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

Nikhil Vyas [*1] **Depen Morwani** [*1] **Rosie Zhao** [*1] **Gal Kaplun** [*1] **Sham Kakade** [1 2] **Boaz Barak** [1]

Abstract

The success of SGD in deep learning has been ascribed by prior works to the *implicit bias* induced by finite batch sizes ("SGD noise"). While prior works focused on *offline learning* (i.e., multiple-epoch training), we study the impact of SGD noise on *online* (i.e., single epoch) learning. Through an extensive empirical analysis of image and language data, we demonstrate that small batch sizes do *not* confer any implicit bias advantages in online learning. In contrast to offline learning, the benefits of SGD noise in online learning are strictly computational, facilitating more cost-effective gradient steps. This suggests that SGD in the online regime can be construed as taking noisy steps along the "golden path" of the noiseless *gradient descent* algorithm. We study this hypothesis and provide supporting evidence in loss and function space. Our findings challenge the prevailing understanding of SGD and offer novel insights into its role in online learning.

1. Introduction

In the field of optimization theory, the selection of hyperparameters, such as learning rate and batch size, plays a significant role in determining the *optimization efficiency*, which refers to the computational resources required to minimize the loss function to a predetermined level. In strongly *convex problems*, altering these hyperparameters does not affect the final solution since all local minima are global. Hence, the final model only depends on the *explicit biases* of architecture and objective function (including any explicit regularizers). In contrast, Deep Learning is *non-convex*, which means that the choices of algorithm and hyperparameters can impact not only optimization efficiency but also introduce an *implicit bias*, i.e., change the regions of the search space explored by the optimization algorithm, consequently impacting the final learned model.

The implicit bias induced by the algorithm and hyperparameter choices can significantly affect the quality of the learned model, including generalization, robustness to distribution shifts, downstream performance, and more. Hence, the implicit bias of *stochastic gradient descent (SGD)* has garnered considerable attention within the research community (Jastrzebski et al., 2017; Lewkowycz et al., 2020; Damian et al., 2021; He et al., 2019; Nacson et al., 2022; HaoChen et al., 2021; Andriushchenko et al., 2022). In particular, this implicit bias emerges due to SGD using finite batch sizes resulting in a *noisy approximation* of the population gradient.

Perhaps counter-intuitively, SGD noise is often deemed *advantageous* for implicit bias. In particular, several works showed that higher learning rates and smaller batch sizes yield *flatter* minima (Keskar et al., 2016; LeCun et al., 1998; Masters & Luschi, 2018; Goyal et al., 2017), which tend to generalize well (Hochreiter & Schmidhuber, 1997; Smith et al., 2020) (see also Figure 1, left). However, these works are limited to the setting of multi-epoch or *offline training*.

In this work, we examine the implicit bias of SGD in the **online learning** setting, in which data is processed through a *single epoch*. Online learning is common in several self-supervised settings, including large language models (LLMs) (Komatsuzaki, 2019; Brown et al., 2020; Hoffmann et al., 2022b;a; Chowdhery et al., 2022). While in online learning, the train and test distributions are identical (and hence there are no generalization considerations), it is still a *non-convex* optimization. So, the inductive bias introduced by algorithm and hyperparameter choices could still potentially play a major role in learning trajectory and model quality. However, we find that the impact of batch sizes in practical settings of online learning is qualitatively similar to their impact on *convex* optimization. Specifically, we undertake an extensive empirical investigation and find that, *in online learning, SGD noise is indeed only "noise"* and offers no implicit benefits beyond optimization efficiency. This can be seen in Figure 1 (right), where we observe that (neglecting computational cost) performance in online learning improves with increasing batch size.

The "Golden Path" Hypothesis. When taking SGD to

[*]Equal contribution [1]SEAS, Harvard University [2]Kempner Institute, Harvard University. Correspondence to: Nikhil Vyas <nikhil@g.harvard.edu>, Depen Morwani <dmorwani@g.harvard.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

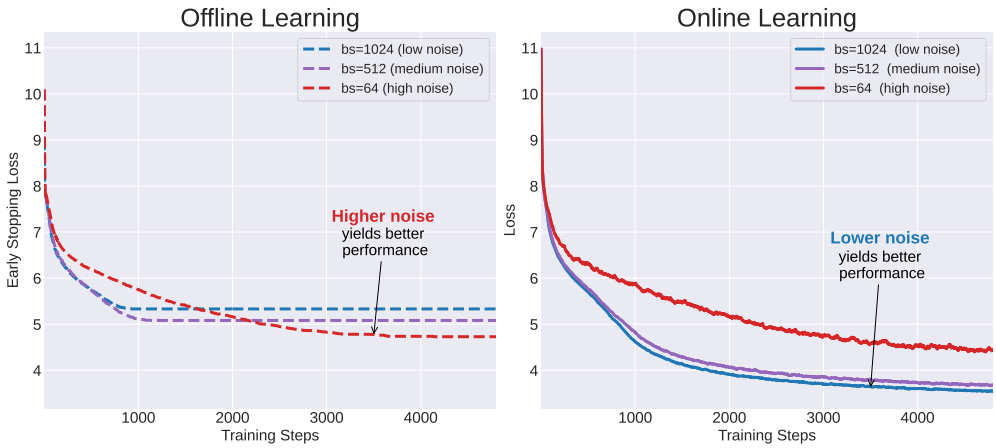


Figure 1. Experiment on offline (**left**) and online (**right**) learning on the C4 dataset across various batch sizes. As shown in prior works, in *offline learning* (left), higher SGD noise (lower batch size) offers an implicit bias advantage and plateaus at a lower loss. In contrast, we show that in *online learning* (right), higher SGD noise does not provide any implicit bias benefit to performance, and lower noise reaches a smaller loss. The y-axis measures early stopping (true) loss. See Section 2 for more details.

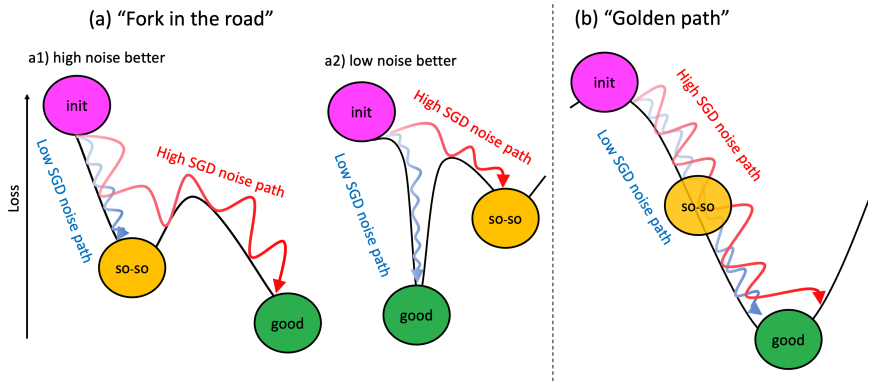


Figure 2. A-priori online learning can exhibit two potential scenarios: **(a)** *"Fork in the Road,"* wherein the selection of batch size leads the optimization algorithm to explore distinct regions of the search space, potentially resulting in different loss outcomes. (a1: better loss for the high-noise path, which is the common case for offline learning, and a2: better loss for the low-noise path). **(b)** *"Golden path,"* wherein the optimization trajectory remains similar for both gradient descent and SGD. In the latter scenario, the noise in SGD primarily influences the algorithm's traversal speed (and stability) along the path. Our research provides evidence supporting the "golden path" scenario for online learning.

the limit of large batch sizes, we get the *noiseless Gradient Descent* (GD) algorithm. Our findings hint that the SGD path is just a noisy version of the underlying noiseless GD path as illustrated in Figure 2. We propose this conjecture as the *golden path* hypothesis, which stands in contrast to the alternative "Fork in the Road" possibility, wherein SGD and GD discover qualitatively different functions. As shown in prior works (as well as in our own experiments), the "fork in the road" scenario (and specifically scenario a1, where high noise leads to a better minima) is the typical case for *offline learning*.

To be precise, our work gives evidence for the following hypothesis:

Golden Path Hypothesis: For natural settings in online learning, a path trained via SGD does not deviate far from one trained via GD, in the following sense:

- **Loss Trajectories**: If SGD noise is dropped[1] from high to low at time t, then shortly after t, the loss curve will "snap" to track the curve of a model that was trained with low noise from initialization, and hence follow the "golden path".
- **Function Space**: After reducing the SGD noise, the resulting function from this path is also similar in functional distance to that of a path with low noise from initialization.

Is the Hypothesis False? Indeed, the "Golden Path" hypothesis does not hold in a variety of settings. One simple toy example is provided in Figure 2 of Wu et al. (2018), demonstrating how higher noise could help in escaping "sharp" minima in the non-convex regime (for further discussion, refer to the related works). However, for real world datasets trained using deep learning, we observe that the SGD path becomes "close" to that of the GD path once the SGD noise has been reduced. As we have stated above, closeness is measured both in terms of the loss trajectories and functional distance. One could also consider a stronger notion of the golden path hypothesis in weight space; however, in its naive formulation, the golden path hypothesis **does not hold in weight space**; this is due to the presence of permutation symmetries, dead neurons, and structural differences (see section 1.1). While we believe that a more careful formulation of the "golden path" hypothesis can hold in weight space, we focus on loss and function space in this paper and defer weight space exploration for future work.

Overall, our work gives evidence to the hypothesis that there is a "noiseless" or "golden" path that Gradient Descent takes, and that batch size plays no role in the choice of the path but only in the computational cost to travel on it, the training stability as well as the level of "variance" along the path. Hence choosing batch sizes should be determined by balancing their negative impact on noise with their positive impact on computation. This is in stark contrast to the role of SGD in *offline learning*, wherein SGD noise can influence not just the *speed* of optimization but also its *journey* and even its final *destination* (i.e., function at convergence).

Contributions and organization. We delineate our contributions as follows:

1. Our first contribution is demonstrating that, unlike in offline learning, SGD noise does not provide any implicit bias advantage in a variety of practical online learning settings. This is presented in Section 2, which contains a systematic investigation of the effects of SGD noise in the online versus offline settings. Our analysis encompasses both vision (ResNet-18 on CIFAR-5m, ConvNext-T on ImageNet) and language tasks (GPT-2-small on C4),

2. A second contribution is to propose and examine the "golden path" hypothesis in the context of online learning. In Section 3, we provide evidence that SGD loss trajectories follow that of gradient descent by showing that the loss curves of high-noise SGD "snap" to those of low-noise SGD when the noise levels are equalized.

3. In Section 4, we further substantiate the "golden path" hypothesis in *function space*. We present evidence that models trained with varying levels of SGD noise learn similar functions, indicating that the differences in noise do not significantly impact the learned representations.

[1]Note that high SGD noise runs have more variance, and to remove this factor we equalize the noise of two trajectories before comparison. See Figure 6 for an example.

Overall, our work sheds new light on the role of batch size in online deep learning, showcasing that its benefits are merely computational. We also provide a pathway for a more unified understanding of training trajectories, by giving evidence that SGD takes noisy steps that approximate the "golden path" taken by gradient descent.

1.1. Related Work

Implicit Bias: A considerable volume of literature has been devoted to examining the impact of batch size on the training of neural networks from both theoretical (Andriushchenko et al., 2022; Lewkowycz et al., 2020; He et al., 2019; Nacson et al., 2022; Paquette et al., 2022a;b) and practical (Geiping et al., 2021; Nado et al., 2021; Masters & Luschi, 2018; Xing et al., 2018; Jastrzebski et al., 2020a; Karpathy, 2019) perspectives. Among practitioners, the consensus revolves around maximizing computational resources: large batch sizes are employed to fully exploit the hardware. However, regarding optimal hyperparameters, it is widely held that smaller batch sizes result in superior minima (Keskar et al., 2016; LeCun et al., 1998; Masters & Luschi, 2018; Goyal et al., 2017). Although some empirical studies (Geiping et al., 2021; Lee & Avestimehr, 2022; Novack et al., 2022; Hoffer et al., 2018) contest this notion by utilizing various techniques, it remains the prevalent intuition within the community. From a theoretical standpoint, several works showed the benefit of SGD noise as yielding a more favorable implicit bias (Damian et al., 2021; HaoChen et al., 2021; Ali et al., 2020; Blanc et al., 2020; Jastrzebski et al., 2020b). These works show that in certain overparameterized settings, higher SGD noise leads to a better generalization.

Within non-convex optimization literature, it is known that SGD can exhibit important optimization effects such as escaping saddle points and sharp minima (Kleinberg et al., 2018). A simple example of such escape behavior is provided in Figure 2 of Wu et al. (2018).

In convex optimization, however, diminishing the stochastic gradient descent (SGD) noise typically leads to enhanced performance. For instance, (Paquette et al., 2022a;b) establish that, under certain assumptions for high-dimensional random features models, a reduced batch size results in a worse test error for a given number of training steps. These works operate within a regime where number of data points scales proportionally with the model size, thereby aligning more closely with the "online learning" paradigm.

Offline vs. Online: One distinction absent from the afore-

mentioned discussion is the difference between the online and offline regimes. For instance, (Smith et al., 2021) clearly investigate the effect of SGD noise for a single epoch training, and show that it has an implicit bias towards reducing gradient norms. In contrast, we observe empirically that in the online setting, implicit bias doesn't affect the network in function space. The Deep Bootstrap framework of (Nakkiran et al., 2021; Ghosh et al., 2021) contrasts the online and offline worlds, revealing that a significant portion of offline training gains can be attributed to its online component. Recent works also demonstrated the detrimental effects of repeating even a small fraction of data (Hernandez et al., 2022; Xue et al., 2023), for LLMs.

Network Evolution. Similar to us, multiple works discuss the similarity of SGD dynamics across hyperparameter choices. This question has been studied from the lenses of example order (Nakkiran et al., 2019b; Shah et al., 2020; Hacohen et al., 2020; Baldock et al., 2021), representation similarity (Kornblith et al., 2019; Bansal et al., 2021), model functionality (Olsson et al., 2022), loss behaviour (Nakkiran et al., 2019a), weight space connectivity (Fort et al., 2020; Frankle et al., 2020) and the structure of the Hessian (Cohen et al., 2021). Our work focuses on the online regime, and as opposed to previous studies, gives evidence to the *Golden Path* conjecture in this regime wherein SGD noise strides (in function space) along the gradient descent trajectory but with noise. As shown by both our work and others, the golden path conjecture does *not* hold for offline learning.

2. The Implicit Bias of SGD in Online Learning

In this section, we present our experimental results on the impact of SGD noise (i.e. batch size) on implicit bias. We show that the effect of this noise can differ significantly between the *offline* and *online* regimes. Since our goal is to study the impact of SGD noise on implicit bias rather than on computational efficiency, in our experiments, we measure loss as a function of the number of gradient steps, and not as function of datapoints seen.

We conduct an experimental evaluation of our claims employing convolutional models in computer vision and Transformer models in natural language processing (NLP). Specifically, we run ResNet-18 on CIFAR-5m (Nakkiran et al., 2021), a synthetically generated version of CIFAR-10 with 5 million examples, ConvNext-T on ImageNet, and GPT-2-small on C4. To imitate the online regime with ImageNet, we only train for 10 epochs with data augmentation. For full experimental details, refer to Appendix A. As we show in Figure 3, we find that

1. In the **offline setting**, consistent with prior work (Keskar et al., 2016), SGD noise can (and often does) lead to better implicit bias for the final models. Specifically, even if runs with smaller noise initially[2] decrease the loss faster, eventually they get "stuck" at a worse local minima than the runs with higher SGD noise (smaller batch size). This is consistent with Scenario a1 of Figure 2 ("fork in the road" with high noise being better), where a higher noise enables escaping from bad local minima.

2. In contrast, in the **online setting**, the implicit bias advantage of SGD noise *completely disappears*, and the main benefit from small batch sizes reduces to being just computational. Specifically, after we control for computation (by measuring number of gradient steps), the low-noise runs consistently outperform the higher noise runs. This is consistent with either Scenario a2 of Figure 2 ("fork in the road" where a lower noise run can explore better minima) or with Scenario b (the "golden path": higher noise follow a similar trajectory but with some degradation due to noise).

First, we theoretically show, that for convex online setting, in particular quadratic loss, SGD path always remains close to the GD path in expectation, and SGD noise simply acts as variance on top of this path, thus supporting the golden path hypothesis (these results are similar in spirit to the results of Paquette et al. (2022a)).

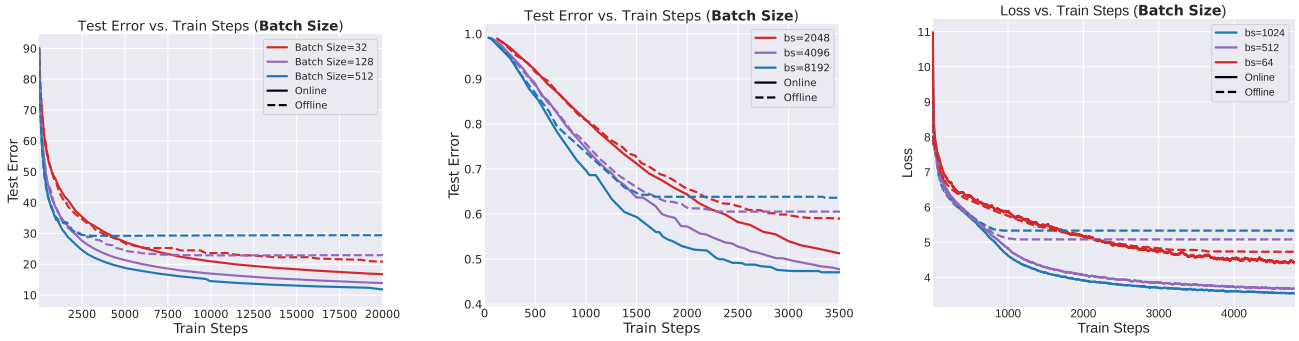
Theorem 2.1. *Consider the quadratic loss function given by $\mathcal{L}(w) = w^\top H w$, where H is a positive semi-definite matrix. With stochastic gradients (denoted by $g(w)$) modeled as additive gaussian noise, i.e, $g(w) = \nabla\mathcal{L}(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and for a fixed learning rate schedule, the following holds:*

- *Consider two different SGD runs, R_1 and R_2, starting from the same initialization w_0 and having noise variances given by σ_1 and σ_2 ($\sigma_1 \geq \sigma_2$). Then $\mathbb{E}[\mathcal{L}(w_{R_1}(t))] \geq \mathbb{E}[\mathcal{L}(w_{R_2}(t))]$.*
- *Let the noiseless GD run from the same initialization w_0 be denoted by $w_{GD}(t)$. Then $\mathbb{E}[w_{R_1}(t)] = \mathbb{E}[w_{R_2}(t)] = w_{GD}(t)$.*

Moreover, in Sections 3 and 4, our additional experimental results give evidence for the "golden path" case in practical scenarios.

Isn't this trivial? One natural intuition is that regularization is not needed in the online learning setup because there is no difference between optimization and generalization (or between train and test). However, we have empirically observed that well-known explicit regularization techniques such as weight decay can yield better performance in the

[2]The curves for offline learning initially track the online learning curves (as predicted by Deep Bootstrap (Nakkiran et al., 2021)) but then plateau at a higher loss for lower SGD noise.



(a) CIFAR-5m (b) ImageNet (c) C4

Figure 3. Test performance for ResNet-18 trained on CIFAR-5m (**left**), ConvNext-T on ImageNet (**middle**), and GPT-2-small on C4 (**right**) across varying batch sizes. Red corresponds to high SGD noise (small batch size), blue to low SGD noise (high batch size), and purple to an intermediate setting. Solid (resp. dotted) lines correspond to runs in the online (resp. offline) setting. For online learning, lower SGD noise runs consistently outperform higher noise runs per given step. Offline learning performance initially matches online performance, eventually runs with higher noise outperform low-noise runs. All experiments are averaged over ≥ 4 runs. See Figure 7, 8 for error bars and more hyperparameter values.

online regime for standard tasks (Figure 9, Appendix E.1), suggesting that this intuition might be flawed and more nuance is required.

3. Snapping Back to the Golden Path

The results of Section 2 show that SGD noise does not provide beneficial implicit bias in online learning. As we discussed in Figure 2, there are two potential explanations for why in online learning (unlike the offline case), decreasing SGD noise steers optimization towards a smaller-loss trajectory. One explanation is Scenario (a2) of the figure. Namely, it may be the case that choosing low SGD noise leads the optimization algorithm to a different (and better) trajectory, that is completely inaccessible to the high SGD-noise runs. The second is the "golden path" hypothesis: higher-noise runs travel on approximately the same path as lower-noise ones, suffering some loss-degradation resulting from the imperfect approximation. To rule out the first explanation, we conduct the following experiment in the online setting (see Figure 4, left):

1. Run two experiments—one with high batch size, and one with small batch size—for T_0 steps.

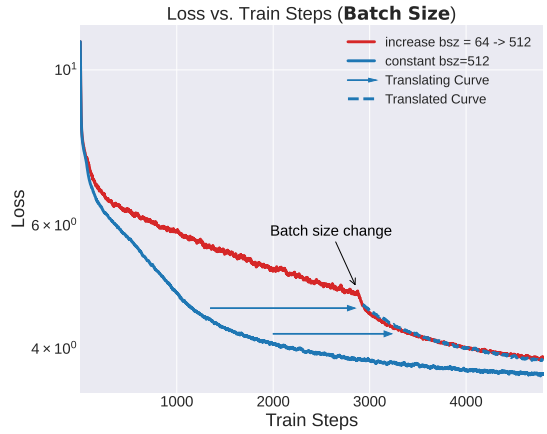
2. After T_0 steps, decrease the SGD noise by increasing the batch size of the second experiment to match the hyperparameters of the first one, and continue both runs.

Under the golden path hypothesis, we expect that shortly after increasing the batch size (i.e., at $T_0 + \tau$ for $\tau \ll T_0$), the loss curve would "snap" to the golden path, and continue following the same trajectory of the model that was trained with low SGD noise. On the other hand, the (a2) scenario of Figure 2 implies that decreasing the noise would not result in any significant change to the loss curve.

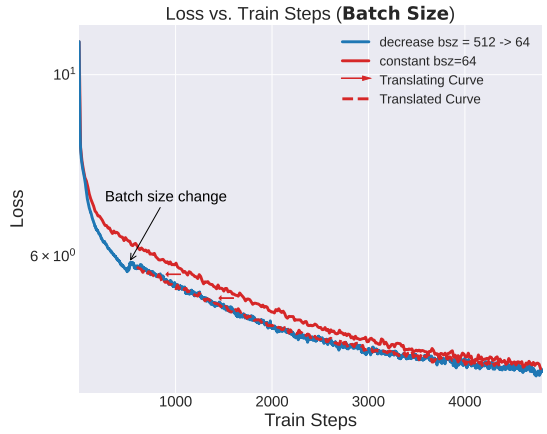
We perform a series of experiments as described above with ResNet-18 on CIFAR-5m and GPT2-small on the C4 dataset in Figure 4 ((a) and (c)). We consistently observe that, after dropping the SGD noise at some time T_0, the loss sharply improves to some value ℓ_0. From this point onward, the loss curve of the model is nearly identical to a right-translation of the loss curve of the model that was trained with low SGD noise from initialization. This phenomenon does not hold generally in offline learning, as it is well known that at convergence, different minima are reached by gradient descent and SGD.

In contrast to exploring the transition from a noisy trajectory towards the "golden path", we can also investigate the effects of introducing SGD noise, thereby deviating away from the optimal trajectory. To this end, we conduct an analogous experiment to the one presented before, but with a crucial difference: instead of reducing the noise in SGD at time T_0, we *increase* it by decreasing the batch size. Looking at the results for ResNet-18 on CIFAR-5m and GPT-2-small on the C4 dataset, we observe an immediate and significant increase in the loss upon introducing additional noise, as illustrated in Figure 4 ((b) and (d)). We see a similar phenomenon as in the experiments shown in Figure 4 (left), where the lower noise loss curve, after noising, follows a translated version of the higher noise curve (translation is not needed for the CIFAR-5m experiment as shown in Figure 4 bottom right).

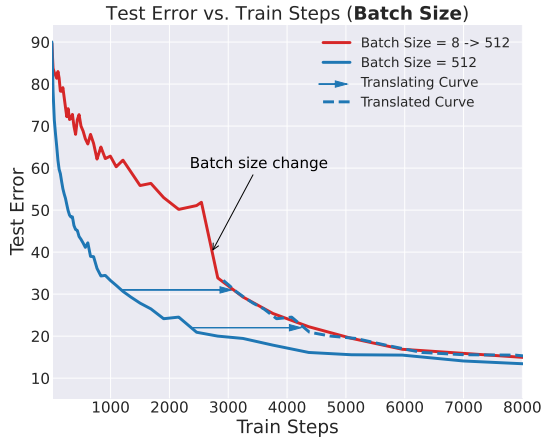
In Appendix E.2, we provide additional plots showing that the phenomenon is consistently observed even when varying the timestep when the batch size is increased.



(a) Increasing Batch Size: C4



(b) Decreasing Batch Size: C4



(c) Increasing Batch Size: CIFAR 5m

(d) Decreasing Batch Size: CIFAR 5m

Figure 4. Changing SGD noise (**left**: increasing batch size, **right**: decreasing batch size) during training for ResNet-18 on CIFAR-5m (bottom) and GPT2-small on C4 dataset (top). The red curves correspond to models trained with high SGD noise from initialization, and the blue curves trained with low SGD noise from initialization. In left plot the batch size is increased after T_0 steps while in right plot the batch size is decreased after T_0 steps. Across both experiments, changing batch size causes the original curve to follow a translated version (dashed) of new batch size curve (except for increasing batch size experiment on CIFAR-5m where no translation is required).

4. Pointwise Prediction Differences Between Trajectories

Section 3 offered empirical substantiation for the golden path hypothesis, but was restricted to claims about the *loss curve*. Now we investigate a stronger version of this hypothesis: the golden path in **Function Space.** Specifically, our goal is to verify whether trajectories using different SGD noise are functionally similar. Due to differing noise levels between the trajectories, we cannot directly compare their functional similarity; instead, we show that lowering the SGD noise causes models to "snap" to the golden path in a functional sense, i.e. when measuring the *pointwise* distance.

To be precise, we investigate the functional distance between models by measuring the average total variation (TV) distance[3] of their softmax probabilities on the test dataset. Figure 5 gives a schematic illustration of our experimental setting. In this model, we have a low noise SGD run to serve as the "ground truth" for the golden path. The blue curve in Figure 5 serves as a baseline for the functional distance (TV) to this "golden path" as it represents another low noise run trained from a different initialization. This is a strong baseline since different initialization seeds are often treated as a nuisance parameter in deep learning. We compare it to the TV distance of the high noise path depicted by the red curve.

As illustrated in Figure 5, under the "fork in the road" hypothesis, we expect that high noise and low noise trajecto-

[3]The TV distance is defined as half of the ℓ_1 distance for probability measures: $TV(p, q) = \frac{1}{2}\sum_i |p_i - q_i|$.

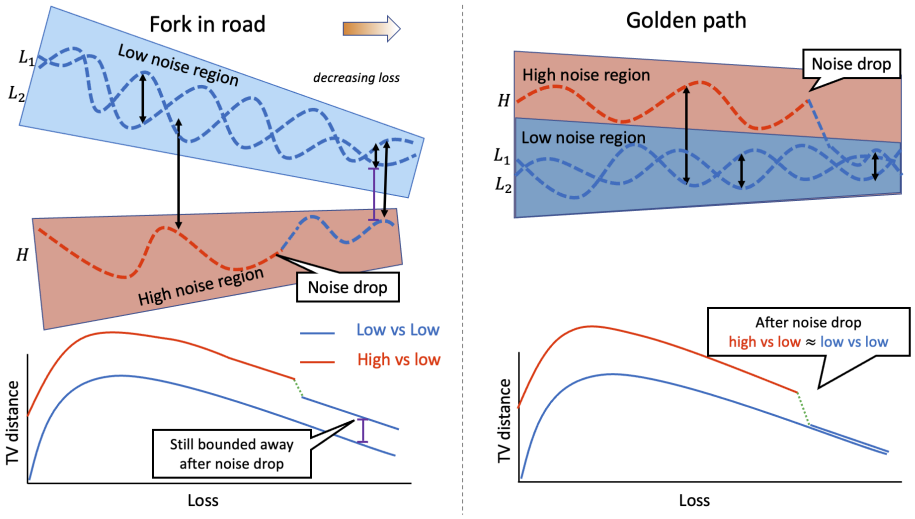
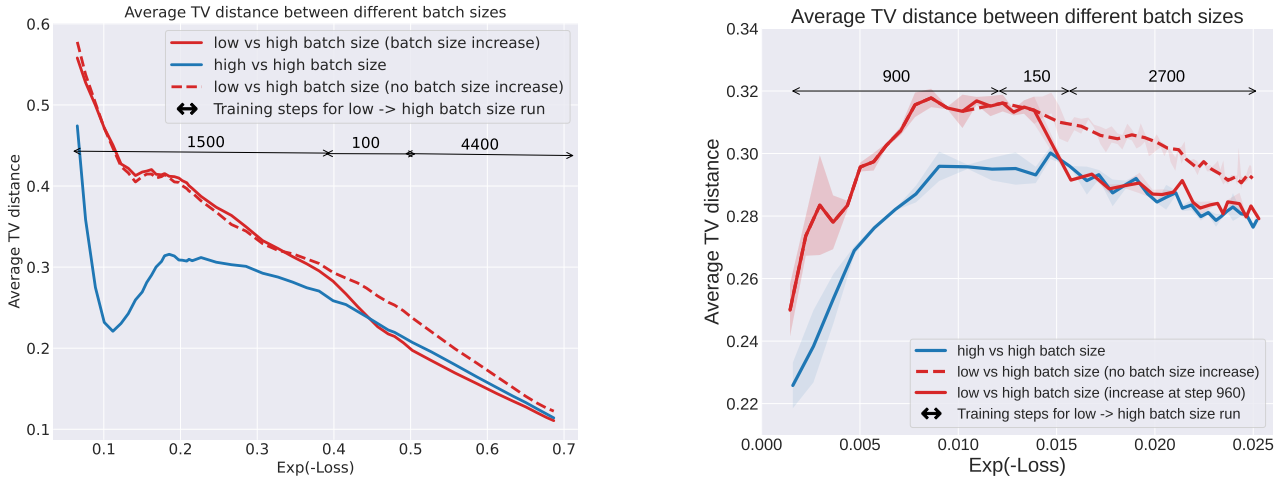


Figure 5. This figure illustrates the potential total variation distance in function space for two scenarios: "fork in the road" (left) and "golden path" (right). In the "fork in the road" scenario, training runs with low and high batch sizes explore different regions of the function space, leading to a consistently high total variation distance, even when the batch size is increased. In the "golden path" scenario, the low batch size run follows a noisy version of the high batch size trajectory. Increasing the batch size causes the trajectory to align with the high batch size path, resulting in a total variation distance similar to two independent high batch size trajectories.



(a) CIFAR-5m

(b) C4

Figure 6. Average TV distance behavior on CIFAR-5m (left) and C4 (right) with and without reducing SGD noise. For CIFAR-5m, we increase the batch size from 32 to 512. For C4, we increase the batch size from 64 to 512. The dashed line represents the run with constant high SGD noise as a baseline. The figure demonstrates that the distance between the model with reduced noise and a low-noise model at the same loss is nearly identical to the distance between two random low-noise models trained with the same hyperparameters. This is in contrast with the distance between a high noise and a low-noise model at the same loss. This observation supports the "golden path hypothesis" and contradicts the "fork in the road" hypothesis.

ries will explore different regions of the search space. Thus, even if dropping the noise improves loss, the baseline (distance between independent runs of low SGD noise) will still be significantly lower than the distance between the 'dropped-noise model' and the low-noise model.

In this section, by an empirical study on CIFAR-5m and C4 dataset, we show that this is not the case, and TV distance behaves as would be expected from the "golden path hypothesis". Specifically, as shown in Figure 6, after dropping the noise, the TV distance between this model and a low noise model at the same loss is virtually identical to the

TV distance between *two random low-noise models trained with identical hyper parameters*. Figure 6 also shows that this is not true for the high noise model, demonstrating that the two paths indeed differ due to the difference in the level of noise. Moreover, in Figure 6, it could also be seen that the high noise model reaches the "golden path" very quickly upon dropping the SGD noise. In Appendix E.3, we provide additional plots showing that this phenomenon holds across multiple different timesteps when the batch size is increased.

5. Discussion and Conclusion

The results of our investigation reveal a striking discrepancy between the online and offline learning regimes, in terms of the implicit bias of SGD. In the offline regime, this bias exhibits a beneficial regularization effect, whereas in the online regime, it merely introduces noise to the optimization path. This critical distinction between the two regimes has largely been overlooked in both theoretical and empirical research, with few studies explicitly addressing the difference. We argue that recognizing and accounting for the online versus offline learning regimes is crucial for understanding various deep learning phenomena and for informing the design of optimization algorithms.

Our experiments cover three practical datasets spanning vision and language, but further research may be needed to fully understand the generalizability of our findings. Although our work represents only an initial exploration into the disparities between the online and offline learning regimes, we can draw several immediate conclusions.

Implications for Practitioners. For online learning, our findings emphasize the relative *simplicity* of hyperparameter tuning, primarily focusing on computational efficiency and stability. In situations where data or computational resources are limited, however, the regularization effects of SGD become more significant, and hyperparameter selection and optimization take on greater importance. For instance, in the low-data regime, it may be crucial to use a smaller batch size, even if it results in not fully utilizing the GPU. In online training, this consideration appears to be consistently irrelevant.

Implications for Theorists. The primary takeaway for theoretical research is that the study of the regularizing effects of high SGD noise should be confined to the offline learning regime, as failing to make this distinction creates tension with practical applications. Theoretical findings that do not account for this difference can not fully capture why SGD is effective for deep learning. Furthermore, our observation that SGD follows a noisy trajectory near the "golden path" of gradient descent in loss and function spaces, coupled with the Deep Bootstrap (Nakkiran et al., 2021) assertion that a substantial portion of offline training can be explained by the online regime, implies that gradient descent may be instrumental in understanding many aspects of deep learning.

In conclusion, given that many large-scale deep learning systems, such as Language Models (LLMs), predominantly operate within the online learning regime, our findings challenge the conventional understanding of deep learning, primarily based on offline learning. We contend that it is necessary to reevaluate our comprehension of various deep learning phenomena in the context of online settings. Moreover, we propose gradient descent as a promising theoretical tool for studying online learning, considering the minimal impact of SGD noise on the functional trajectory.

Acknowledgements

We thank Preetum Nakkiran, Ben L. Edelman and Eran Malach for helpful comments on the draft.

NV, DM, RZ, GK and BB are supported by a Simons Investigator Fellowship, NSF grant DMS-2134157, DARPA grant W911NF2010021,and DOE grant DE-SC0022199. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. SK, DM, and RZ acknowledge funding from the Office of Naval Research under award N00014-22-1-2377 and the National Science Foundation Grant under award #CCF-2212841.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ali, A., Dobriban, E., and Tibshirani, R. J. The implicit regularization of stochastic gradient flow for least squares, 2020.

Andriushchenko, M., Varre, A., Pillaud-Vivien, L., and Flammarion, N. Sgd with large step sizes learns sparse features, 2022.

Baldock, R., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10876–10889. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5a4b25aaed25c2ee1b74de72dc03c14e-Paper.pdf.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. *CoRR*, abs/2106.07682, 2021. URL https://arxiv.org/abs/2106.07682.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/blanc20a.html.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Damian, A., Ma, T., and Lee, J. D. Label noise SGD provably prefers flat global minimizers. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=x2TMPhseWAW.

Fort, S., Dziugaite, G. K., Paul, M., Kharaghani, S., Roy, D. M., and Ganguli, S. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3259–3269. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/frankle20a.html.

Geiping, J., Goldblum, M., Pope, P. E., Moeller, M., and Goldstein, T. Stochastic training is not necessary for generalization. *CoRR*, abs/2109.14119, 2021. URL https://arxiv.org/abs/2109.14119.

Ghosh, N., Mei, S., and Yu, B. The three stages of learning dynamics in high-dimensional kernel methods, 2021.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017. URL http://arxiv.org/abs/1706.02677.

Hacohen, G., Choshen, L., and Weinshall, D. Let's agree to agree: Neural networks share classification order on real datasets. In *International Conference on Machine Learning*, pp. 3950–3960. PMLR, 2020.

HaoChen, J. Z., Wei, C., Lee, J., and Ma, T. Shape matters: Understanding the implicit bias of the noise covariance. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2315–2357. PMLR, 15–19 Aug 2021.

He, F., Liu, T., and Tao, D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/

dc6a70712a252123c40d2adba6a11d84-Paper.pdf.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9:1–42, 1997.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks, 2018.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022a.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., and Sifre, L. An empirical analysis of compute-optimal large language model training. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=iBBcRUlOAPR.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. J. Three factors influencing minima in SGD. *CoRR*, abs/1711.04623, 2017. URL http://arxiv.org/abs/1711.04623.

Jastrzebski, S., Arpit, D., Åstrand, O., Kerg, G., Wang, H., Xiong, C., Socher, R., Cho, K., and Geras, K. J. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. *CoRR*, abs/2012.14193, 2020a. URL https://arxiv.org/abs/2012.14193.

Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., and Geras, K. J. The break-even point on optimization trajectories of deep neural networks. *CoRR*, abs/2002.09572, 2020b. URL https://arxiv.org/abs/2002.09572.

Karpathy, A. A recipe for training neural networks, 2019. URL http://karpathy.github.io/2019/04/25/recipe/.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016. URL http://arxiv.org/abs/1609.04836.

Kleinberg, B., Li, Y., and Yuan, Y. An alternative view: When does SGD escape local minima? In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2698–2707. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kleinberg18a.html.

Komatsuzaki, A. One epoch is all you need. *arXiv preprint arXiv:1906.06669*, 2019.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. E. Similarity of neural network representations revisited. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL http://proceedings.mlr.press/v97/kornblith19a.html.

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K. R. *Efficient BackProp*, pp. 9–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-540-49430-0. doi: 10.1007/3-540-49430-8_2. URL https://doi.org/10.1007/3-540-49430-8_2.

Lee, S. and Avestimehr, S. Achieving small-batch accuracy with large-batch scalability via adaptive learning rate adjustment, 2022. URL https://openreview.net/forum?id=39Q__qgCpAH.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism. *CoRR*, abs/2003.02218, 2020. URL https://arxiv.org/abs/2003.02218.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL https://arxiv.org/abs/2201.03545.

Masters, D. and Luschi, C. Revisiting small batch training for deep neural networks. *CoRR*, abs/1804.07612, 2018. URL http://arxiv.org/abs/1804.07612.

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. Implicit bias of the step size in linear diagonal neural networks. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16270–16295. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nacson22a.html.

Nado, Z., Gilmer, J., Shallue, C. J., Anil, R., and Dahl, G. E. A large batch optimizer reality check: Traditional, generic optimizers suffice across batch sizes. *CoRR*, abs/2102.06356, 2021. URL https://arxiv.org/abs/2102.06356.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *CoRR*, abs/1912.02292, 2019a. URL http://arxiv.org/abs/1912.02292.

Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., and Barak, B. SGD on neural networks learns functions of increasing complexity. *CoRR*, abs/1905.11604, 2019b. URL http://arxiv.org/abs/1905.11604.

Nakkiran, P., Neyshabur, B., and Sedghi, H. The deep bootstrap framework: Good online learners are good offline generalizers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=guetrIHLFGI.

Novack, Z., Kaur, S., Marwah, T., Garg, S., and Lipton, Z. C. Disentangling the mechanisms behind implicit regularization in sgd, 2022.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads, 2022.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Implicit regularization or implicit conditioning? exact risk trajectories of SGD in high dimensions. In *NeurIPS*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/e9d89428e0ef0a70913845b3ae812ee0-Abstract-Conference.html.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Homogenization of sgd in high-dimensions: Exact dynamics and generalization properties, 2022b.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *CoRR*, abs/2006.07710, 2020. URL https://arxiv.org/abs/2006.07710.

Smith, S. L., Elsen, E., and De, S. On the generalization benefit of noise in stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9058–9067. PMLR, 2020. URL http://proceedings.mlr.press/v119/smith20a.html.

Smith, S. L., Dherin, B., Barrett, D. G. T., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=rq_Qr0c1Hyo.

Wu, L., Ma, C., and E, W. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf.

Xing, C., Arpit, D., Tsirigotis, C., and Bengio, Y. A walk with sgd, 2018.

Xue, F., Fu, Y., Zhou, W., Zheng, Z., and You, Y. To repeat or not to repeat: Insights from scaling llm under token-crisis, 2023.

A. Experimental Details

A.1. CIFAR-5m

In our CIFAR-5m experiments, we trained ResNet-18, on normalized (across channels) images and using the SGD optimizer with 0.9 momentum.

Section 2: For Figure 3 (a) we trained without any data augmentations. For both offline and online learning, we used a learning rate of 0.025. For offline learning we trained on a random subset of 50k samples (class balanced). We also used exponential moving average with coefficient of .8.

Section 3: For Figure 4 (c,d) we trained with standard CIFAR data augmentation of a random crop (`RandomCrop(32, padding=4)` in Pytorch) and horizontal flip (`RandomHorizontalFlip()` in PyTorch), and used a learning rate of 0.025.

Section 4: For Figure 6 (a), we trained on CIFAR-5m without any augmentations so as to remove the pointwise difference due to different augmentations. We trained two networks with learning rate 0.05, one with batch size 32 and the other with 512, both for 12000 steps (~ 1 epoch for batch size 512) and the one with batch size 32 was changed to 512, after 1500 training steps.

A.2. ImageNet

For the ImageNet experiments, we used ConvNext-T (Liu et al., 2022) and unless specified otherwise, used a batch size of 2048 and learning rate of $1e$-4 with the AdamW optimizer with weight decay 0.005. For all experiments, we use cosine decay scaling of the learning rate with respect to training steps (not epochs). We used the `RandomHorizontalFlip` and `RandomCrop` augmentations and also preprocessed the dataset to be resized to 256×256 using OpenCV before training for speed purposes. For the offline results, we downsample the dataset by a factor of 10, i.e., use 128k examples.

A.3. C4

For all experiments we trained GPT-2-small (124m parameters) on the C4 dataset with sequence length 2048. The optimizer we use is Adam without weight decay and a constant learning rate of 6×10^{-4}.

Section 2: For Figure 3 (**c**) we used a learning rate of 6×10^{-4}. For offline learning we trained on a random subset of roughly 100 million tokens, and all hyperparameters are otherwise identical with the analogous run in the online setting.

Section 3: In Figures 4a and 4b, we change the batch size from 64 to 512 and vice versa. When changing the batch size we keep the learning rate fixed to 6×10^{-4}. For Figure 4a, the change occurs after 60% of the training duration has elapsed (i.e. after 2880 steps) and for Figure 4b the change occurs after 960 steps.

Section 4: In Figure 6b we plot the average TV distance averaged across three pairs of different seeds for the (high vs high) and (low vs high) batch size runs. The average TV distance is computed over 500 batches of size 64 (32000 samples). The plotting begins after 50 training steps have elapsed until the end of the training duration.

B. Discussion of Results in Section 2

In Figure 3, we showed the stark contrast in offline vs online learning regarding their interaction with SGD noise. For the offline setting, higher SGD noise generally leads to a better performance, while in online learning, SGD noise only hurts performance.

However, upon closer observation, we can see that in all the plots of Figure 3, offline learning performance closely follows the online learning performance throughout the majority of the training period before reaching a plateau. This is in agreement with the results of (Nakkiran et al., 2021). They empirically demonstrated that, across a variety of scenarios, a major part of offline training can be explained away by online learning.

This result, combined with our observation about the SGD trajectory being a noisy version of the "golden path" exhibited by

gradient descent in online learning, shows that gradient descent is a useful theoretical tool even for studying offline learning. In particular, given that in practice, we choose hyperparameters to maximize test performance, this means that we move as far along the online trajectory as possible. The "denoised" version of this online trajectory is given by gradient descent on the population loss. Thus, our work proposes studying "gradient descent on population loss" as an alternative (or a "denoised" version) of studying SGD in offline learning.

C. Discussion of Results in Section 3

As exhibited in Figure 4, we observe the loss snapping phenomenon even when we increase SGD noise during training. Intuitively, the model that initially has a low amount of SGD noise (high batch size) has progressed "far along" the path, and thus the additional noise results in progress lost on the path and an increase in loss. One plausible reason for this could be the recent phenomenon of Edge-of-Stability (Cohen et al., 2021), where the authors showed that SGD dynamically leads the model to an "edge-of-stability" curvature, i.e, increasing SGD noise leads to an increase in loss in the successive steps. However, our paper shows that, following this brief increase in loss, the model is able to recover and continue training, albeit with higher SGD noise; thus the performance follows that of a noisier trajectory.

D. Proof of Theorem 2.1

Theorem. Consider the quadratic loss function given by $\mathcal{L}(w) = w^\top H w$, where H is a positive semi-definite matrix. With stochastic gradients ($g(w)$) modeled as additive gaussian noise, i.e, $g(w) = \nabla\mathcal{L}(w) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and for a fixed learning rate schedule, the following holds:

- Consider two different SGD runs, R_1 and R_2, starting from the same initialization w_0 and having noise variances given by σ_1 and σ_2 ($\sigma_1 \geq \sigma_2$). Then $\mathbb{E}[\mathcal{L}(w_{R_1}(t))] \geq \mathbb{E}[\mathcal{L}(w_{R_2}(t))]$.
- Let the noiseless GD run from the same initialization w_0 be denoted by $w_{GD}(t)$. Then $\mathbb{E}[w_{R_1}(t)] = \mathbb{E}[w_{R_2}(t)] = w_{GD}(t)$.

Proof. Consider the stochastic gradient descent update given by

$$w(t+1) = w(t) - \eta(t) g(w(t)) = (I - \eta(t)H)w(t) - \eta(t)\xi(t)$$

Unrolling the update, we get

$$w(t+1) = \left[\Pi_{i=0}^{t}(I - \eta(i)H)\right] w(0) - \sum_{i=0}^{t-1} \eta(i) \left[\Pi_{j=i+1}^{t}(I - \eta(j)H)\right] \xi(i) - \eta(t)\xi(t)$$

Taking expectation on both the sides, we get

$$\mathbb{E}[w(t+1)] = \left[\Pi_{i=0}^{t}(I - \eta(i)H)\right] w(0)$$

This shows that $\mathbb{E}[w_{R_1}(t)] = \mathbb{E}[w_{R_2}(t)] = w_{GD}(t)$.

Now, consider the loss given by $\mathcal{L}(w(t+1)) = w(t+1)^\top H w(t+1)$. Expanding it, we get three terms.

The first one is the loss of the GD trajectory given by

$$\left(\left[\Pi_{i=0}^{t}(I - \eta(i)H)\right] w(0)\right)^\top H \left(\left[\Pi_{i=0}^{t}(I - \eta(i)H)\right] w(0)\right)$$

The second term is the cross term between the noise and $w(0)$ term, given by

$$2\left(-\sum_{i=0}^{t-1} \eta(i) \left[\Pi_{j=i+1}^{t}(I - \eta(j)H)\right] \xi(i) - \eta(t)\xi(t)\right)^\top H \left(\left[\Pi_{i=0}^{t}(I - \eta(i)H)\right] w(0)\right)$$

Clearly the expectation of this term is zero as $\mathbb{E}[\xi(i)] = 0$ for all i.

The third term is the cross term between the noise terms given by

$$\left(\sum_{i=0}^{t-1} \eta(i) \left[\Pi_{j=i+1}^{t}(I - \eta(j)H)\right] \xi(i) + \eta(t)\xi(t)\right)^{\top} H \left(\sum_{i=0}^{t-1} \eta(i) \left[\Pi_{j=i+1}^{t}(I - \eta(j)H)\right] \xi(i) + \eta(t)\xi(t)\right)$$

Here, any term of the form $\xi(i)^{\top} M \xi(j)$ for some matrix M where $i \neq j$ has 0 expectation by independence and zero mean of $\xi(i)$ and $\xi(j)$. For the terms of the form $\xi(i)^{\top} M \xi(i)$, expanding $\xi(i)$ in the eigenbasis of H, we get that expectation of the third term scales with σ^2.

Thus, we can see, for two different runs R_1 and R_2 satisfying $\sigma_1 \geq \sigma_2$, $\mathbb{E}[\mathcal{L}(w_{R_1}(t))] \geq \mathbb{E}[\mathcal{L}(w_{R_2}(t))]$.

□

E. Additional Plots

E.1. Additional Plots for Section 2

In Figures 7 and 8 we show a subset of figures from Section 2 with error bars. Note that all figures in Section 2 were already averaged over ≥ 4 runs.

In Figure 9 we show an example in which explicit regularization helps in online learning. This is unlike the implicit regularization of SGD noise which does not help in our experiments in online learning.

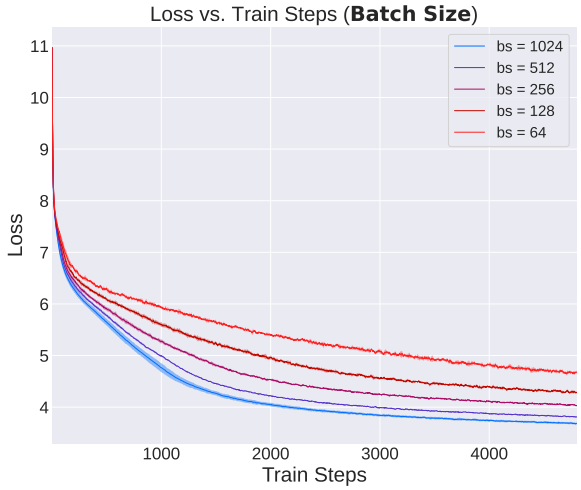


Figure 7. Experiments for online learning on the C4 dataset across several batch sizes with error bars. Consistent with our findings, runs with a lower amount of SGD noise yield better performance.

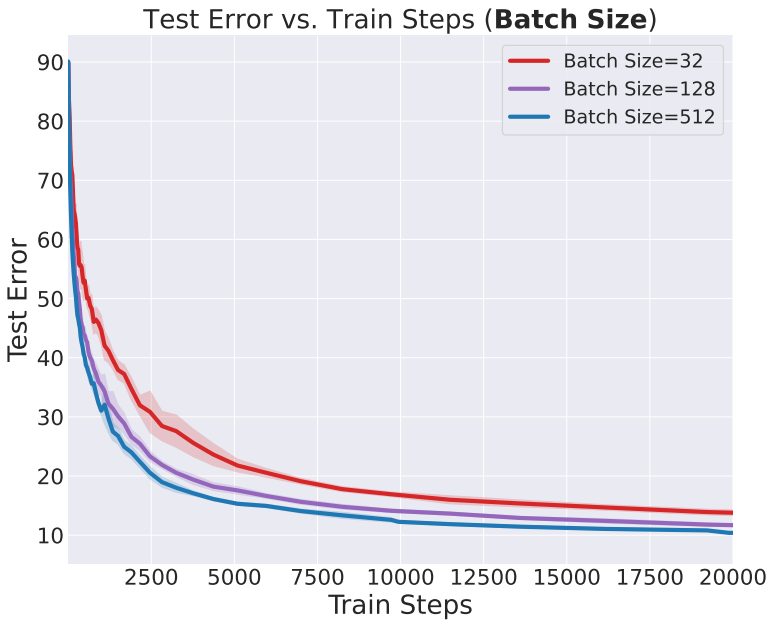


Figure 8. Test performance with error bars for CIFAR-5m in the online setting varying batch size. Note that all figures in Section 2 were already averaged over ≥ 4 runs.

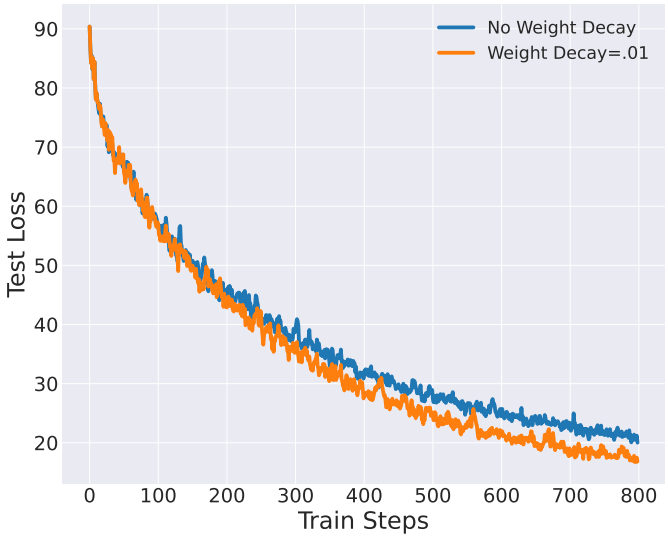


Figure 9. Explicit regularization can help in online learning. Specifically when training on the CIFAR 5m dataset, the test loss for a run using a small amount of weight decay is lower than that of a run with no weight decay.

In Figure 10 we plot Figure 3a but with x-axis as number of samples seen by the model rather than number of gradient steps. In this case we observe that larger batch sizes are worse.

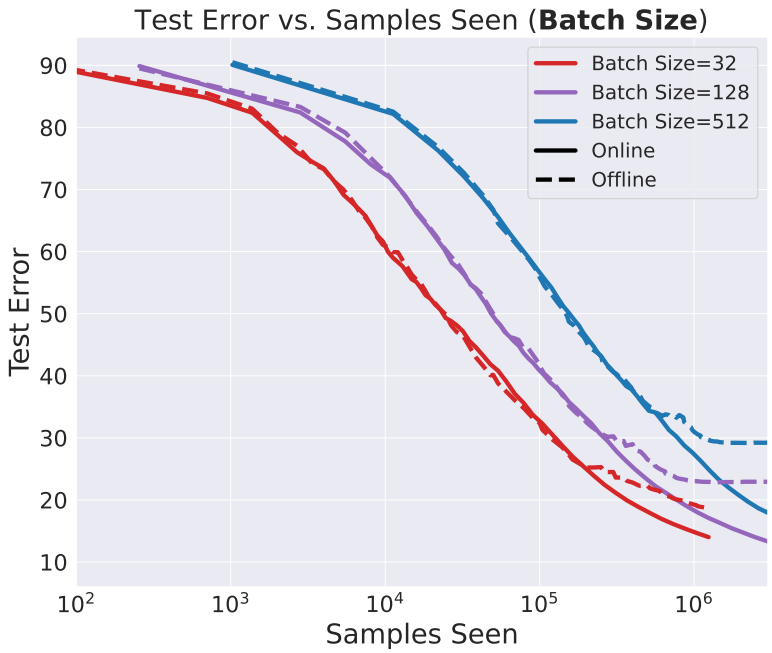


Figure 10. Test performance for Resnet18 on CIFAR-5m across varying batch sizes, with the number of examples seen in offline and online regime. The trend with number of examples remains similar in offline and online regime.

E.2. Additional Plots for Section 3

In Figure 11 we perform the same experiments in Section 3 on C4 but sweeping across the timestep when the batch size is increased from 64 to 512. We see that the loss curves shortly after increasing the batch size match a translated curve of the run with a constant batch size of 512.

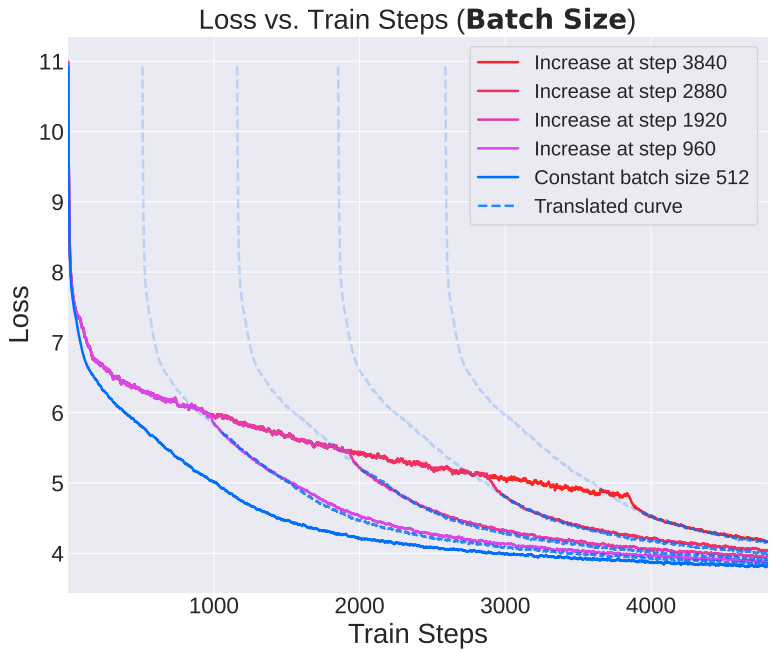


Figure 11. Experiments when switching the batch size from 64 to 512 at multiple points in training. As detailed in Section 4, matching losses shortly after increasing the batch size results in a translated curve corresponding to the constant high batch size run.

In Figure 12 we do the batch size increasing experiment for ImageNet.

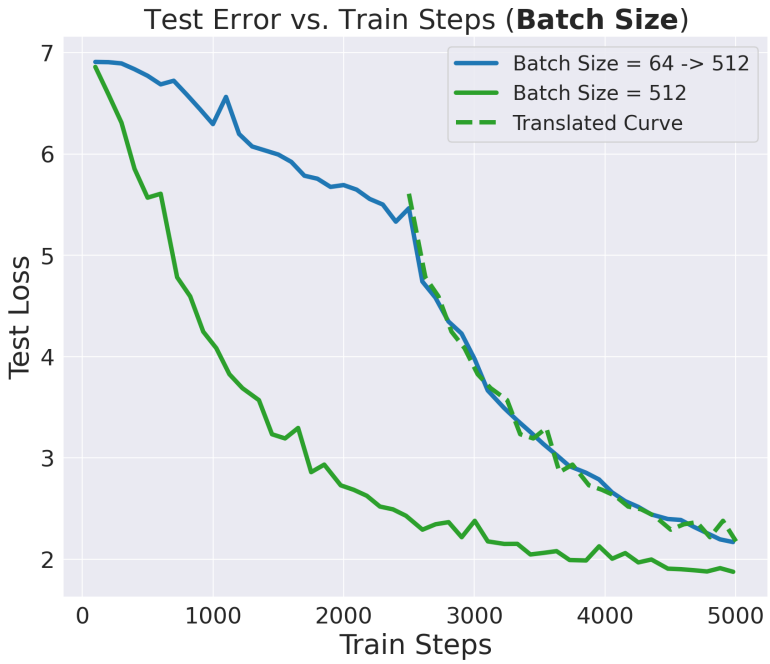


Figure 12. Changing SGD noise - increasing batch size during training for ConvNext on Imagenet. The red curves correspond to models trained with high SGD noise from initialization, and the blue curves trained with low SGD noise from initialization. In left plot the batch size is increased after T_0 steps while in right plot the batch size is decreased after T_0 steps. Across both experiments, changing batch size causes the original curve to follow a translated version (dashed) of new batch size curve

E.3. Additional Plots for Section 4

As in Appendix E.2, we perform the same experiments from Section 4 across different timesteps for when the batch size is increased. The lines are averaged across three pairs of different seeds. As seen in Figure 13, the results shown in Section 4 are consistent across all different points in which we increase the batch size during training.

Moreover, in Figure 15 and 14, we repeat the experiments of Section 4 for CIFAR-5m, but with different runs starting from the same initialization. This removes the initialization noise, and purely takes SGD noise into account.

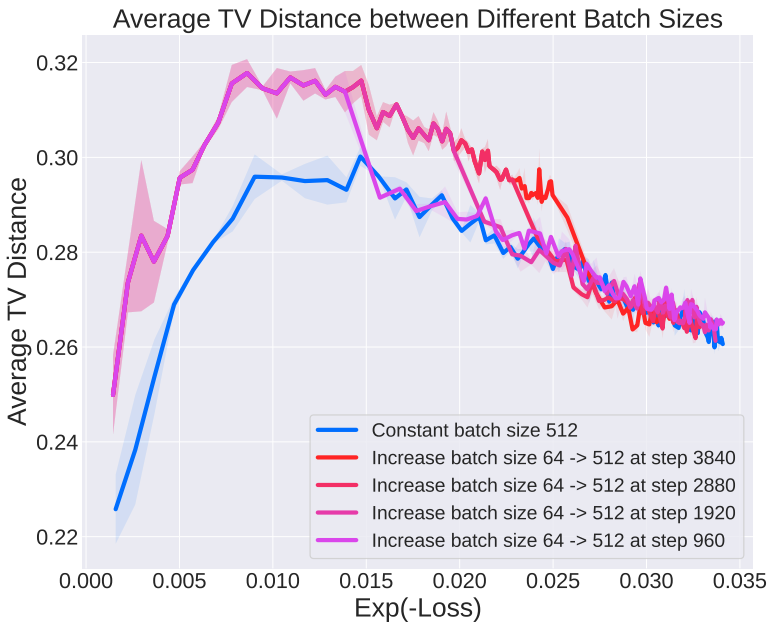


Figure 13. Examining the average TV distance between models when switching the batch size from 64 to 512 at multiple points in training. We can see that shortly after the increase in batch size, the average TV distance between this model and a model which has been trained with a constant batch size of 512 drops to match the average TV distance between two models trained with batch size 512 on two different seeds.

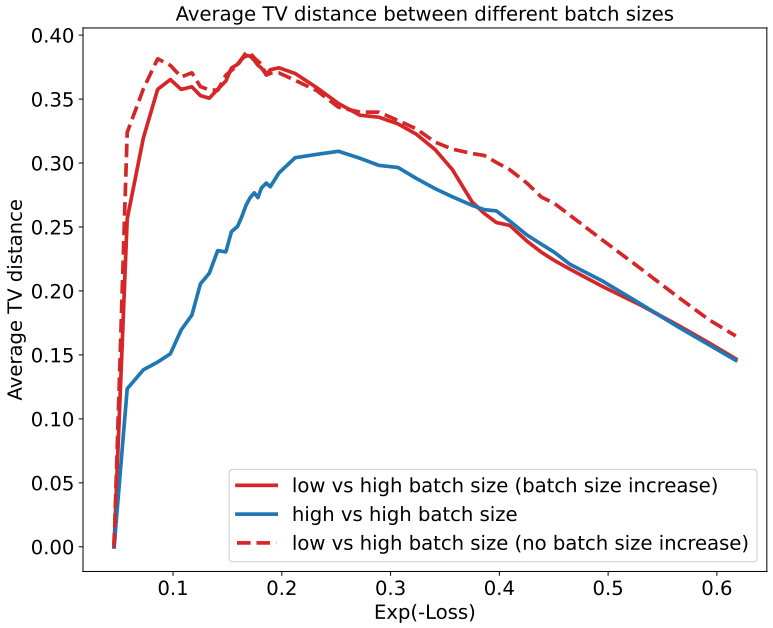


Figure 14. Average TV distance behavior on CIFAR-5m with and without reducing SGD noise, starting from the same initialization. For CIFAR-5m, we increase the batch size from 32 to 1024. The dashed line represents the run with constant high SGD noise as a baseline. The figure demonstrates that the distance between the model with reduced noise and a low-noise model at the same loss is nearly identical to the distance between two random low-noise models trained with the same hyperparameters. This is in contrast with the distance between a high noise and a low-noise model at the same loss. This observation supports the "golden path hypothesis" and contradicts the "fork in the road" hypothesis.

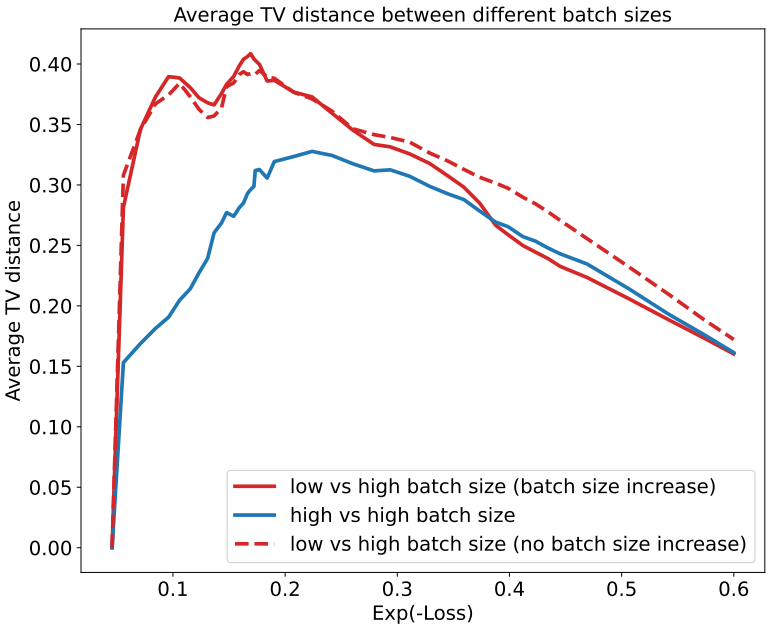


Figure 15. Average TV distance behavior on CIFAR-5m with and without reducing SGD noise, starting from the same initialization. For CIFAR-5m, we increase the batch size from 32 to 512. The dashed line represents the run with constant high SGD noise as a baseline. The figure demonstrates that the distance between the model with reduced noise and a low-noise model at the same loss is nearly identical to the distance between two random low-noise models trained with the same hyperparameters. This is in contrast with the distance between a high noise and a low-noise model at the same loss. This observation supports the "golden path hypothesis" and contradicts the "fork in the road" hypothesis.