

A Space-Time Approach to Analyzing SGD

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 sharp 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. 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.

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 inform practitioners confronting 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 η starting at weight θ_0 , then by Taylor expansion we may express the expected test loss of the final weight θ_T in terms of statistics of the loss landscape evaluated at θ_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 η, 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 θ 's loss on datapoint x . We define a tensor \leftrightarrow diagram dictionary:

$$\begin{aligned} G &\triangleq \mathbb{E}_x [\nabla l_x(\theta)] \triangleq \text{↘} \\ H &\triangleq \mathbb{E}_x [\nabla \nabla l_x(\theta)] \triangleq \text{↘↘} \quad C \triangleq \mathbb{E}_x [(\nabla l_x(\theta) - G)^2] \triangleq \text{↘↘} \\ J &\triangleq \mathbb{E}_x [\nabla \nabla \nabla l_x(\theta)] \triangleq \text{↘↘↘} \quad S \triangleq \mathbb{E}_x [(\nabla l_x(\theta) - G)^3] \triangleq \text{↘↘↘} \end{aligned}$$

Here, G, H, J denote the loss's derivatives w.r.t. θ , 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 higher-degree analogues) to obtain *diagrams*.¹ E.g., we may join $C = \text{↘↘}$ with $H = \text{↘↘}$ to get . As another example, we may join two copies of $G = \text{↘}$ with two copies of $H = \text{↘↘}$ to get .

Example 1. Does non-Gaussian noise affect SGD?² 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 $S = \text{↘↘↘}$. In this case, there is one fewest-edged diagram — ; it results from joining S with $J = \text{↘↘↘}$. To evaluate a diagram, we multiply its components (here, S, J) with exponentiated ηH 's, one for each edge:

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

This is S 's leading order contribution to SGD's test loss written in an eigenbasis of ηH .

Remark 1. For large T and isotropic ηH , (1) 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. θ . So non-gaussian noise affects SGD proportion to the logarithmic derivative of curvature.

In general, each diagram intuitively represents the net effect of a certain combination of gradients (G), noise (C, S, \dots) and curvature (H, J, \dots).

¹ 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.

² Here but not in our Theorems, we assume that θ_0 minimizes l and that we run SGD for 1 epoch with batch size 1. This simplifies our analysis by limiting the number of relevant diagrams (see Proposition 1).

1.2 Background, Notation, and Assumptions

We sometimes implicitly sum repeated Greek indices: if a covector A and a vector B ¹ have coefficients A_μ, B^μ , then $A_\mu B^\mu \triangleq \sum_\mu A_\mu \cdot B^\mu$. To expedite dimensional analysis, 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 η 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 η is a tensor, we may still define $o(\eta^d)$: a quantity q *vanishes to order* η^d when $\lim_{\eta \rightarrow 0} q/p(\eta) = 0$ for some homogeneous degree- d polynomial p . We then write $q \in o(\eta^d)$.

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 regularity conditions, e.g. that l, l_x are analytic and that all 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.

SGD performs η -steepest descent on the estimates l_n . We describe SGD in terms of N, T, B, E, M : N counts training points, T counts updates, B counts points per batch, $E = TN/B$ counts epochs, and $M = E/B = T/N$ counts updates per point. So SGD performs $T = NM$ updates $\theta^\mu := \theta^\mu - \eta^{\mu\nu} \nabla_\nu \sum_{n \in \mathcal{B}_t} l_n(\theta)/B$ where \mathcal{B}_t is the t th batch.

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 η^1 and $T = 2$.

Our theory is vacuous for large η . Other analyses treat large- η 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 η^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 η . 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 ,² so we may model VAEs.

¹ Vectors/covectors are also called column/row vectors.

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

2 Diagram Calculus

Theorem 1 expresses SGD’s test loss as a sum over diagrams. Recalling that a diagram with d edges is $O(\eta^d)$, we read the following as a Taylor series in η . In practice, we truncate the series to small d , thus focusing on the few-edged diagrams.

Theorem 1. *For any T : for η sufficiently small, SGD’s expected test loss is*

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

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 graph automorphisms of D that preserve f ’s assignment of nodes to (n, t) pairs. As a reminder, B is batch size.

Though the combinatorics of embeddings and graph automorphisms may seem forbidding, our focus on few-edged diagrams will make this counting nearly trivial.

Theorem 2 (Long-Term Behavior near a Local Minimum). *If θ_\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 θ_\star , the d th-order truncation of Theorem 1 converges as T diverges.*

Remark 2. We may approximate sums over embeddings by integrals over times and $(I - \eta H)^t$ by $\exp(-\eta H t)$, incurring a multiplicative error of $1 + o(\eta)$. We thus reduce to routine integration of exponentials. Sometimes, we may 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 guarantee no longer applies **PLOT**.



Remark 3. 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 2 have simple analogues for each of these 2^3 possibilities, which we discuss in the appendix.

2.1 Insights from the Formalism

2.1.1 SGD descends on a C -smoothed landscape and favors minima flat w.r.t. C .

Corollary 1. *Initialized at a test minimum, and run for long times $T \gg 1/\eta H$ single-epoch singleton-batch SGD’s weight moves with an expected average velocity of*

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

This corollary results from evaluating  (call this D). Intuitively, D contains a subdiagram  $= CH$; by a routine check, CH is the leading-order loss increase upon convolve l with a C -shaped Gaussian. Since D connects the subdiagram CH to **to the test measurement** via 1 edge, it couples CH to l ’s linear part, so it represents a displacement

of θ away from high CH . In short, *SGD descends on a covariance-smoothed landscape*. That is, SGD prefers minima that are flat w.r.t. C (Figure 1, left).

Yaïda [2019b] reports a small- T version of this result that scales with η^3 . Meanwhile, Corollary 1’s large- T analysis scales with η^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).

2.1.2 Both flat and sharp minima overfit less

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

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

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.

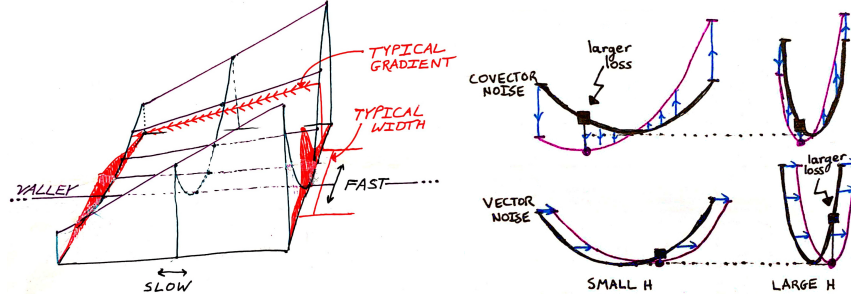


Figure 1: **Re-summation reveals novel phenomena.** **Left:** The entropic force mechanism: gradient noise induces a flow toward minima *with respect to the covariance*. Though our analysis assumes neither thermal equilibrium nor fast-slow mode separation, we label “fast and slow directions” to ease comparison with Wei and Schwab [2019]. Here, red densities denote the spread predicted by a re-summed $C^{\mu\nu}$, and the spatial variation of curvature corresponds to $J_{\mu\nu\lambda}$. **Right:** Noise structure determines how curvature affects overfitting. Geometrically, for (empirical risk minimization on) 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- ηT descent interpolates between the two rows.

2.1.3 Epochs and batch size

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

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

2.1.4 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 Differs from ODE, SDE). *The test loss of single-epoch, singleton-batch SGD deviates from that of ODE and SDE by $\frac{T}{2}C_{\mu\nu}H^{\mu\nu} + o(\eta^2)$. The deviation from SDE due to non-Gaussian noise is $-(T/6)\text{[diagram]} + o(\eta^3) = -(T/6)S_{\mu\nu\lambda}J^{\mu\nu\lambda} + o(\eta^3)$.¹*

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

3 Applying the Theory

3.1 Experiments

We run experiments whose rejection of the null hypothesis is so drastic as to be visually clear. E.g., in Figure ??, [Chaudhari and Soatto, 2018] predicts a velocity of 0 while we predict a velocity of $\eta^2/6$. We use I bars and + signs to mark a 95% confidence interval based on the standard error of the mean. Appendix ?? lists architectures and procedure.

3.1.1 High- C regions repel few-epoch and small-batch SGD

We test Theorem 1 on smooth convnets for CIFAR-10 and Fashion-MNIST. Our order η^3 predictions, simplified via Remark 2, agree with experiment up to $\eta T \approx 10^0$ (Figure 2, left). Also, Corollary 3 correctly predicts the effect of multi-epoch training (Appendix ??) for $\eta T \approx 10^{-1/2}$. **FIGURE** tests Corollary 4, supporting our 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**). Overall, these tests verify that our proofs hide no mistakes of proportionality or sign.

¹ This expression differs from the more exact expression of Example 1 because here we use Remark 2's substitution. One may check that the two expressions agree to leading order.

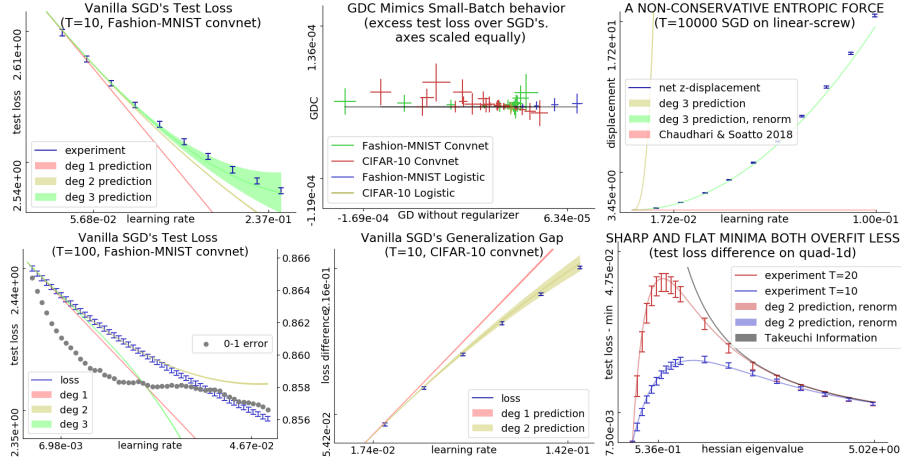


Figure 2: **Left: Perturbation models SGD for small ηT .** Fashion-MNIST convnet's test loss vs learning rate; un-re-summed predictions. \blacksquare : For all init.s tested (1 shown, 11 unshown), our degree-3 prediction agrees with experiment through $\eta T \approx 10^0$, corresponding to a decrease in 0-1 error of $\approx 10^{-3}$. \blacksquare : For large ηT , 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.** \blacksquare : 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 η ($\approx 10^{-3} - 10^{-1}$) and init.s (zero and several Xavier-Glorot trials) for logistic regression and convnets. Here, $T = 10$. \blacksquare : 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 ηT .** \blacksquare : 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 η , 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 2π . \blacksquare : 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 0.¹ 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

¹ Indeed, our velocity is η -perpendicular to the image of ηC .




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 2π (Figure ??, left).

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 1, right).

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. Experiments on MEAN ESTIMATION recommend STIC for model selection when $H \ll C/N$ as in the noisy, small- N regime Appendix. Since matrix exponentiation takes time cubic in dimension, exact STIC is most useful for small models on noisy, limited data.

3.2 Conclusion

We presented a diagram-based method for studying stochastic optimization on short timescales or near minima. Our theory yields several corollaries. Analyzing , we show 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 extend 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 show 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].

Corollaries 1 and 2 together offer insight into SGD’s success in training deep networks: SGD senses curvature, and curvature controls generalization.

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 ηT) near an isolated minimum or for short times (small ηT) in general. E.g., our theory might especially illuminate meta-learners such as MAML [Finn et al., 2017], which seek models initialized near minima and tunable to new data via few updates.

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.