A Space-Time Approach to Analyzing Stochastic Gradient Descent

Anonymous Authors[1]

Abstract

We present a diagrammatic calculus for reasoning about the behavior, at small learning rates, of SGD and its variants. We interpret the diagrams as histories of scattering events, thus offering a new physical analogy for descent. Illustrating this technique, we construct a regularizing term that causes large-batch GD to emulate small-batch SGD, present a model-selection heuristic that depends only on statistics measured before optimization, and exhibit a counter-intuitive loss landscape wherein SGD eternally cycles counterclockwise around a circle of minima.

IDEA: ASCENT?

Fashion Mnist and CIFAR 10

1. Introduction

Stochastic gradient descent (SGD) decreases an unknown objective l by performing discrete-time steepest descent on noisy estimates of l. A key question is how the noise affects the final objective value. We connect SGD dynamics to physical scattering theory, thus providing a quantitative and qualitative toolkit for answering this question.

Specifically, we derive a diagram-based formalism for reasoning about SGD via a path integral over possible interactions between weights and data. The formalism permits perturbative analysis, leading to predictions of learning curves for small η. Unlike the continuous-time limits of previous work, this framework models discrete time, and with it, the potential **non-Gaussianity** of noise. We thus obtain new results quantifying the **effect of epoch number, batch size, and momentum** on SGD test loss. We also contrast SGD against popular continuous-time approximations such as ordinary or stochastic differential equations (ODE, SDE).

Path integrals offer not only quantitative predictions but also an exciting new viewpoint — that of iterative optimization as a **scattering process**. Much as individual Feynman diagrams (see Dyson (1949a)) depict how local particle interactions compose into global outcomes, our diagrams depict how individual SGD updates influence each other before affecting a final test loss. In fact, we import from physics tools such as **crossing symmetries** (see Dyson (1949b)) and **re-normalization** (see Gell-Mann & Goldberger (1954)) to simplify our calculations and refine our estimates. The diagrams' combinatorial properties immediately yield several precise qualitative conclusions as well, for instance that to order η^2, **inter-epoch** shuffling does not affect expected test loss.

1.1. Related Work

It was Kiefer & Wolfowitz (1952) who, in uniting gradient descent (Cauchy, 1847) with stochastic approximation (Robbins & Monro, 1951), invented SGD. Since the development of back-propagation for efficient differentiation (Werbos, 1974), SGD and its variants have been used to train connectionist models including neural networks (Bottou, 1991), in recent years to remarkable success (LeCun et al., 2015).

Several lines of work quantify the overfitting of SGD-trained networks (Neyshabur et al., 2017a). For instance, Bartlett et al. (2017) controls the Rademacher complexity of deep hypothesis classes, leading to generalization bounds that are post hoc or optimizer-agnostic. However, since deep networks trained via SGD generalize despite their seeming ability to shatter large sets (Zhang et al., 2017), one infers that generalization arises from the aptness to data of not only architecture but also optimization (Neyshabur et al., 2017b). Others have focused on the implicit regularization of SGD itself, for instance by modeling descent via stochastic differential equations (SDEs) (e.g. Chaudhari & Soatto (2018)). However, as explained by Yaida (2019), such continuous-time analyses cannot treat covariance correctly, and so they err when interpreting results about SDEs as results about SGD for finite trainset sizes.

Following Roberts (2018), we avoid making a continuous-time approximation by instead Taylor-expanding around the learning rate $\eta = 0$. In fact, we develop a diagrammatic language for evaluating each Taylor term that is similar

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

to and inspired by the field theory methods popularized by Dyson (1949a). Using this technique, we quantify the overfitting effects of batch size and epoch number, and based on this analysis, propose a regularizing term that causes large-batch GD to emulate small-batch SGD, thus establishing a precise version of the Covariance-BatchSize-Generalization relationship conjectured in Jastrzębski et al. (2018).

While we make rigorous, architecture-agnostic predictions of learning curves, these predictions become vacuous for large η. In particular, while our work does not assume convexity of the loss landscape, it also is blind to large-ηT convergence of SGD. Other discrete-time dynamical analyses allow large η by treating deep generalization phenomenologically, whether by fitting to an empirically-determined correlate of Rademacher bounds (Liao et al., 2018), by exhibiting generalization of local minima **flat** with respect to the standard metric (see Hoffer et al. (2017), Keskar et al. (2017), citetwa18), or by exhibiting generalization of local minima **sharp** with respect to the standard metric (see Stein (1956), Dinh et al. (2017), Wu et al. (2018)). Our work, which makes explicit the dependence of generalization on the underlying metric and on the form of gradient noise, reconciles those latter, seemingly clashing claims.

Others have imported the perturbative methods of physics to analyze descent dynamics: Dyer & Gur-Ari (2019) perturb in inverse network width, employing 't Hooft diagrams to compute deviations of non-infinitely-wide deep learning from Gaussian processes. Meanwhile, (Chaudhari & Soatto, 2018) and Li et al. (2017) perturb in learning rate to second order by approximating noise between updates as gaussian and uncorrelated. This approach does not generalize to higher orders, and, because correlations and heavy tails are essential obstacles to concentration of measure and hence of generalization, it does not model the generalization behavior of SGD. By contrast, we use Penrose diagrams to compute test and train losses to arbitrary order in learning rate, quantifying the effect of non-gaussian and correlated noise. We hence extend Roberts (2018) beyond leading order and beyond 2 time steps, allowing us to compare, for instance, the expected test losses of multi-epoch and single-epoch SGD.

2. Background and Notation

2.1. A Smooth Stage: Tensor Conventions

We adopt summation notation for Greek but not Roman indices, suppressing indices when convenient and clear. To expedite dimensional analysis, we follow (Bonnabel, 2013) in considering the learning rate as an inverse metric $\eta^{\mu\nu}$ that converts a gradient (row vector) into a displacement (column vector). Viewing η^{-1} as the only available flat metric, we will use η to raise indices; for example, with C denoting the covariance of gradients, its "trace" will be $C^{\mu}_{\mu} = \eta^{\mu\nu} C_{\mu\nu}$. The standard syntactic constraints on indexed expressions then give a strong check on which expressions transform naturally with respect to optimization dynamics.

We assume that every all moments of the 0th and higher derivatives of the losses l_n, considered as random functions on weight space, exist and are infinitely differentiable.

Kolář gives a careful introduction to these differential geometric ideas (1993).

2.2. Combinatorial Costumes: Structure Sets

We make use of *structure sets*, i.e. sets S equipped with a preorder $\leq$ and an equivalence relation $\sim$. The morphisms of structure sets are non-decreasing maps that preserve $\sim$ and its negation. A structure set is *pointed* if it has a unique maximum element and this element forms a singleton $\sim$-class. The categories $\mathcal{S}$ of structure sets and $\mathcal{P}$ of pointed structure sets enjoy a free-forgetful adjunction $\mathcal{F}, \mathcal{G}$. Modding out a structure set S by its $\sim$ yields another structure set $\mathcal{M}(S)$.

A *diagram* is a rooted tree equipped with an equivalence relation on nodes. We draw the tree of $\leq$ by thin edges, with the root at the far right, and we draw the equivalence relation $\sim$ by fuzzy ties. By reading the tree as a Hasse graph, we see that each diagram D induces a pointed structure set, by abuse of notation also named D. A map from this induced D to a total order with finest $\sim$ is an *ordering* of D.

Fong gives a swift introduction to these category theoretic and diagrammatic ideas (2019).

2.3. The Parameterized *Personae*: Forms of SGD

SGD decreases an objective l by updating on smooth, unbiased i.i.d. estimates $(l_n : 0 \leq n < N)$ of l. The pattern of updates is determined by a structure set § whose preorder is a total preorder: for a map $\pi : S \to [N]$ that induces $\sim$, we define SGD inductively as $\mathrm{SGD}_S(\theta) = \theta$ when S is empty and otherwise

$$\mathrm{SGD}_S(\theta) = \mathrm{SGD}_{S \setminus M}(\theta^{\mu} - \eta^{\mu\nu} \nabla_{\nu} l_M(\theta))$$

where $M = \min S \subseteq S$ specifies a batch and $l_M = \frac{1}{M} \sum_{m \in M} l_{\pi(m)}$ is a batch average. Since the distribution of l_n is permutation invariant, the non-canonical choice of π does not affect the distribution of output θs.

Of special interest are structure sets that divide into $M \times B$ many *epochs* each with N/B many disjoint *batches* of size B. An SGD instance is then determined by N, B, M, and an *inter-epoch shuffling scheme*. The cases $B = 1$ and $B = N$ we call *pure SGD* and *pure GD*.

2.4. The Tempting Tool: Taylor Series

Intuitively, each descent step displaces θ by $-\eta\nabla l$ and hence decreases the loss $l(\theta)$ by $\eta(\nabla l)^2$; thus, we expect after T steps a net decrease of $T\eta(\nabla l)^2$:

$$l(\theta_T) \approx l(\theta_0) - T\cdot\eta\cdot(\nabla l(\theta_0))^2 \qquad (1)$$

This intuition fails to capture two crucial facts: **curvature** — that as θ changes during training, so may $\nabla l(\theta)$ — and **noise** — that l_n and l may differ.

To account for noise, we should replace each $(\nabla l_t)(\nabla l)$ by an expectation. If we are interested in train instead of test loss, We get some expectations of the form $(\nabla l_t)(\nabla l_t)$, and hence obtain a different result than for test loss.

To account for curvature, FILL IN

3. Diagram Calculus for SGD

3.1. Role of Diagrams

Suppose s is smooth on weight space; for example, s may be a test or train loss. We may track $s(\theta)$ as θ is updated by SGD as follows:

Key Lemma 1. *The formal Maclaurin series of $s(\theta_T)$ with respect to η is:*

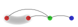
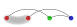
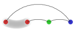
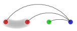
$$\sum_{0\le d<\infty} (-\eta)^d \sum_{\substack{(d_t:0\le t<T)\\ \sum_t d_t=d}} \left(\prod_{0\le t<T} \left.\frac{(g\nabla)^{d_t}}{d_t!}\right|_{g=\nabla l_t(\theta)}\right) s(\theta_0)$$

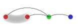
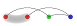
In averaging over training sets (and hence over the sequence $(l_t : 0\le t<T)$ considered as a random variable), we may factor the expectation of the above product according to independence relations between the l_t. We view various training procedures (e.g. GD, SGD with(out) inter-epoch shuffling) as **prescribing different independence relations** that lead to different factorizations and hence to potentially different generalization behavior at each order of η.

An instance of the above product (for $s = l_a$ drawn from a test set and $0\le c\le b<T$) is $\eta^3(\nabla l_c\nabla)^2(\nabla l_b\nabla)l_a$, which is

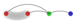
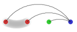
$$(\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla_\lambda\nabla_\mu\nabla^\nu l_b)(\nabla_\nu l_a) + (\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla_\lambda\nabla^\nu l_b)(\nabla_\mu\nabla_\nu l_a)$$
$$+(\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla_\mu\nabla^\nu l_b)(\nabla_\lambda\nabla_\nu l_a) + (\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla^\nu l_b)(\nabla_\lambda\nabla_\mu\nabla_\nu l_a)$$

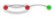
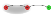
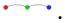
To reduce clutter, we adapt the string notation of Penrose (1971). Then, in expectation over (l_c, l_b, l_a) drawn i.i.d.:

$\cdots =$  $+$  $+$  $+$  (2)

$=$ 2  $+$ 2  (3)

$2\ \mathbb{E}[(\nabla l)(\nabla l)]\ \mathbb{E}[\nabla\nabla\nabla l]\ \mathbb{E}[\nabla l]$ $\quad$ $2\ \mathbb{E}[(\nabla l)(\nabla l)]\ \mathbb{E}[\nabla\nabla l]\ \mathbb{E}[\nabla\nabla l]$

Above, each node corresponds to a loss function (here, red for l_c, green for l_b, blue for l_a), differentiated d times for a degree-d node (for instance, l_b is differentiated thrice in the first diagram and twice in the second). **Thin "edges"** mark contractions by η. **Fuzzy "ties"** denote independence relationships by connecting identical loss functions (here, l_c with l_c): nodes not connected by a path of fuzzy ties are independent. The colors are redundant with the fuzzy ties and used only so that we may concisely refer to a specific node in prose. The value of a diagram is the expected value of the corresponding tensor expression. Crucially, for a fixed, i.i.d. distribution over (l_c, l_b, l_a), **the topology of a diagram determines its value**. For instance,  $=$  because both are trees with two leaves tied. Thus follows the simplification on the second line above. As shown with braces, we may convert back to explicit tensor expressions, invoking independence between untied nodes to factor the expression. However, as we will see, the diagrams offer physical intuition, streamline computations, and determine useful unbiased estimators of the statistics they represent.

We define a diagram with fuzzy outlines instead of fuzzy ties to be the difference between the fuzzy tied version and the completely untied version:  $=$  $-$ .

The recipes for writing down test (or train) losses of SGD and its variants are straight-forward in the diagram notation because they reduce the problem of evaluating the previous dynamical expressions to the problem of counting isomorphic graphs. The more complicated the direct computation, the greater the savings of using diagrams. An appendix provides details and proofs for a variety of situations. For now, we focus on the test loss of SGD.

3.2. Recipe for the Test Loss of SGD

Our results all follow from this theorem and its analogues:

Theorem 1. *The order η^d contribution to the expected test loss of SGD is:*

$$(-1)^d \sum_D \sum_{f:D\to\mathcal{F}(S)} \prod_{i\in S} \frac{1}{|f^{-1}(i)|!} D \qquad (4)$$

where D ranges over (isomorphism classes of) diagrams with d edges and f ranges over morphisms in $\mathcal{P}$.

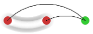
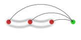
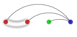
In the special case of $B = 1, M = 1$:

Proposition 1. *The order η^d contribution to the expected test loss of one-epoch SGD with singleton batches is:*

$$\frac{(-1)^d}{d!} \sum_D |\mathcal{P}(D\to[P])| \binom{N}{P-1} \binom{d}{d_0,\cdots,d_{P-1}} D \qquad (5)$$

where D ranges over d-edged diagrams whose equivalence classes are each totally disconnected (else, the coefficient is 0) and have sizes $d_p : 0\le p\le P$, with $d_P = 1$.

A P-part, d-edged diagram then contributes $\Theta\left((\eta N)^d N^{P-d-1}\right)$ to the loss. For example, there are six diagrams to third order, and they have $(4+2)+(2+2+3)+(1)$ many orderings. See Table 1. Intuitively, ηN measures the **physical time** of optimization, and $1/N$ measures **coarseness** of time discretization. More precisely, we have a double-series in $(\eta N)^d N^{P-d-1}$, where d counts thin edges and $d+1-P$ counts fuzzy ties; the $P = d+1$ terms correspond to a discretization-agnostic (hence continuous-time, noiseless) ODE approximation to SGD, while $P \leq d$ gives correction terms modeling time-discretization and hence noise.

Corollary 1. *For one-epoch SGD on singleton batches through fixed physical time T: the order N^{-1} deviation of SGD's test loss from ODE's is* $\frac{T^2 N^{-1}}{2}$  *. The order N^{-2} deviation of SGD's test loss due to non-gaussian noise is* $\frac{T^3 N^{-2}}{6}\Big($  $-\,3$  $\Big)$.

For finite N, these effects make SDE different from SGD. SDE also fails to model the correlations between updates in multiepoch SGD. On the other hand, in the $N = \infty$ limit for which SDE matches SGD, optimization and generalization become computationally intractable and trivial, respectively.

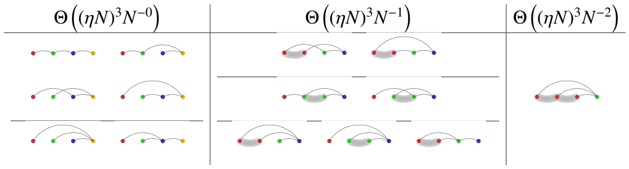
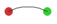
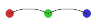
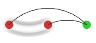
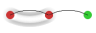

$\Theta\left((\eta N)^3 N^{-0}\right)$
$\Theta\left((\eta N)^3 N^{-1}\right)$
$\Theta\left((\eta N)^3 N^{-2}\right)$

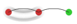
Table 1. Degree-3 scattering diagrams for $B = M = 1$ SGD's test loss. **Left:** $(d, P) = (3, 3)$. Diagrams for ODE behavior. **Center:** $(d, P) = (3, 2)$. 1st order deviation of SGD away from ODE. **Right:** $(d, P) = (3, 1)$. 2nd order deviation of SGD from ODE with appearance of non-Gaussian statistics.

Proposition 2. *To second order in η, the test loss of SGD — on N samples for M epochs with batch size B dividing N and with any shuffling scheme — has expectation*

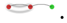
$-\,MN$  $+\,MN\left(MN-\frac{1}{2}\right)$ 

$+\,MN\left(\frac{M}{2}\right)$  $+\,MN\left(\frac{M-\frac{1}{B}}{2}\right)$ 

Corollary 2. *To second order in η, inter-epoch shuffling doesn't affect SGD's expected test loss.*

Corollary 3. *To second order in η, one-epoch SGD has* $\left(\frac{M-1}{M}\right)\left(\frac{B+1}{B}\right)\left(\frac{N}{2}\right)$  *less test loss than M-epoch SGD with learning rate η/M.*

Given an unbiased estimator $\hat{C}$ of gradient covariance, we may get GD to mimic SGD:

Corollary 4. *The expected test loss of pure SGD is, to second order in η, less than that of pure GD by* $\left(\frac{M}{2}\right)\left(\frac{N-1}{N}\right)$ . *Moreover, GD on a modified loss $\tilde{l}_n = l_n + \left(\frac{N-1}{4N^2}\right)\hat{C}_\nu^\nu(\theta)$ has an expected test loss that agrees with SGD's to second order.*

3.3. Renormalization

An important idea is that of renormalization, i.e. the summarization of myriad small-scale interactions into an effective large-scale theory. We can use this two ways: (**A**) to refine our computations if we know the hessian; (**B**) to refine our computations if we know the "effective propagator". Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

3.4. Descent as Scattering

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

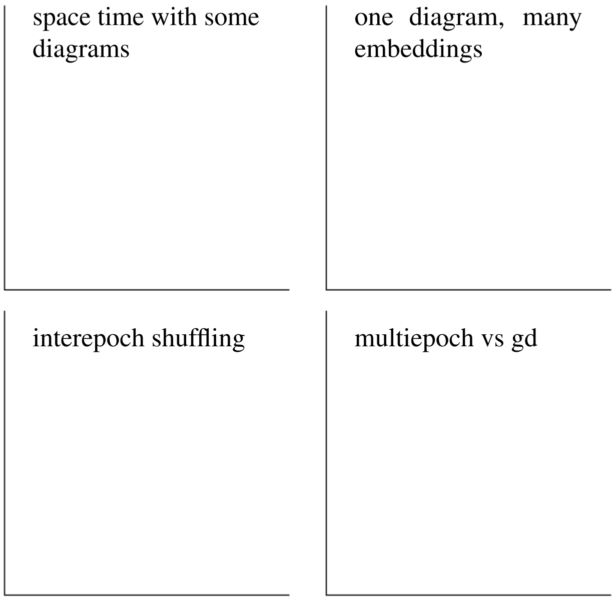


Figure 1. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4. Consequences and Applications

4.1. Vanilla SGD

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4.2. Emulating Small Batches with Large Ones

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

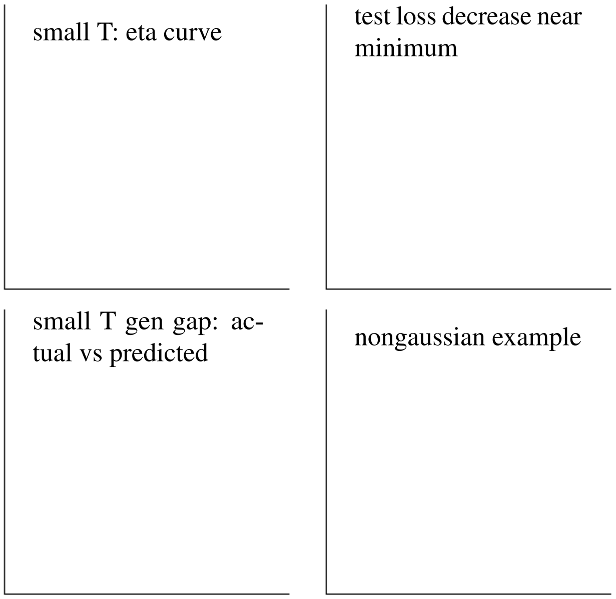


Figure 2. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4.3. Analyzing Second Order Methods

We demonstrate how our approach extends to more sophisticated optimizers by analyzing momentum and a hessian-based method.

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

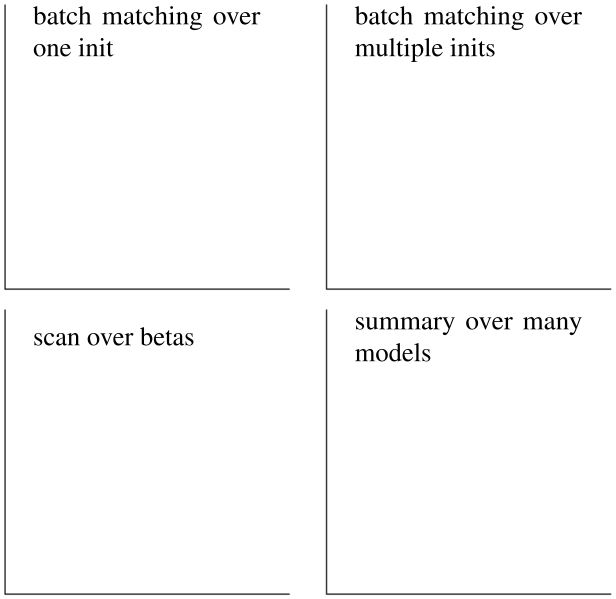


Figure 3. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

...sunt in culpa qui officia deserunt mollit anim id est laborum. momentum

Now consider a hessian-based update parameterized by a scalar λ:

$$\theta \leftsquigarrow \theta - (\eta^{-1} + \lambda \nabla\nabla l_t(\theta))^{-1} \nabla l_t(\theta)$$

invhess

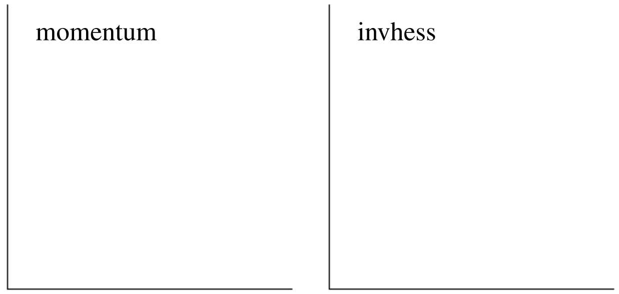


Figure 4. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4.4. Epochs and Overfitting

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

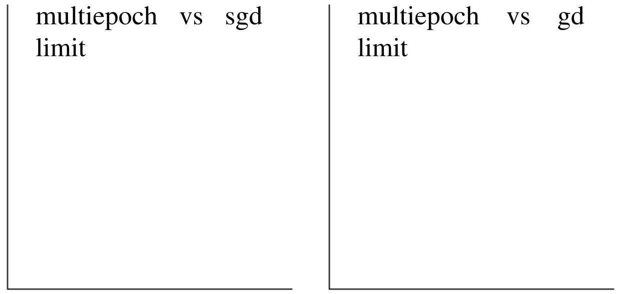


Figure 5. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4.5. Myopic Model Selection

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

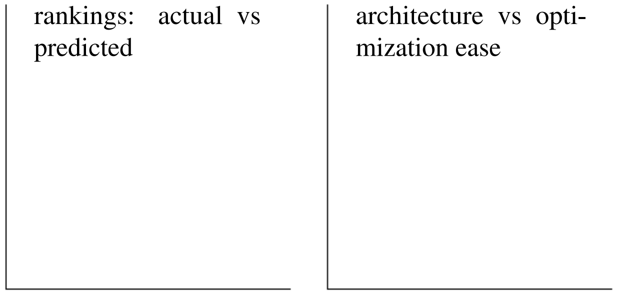


Figure 6. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4.6. Comparison to Continuous Time

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Also, sgd interepoch correlations

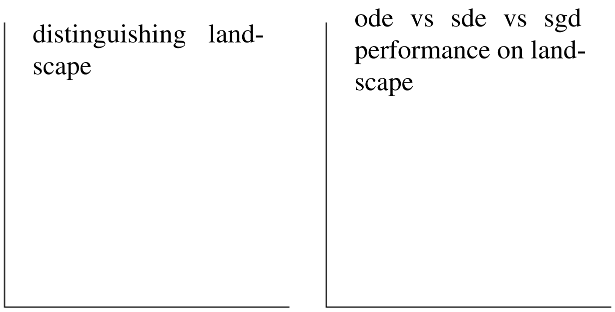


Figure 7. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

4.7. Thermodynamic Engine

We clarify Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. We constructed a counter-intuitive loss landscape wherein, for arbitrarily small learning rates, SGD cycles counterclockwise around a circle of minima. Our mechanism differs from that discovered by Chaudhari & Soatto (2018) discuss the thermodynamic significance of both

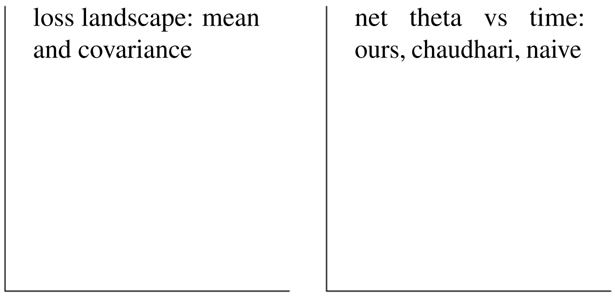


Figure 8. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

5. Conclusion

We presented a novel diagrammatic tool for analyzing gradient-based descent. Via a new regularizing term, we showed that **large-batch GD can be made to emulate small-batch SGD**, thus completing a project suggested by Roberts (2018). This is significant because, while small batch sizes can lead to better generalization (Bottou, 1991), modern infrastructure increasingly rewards large batch sizes (Goyal et al., 2018). We showed also that in multi-epoch SGD, inter-epoch shuffling induces only a 3rd order effect on test loss. Intuitively, we proved that **the hessian matters asymptotically more than shuffling order**.

The diagram method is also a rich source of intuitions and physical analogies. For example, it offers a clearer understanding of the empirically verified limit cycles found in Chaudhari. As our physical analogy emphasizes the underlying metric, it reconciles competing views of whether sharp or flat minima generalize. Further exploration of this bridge to particle physics, especially within the framework of renormalization theory, pose a promising direction for future research.

Variances

5.1. Acknowledgements

We thank Dan A. Roberts and Sho Yaida for patient introductions to their work and for precisely posing several of the questions we answer here. We feel deeply grateful to Sho Yaida and Josh B. Tenenbaum for their compassionate guidance. We appreciate the generosity of Andrzej Banburski and Wenli Zhao in offering crucial feedback on writing.

References

Bartlett, P., Foster, D., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 2017.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 2013.

Bottou, L. Stochastic gradient learning in neural networks. *Neuro-Nîmes*, 1991.

Cauchy, A.-L. Méthode générale pour la résolution des systémes d'équations simultanées. *Comptes rendus de l'Académie des Sciences*, 1847.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *ICLR*, 2018.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. *ICLR*, 2017.

Dyer, E. and Gur-Ari, G. Asymptotics of wide networks from feynman diagrams. *ICML Workshop*, 2019.

Dyson, F. The radiation theories of tomonaga, schwinger, and feynman. *Physical Review*, 1949a.

Dyson, F. The s matrix in quantum electrodynamics. *Physical Review*, 1949b.

Fong, B. and Spivak, D. An invitation to applied category theory. *Cambridge University Press*, 2019.

Gell-Mann, M. and Goldberger, M. Scattering of low-energy photons by particles of spin $\frac{1}{2}$. *Physical Review*, 1954.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd. *Data @ Scale*, 2018.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better. *NeurIPS*, 2017.

Jastrzębski, S., Kenton, Z., Arpit, D., N., B., Fischer, A., Y., B., and A., S. Three factors influencing minima in sgd. *Arxiv Preprint*, 2018.

Keskar, N., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*, 2017.

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 1952.

Kolář, I., Michor, P., and Slovák, J. Natural operations in differential geometry. *Springer*, 1993.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 2015.

Li, Q., Tai, C., and E, W. Stochastic modified equations and adaptive stochastic gradient algorithms i. *PMLR*, 2017.

Liao, Q., Miranda, B., Banburski, A., Hidary, J., and Poggio, T. A surprising linear relationship predicts test performance in deep networks. *Center for Brains, Minds, and Machines Memo 91*, 2018.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *NeurIPS*, 2017a.

Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. Geometry of optimization and implicit regularization in deep learning. *Chapter 4 from Intel CRI-CI: Why and When Deep Learning Works Compendium*, 2017b.

Penrose, R. Applications of negative dimensional tensors. *Combinatorial Mathematics and its Applications*, 1971.

Robbins, H. and Monro, S. A stochastic approximation method. *Pages 400-407 of The Annals of Mathematical Statistics.*, 1951.

Roberts, D. Sgd implicitly regularizes generalization error. *NeurIPS: Integration of Deep Learning Theories Workshop*, 2018.

Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Berkeley Symposium on Mathematical Probability*, 1956.

Werbos, P. Beyond regression: New tools for prediction and analysis. *Harvard Thesis*, 1974.

Wu, L., C., M., and E, W. How sgd selects the global minima in over-parameterized learning. *NeurIPS*, 2018.

Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. *ICLR*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

A. Derivation of Diagram Rules

5.2. Dyson Series for Iterative Optimizers

If a density ρ governs a point θ in weight space, then after a sequence of updates $\theta \mapsto \theta - \eta^{\mu\nu}\nabla_\mu l(\theta)$ on losses $(l_t : 0 \leq t < T)$, the following density (up to an error term whose Taylor series vanishes; all perturbative results will implicitly carry such terms) will govern the new point:

$$\exp\left(+\eta^{\mu\nu}\nabla_\mu l_{T-1}(\theta)\nabla_\nu\right)\cdots\exp\left(+\eta^{\mu\nu}\nabla_\mu l_0(\theta)\nabla_\nu\right)\rho \qquad (6)$$

or $\prod \exp\left(+\eta\nabla l\nabla\right)\rho$ for short. The exponent above is a linear operator that acts on a space of sufficiently smooth maps; in particular, the ∇_ν does not act on the $\nabla_\mu l(\theta)$ with which it pairs. Integrating by parts, we write the expectation over initial values after T steps of a function s of weight space (e.g. s may be test or train loss) as:

$$\int_\theta \rho(\theta)\left(\prod_{0\leq t\leq T}\exp\left(-\eta^{\mu\nu}\nabla_\mu l(\theta)\nabla_\nu\right)s\right)(\theta) \qquad (7)$$

Since the exponentials above might not commute, we may not compose the product of exponentials into an exponential of a sum. We instead compute an expansion in powers of η. Setting the initialization $\rho(\theta) = \delta(\theta - \theta_0)$ to be deterministic, and labeling as θ_t the weight after t steps, we find:

$$s(\theta_T) = \sum_{0\leq d<\infty}(-\eta)^d \sum_{\substack{(d_t:0\leq t<T)\\ \sum_t d_t=d}}\left(\prod_{0\leq t<T}\frac{(\nabla l_t(\theta)\nabla)^{d_t}}{d_t!}\right)s(\theta_0) \qquad (8)$$

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

B. Tutorial on Diagram Rules

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

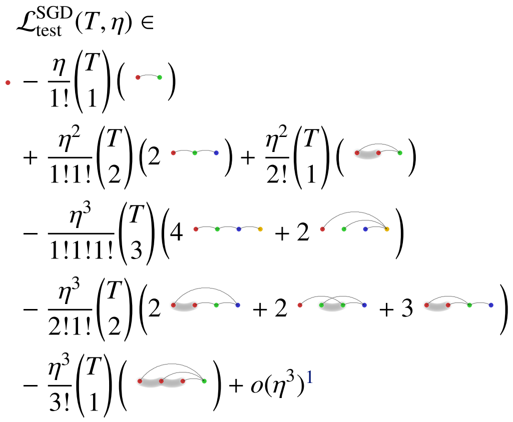
...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

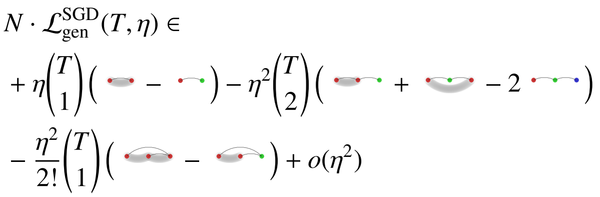
C. Derivations of Perturbative Results

For single-epoch SGD with singleton batches, we sum all relevant diagrams through order 3; the coefficients $4, 2; 2, 2, 3; 1$ come from counting the elements of Table 1, and the other coefficients come from analogous tables. This yields:

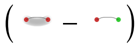


$$\mathcal{L}^{\mathrm{SGD}}_{\mathrm{test}}(T,\eta) \in - \frac{\eta}{1!}\binom{T}{1}\left(\quad \right) + \frac{\eta^2}{1!1!}\binom{T}{2}\left(2 \quad \right) + \frac{\eta^2}{2!}\binom{T}{1}\left(\quad \right) - \frac{\eta^3}{1!1!1!}\binom{T}{3}\left(4 \quad + 2 \quad \right) - \frac{\eta^3}{2!1!}\binom{T}{2}\left(2 \quad + 2 \quad + 3 \quad \right) - \frac{\eta^3}{3!}\binom{T}{1}\left(\quad \right) + o(\eta^3)$$ [1]

By contrast, the generalization gap $\mathcal{L}^{\mathrm{SGD}}_{\mathrm{gen}} = \mathcal{L}^{\mathrm{SGD}}_{\mathrm{test}} - \mathcal{L}^{\mathrm{SGD}}_{\mathrm{train}}$ is suppressed by a factor $1/N$ $(N \leq T)$:



$$N \cdot \mathcal{L}^{\mathrm{SGD}}_{\mathrm{gen}}(T,\eta) \in + \eta\binom{T}{1}\left(\quad - \quad \right) - \eta^2\binom{T}{2}\left(\quad + \quad - 2 \quad \right) - \frac{\eta^2}{2!}\binom{T}{1}\left(\quad - \quad \right) + o(\eta^2)$$

The leading order term is $N \cdot \mathcal{L}^{\mathrm{SGD}}_{\mathrm{gen}}(T,\eta) \approx \eta T$  $= T \cdot \eta^{\lambda\mu} C_{\lambda\mu}$, where C is the covariance of gradients. We thus recover a main result of Roberts (2018).

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

[1] We use little-$o(\eta^d)$ instead of big-$O(\eta^{d+1})$ to avoid specializing to analytic functions. Error terms depend on the loss landscape and on T. When gradients are uniformly bounded, the T dependence is at most linear.

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

D. Diagram Rules vs Direct Perturbation

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

E. The η-Series’ Domain of Convergence

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

F. Autonomous ODE Fitting

We fit a . These have the benefit In particular:

$$y\prime(t) = a \rightarrow \qquad y(t) = y(0) + at \tag{9}$$

$$y\prime(t) = by + a \rightarrow \quad y(t) = (y(0) - (a/b))\exp(bt) + (a/b) \tag{10}$$

$$y\prime(t) = cy^2 + by + a \rightarrow \tag{11}$$

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

G. Generalized Bessel Factors

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.