\mathbf{p_i}$

$$\tau_i := \frac{2\sigma^2_{\theta,i}}{\sigma^2_{v,i}} = \frac{\sum_{h=1}^{\infty} h \cdot \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)}{\sum_{h=1}^{\infty} \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)} \qquad (18)$$

Velocity Correlation Time

- more comprehensive understanding with velocity-correlation-time τ_i $\begin{pmatrix} \theta_{k,i} := \boldsymbol{\theta}_k \cdot \mathbf{p_i} \\ v_{k,i} := \mathbf{v}_k \cdot \mathbf{p_i} \end{pmatrix}$ for common eigenvector $\mathbf{p_i}$

$$\tau_i := \frac{2\sigma^2_{\theta,i}}{\sigma^2_{v,i}} = \frac{\sum_{h=1}^{\infty} h \cdot \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)}{\sum_{h=1}^{\infty} \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)} \qquad (18)$$

- two distinct regimes of Hessian eigenvalues (EVs), separated by $\lambda_{\mathrm{cross}} := \frac{3(1-\beta)S}{\eta N}$

$\lambda_i > \lambda_{\mathrm{cross}}$: can neglect anti-correlations

$$\rightarrow \tau_i \approx \frac{1+\beta}{\eta \lambda_i}$$

Velocity Correlation Time

- more comprehensive understanding with velocity-correlation-time τ_i $\left(\begin{smallmatrix}\theta_{k,i}:=\boldsymbol{\theta}_k\cdot\mathbf{p_i}\\ v_{k,i}:=\mathbf{v}_k\cdot\mathbf{p_i}\end{smallmatrix}\right.$ for common eigenvector $\mathbf{p_i}$$\left.\right)$

$$\tau_i := \frac{2\sigma^2_{\theta,i}}{\sigma^2_{v,i}} = \frac{\sum_{h=1}^{\infty} h\cdot \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)}{\sum_{h=1}^{\infty} \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)} \qquad (18)$$

- two distinct regimes of Hessian eigenvalues (EVs), separated by $\lambda_{\mathrm{cross}} := \frac{3(1-\beta)S}{\eta N}$

$\lambda_i > \lambda_{\mathrm{cross}}$: can neglect anti-correlations

$$\rightarrow \tau_i \approx \frac{1+\beta}{\eta\lambda_i}$$

$\lambda_i < \lambda_{\mathrm{cross}}$: anti-correlations important

$$\rightarrow \tau_i \approx \frac{M}{3}\frac{1+\beta}{1-\beta} =: \tau_{\mathrm{SGD}}$$

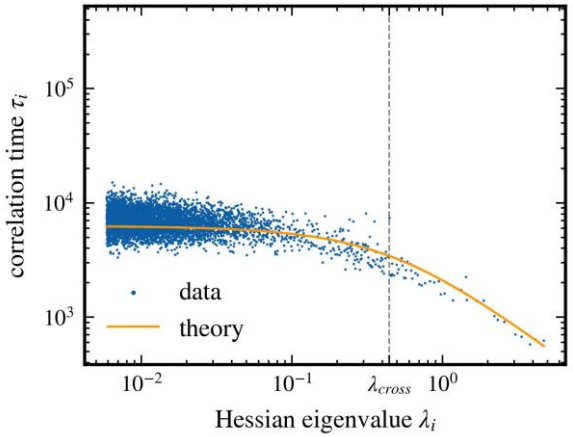


FIG 18. Correlation time against Hessian eigenvalue -Without replacement

Velocity Correlation Time

- more comprehensive understanding with velocity-correlation-time τ_i $\left(\begin{smallmatrix}\theta_{k,i}:=\boldsymbol{\theta}_k\cdot\mathbf{p_i}\\ v_{k,i}:=\mathbf{v}_k\cdot\mathbf{p_i}\end{smallmatrix}\right.$ for common eigenvector $\mathbf{p_i}$ $\left.\right)$

$$\tau_i := \frac{2\sigma^2_{\theta,i}}{\sigma^2_{v,i}} = \frac{\sum_{h=1}^{\infty} h\cdot \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)}{\sum_{h=1}^{\infty} \mathrm{cov}\left(v_{k,i}, v_{k+h,i}\right)} \qquad (18)$$

- two distinct regimes of Hessian eigenvalues (EVs), separated by $\lambda_{\mathrm{cross}} := \frac{3(1-\beta)S}{\eta N}$

$\lambda_i > \lambda_{\mathrm{cross}}$: can neglect anti-correlations

$$\rightarrow \tau_i \approx \frac{1+\beta}{\eta\lambda_i}$$

$\lambda_i < \lambda_{\mathrm{cross}}$: anti-correlations important

$$\rightarrow \tau_i \approx \frac{M}{3}\frac{1+\beta}{1-\beta} =: \tau_{\mathrm{SGD}}$$

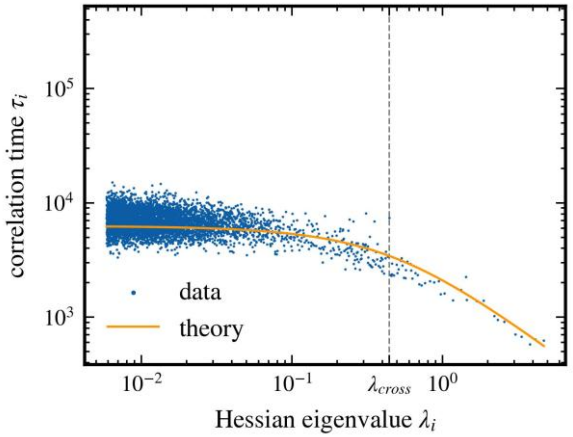


FIG 18. Correlation time against Hessian eigenvalue -Without replacement

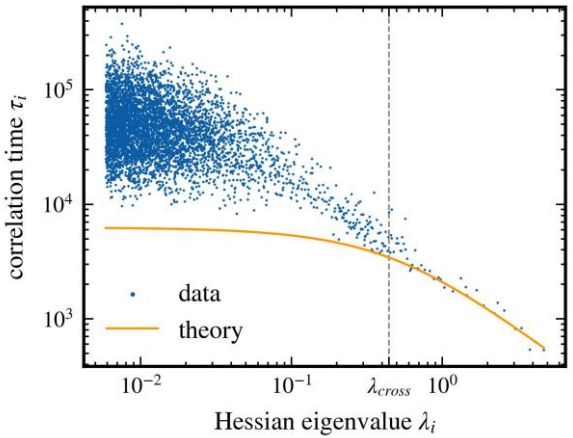


FIG 19. Correlation time against Hessian eigenvalue -With replacement

Approximate Weight Fluctuations

- two distinct regimes separated by $\lambda_{\text{cross}} := \frac{3(1-\beta)S}{\eta N}$
- **Large Hessian EVs ($\lambda_i > \lambda_{\text{cross}}$):** can neglect anti-correlations $\rightarrow \tau_i \approx \frac{1+\beta}{\eta\lambda_i}$ and

$$\sigma^2_{\theta,i} \approx \frac{\eta^2\sigma^2_{\delta g,i}}{2(1-\beta)(1+\beta)} \cdot \frac{1+\beta}{\eta\lambda_i}\,,\ \sigma^2_{v,i} \approx \frac{\eta^2\sigma^2_{\delta g,i}}{(1-\beta)(1+\beta)} \tag{19}$$

Approximate Weight Fluctuations

- two distinct regimes separated by $\lambda_{\text{cross}} := \frac{3(1-\beta)S}{\eta N}$
- **Large Hessian EVs ($\lambda_i > \lambda_{\text{cross}}$):** can neglect anti-correlations $\rightarrow \tau_i \approx \frac{1+\beta}{\eta\lambda_i}$ and

$$\sigma^2_{\theta,i} \approx \frac{\eta^2\sigma^2_{\delta g,i}}{2(1-\beta)(1+\beta)} \cdot \frac{1+\beta}{\eta\lambda_i}\,,\ \sigma^2_{v,i} \approx \frac{\eta^2\sigma^2_{\delta g,i}}{(1-\beta)(1+\beta)} \tag{19}$$

- **Small Hessian EVs ($\lambda_i < \lambda_{\text{cross}}$):** anti-correlations important, $\tau_i \approx \frac{M}{3}\frac{1+\beta}{1-\beta} =: \tau_{\text{SGD}}$ and

$$\sigma^2_{\theta,i} \approx \frac{\eta^2\sigma^2_{\delta g,i}}{2(1-\beta)(1+\beta)} \cdot \frac{M}{3}\frac{1+\beta}{1-\beta}\,,\ \sigma^2_{v,i} \approx \frac{\eta^2\sigma^2_{\delta g,i}}{(1-\beta)(1+\beta)} \tag{20}$$

Approximate Weight Fluctuations

- approximate relation:

$$\sigma^2_{v,i} \propto \sigma^2_{\delta g,i} \quad \text{or} \quad \mathbf{\Sigma_v} \propto \mathbf{C}\ (\propto \mathbf{H})$$

$$\sigma^2_{\theta,i} = \frac{1}{2}\tau_i \sigma^2_{v,i} \quad \text{or} \quad \mathbf{\Sigma} = \frac{1}{2}\mathrm{diag}(\tau_1, \ldots, \tau_d)\mathbf{\Sigma_v}$$

(in corresponding eigenbasis)

Approximate Weight Fluctuations

- approximate relation:

$$\sigma^2_{v,i} \propto \sigma^2_{\delta g,i} \quad \text{or} \quad \mathbf{\Sigma_v} \propto \mathbf{C}\ (\propto \mathbf{H})$$

$$\sigma^2_{\theta,i} = \frac{1}{2}\tau_i \sigma^2_{v,i} \quad \text{or} \quad \mathbf{\Sigma} = \frac{1}{2}\mathrm{diag}(\tau_1, \ldots, \tau_d)\mathbf{\Sigma_v}$$

(in corresponding eigenbasis)

- Hessian eigenvector subspace with $\lambda_i > \lambda_{\mathrm{cross}}$ $(\tau_i \propto \lambda_i^{-1})$:

$$\mathbf{\Sigma}^{(>)} \propto \mathbb{1}$$

- Hessian eigenvector subspace with $\lambda_i < \lambda_{\mathrm{cross}}$ $(\tau_i = \tau_{\mathrm{SGD}})$:

$$\mathbf{\Sigma}^{(<)} \propto \mathbf{H}$$

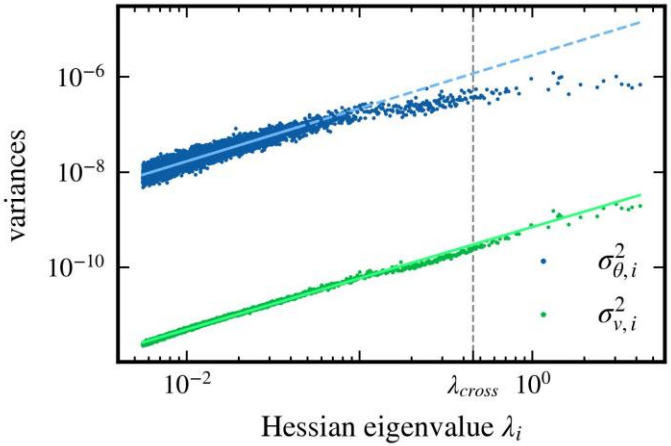


FIG 17. Empirical weight variance and velocity variance against Hessian eigenvalue

Approximate Weight Fluctuations

- approximate relation:

$$\sigma^2_{v,i} \propto \sigma^2_{\delta g,i} \quad \text{or} \quad \mathbf{\Sigma_v} \propto \mathbf{C}\ (\propto \mathbf{H})$$

$$\sigma^2_{\theta,i} = \frac{1}{2}\tau_i \sigma^2_{v,i} \quad \text{or} \quad \mathbf{\Sigma} = \frac{1}{2}\mathrm{diag}(\tau_1, \ldots, \tau_d)\mathbf{\Sigma_v}$$

(in corresponding eigenbasis)

- Hessian eigenvector subspace with $\lambda_i > \lambda_{\mathrm{cross}}$ $(\tau_i \propto \lambda_i^{-1})$:

$$\mathbf{\Sigma}^{(>)} \propto \mathbb{1}$$

- Hessian eigenvector subspace with $\lambda_i < \lambda_{\mathrm{cross}}$ $(\tau_i = \tau_{\mathrm{SGD}})$:

$$\mathbf{\Sigma}^{(<)} \propto \mathbf{H}$$

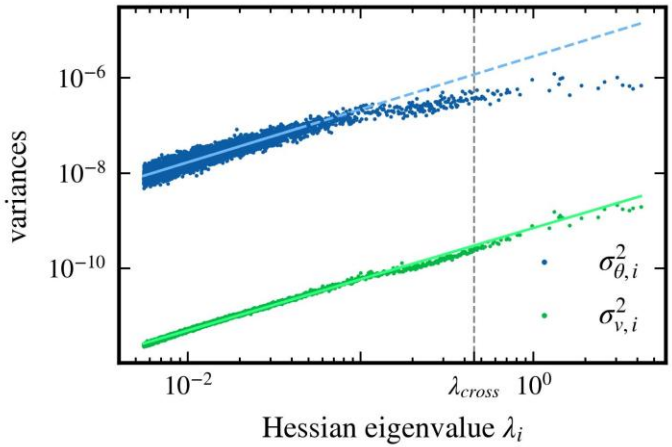


FIG 17. Empirical weight variance and velocity variance against Hessian eigenvalue

- previous observation by Feng and Tu (2022) closer to $\mathbf{\Sigma} \propto \mathbf{H}^2$
- principal component analysis of $\mathbf{\Sigma}$ leads to finite size effects $\rightarrow \mathbf{H}$ eigenbasis analysis superior

Varying Hyperparameters

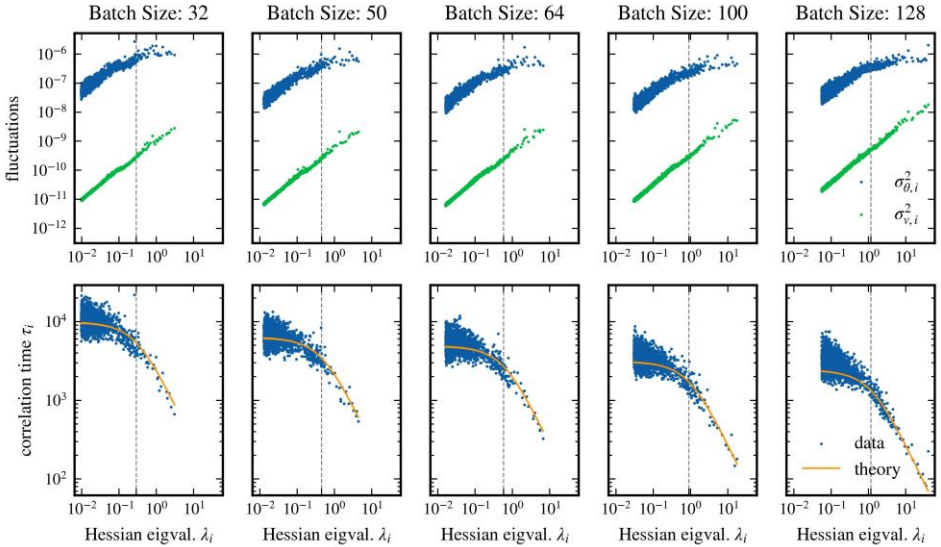


FIG 20. Varying Batch Size: Empirical weight variance, velocity variance and correlation time against Hessian eigenvalue

Varying Hyperparameters

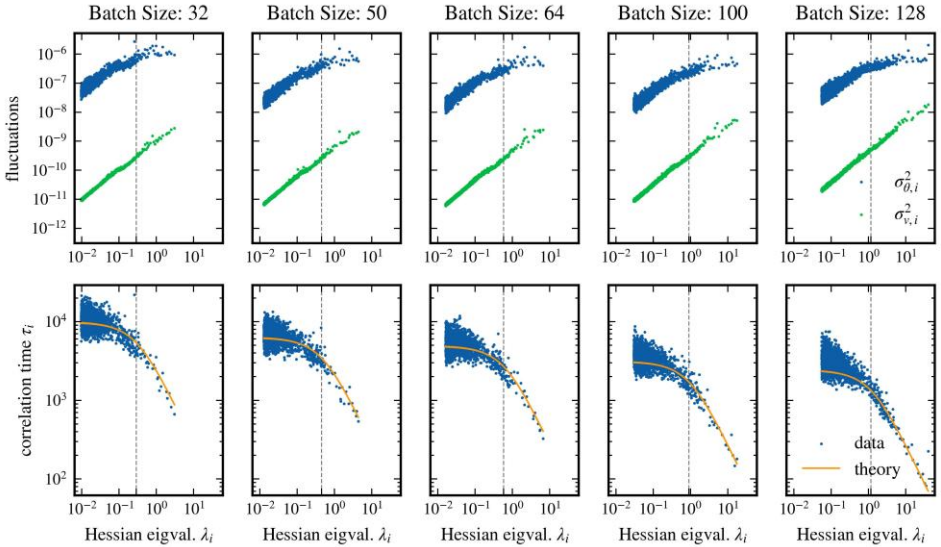
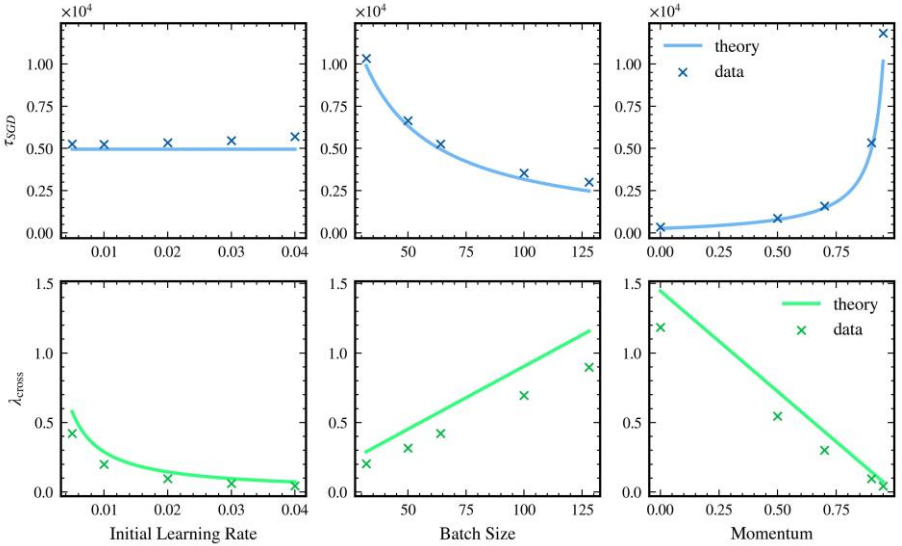


FIG 20. Varying Batch Size: Empirical weight variance, velocity variance and correlation time against Hessian eigenvalue

FIG 21. Empirical maximum correlation time and crossover value vs. theory



Different Network

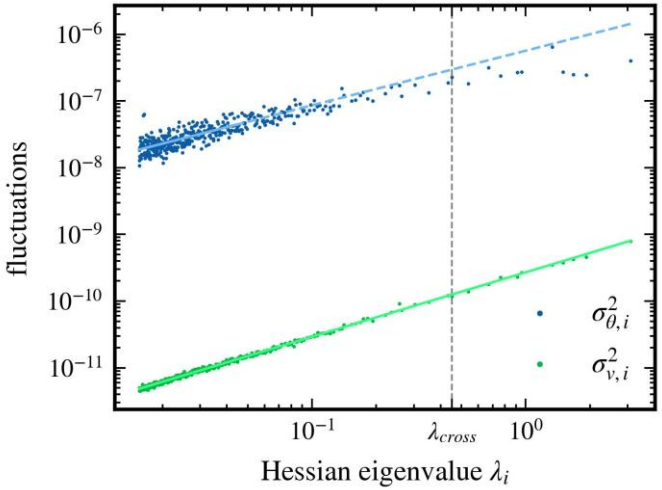


FIG 22. Improved Architecture - ResNet instead of LeNet: Empirical weight variance and velocity variance against Hessian eigenvalue

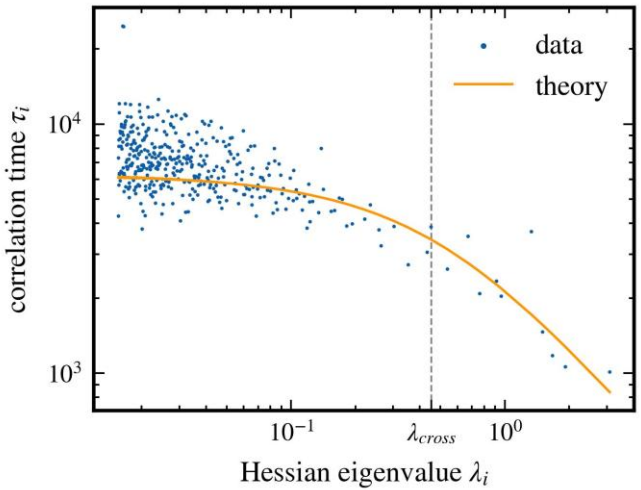


FIG 23. Improved Architecture - ResNet instead of LeNet: Correlation time against Hessian eigenvalue

Questions about Weight Fluctuations?

Questions about Weight Fluctuations?

- **Implications for Training**

Additional Error due to Fluctuations

$$L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\theta}$$

$$\left\langle L(\boldsymbol{\theta}_k)\right\rangle_k = \left\langle \sum_{i=1}^{d} \frac{1}{2}\theta_{i,k}\lambda_i\theta_{i,k} \right\rangle_k$$

$$= \sum_{i=1}^{d} \frac{1}{2}\lambda_i\sigma^2_{\theta,i}$$

$$\sigma^2_{\theta,i} \begin{cases} = \sigma^2_{\theta,\max} & \text{if } \lambda_i > \lambda_{\mathrm{cross}} \\ \propto \lambda_i \text{ , and } < \sigma^2_{\theta,\max} & \text{if } \lambda_i < \lambda_{\mathrm{cross}} \end{cases}$$

Additional Error due to Fluctuations

$$L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\theta}$$

$$\left\langle L(\boldsymbol{\theta}_k)\right\rangle_k = \left\langle \sum_{i=1}^{d} \frac{1}{2}\theta_{i,k}\lambda_i\theta_{i,k} \right\rangle_k$$

$$= \sum_{i=1}^{d} \frac{1}{2}\lambda_i\sigma^2_{\theta,i}$$

$$\sigma^2_{\theta,i} \begin{cases} = \sigma^2_{\theta,\max} & \text{if } \lambda_i > \lambda_{\mathrm{cross}} \\ \propto \lambda_i \text{ , and } < \sigma^2_{\theta,\max} & \text{if } \lambda_i < \lambda_{\mathrm{cross}} \end{cases}$$

- Reduction of weight fluctuations leads to 62% reduction of loss fluctuations

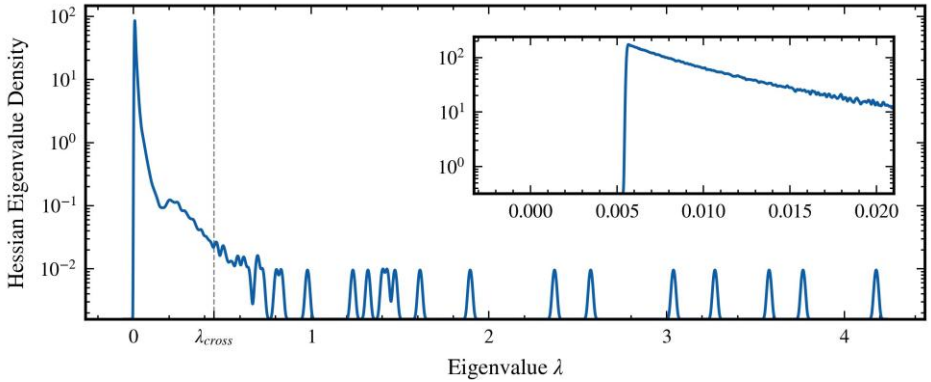


FIG 24. Experimental weight variance and velocity variance against Hessian eigenvalue

Improved Generalization

- test accuracy of LeNet is 0.7% ± 0.2% higher for drawing examples in SGD without replacement (64.5%) than for SGD with replacement (63.8%)

Improved Generalization

- test accuracy of LeNet is 0.7% ± 0.2% higher for drawing examples in SGD without replacement (64.5%) than for SGD with replacement (63.8%)

- in agreement with Orvieto et al. (2022), where anti-correlated perturbed gradient descent (Anti-PGD) was found beneficial for test accuracy and led to flatter minima

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{\nabla} L(\boldsymbol{\theta}_k) + (\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k) \ \ldots \text{ Anti-PGD}$$

$$\xi_{k,i} \sim \mathcal{N}(0,\, \sigma^2)$$

Improved Generalization

- test accuracy of LeNet is 0.7% ± 0.2% higher for drawing examples in SGD without replacement (64.5%) than for SGD with replacement (63.8%)

- in agreement with Orvieto et al. (2022), where anti-correlated perturbed gradient descent (Anti-PGD) was found beneficial for test accuracy and led to flatter minima

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{\nabla} L(\boldsymbol{\theta}_k) + (\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k) \quad \ldots \text{ Anti-PGD}$$

$$\xi_{k,i} \sim \mathcal{N}(0,\, \sigma^2)$$

$$\mathbb{E}[\mathbf{z}_{k+1} | \mathbf{z}_k] \approx \mathbf{z}_k - \eta \boldsymbol{\nabla} \tilde{L}(\mathbf{z}_k) \qquad (\mathbf{z}_k = \boldsymbol{\theta}_k - \boldsymbol{\xi}_k)$$

$$\tilde{L}(\mathbf{z}) \coloneqq L(\mathbf{z}) + \frac{\sigma^2}{2} \mathrm{Tr}\big(\mathbf{H}(\mathbf{z})\big)$$

Improved Generalization

- test accuracy of LeNet is 0.7% ± 0.2% higher for drawing examples in SGD without replacement (64.5%) than for SGD with replacement (63.8%)

- in agreement with Orvieto et al. (2022), where anti-correlated perturbed gradient descent (Anti-PGD) was found beneficial for test accuracy and led to flatter minima

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{\nabla} L(\boldsymbol{\theta}_k) + (\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k) \quad \ldots \text{ Anti-PGD}$$

$$\xi_{k,i} \sim \mathcal{N}(0,\, \sigma^2)$$

$$\mathbb{E}[\mathbf{z}_{k+1} | \mathbf{z}_k] \approx \mathbf{z}_k - \eta \boldsymbol{\nabla} \tilde{L}(\mathbf{z}_k) \qquad (\mathbf{z}_k = \boldsymbol{\theta}_k - \boldsymbol{\xi}_k)$$

$$\tilde{L}(\mathbf{z}) := L(\mathbf{z}) + \frac{\sigma^2}{2} \mathrm{Tr}\big(\mathbf{H}(\mathbf{z})\big)$$

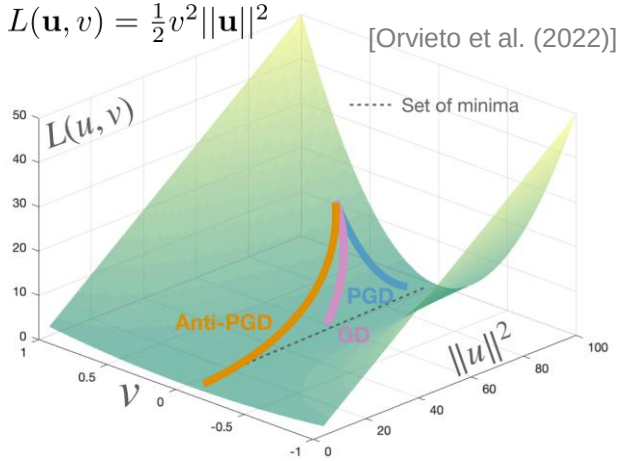


FIG 25. Gradient Descent with anti-correlated noise in a widening valley

UNIVERSITÄT
LEIPZIG

Conclusion

- SGD without replacement induces anti-correlated noise

UNIVERSITÄT
LEIPZIG

Conclusion

- SGD without replacement induces anti-correlated noise
- Results in lower-than-expected weight variance in Hessian eigendirections with small EVs

UNIVERSITÄT
LEIPZIG

Conclusion

- SGD without replacement induces anti-correlated noise
- Results in lower-than-expected weight variance in Hessian eigendirections with small EVs
- beneficial because gradients in flat directions then dominate fluctuations and lead network towards even flatter minima with improved generalization performance

UNIVERSITÄT
LEIPZIG

Conclusion

- SGD without replacement induces anti-correlated noise
- Results in lower-than-expected weight variance in Hessian eigendirections with small EVs
- beneficial because gradients in flat directions then dominate fluctuations and lead network towards even flatter minima with improved generalization performance

Thank You!

Marcel Kühn and Bernd Rosenow
Institute for Theoretical Physics

mkuehn@itp.uni-leipzig.de

References

Berner et al. (2022)

Berner J, Grohs P, Kutyniok G and Petersen P 2022 The Modern Mathematics of Deep Learning. In: Grohs P, Kutyniok G, editors. Mathematical Aspects of Deep Learning. Cambridge: Cambridge University Press pp 1–111

Orvieto et al. (2022)

Orvieto A, Kersting H, Proske F, Bach F and Lucchi A 2022 Anticorrelated Noise Injection for Improved Generalization *Proceedings of the 39th International Conference on Machine Learning* vol 162 (PMLR) pp 17094–17116

Clark et al. (2019)

Clark C, Lee K, Chang M, Kwiatkowski T, Collins M, Toutanova K 2019 BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* vol 1 pp 2924--2936

Krizhevsky and Hinton (2009)

Krizhevsky A 2009 Learning multiple layers of features from tiny images Tech. rep. University of Toronto

Feng and Tu (2021)

Feng Y and Tu Y 2021 *Proceedings of the National Academy of Sciences* **118**

He et al. (2016)

He K, Zhang X, Ren S and Sun J 2016 Deep Residual Learning for Image Recognition *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp 770–778

Jastrzębski et al. (2018)

Jastrzębski S, Kenton Z, Arpit D, Ballas N, Fischer A, Bengio Y and Storkey A 2018 Three Factors Influencing Minima in SGD (Arxiv Preprint 1711.04623)

Liu et al. (2021)

Liu K, Ziyin L and Ueda M 2021 Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent *Proceedings of the 38th International Conference on Machine Learning* vol 139 (PMLR) pp 7045–7056

Training evolution

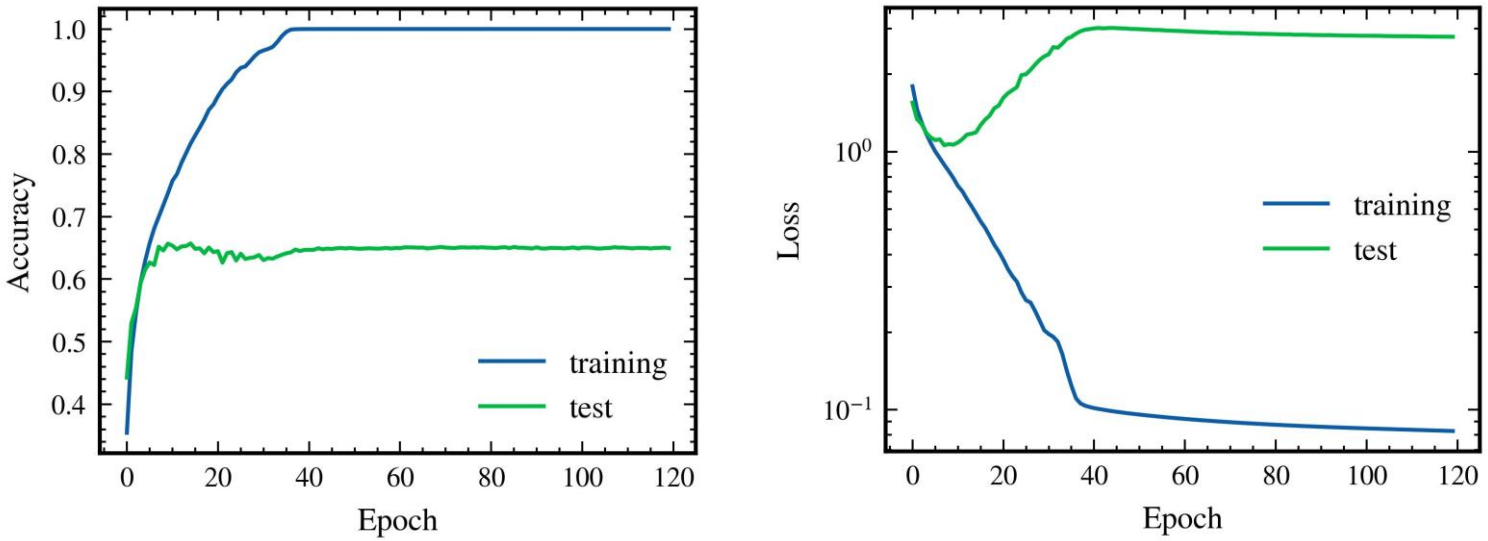


FIG 26. Evolution of loss (left) and accuracy (right) during training of Lenet