

# A Perturbative Analysis of Stochastic Descent

**Samuel C. Tenka**

Computer Science and AI Lab  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
`colimit@mit.edu`


June 1, 2020

## Abstract

We analyze Stochastic Gradient Descent (SGD) at small learning rates. Unlike prior analyses based on stochastic differential equations, our theory models discrete time and hence non-Gaussian noise. We prove that gradient noise systematically pushes SGD toward flatter minima. We characterize when and why flat minima overfit less than other minima. We generalize the Akaike Info. Criterion (AIC) to a smooth estimator of overfitting, hence enabling gradient-based model selection. We show how non-stochastic GD with a modified loss function may emulate SGD. We verify our predictions on convnets for CIFAR-10 and Fashion-MNIST.

## 1 Introduction

Practitioners benefit from the intuition that SGD approximates noiseless GD Bottou [1991]. In this paper, we refine that intuition by showing how gradient noise biases learning toward certain areas of weight space. Departing from prior work, we model discrete time and hence non-Gaussian noise. Indeed, we derive corrections to continuous-time, Gaussian-noise approximations such as ordinary and stochastic differential equations (ODE, SDE). For example, we construct a loss landscape on which SGD eternally cycles counterclockwise, a phenomenon impossible with ODEs. Leaving the rigorous development of our general theory to **APPENDIX**, our paper body highlights and intuitively discusses the theory’s main corollaries.

Our work offers a novel viewpoint of SGD as many concurrent interactions between weights and data. Diagrams such as , analogous to those of Feynman [1949], Penrose [1971], depict these interactions. In the appendix, we discuss this bridge to physics — and its relation to Hessian methods and natural GD — as topics for future research. We also discuss how this work may ameliorate or exacerbate the learning community’s disproportionate contribution to climate change. More broadly, our work adds to the body of theory on optimization in the face of uncertainty, theory that may one day inform solutions to emerging issues in user privacy and pedestrian safety.

## 1.1 Example of diagram-based computation of SGD's test loss



If we run SGD for  $T$  gradient steps with learning rate  $\eta$  starting at weight  $\theta_0$ , then by Taylor expansion we may express the expected test loss of the final weight  $\theta_T$  in terms of statistics of the loss landscape evaluated at  $\theta_0$ . Our technical contribution is to organize the computation of this Taylor series via combinatorial objects we call *diagrams*:


**Main Idea** (Informal). We can enumerate all diagrams, and assign to each diagram a number depending on  $\eta, T$ , such that summing these numbers over all diagrams yields SGD's expected test loss. Restricting to diagrams with  $\leq d$  edges leads to  $o(\eta^d)$  error.

Deferring details to later sections and appendices, we illustrate this work flow. First, let  $l_x(\theta)$  be weight  $\theta$ 's loss on datapoint  $x$ . We define a tensor  $\leftrightarrow$  diagram dictionary:

$$\begin{aligned} G &\triangleq \mathbb{E}_x [\nabla l_x(\theta)] \triangleq \text{red arrow} \\ H &\triangleq \mathbb{E}_x [\nabla \nabla l_x(\theta)] \triangleq \text{red double arrow} & C &\triangleq \mathbb{E}_x [(\nabla l_x(\theta) - G)^2] \triangleq \text{red arrow loop} \\ J &\triangleq \mathbb{E}_x [\nabla \nabla \nabla l_x(\theta)] \triangleq \text{red triple arrow} & S &\triangleq \mathbb{E}_x [(\nabla l_x(\theta) - G)^3] \triangleq \text{red arrow loop with tail} \end{aligned}$$

Here,  $G, H, J$  denote the loss's derivatives w.r.t.  $\theta$ , and  $G, C, S$  denote the gradient's cumulants w.r.t. the randomness in  $x$ . Each  $\nabla^d l_x$  corresponds to a node with  $d$  edges emanating, and fuzzy outlines group nodes that occur within the same expectation.

We may pair together the loose ends of the above (and of analogues with more edges) to obtain *diagrams*.<sup>1</sup> E.g., we may join  $C = \text{red arrow loop}$  with  $H = \text{red double arrow}$  to get . As another example, we may join two copies of  $G = \text{red arrow}$  with two copies of  $H = \text{red double arrow}$  to get . Intuitively, each diagram represents the interaction of its parts: of gradients ( $G$ ), noise ( $C, S, \dots$ ) and curvature ( $H, J, \dots$ ). **APPENDIX** interprets these diagrams physically.

**Example 1.** Does non-Gaussian noise affect SGD? Specifically, since the skew  $S$  measures non-gaussianity, let's compute how  $S$  affects test loss. The recipe is to identify the fewest-edged diagrams containing  $S = \text{red arrow loop with tail}$ . In this case, there is one fewest-edged diagram — ; it results from joining  $S$  with  $J = \text{red triple arrow}$ . To evaluate a diagram, we multiply its components (here,  $S, J$ ) with exponentiated  $\eta H$ 's, one for each edge:

$$-\frac{\eta^3}{3!} \sum_{\mu\nu\lambda} S_{\mu\nu\lambda} \frac{1 - \exp(-T\eta(H_{\mu\mu} + H_{\nu\nu} + H_{\lambda\lambda}))}{\eta(H_{\mu\mu} + H_{\nu\nu} + H_{\lambda\lambda})} J_{\mu\nu\lambda}$$

This is  $S$ 's leading order contribution to SGD's test loss written in an eigenbasis of  $\eta H$ .

**Remark 1.** For large  $T$  and isotropic  $\eta H$ , this becomes  $-(\eta^3/3!) \sum_{\mu\nu\lambda} S_{\mu\nu\lambda} J_{\mu\nu\lambda} / 3\eta |H|$ . Since  $J = \nabla H$ ,  $J/|H|$  measures the relative change in curvature  $H$  w.r.t.  $\theta$ . So non-gaussian noise affects SGD proportion to the logarithmic derivative of curvature.

<sup>1</sup> A diagram's colors and geometric layout lack meaning: we **color** only for convenient reference, e.g. to a diagram's "green nodes". Only the topology of a diagram — not its size or angles — appear in our theory.

## 1.2 Background, Notation, and Assumptions

We sometimes implicitly sum repeated Greek indices: if a covector  $A$  and a vector  $B$ <sup>1</sup> have coefficients  $A_\mu, B^\mu$ , then  $A_\mu B^\mu \triangleq \sum_\mu A_\mu \cdot B^\mu$ . We regard the learning rate as an inverse metric  $\eta^{\mu\nu}$  that converts gradient covectors to displacement vectors [Bonnabel, 2013]. We use the learning rate  $\eta$  to raise indices: e.g.,  $H^\mu_\lambda \triangleq \eta^{\mu\nu} H_{\nu\lambda}$  and  $C^\mu_\mu \triangleq \sum_{\mu\nu} \eta^{\mu\nu} \cdot C_{\nu\mu}$ . Though  $\eta$  is a tensor, we may still define  $o(\eta^d)$ : a quantity  $q$  *vanishes to order*  $\eta^d$  when  $\lim_{\eta \rightarrow 0} q/p(\eta) = 0$  for some homogeneous degree- $d$  polynomial  $p$ .

We fix a loss function  $l : \mathcal{M} \rightarrow \mathbb{R}$  on a space  $\mathcal{M}$  of weights. We fix a distribution  $\mathcal{D}$  from which unbiased estimates of  $l$  are drawn. We write  $l_x$  for a generic sample from  $\mathcal{D}$  and  $(l_n : 0 \leq n < N)$  for a training sequence drawn i.i.d. from  $\mathcal{D}$ . We refer both to  $n$  and to  $l_n$  as *training points*. We assume Appendix FILL IN’s hypotheses, e.g. that  $l, l_x$  are analytic and that moments exist. E.g., our theory models tanh networks with cross entropy loss on bounded data — with arbitrary weight sharing, skip connections, soft attention, dropout, and weight decay.

Our general theory describes SGD with any number  $N$  of training points,  $T$  of updates, and  $B$  of points per batch. SGD then runs  $T$  many updates (i.e.  $E = TN/B$  epochs, i.e.  $M = T/N$  updates per point)  $\theta^\mu := \theta^\mu - \eta^{\mu\nu} \nabla_\nu \sum_{n \in \mathcal{B}_t} l_n(\theta)/B$ , where  $\mathcal{B}_t$  is the  $t$ th batch. Our paper’s body — but not appendices — will assume **SGD has**  $E = B = 1$  **and GD has**  $T = B = N$  unless otherwise stated.

## 1.3 Related Work

Several research programs treat the overfitting of SGD-trained networks [Neyshabur et al., 2017a]. E.g., Bartlett et al. [2017] controls the Rademacher complexity of deep hypothesis classes, leading to optimizer-agnostic generalization bounds. Yet SGD-trained networks generalize despite their ability to shatter large sets [Zhang et al., 2017], so generalization must arise from not only architecture but also optimization [Neyshabur et al., 2017b]. Others approximate SGD by SDE to analyze implicit regularization (e.g. Chaudhari and Soatto [2018]), but, per Yaida [2019a], such continuous-time analyses cannot treat SGD noise correctly. We avoid these pitfalls by Taylor expanding around  $\eta = 0$  as in Roberts [2018]; unlike that work, we generalize beyond order  $\eta^1$  and  $T = 2$ .


Our theory is vacuous for large  $\eta$ . Other analyses treat large- $\eta$  learning phenomenologically, whether by finding empirical correlates of gen. gap [Liao et al., 2018], by showing that *flat* minima generalize (Hoffer et al. [2017], Keskar et al. [2017], Wang et al. [2018]), or by showing that *sharp* minima generalize (Stein [1956], Dinh et al. [2017], Wu et al. [2018]). Our theory reconciles these clashing claims.

Prior work analyzes SGD perturbatively: Dyer and Gur-Ari [2019] perturb in inverse network width, using ’t Hooft diagrams to correct the Gaussian Process approximation for specific deep nets. Perturbing to order  $\eta^2$ , Chaudhari and Soatto [2018] and Li et al. [2017] assume uncorrelated Gaussian noise, so they cannot describe SGD’s gengap. We use Penrose diagrams to compute test losses to arbitrary order  $\eta$ . We allow for correlated, non-Gaussian noise and thus *any* smooth architecture. E.g., we do not assume information-geometric relationships between  $C$  and  $H$ ,<sup>2</sup> so we may model VAEs.

<sup>1</sup> Vectors/covectors are also called column/row vectors.

<sup>2</sup> Disagreement of  $C$  and  $H$  is typical in modern learning [Roux et al., 2012, Kunstner et al., 2019].


## 2 Theory, Specialized to $E = B = 1$ SGD's Test Loss

A *diagram* is a finite rooted tree equipped with a partition of its nodes obeying the *path condition*: no path from leaf to root may encounter any part more than once. We specify the root by drawing it rightmost. We draw the parts of the partition by grouping the nodes within each part via fuzzy outlines. A diagram is *irreducible* when each of its degree-2 nodes is in a part of size one. An *embedding*  $f$  of a diagram  $D$  is an injection from the diagram's parts to (integer) times  $0 \leq t \leq T$  that sends the root to  $T$  and such that, for each path from leaf to root, the corresponding sequence of times is increase. E.g.,  $f$  might send 's

red part to  $t = 3$  and its green part to  $t = 4$ , but not vice versa. Let  $|\text{Aut}_f(D)|$  count the graph automorphisms of  $D$  that commute with  $f$ .

Up to unbiasing terms,<sup>1</sup> the *re-summed value*  $\text{rvalue}_f(D)$  is constructed as follows.

**Node rule:** insert a factor a  $\nabla^d l_x$  for each degree  $d$  node. **Edge rule:** for each edge whose endpoints  $f$  sends to times  $t, t'$ , insert a factor of  $K^{|t'-t|-1}\eta$  where  $K \triangleq (I - \eta H)$ .

**Outline rule:** group the nodes in each part within expectation brackets  $\mathbb{E}_x$ . E.g., if  $f$  maps 's red part to time  $t = T - \Delta t$ , then (the red part gives  $S$ ; the green part,  $J$ ):

$$\text{rvalue}_f \left( \text{diagram} \right) = S_{\mu\lambda\rho} (K^{\Delta t-1}\eta)^{\mu\nu} (K^{\Delta t-1}\eta)^{\lambda\sigma} (K^{\Delta t-1}\eta)^{\rho\pi} J_{\nu\sigma\pi}$$

In fact, we may integrate this expression per Remark 2 to recover Example 1.

### 2.1 Main result


Theorem 1 expresses SGD's test loss as a sum over diagrams. A diagram with  $d$  edges scales as  $O(\eta^d)$ , so the following is a series in  $\eta$ . We will truncate the series to small  $d$ , thus focusing on few-edged diagrams and easing the combinatorics of embeddings.

**Theorem 1** (Special Case). *For any  $T$ : for  $\eta$  small enough, SGD has expected test loss*

$$\sum_{\substack{\text{irreducible} \\ \text{diagrams } D}} \sum_{f \text{ of } D} \frac{(-1)^{|\text{edges}(D)|}}{|\text{Aut}_f(D)|} \text{rvalue}_f(D)$$


**Theorem 2** (Long-Term Behavior near a Local Minimum). *If  $\theta_\star$  locally minimizes  $l$  and for some positive form  $Q$ ,  $Q < \nabla^2 l_x(\theta_\star)$  for all  $x$ , then when we initialize SGD sufficiently close to  $\theta_\star$ , the  $d$ th-order truncation of Theorem 1 converges as  $T$  diverges.*

**Remark 2.** We may approximate sums by integrals and  $(I - \eta H)^t$  by  $\exp(-\eta H t)$ , reducing to a routine integration of exponentials at the cost of an error factor  $1 + o(\eta)$ .



<sup>1</sup> E.g., we actually define  to be the cumulant  $C = \mathbb{E}_x[(\nabla l_x(\theta) - G)^2]$ , not the moment  $\mathbb{E}_x[(\nabla l_x(\theta))^2]$ . This centering is routine (see APPENDIX), tedious to keep notating, and un-germane, so we ignore it.

## 2.2 Insights from the Formalism

### 2.2.1 SGD descends on a $C$ -smoothed landscape and prefers minima flat w.r.t. $C$ .



**Corollary 1** (Computed from ). *Initialized at a test minimum, and run for long times  $T \gg 1/\eta H$ , SGD drifts with an expected time-averaged velocity of*

$$v^\lambda = \frac{\eta^3}{T} \sum_{\mu\nu} C_{\mu\nu} \frac{1}{\eta(H_{\mu\mu} + H_{\nu\nu})} J_{\mu\nu\lambda} \frac{1}{H_{\lambda\lambda}} + o(\eta^2) \quad \text{in an eigenbasis for } \eta H$$

Intuitively,  $D = \text{$  contains a subdiagram  $\text{$   $= (K\eta)^2 CH$ . By a routine check,  $CH + o(\eta^2)$  is the loss increase upon convolving  $l$  with a  $C$ -shaped Gaussian. Since  $D$  connects the subdiagram to **to the test measurement** via 1 edge, it couples  $CH$  to  $l$ 's linear part, so it represents a displacement of  $\theta$  away from high  $CH$ . In short, *SGD descends on a covariance-smoothed landscape*. That is, SGD prefers from among a valley's minima those that are flat w.r.t.  $C$ . Figure 1 (left) illustrates this intuition.

Yaïda [2019b] reports a small- $T$  version of this result that scales with  $\eta^3$ . Meanwhile, Corollary 1's large- $T$  analysis scales with  $\eta^2$ . Our analysis integrates the noise over many updates, hence amplifying the contribution of  $C$ , and experiments verify this scaling law. We do not make Wei and Schwab [2019]'s assumptions of thermal equilibrium, fast-slow mode separation, or constant covariance. This generality reveals novel dynamics: that the velocity field above is generically non-conservative (Section 3.1.2).

### 2.2.2 Both flat and sharp minima overfit less

**Corollary 2** (from , ). *Initialize GD at a test minimum. The test-loss-increase and the gen.-gap (test minus train loss) due to training are, with errors  $o(\eta^2)$  and  $o(\eta^1)$ :*

$$\frac{C_{\mu\nu}}{2N} \left( (I - \exp(-\eta TH))^{\otimes 2} \right)_{\rho\lambda}^{\mu\nu} (H^{-1})^{\rho\lambda} \quad \text{and} \quad \frac{C_{\mu\nu}}{N} (I - \exp(-\eta TH))_{\lambda}^{\nu} (H^{-1})^{\lambda\mu}$$

This gen. gap tends with large  $T$  to  $C_{\mu\nu}(H^{-1})^{\mu\nu}/N$ . For maximum likelihood (ML) estimation in well-specified models near the “true” minimum,  $C = H$  is the Fisher metric, so we recover AIC: (number of parameters)/ $N$ . Unlike AIC, our more general expression is descendably smooth, may be used with MAP or ELBO tasks instead of just ML, and does not assume a well-specified model.

### 2.2.3 High- $C$ regions repel small- $E$ , $B$ SGD

**Corollary 3** (Epoch Number). *To order  $\eta^2$ ,  $M = 1$  SGD with learning rate  $\eta$  has  $\left(\frac{M-1}{M}\right)\left(\frac{B+1}{B}\right)\left(\frac{N}{2}\right)(\nabla_{\mu} C_{\nu}^{\nu}) G^{\mu}/2$  less test loss than  $M = M$  SGD with learning rate  $\eta/M$ .*

**Corollary 4** (Batch Size). *The expected test loss of pure SGD is, to order  $\eta^2$ , less than that of pure GD by  $\frac{M(N-1)}{2} (\nabla_{\mu} C_{\nu}^{\nu}) G^{\mu}/2$ . Moreover, if  $\hat{C}$  is a smooth unbiased estimator of  $C$ , then GD on a modified loss  $\tilde{l}_n = l_n + \frac{N-1}{4N} \hat{C}_{\nu}^{\nu}(\theta)$  has an expected test loss that agrees with SGD's to second order. We call this method GDC.*

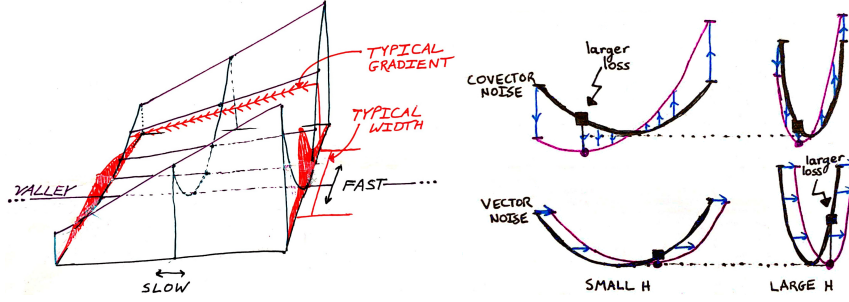




Figure 1: **Novel phenomena.** **Left:** Gradient noise induces a flow toward minima *with respect to the covariance*. Our analysis assumes neither thermal equilibrium nor fast-slow mode separation, but we label “fast and slow directions” to ease comparison with Wei and Schwab [2019]. Red densities depict a typical locations for  $\theta$  in each cross-section of the valley, and the spatial variation of curvature corresponds to  $J_{\mu\nu\lambda}$ . **Right:** Noise structure determines how curvature affects overfitting. Geometrically, for a vector-perturbed landscape, small Hessians are favored (top row), while for covector-perturbed landscapes, large Hessians are favored (bottom row). Corollary 2 shows how the implicit regularization of fixed- $\eta T$  descent interpolates between the two rows.


## 2.2.4 Non-Gaussian noise affects SGD but not SDE

Stochastic Differential Equations (SDE: see Liao et al. [2018]) are a popular theoretical approximation to SGD, but SDE and SGD differ in several ways. For instance, the inter-epoch noise correlations in multi-epoch SGD measurably affect SGD’s final test loss (Corollary 3), but SDE assumes uncorrelated gradient updates. Even if we restrict to single-epoch SDE, differences arise due to time discretization and non-gaussian noise.

**Corollary 5** (  ,  ). SGD’s test loss is  $\frac{T}{2} C_{\mu\nu} H^{\mu\nu} + o(\eta^2)$  more than ODE’s and SDE’s. The deviation from SDE due to non-Gaussian noise is  $-(T/6) S_{\mu\nu\lambda} J^{\mu\nu\lambda} + o(\eta^3)$ .<sup>1</sup>

## 3 Applying the Theory

### 3.1 Experiments

Our experiments’ rejection of the null hypothesis is sometimes drastic. E.g., in Figure 2  , [Chaudhari and Soatto, 2018] predicts a velocity of 0 while we predict a velocity of  $\eta^2/6$ . I bars and + signs to mark a 95% confidence interval based on the standard error of the mean. Appendix ?? lists architectures, procedure, and further tests.

#### 3.1.1 Training time, epochs, and batch size

We test Theorem 1’s order  $\eta^3$  truncation on smooth convnets for CIFAR-10 and Fashion-MNIST. Theory agrees with experiment through on timescales long enough for accuracy

<sup>1</sup> This is Example 1’s more exact expression for  $\eta \ll 1$ : they agree to leading order in  $\eta$ .

to increase by 0.5% (Figure 2  $\square\square\square$ ,  $\square\square\square$ ). Figure 2  $\square\square\square$  tests Corollary 4's claim that high- $C$  regions repel SGD more than GD. This is significant because  $C$  controls the rate at which the gengap (test minus train loss) grows (Corollary 2, Figure 2  $\square\square\square$ ).

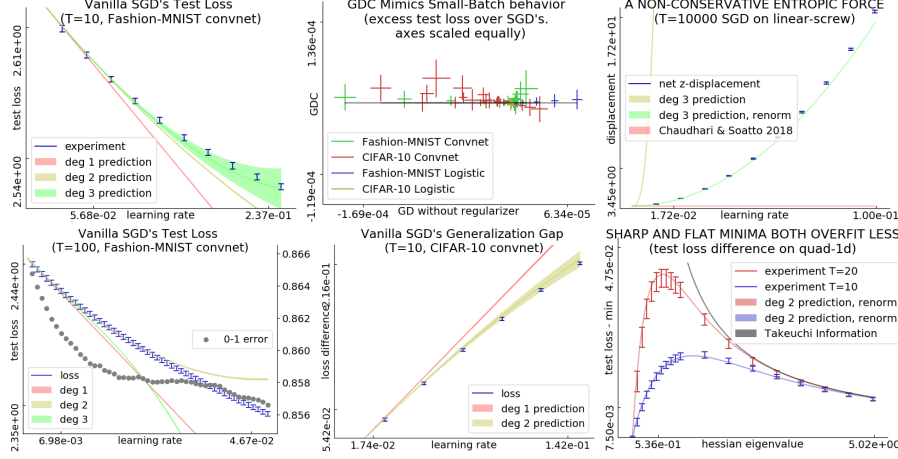



Figure 2: **Left: Perturbation models SGD for small  $\eta T$ .** Fashion-MNIST convnet's test loss vs learning rate; un-re-summed predictions.  $\square\square\square$ : For all init.s tested (1 shown, 11 unshown), our degree-3 prediction agrees with experiment through  $\eta T \approx 10^0$ , corresponding to a decrease in 0-1 error of  $\approx 10^{-3}$ .  $\square\square\square$ : For large  $\eta T$ , our predictions break down. Here, the order-3 prediction holds until the 0-1 error improves by  $5 \cdot 10^{-3}$ . **Center:  $C$  controls gen. gap and distinguishes GD from SGD.**  $\square\square\square$ : With equal-scaled axes, this plot shows that GDC matches SGD (small vertical variance) better than GD matches SGD (large horizontal variance) in test loss for a range of  $\eta$  ( $\approx 10^{-3} - 10^{-1}$ ) and init.s (zero and several Xavier-Glorot trials) for logistic regression and convnets. Here,  $T = 10$ .  $\square\square\square$ : CIFAR-10 generalization gaps. For all init.s tested (1 shown, 11 unshown), the degree-2 prediction agrees with experiment through  $\eta T \approx 5 \cdot 10^{-1}$ . **Right: Re-summed predictions excel even for large  $\eta T$ .**  $\square\square\square$ : On ARCHIMEDES, SGD travels the valley of global minima in the positive  $z$  direction. Since  $H$  and  $C$  are bounded and the effect appears for all small  $\eta$ , the effect is not a pathology of well-chosen learning rate or divergent noise. The net displacement of  $\approx 10^{1.5}$  well exceeds the  $z$ -period of  $2\pi$ .  $\square\square\square$ : For MEAN ESTIMATION with fixed  $C$  and a range of  $H$ s, initialized at the truth, the test losses after fixed- $T$  optimization are smallest for very small and very large curvatures. As predicted: both sharp and flat minima overfit less.


### 3.1.2 Minima that are flat with respect to $C$ attract SGD

To test Corollary 1, we construct a counter-intuitive loss landscape wherein SGD steadily moves in a direction of 0 test gradient. Our mechanism differs from that of Chaudhari and Soatto [2018]'s approximate analysis, which in this case predicts a velocity of






0.<sup>1</sup> Specifically, the ARCHIMEDES landscape has weights  $\theta = (u, v, z) \in \mathbb{R}^3$ , data points  $x \sim \mathcal{N}(0, 1)$ , and loss:  $l_x(w) \triangleq \frac{1}{2}H(\theta) + x \cdot S(\theta)$ , where  $H(\theta) = u^2 + v^2 + (\cos(z)u + \sin(z)v)^2$  and  $S(\theta) = \cos(z - \pi/4)u + \sin(z - \pi/4)v$ . Note that for fixed  $z$ ,  $H$  is quadratic and  $S$  is linear. Also, since  $x \sim \mathcal{N}(0, 1)$ ,  $xS(\theta)$  has expectation 0. ARCHIMEDES thus has valley of global minima on the line  $x = y = 0$ . For SGD initialized at  $\theta = 0$ , Corollary 1 predicts a  $z$ -velocity of  $+\eta^2/6$  per timestep. The prediction agrees with experiment even as the net displacement exceeds the the landscape’s natural length scale of  $2\pi$  (Figure 2 ).

### 3.1.3 Sharp and flat minima both overfit less than medium minima

Prior work finds both that *sharp* minima overfit less (for,  $l^2$  regularization sharpens minima) or that *flat* minima overfit less (for, flat minima are robust to small displacements). In fact, generalization’s relationship to curvature depends on the landscape’s noise structure (Corollary 2, Figure 2 ). To combat overfitting, we may add Corollary 2’s expression for gen. gap to  $l$ . Unlike AIC, which it subsumes, this regularizer is continuous and thus liable to descent. We call this regularizer *STIC* (APPENDIX). By descending on STIC, we may tune smooth hyperparameters such as  $l_2$  regularization coefficients in the noisy, small- $N$  regime  $H \ll C/N$ . Since matrix exponentiation takes time cubic in dimension, exact STIC is most useful for small models.

## 3.2 Conclusion

We presented a diagram-based method for studying stochastic optimization on short timescales or near minima. Corollaries 1 and 2 together offer insight into SGD’s success in training deep networks: SGD senses curvature, and curvature controls generalization. Analyzing , we proved that **flat and sharp minima both overfit less than medium minima**. Intuitively, flat minima are robust to vector noise, sharp minima are robust to covector noise, and medium minima robust to neither. We thus propose a smooth analogue of AIC enabling gradient-based hyperparameter tuning. Inspecting , we extended Wei and Schwab [2019] to nonconstant, nonisotropic covariance to reveal that **SGD descends on a landscape smoothed by the current covariance  $C$** . As  $C$  evolves, the smoothed landscape evolves, resulting in non-conservative dynamics. Examining , we showed that **GD may emulate SGD**, as conjectured 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].

Since our predictions depend only on loss data near initialization, they break down after the weight moves far from initialization. Our theory thus best applies to small-movement contexts, whether for long times (large  $\eta T$ ) near an isolated minimum or for short times (small  $\eta T$ ) in general. E.g., our theory might especially illuminate meta-learners that are based on fine-tuning (e.g. Finn et al. [2017]’s MAML).

Much as meteorologists understand how warm and cold fronts interact despite long-term forecasting’s intractability, we quantify the counter-intuitive dynamics governing each short-term interval of SGD’s trajectory. Equipped with our theory, practitioners may now refine intuitions (e.g. that SGD descends on the train loss) to account for noise.

<sup>1</sup> Indeed, our velocity is  $\eta$ -perpendicular to the image of  $(\eta C)_v^\mu$ .



## Broader Impacts

Though machine learning has the long-term potential for vast improvements in world-wide quality of life, it is today a source of enormous carbon emissions [Strubell et al., 2019]. Our analysis of SGD may lead to a reduced carbon footprint in three ways.

First, Section 2.2.3 shows how to modify the loss landscape so that large-batch GD enjoys the stochastic regularizing properties of small-batch SGD, or (symmetrically) so that small-batch SGD enjoys the stability of large-batch GD. By unchaining the effective batch size from the actual batch size, we raise the possibility of training neural networks on a wider range of hardware than currently practical. For example, asynchronous concurrent small-batch SGD (e.g., Niu et al. [2011]) might require less inter-GPU communication and therefore less power.

Second, Section 3.2 discusses an application to meta-learning, which has the potential to decrease the per-task sample complexity and hence carbon footprint of modern ML.

Third, the generalization of AIC developed in Sections 2.2.2 and 3.1.3 permits certain forms of model selection by gradient descent rather than brute force search. This might drastically reduce the energy consumed during model selection.

That said, insofar as our theory furthers practice, it may instead contribute to the rapidly growing popularity of GPU-intensive learning, thus negating the aforementioned benefits and accelerating climate change.

More broadly, this paper analyzes optimization in the face of uncertainty. As ML systems deployed today must increasingly address *user privacy*, *pedestrian safety*, and *dataset diversity*, it becomes important to recognize that training sets and test sets differ. Toward this end, theoretical work relating to non-Gaussian noise may assist practitioners in building provably non-discriminatory, safe, or private models (e.g., Dwork et al. [2006]). By quantifying how correlated, non-Gaussian gradient noise affects descent-based learning, this paper contributes to such broader theory.

## Acknowledgements

We feel deep gratitude to SHO YAJIDA, DAN A. ROBERTS, and JOSH TENENBAUM for posing some of the problems this work resolves and for their patient guidance. We appreciate the generosity of ANDY BANBURSKI, BEN R. BRAY, JEFF LAGARIAS, and WENLI ZHAO in critiquing our drafts. Without the encouragement of JASON CORSO, CHLOE KLEINFELDT, ALEX LEW, ARI MORCOS, and DAVID SCHWAB, this paper would not be. Finally, we thank our anonymous reviewers for inspiring an improved presentation. This work was funded in part by MIT’s Jacobs Presidential Fellowship and in part by Facebook AI Research.




## References

- P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds, chapter 4. *Princeton University Press*, 2007.
- S.-I. Amari. Natural gradient works efficiently. *Neural Computation*, 1998.

- P.L. Bartlett, D.J. Foster, and M.J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 2017.
- S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 2013.
- L. Bottou. Stochastic gradient learning in neural networks. *Neuro-Nîmes*, 1991.
- P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *ICLR*, 2018.
- Laurent Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. *ICLR*, 2017.
- C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 2006.
- E. Dyer and G. Gur-Ari. Asymptotics of wide networks from feynman diagrams. *ICML Workshop*, 2019.
- R.P. Feynman. A space-time approach to quantum electrodynamics. *Physical Review*, 1949.
- C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017.
- P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd. *Data @ Scale*, 2018.
- E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better. *NeurIPS*, 2017.
- N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P.T.P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*, 2017.
- F. Kunstner, P. Hennig, and L. Balles. Limitations of the empirical fisher approximation for natural gradient descent. *NeurIPS*, 2019.
- L.D. Landau and E.M. Lifshitz. The classical theory of fields. *Addison-Wesley*, 1951.
- L.D. Landau and E.M. Lifshitz. Mechanics. *Pergamon Press*, 1960.
- Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms i. *PMLR*, 2017.
- Qianli Liao, B. Miranda, A. Banburski, J. Hidary, and T. Poggio. A surprising linear relationship predicts test performance in deep networks. *Center for Brains, Minds, and Machines Memo 91*, 2018.
- B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *NeurIPS*, 2017a.

- B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *Chapter 4 from Intel CRI-CI: Why and When Deep Learning Works Compendium*, 2017b.
- M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. *ICML*, 2017.
- Feng Niu, B. Recht, C. Ré, and S.J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *NeurIPS*, 2011.
- R. Penrose. Applications of negative dimensional tensors. *Combinatorial Mathematics and its Applications*, 1971.
- D.A. Roberts. Sgd implicitly regularizes generalization error. *NeurIPS: Integration of Deep Learning Theories Workshop*, 2018.
- G.-C. Rota. Theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1964.
- N.L. Roux, Y. Bengio, and A. Fitzgibbon. Improving first and second-order methods by modeling uncertainty. *Book Chapter: Optimization for Machine Learning, Chapter 15*, 2012.
- C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Berkeley Symposium on Mathematical Probability*, 1956.
- E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *ACL*, 2019.
- Huan Wang, N.S. Keskar, Caiming Xiong, and R. Socher. Identifying generalization properties in neural networks. *Arxiv Preprint*, 2018.
- Mingwei Wei and D.J. Schwab. How noise affects the hessian spectrum in overparameterized neural networks. *Arxiv Preprint*, 2019.
- Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning. *NeurIPS*, 2018.
- Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. *ICLR*, 2019a.
- Sho Yaida. A first law of thermodynamics for stochastic gradient descent. *Personal Communication*, 2019b.
- Chiyuan Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.
- Hongyi Zhang, S.J. Reddi, and S. Sra. Fast stochastic optimization on riemannian manifolds. *NeurIPS*, 2016.

### 3.3 Diagrams and embeddings

**Definition 1** (Diagrams). A *diagram* is a finite rooted tree equipped with a partition of its nodes. We draw the tree using thin edges. By convention, we draw each node to the right of its children; the root is thus always rightmost. We draw the partition by connecting the nodes within each part via fuzzy ties. For example,  has 2 parts. We insist on using as few fuzzy ties as possible so that, if  $d$  counts edges and  $c$  counts ties, then  $d + 1 - c$  counts the parts of the partition. There may be multiple ways to draw a single diagram, e.g.  = .

**Definition 2** (Spacetime). The *spacetime* associated with an SGD run is the set of pairs  $(n, t)$  where the  $n$ th datapoint participates in the  $t$ th gradient update. Spacetimes thus encode batch size, training set size, and epoch number.

**Definition 3** (Embedding Diagrams into Spacetime). An *embedding* of a diagram  $D$  into a spacetime is an assignment of  $D$ 's non-root nodes to pairs  $(n, t)$  such that each node occurs at a time  $t'$  strictly after each of its children and such that two nodes occupy the same row  $n$  if and only if they inhabit the same part of  $D$ 's partition.

To visualize embeddings, we draw the  $(n, t)$  pairs of a space-time as shaded cells in an  $N \times T$  grid. A diagram embedding is then an assignment of nodes to shaded cells. The  $t < t'$  constraint forbids intra-cell edges (Figure 3 left), and we may interpret each edge as an effect of the past on the future (right).

**Definition 4** (A Diagram's Un-resummed Value). The *un-resummed value* of a diagram  $D$  is the product of the values of each part  $p$  in its partition. The value of a part  $p$  with  $|p|$  many nodes is the expectation  $\mathbb{E}_x \left[ (\nabla l_x(\theta))^{|p|} \right]$ . The edges of  $D$ 's tree indicate how to multiply the values of these parts: each edge indicates a contraction. For instance, since the training points are independent:

$$\text{Diagram} \triangleq \mathbb{E}_{n, n', n''} \left[ (\nabla_\mu l_n) (\nabla_\nu l_{n'}) (\nabla^\mu \nabla^\nu \nabla_\lambda l_{n''}) (\nabla^\lambda l_{n''}) \right]$$

Implicit in the three raised indices are three factors of  $\eta$ . We denote  $D$ 's un-resummed value by  $\text{uvalue}(D)$ , or by  $D$  when clear.

**Definition 5** (An Embedding's Re-summed Value). The *re-summed value*  $\text{rvalue}_f(D)$  of an embedding  $f$  of a diagram  $D$  is the same as the un-resummed value of  $D$ , save for one change having to do with edges. Consider an edge between two nodes embedded to  $(n, t)$  and  $(n', t + \Delta t)$ . Whereas  $\text{uvalue}(D)$  has a factor of  $\eta^{\mu\nu}$  for this edge,  $\text{rvalue}_f(D)$  instead has a factor of  $((I - \eta H)^{\Delta t - 1})_\lambda^\mu \eta^{\lambda\nu}$ . Here,  $1 \leq \Delta t$  is the temporal distance between the two nodes' embeddings.





We will often seek *differences*, e.g. between ODE's and SGD's test loss or between a test loss and a train loss. We thus define a compact notation for differences of diagrams:

**Definition 6** (Fuzzy Outlines Denote Noise's Net Effect). A diagram drawn with one *fuzzy outline* denotes the difference between the versions with and without fuzzy ties. E.g.:

$$\text{Diagram with fuzzy tie} \triangleq \text{Diagram without fuzzy tie} - \text{Diagram with fuzzy tie}$$

We define a diagram drawn with more than one fuzzy outline as the fully tied version minus all the versions with fewer fuzzy outlines (these are the Möbius sums of Rota [1964]):

$$\begin{aligned} & \text{Diagram 1} \triangleq \text{Diagram 2} - \text{Diagram 3} - \text{Diagram 4} - \text{Diagram 5} - \text{Diagram 6} - \text{Diagram 7} \\ & \triangleq \text{Diagram 8} - \text{Diagram 9} - \text{Diagram 10} - \text{Diagram 11} + 2 \text{Diagram 12} \end{aligned}$$

**Definition 7** (Irreducible Diagrams). A diagram, drawn with fuzzy outlines instead of ties, is *irreducible* when none of its degree-2 non-root nodes participates in fuzzy outlines. So , , are irreducible, but not , .

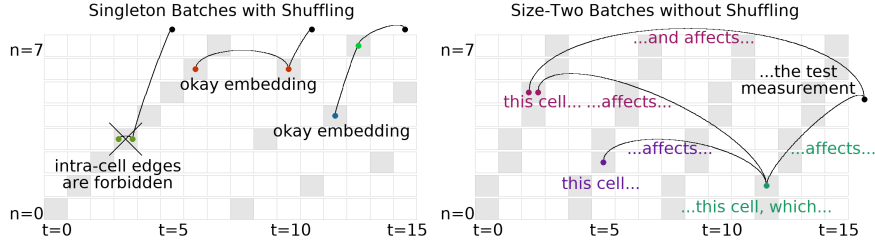





Figure 3: **Diagrams in Spacetime Depict SGD's Subprocesses.** Two spacetimes with  $N = 8, T = 16$ . **Left:** Batchsize  $B = 1$  with inter-epoch shuffling. Embeddings, legal and illegal, of , , and . **Right:** Batchsize  $B = 2$  without inter-epoch shuffling. Interpretation of an order  $\eta^4$  diagram embedding.








$\Theta((\eta T)^3 T^{-0})$	$\Theta((\eta T)^3 T^{-1})$	$\Theta((\eta T)^3 T^{-2})$
		
		
		

Table 1: **Degree-3 diagrams for  $B = M = 1$  SGD's test loss.** The 6 diagrams have  $(4 + 2) + (2 + 2 + 3) + (1)$  total orderings relevant to Proposition 1. **Left:**  $(d, c) = (3, 0)$ . Diagrams for ODE behavior. **Center:**  $(d, c) = (3, 1)$ . 1st order deviation of SGD away from ODE. **Right:**  $(d, c) = (3, 2)$ . 2nd order deviation of SGD from ODE with appearance of non-Gaussian statistics.

### 3.4 Comparison to continuous time

Consider fitting a centered normal  $\mathcal{N}(0, \sigma^2)$  to data  $x$  drawn i.i.d. from a centered standard normal. We parameterize the landscape by  $h = \log(\sigma^2)$  so that the Fisher

information matches the standard dot product [Amari, 1998]. The gradient at sample  $x$  and weight  $h$  is then  $g_x(h) = (1 - x^2 \exp(-h))/2$ . Since  $x \sim \mathcal{N}(0, 1)$ ,  $g_x(h)$  will be affinely related to a chi-squared and in particular non-Gaussian. **FIGURE** shows that even for this simple learning problem, SGD and SDE differ as predicted.

It is routine to check that, by stitching together copies of this example, we may cause SGD to travel along paths that are closed loops or unbounded curves. We may even add a small linear component so that SGD steadily climbs uphill.

So, even in a valley of global minima, SGD will move away from minima whose Hessian aligns with the current covariance. However, by the time it moves, the new covariance might differ from the old one, and SGD will be repelled by different Hessians than before. Setting the covariance to lag the Hessian by a phase, we construct a landscape in which this entropic force dominates. This “*linear screw*” landscape has

### 3.5 Questions

The diagram method opens the door to exploration of Lagrangian formalisms and curved backgrounds<sup>1</sup>:

**Question 1.** *Does some least-action principle govern SGD; if not, what is an essential obstacle to this characterization?*

Lagrange’s least-action formalism intimately intertwines with the diagrams of physics. Together, they afford a modular framework for introducing new interactions as new terms or diagram nodes. In fact, we find that some *higher-order* methods — such as the Hessian-based update  $\theta \leftarrow \theta - (\eta^{-1} + \lambda \nabla \nabla l_t(\theta))^{-1} \nabla l_t(\theta)$  parameterized by small  $\eta, \lambda$  — admit diagrammatic analysis when we represent the  $\lambda$  term as a second type of diagram node. Though diagrams suffice for computation, it is Lagrangians that most deeply illuminate scaling and conservation laws.

**Conjecture 1** (Riemann Curvature Regularizes). *For small  $\eta$ , SGD’s gen. gap decreases as sectional curvature grows.*

Though our work so far assumes a flat metric  $\eta^{\mu\nu}$ , it generalizes to curved weight spaces<sup>2</sup>. Curvature finds concrete application in the *learning on manifolds* paradigm of Absil et al. [2007], Zhang et al. [2016], notably specialized to Amari [1998]’s *natural gradient descent* and Nickel and Kiela [2017]’s *hyperbolic embeddings*. We are optimistic our formalism may resolve conjectures such as above.

For SGD with 1 epoch and batch size 1, Theorem 1 then specializes to:

**Proposition 1.** *Single-epoch singleton-batch SGD has expected test loss*

$$\sum_{0 \leq d < \infty} \frac{(-1)^d}{d!} \sum_D |\text{ords}(D)| \binom{N}{P-1} \frac{d!}{\prod d_p!} \text{uvalue}(D)$$

where  $D$  has  $P$  parts with sizes  $d_p$ . Here,  $D$  ranges over  $d$ -edged diagrams none of whose parts contains any of its nodes’ ancestors, and  $|\text{ords}(D)|$  counts the total orderings of  $D$ ’s nodes s.t. children precede parents and parts are contiguous.

<sup>1</sup> Landau and Lifshitz [1960, 1951] introduce these concepts.

<sup>2</sup> One may represent the affine connection as a node, thus giving rise to non-tensorial and hence gauge-dependent diagrams.

A diagram with  $d$  thin edges and  $c$  fuzzy ties (hence  $d + 1 - c$  parts) thus contributes  $\Theta((\eta T)^d T^{-c})$  to SGD's test loss. Intuitively,  $\eta T$  measures the physical time of descent and  $T^{-1}$  measures the coarseness of time discretization. We thus regard Proposition 1 as a double series in  $(\eta T)^d T^{-c}$ , where each term isolates the  $d$ th order effect of time and the  $c$ th order effect of noise. Indeed,  $c$  counts fuzzy ties and hence the  $c = 0$  terms do not model correlations and hence do not model noise. That is, the  $c = 0$  terms give an ODE approximation to SGD. The remaining terms give the corrections due to noise. See Table 1.

**APPENDIX**'s estimators for  $C$  et al. take time comparable to backpropagation.

Our experiments on image classifiers show that even a single evaluation of our force laws may predict SGD's motion through macroscopic timescales, e.g. long enough to decrease error by 0.5 percentage points.

**Theorem 3.** *For any  $T$ : for  $\eta$  small enough, SGD has expected test loss*

$$\sum_{\substack{D \\ \text{irreducible}}} \sum_{\substack{\text{embeddings} \\ f}} \frac{1}{|\text{Aut}_f(D)|} \frac{\text{rvalue}_f(D)}{(-B)^{|\text{edges}(D)|}}$$

Here,  $D$  ranges through irreducible outlined diagrams,  $f$  ranges through embeddings of  $D$  into the SGD's spacetime, and  $|\text{Aut}_f(D)|$  counts the automorphisms of  $D$  that preserve  $f$ 's assignment of nodes to  $(n, t)$  pairs.

**Remark 3.** Sometimes, we prefer  $\text{uvalue}(D)$  to  $\text{rvalue}_f(D)$  for its simplicity. Theorem 1 persists if we replace each  $\text{rvalue}_f(D)$  by  $\text{uvalue}(D)$  and sum all tied diagrams instead of irreducible outlined diagrams. But the large- $T$  convergence no longer holds **PLOT**.

**Remark 4.** The above gives SGD's expected loss on the test set. How about the train set? Or weight displacements? Or variances? Theorem 1 and Remark ?? have simple analogues for each of these  $2^3$  possibilities, which we discuss in the appendix.

We may approximate  $\text{rvalue}_f(D)$  by simpler *un-resummed values* at the cost of losing Theorem 2's large- $T$  convergence. We state the theorems for expected loss on the test set, but they generalize to variances, weight displacements, and training instead of testing curves **APPENDIX**.

For finite  $N$ , Corollary 5 separates SDE from SGD. Conversely, as  $N \rightarrow \infty$  with  $\eta N$  fixed and  $C$  scaling with  $\sqrt{N}$ , SGD converges to SDE, but generalization and optimization respectively become trivial and computationally intractable.

Overall, these tests verify that our proofs hide no mistakes of proportionality or sign.