Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning

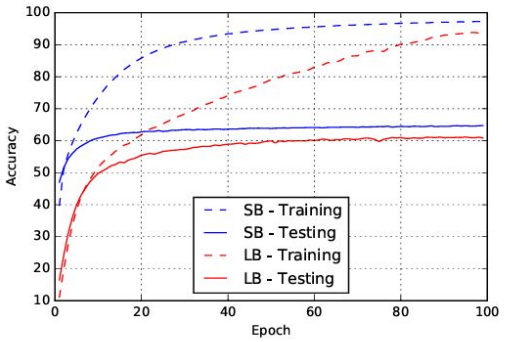
Nikhil Vyas, Depen Morwani*, Rosie Zhao*, Gal Kaplun*, Sham Kakade, Boaz Barak*

ICML 2024 (Spotlight)

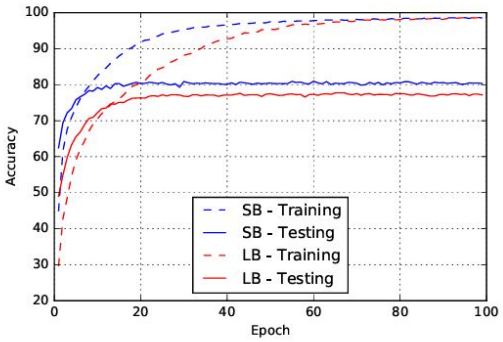
Harvard John A. Paulson
School of Engineering
and Applied Sciences

Background: Implicit Bias of SGD Noise

- In deep learning, hyperparameters and algorithm choice can influence the search space explored by the optimization algorithm (**implicit bias**)
- Existing analyses of implicit bias have mostly focused on offline learning regimes, but deep learning is undergoing a paradigm shift whereby models are often trained in an **online (single-epoch) learning** regime with self-supervised objectives - we study the effect of **SGD noise** in this regime



(a) Network F_2



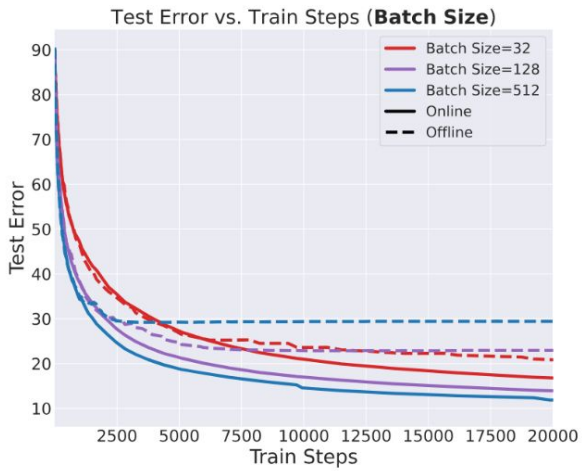
(b) Network C_1

Keskar et al. (2017): SB = small batch, LB = large batch.

Previous work: SGD noise **advantageous** for implicit bias (leads to **flatter minima** which generalize better)

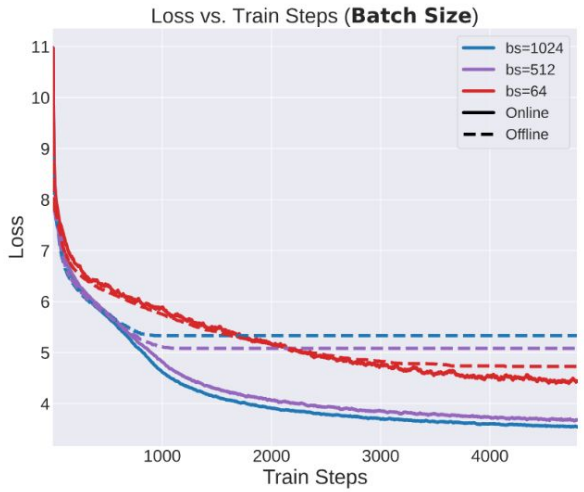
We find that SGD noise does not provide any implicit bias advantage when learning online.

- Our setup: ResNet-18 on CIFAR-5m, ConvNext-T on ImageNet, GPT-2 small on C4.
- **Offline:** SGD noise leads to better implicit bias (**small batch size = better**)
- **Online:** SGD noise has no implicit bias advantage (**large batch size = better**)
- (* Explicit regularization still has an effect in the online setting, so this isn't necessarily a trivial result!)



(a) CIFAR-5m

(b) ImageNet

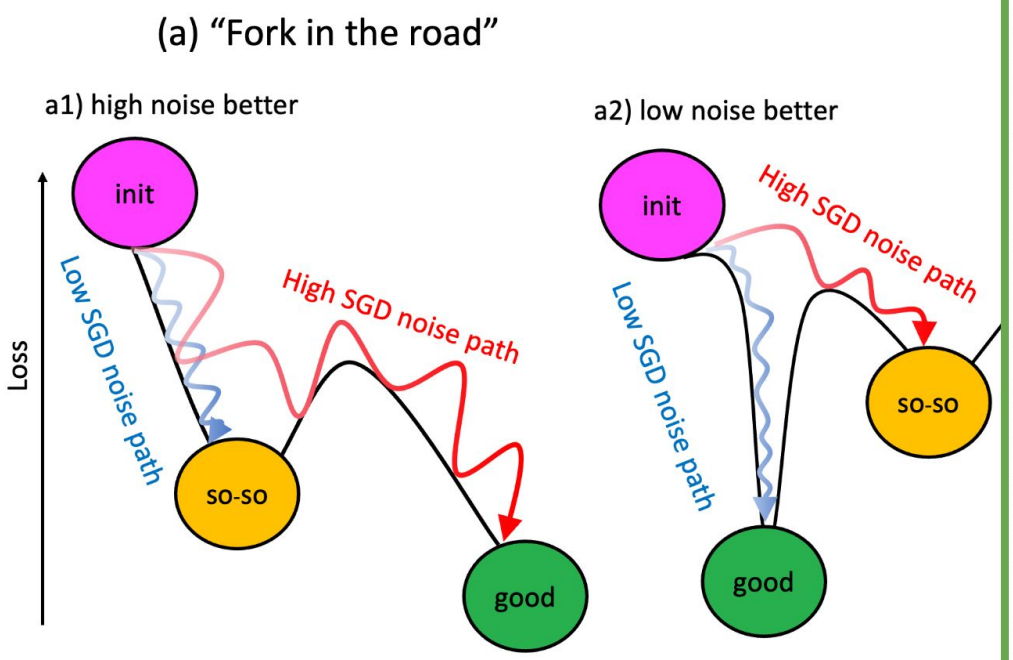


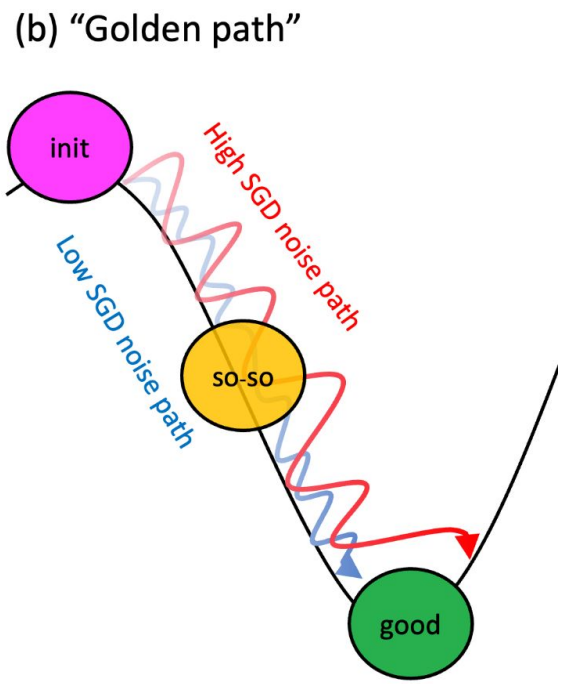
(c) C4

Vyas, Morwani, Zhao, Kaplun, Kakade, Barak. Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning (ICML 2024).

Golden Path Hypothesis

*for natural settings!





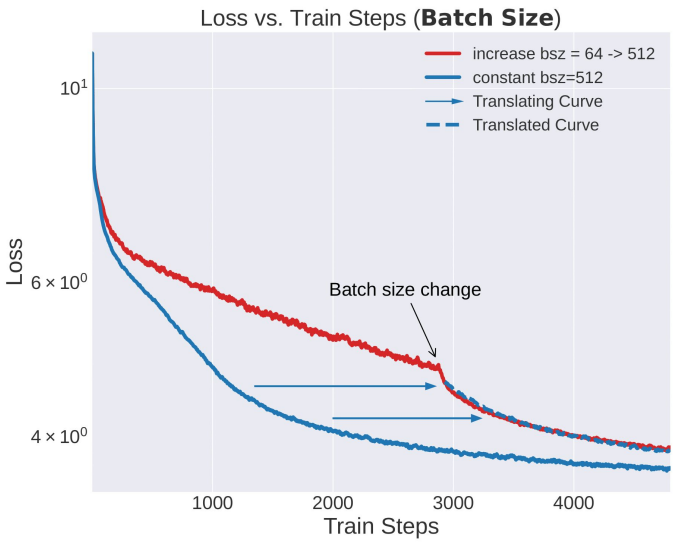
Different levels of SGD noise results in the algorithm exploring distinct parts of the search space

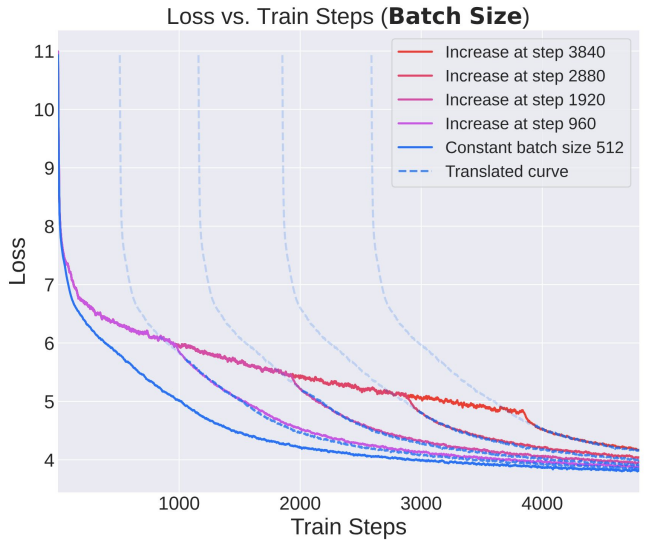
Optimization is similar for both SGD and limiting GD trajectory

Vyas, Morwani, Zhao, Kaplun, Kakade, Barak. Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning (ICML 2024).

Golden Path in Loss Space

- We want to rule out the possibility that low SGD noise leads us to **better, but still distinct** parts of the search space
- We run the following experiment:
  1. Start two runs– one with high batch size, and one with small batch size–for t steps.
  2. After t steps, **increase the batch size of the second experiment** to match the hyperparameters of the first one, and continue both runs.



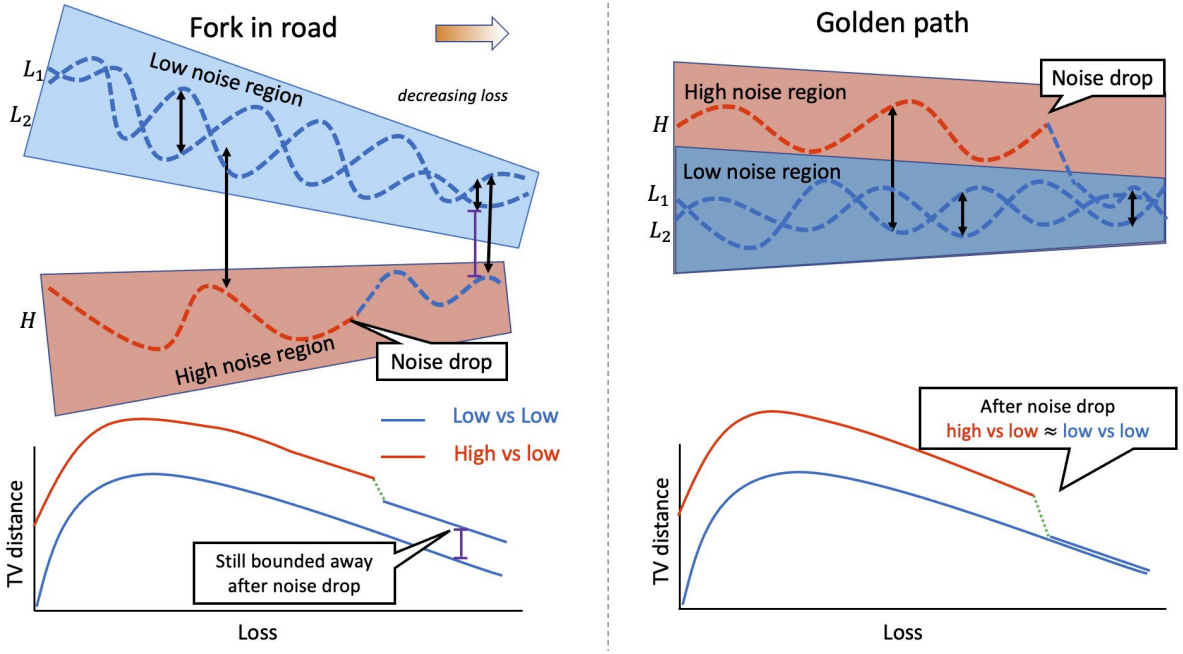


Loss curve of low noise run 'snaps' to high -> low noise run after the change.

Vyas, Morwani, Zhao, Kaplun, Kakade, Barak. Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning (ICML 2024).

Golden Path in Function Space

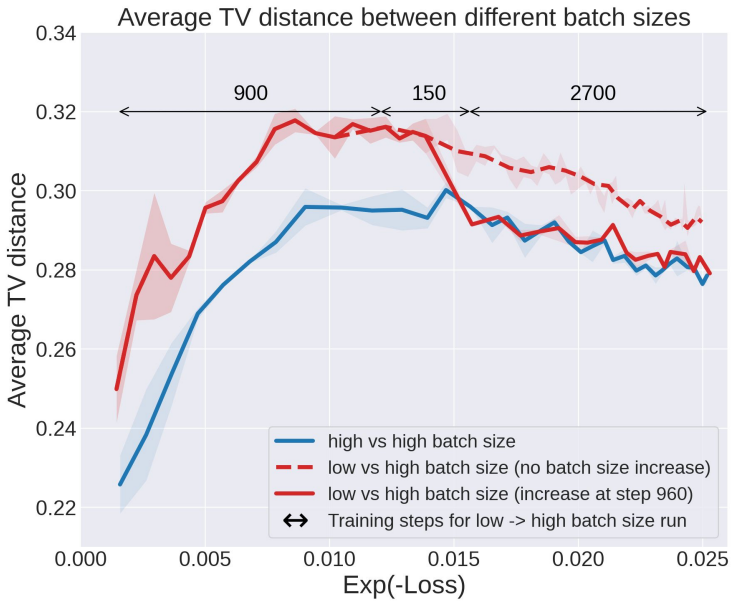
- Are trajectories with different SGD noise **functionally** similar?
- **Fork in the road:** high noise and low noise trajectories explore **different regions**, so functional distance **should remain high** even after dropping the noise
- **Golden Path:** high noise trajectory follows a 'noisy' version of the low noise trajectory, so functional distance **should decrease** after dropping the noise

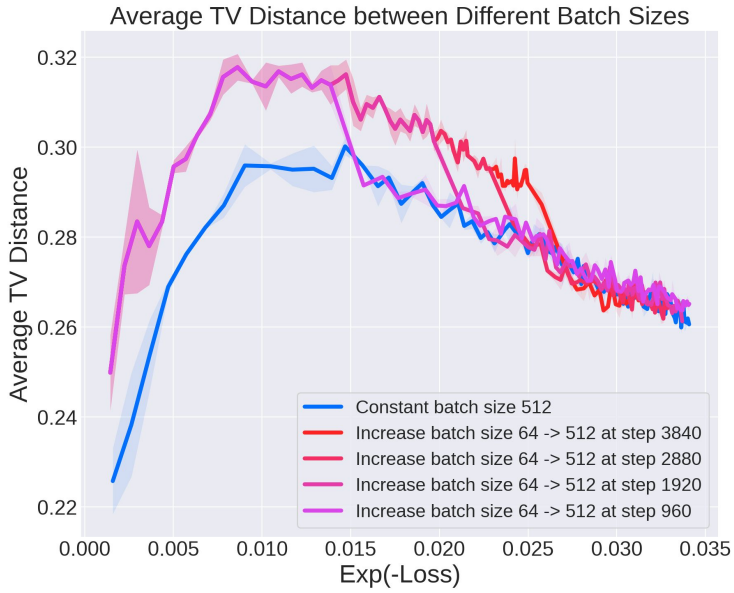


Vyas, Morwani, Zhao, Kaplun, Kakade, Barak. Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning (ICML 2024).

Golden Path in Function Space

- Are trajectories with different SGD noise functionally similar?
- We look at **average total variation (TV) distance** of models' softmax probabilities on the test dataset
- Take the two runs from loss space experiment (one **low noise**, one **high -> low noise**) and report TV distance between two models at the **same loss value**
  - Compare to baseline of **two low noise** runs trained on **different seeds**





Vyas, Morwani, Zhao, Kaplun, Kakade, Barak. Beyond Implicit Bias: The Insignificance of SGD Noise in Online Learning (ICML 2024).

Key Takeaways

- We show there is a striking discrepancy between offline and online training regimes: **small batch sizes are not advantageous in terms of implicit bias in online learning**
- We provide evidence in loss and function space that **SGD in online learning settings follows a trajectory similar to full-batch GD**, up to deviations due to noise
- It may be necessary to reevaluate our comprehension of various deep learning phenomena in the context of online settings