SGD at Small Learning Rates

Samuel C. Tenka   C O L I@M I T . E D U
MIT, CSAIL

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We quantify how gradient noise shapes the dynamics of stochastic gradient descent (SGD) by taking Taylor series in the learning rate. We present in particular a new diagram-based notation that permits resummation to convergent results. We employ our theory to contrast SGD against two popular approximations: deterministic descent and stochastic differential equations. We find that SGD's trajectory avoids both gradient noise and minima that are sharp with respect to gradient noise. Paired with our results that connect overfitting to the width of minima, these repulsions suggest a mechanism for the unexpected generalization of overparameterized learning scenarios.

Keywords: SGD, learning rates, generalization, gradient noise.

1. Introduction

1.1. Intuitions about SGD

Practitioners benefit from the intuition that stochastic gradient descent (SGD) approximates deterministic gradient descent (GD) (Bottou, 1991). This paper refines 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). Our corrections lead to qualitative differences in dynamics: for example, we construct a non-pathological [1] loss landscape with a discrete translational symmetry *violated* by SGD's trajectory. Adding a linear term to this landscape leads SGD to perform perpetual *ascent*. These examples, though artificial, demonstrate that standard intuitions about SGD are wrong. We argue that our theory's quantitative results usefully enhance the intuition of practioners, and we we verify our theory on convolutional CIFAR-10 and Fashion-MNIST loss landscapes.

Our analysis offers a novel interpretation of SGD as a superposition sum of several concurrent processes, each a basic way that data may influence weights. Diagrams such as , analogous to those of Feynman (1949) and Penrose (1971), depict these interactions. §**??** discusses this bridge to physics — and its relation to Hessian methods and natural GD — as topics for future research.

---

[1] All higher derivatives exist and quadratically bounded; the gradient noise at each weight vector is 1-subgaussian.

1.2. Background, notation, assumptions

We fix a *testing loss* function $l : \mathcal{M} \to \mathbb{R}$ on a space $\mathcal{M}$ of weights θ. We fix a distribution $\mathcal{D}$ from which unbiased estimates l_x of l are drawn. We write $(l_n : 0 \le n < N)$ for a training sequence drawn i.i.d. from $\mathcal{D}$. We refer to n and to l_n as *training points*. We assume §**??**'s hypotheses, e.g. that l, l_x are analytic and that all moments exist. For instance, our theory models tanh networks with cross entropy loss on bounded data — and with weight sharing, skip connections, soft attention, dropout, and weight decay. But it does not model ReLU networks.

Our analysis makes heavy use of the tensors defined to the right: $G, H, J; C, S$ have $1, 2, 3; 2, 3$ indices, respectively. We shall implicitly sum repeated Greek indices: if a covector U and a vector V [1] have coefficients U_μ, V^μ, then $U_\mu V^\mu \triangleq \sum_\mu U_\mu \cdot V^\mu$. 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 η to raise indices; thus, $H^\mu_\lambda \triangleq \sum_\nu \eta^{\mu\nu} H_{\nu\lambda}$ and $C^\mu_\mu \triangleq \sum_{\mu\nu} \eta^{\mu\nu} \cdot C_{\nu\mu}$. A quantity q *vanishes to order* η^d when $\lim_{\eta \to 0} q/p(\eta) = 0$ for some homogeneous degree-d polynomial p; we then say $q \in o(\eta^d)$.

$$
\begin{aligned}
G &= \mathbb{E}_x[\nabla l_x(\theta)] &&= \\
H &= \mathbb{E}_x[\nabla\nabla l_x(\theta)] &&= \\
J &= \mathbb{E}_x[\nabla\nabla\nabla l_x(\theta)] &&= \\
C &= \mathbb{E}_x\left[(\nabla l_x(\theta) - G)^2\right] &&= \\
S &= \mathbb{E}_x\left[(\nabla l_x(\theta) - G)^3\right] &&=
\end{aligned}
$$

Above: Named tensors. §2.3 explains how tensors correspond to diagrams.

SGD performs η-steepest descent on the estimates l_n. 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. $E = TB/N$ epochs, i.e. $M = T/N$ updates per point) of the form

$$
\theta^\mu := \theta^\mu - \eta^{\mu\nu}\nabla_\nu \sum_{n \in \mathcal{B}_t} l_n(\theta)/B
$$

where in each epoch, we sample the tth batch $\mathcal{B}_t$ without replacement from the training sequence. Thus, to each initialization $\theta_0 \in \mathcal{M}$ there corresponds a joint distribution over learning trajectories $(\theta_t : 0 \le t \le T)$. The expectation of $l(\theta_T)$ especially interests us.

To illustrate our notation, we quote a well-known proposition (Nesterov (2004)):

Proposition *G controls the leading order loss decrease:* $\mathbb{E}l(\theta_T) - l(\theta_0) = TG_\mu\eta^{\mu\nu}G_\nu|_{\theta=\theta_0} + o(\eta^1)$.

One proves this estimate by induction on T. When the loss landscape is noiseless and linear (that is, $\nabla l_x(\theta)$ depends on neither x nor *theta*), this estimate is exact.

---

[1] Vectors/covectors, a.k.a. column/row vectors, represent distinct geometric concepts (Kolář et al., 1993).

1.3. Related work

It was Kiefer and Wolfowitz (1952) who, in uniting gradient descent (Cauchy, 1847) with stochastic approximation (Robbins and Monro, 1951), invented SGD. Since the development of backpropagation for efficient differentiation (Werbos, 1974), SGD has been used to train connectionist models, e.g. neural networks (Bottou, 1991), recently to remarkable success (LeCun et al., 2015).

Several lines of work treat the overfitting of SGD-trained networks (Neyshabur et al., 2017a). For example, 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 $T = 2$,. To support this generalization, we develop new summation techniques with improved large-T convergence. Our interpretation of the resulting terms offers a new qualitative picture of SGD as a superposition of several simpler information-flow processes.

Our predictions are 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 reveals that SGD's implicit regularization mediates between these seemingly clashing intuitions.

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) are forced to assume uncorrelated Gaussian noise. By contrast, we use Penrose diagrams to compute test losses to arbitrary order in η. We allow correlated, non-Gaussian noise and thus *any* smooth architecture. For instance, we do not assume information-geometric relationships between C and H, [1] so we may model VAEs.

2. Perturbative theory of SGD

2.1. Trivial example

2.2. Perturbation as technique

2.3. The necessity and role of diagrams

2.4. Insights from diagrams

2.5. Resummation

A *diagram* is a finite rooted tree equipped with a partition of its nodes that obeys 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 each part's nodes inside fuzzy outlines. A diagram is *irreducible* when each of its degree-2 nodes is in a part of size

Examples: The diagrams ⌐, ⌐̂ each have 2 parts; ⌐̂, ⌐̄ have 3. Corollaries 2, 4, 3 have $E \neq 1 \neq B$, so they feature ⌒ and ⌐̄, generalized diagrams that violate the path condition. Diagrams ⌐, ⌐̂ are irreducible; due to their green nodes, ⌒̄, ⌐̂ are not. For all f, $|\mathrm{Aut}_f(\overset{\frown}{\frown})| = 1$ and

---

[1] Disagreement of C and H is typical in modern learning (Roux et al., 2012; Künstner et al., 2019)

one. An *embedding* f of a diagram D is an injection from D's parts to (integer) times $0 \leq t \leq T$ that sends the root to T and s.t., for each path from leaf to root, the corresponding sequence of times increases. So f might send ⌢'s red part to $t = 3$ and its green part to $t = 4$, but — because the green node has a red child — not vice versa. Let $|\text{Aut}_f(D)|$ count automorphisms of D that preserve f. Up to unbiasing terms, [1] we construct the *re-summed value* $\text{rvalue}_f(D)$ as follows:

Node rule: insert a factor a $\nabla^d l_x$ for each degree d node.

Outline rule: group each part's nodes within brackets $\mathbb{E}_x[]$.

Edge rule: if f sends an edge's endpoints to times t, t', insert a factor of $K^{|t'-t|-1}\eta$, where $K \triangleq (I - \eta H)$.

So if f maps ⌢'s red part to time $t = T - \Delta t$, then (the red part gives S; the green part, J):

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

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

2.6. 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 η. We later truncate the series to small d, thus focusing on few-edged diagrams and simplifying the combinatorics of embeddings.

Theorem 1 (Special case of $E = B = 1$) *For any T: for η small enough, SGD has expected test loss*

$$\sum_{\substack{D \text{ an irreduc-} \\ \text{-ible diagram}}} \sum_{\substack{f \text{ an embed-} \\ \text{-ding of } D}} \frac{(-1)^{|\text{edges}(D)|}}{|\text{Aut}_f(D)|} \text{rvalue}_f(D)$$

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

Theorem 2 *If $\theta_\star$ is a non-degenerate local minimum of l (i.e. $G(\theta_\star) = 0$ and $H(\theta_\star) > 0$), then for SGD initialized sufficiently close to $\theta_\star$, the dth-order truncation of Theorem 1 converges as $T \to \infty$.*

Caution: the $T \to \infty$ limit in Theorem 2 might not measure any well-defined limit of SGD, since the limit might not commute with the infinite sum. We have not seen such pathologies in practice, so we will freely speak of "SGD in the large-T limit" as informal shorthand when referencing this Theorem.

---

[1] For example, we actually define ⌣ to be the cumulant $C = \mathbb{E}[(\nabla l_x(\theta) - G)^2]$, not the moment $\mathbb{E}[(\nabla l_x(\theta))^2]$. This centering is routine (see §**??**), tedious to notate, and un-germane, so we ignore it in the paper body.

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

Corollary 1 (Computed from ⌢⌣ **)** *Run SGD for $T \gg 1/\eta H$ from a non-degenerate test minimum. Written in an eigenbasis of ηH, θ has an expected displacement of*

$$-\frac{\eta^3}{2} \sum_{\mu\nu} C_{\mu\nu} \frac{1}{\eta(H_{\mu\mu} + H_{\nu\nu})} J_{\mu\nu\lambda} \frac{1}{H_{\lambda\lambda}} + o(\eta^2)$$

Intuitively, $D = $ ⌢⌣ connects the subdiagram ⌢ $\propto CH$, via an extra edge on the green node (an extra ∇ on H), to D's degree-1 root, G. By l'Hôpital, [1] the displacement is $\propto -C\nabla H$. That is, SGD moves toward minima that are flat *with respect to C* (Figure 1 ■⊞). Taking limits to drop the non-degeneracy hypothesis, we expect *sustained* motion toward flat regions in a valley of minima. By avoiding Wei and Schwab (2019)'s assumptions of constant C, we find that SGD's velocity field is typically non-conservative, i.e. has curl (§4.2). Indeed, $\nabla(CH)$ is a total derivative but $C\nabla H$ is not. Since, by low-pass filter theory, $CH/2 + o(C)$ is the loss increase upon convolving l with a C-shaped Gaussian, we say that SGD descends on a C-smoothed landscape that changes as C does. Our $T \gg 1$ result is $\Theta(\eta^2)$, while Yaida (2019b)'s similar $T = 2$ result is $\Theta(\eta^3)$. Indeed, our analysis integrates the noise over many updates, hence amplifying C's effect. Experiments verify our law.

2.8. Both flat and sharp minima overfit less

Intuitively, sharp minima are robust to slight changes in the average *gradient* and flat minima are robust to slight *displacements* in weight space (Figure 1 □⊞). However, as SGD by definition equates displacements with gradients, it may be unclear how to reason about overfitting in the presence of curvature. Our theory, by (automatically) accounting for the implicit regularization of fixed-T descent, shows that both effects play a role. In fact, by routine calculus on the left hand side of Corollary 2, overfitting is maximized for medium minima with curvature $H \sim (\eta T)^{-1}$.

Corollary 2 (from ⌢⌣ **,** ⌣ **)** *Initialize GD at a non-degenerate test minimum $\theta_\star$. The overfitting (test loss minus $l(\theta_\star)$) and gen. gap (test minus train loss) due to training are:*

$$\left(\frac{C/N}{2H}\right)^{\rho\lambda}_{\mu\nu} \left((I - \exp(-\eta TH))^{\otimes 2}\right)^{\mu\nu}_{\rho\lambda} + o(\eta^2) \quad ; \quad \left(\frac{C/N}{H}\right)^{\mu\lambda}_{\mu\nu} (I - \exp(-\eta TH))^{\nu}_{\lambda} + o(\eta)$$

The gen. gap tends to $C_{\mu\nu}(H^{-1})^{\mu\nu}/N$ as $T \to \infty$. For maximum likelihood (ML) estimation in well-specified models near the "true" minimum, $C = H$ is the Fisher metric, so we recover AIC: (model dimension)/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.

3. Consequences of the theory

3.1. High-C regions repel small-(E, B) SGD more than large-(E, B) SGD

---

[1] Roughly: if a displacement $\Delta\theta$ grows loss by $GC\nabla H$ nats, and by G nats per foot, then $\Delta\theta$ is $C\nabla H$ feet.
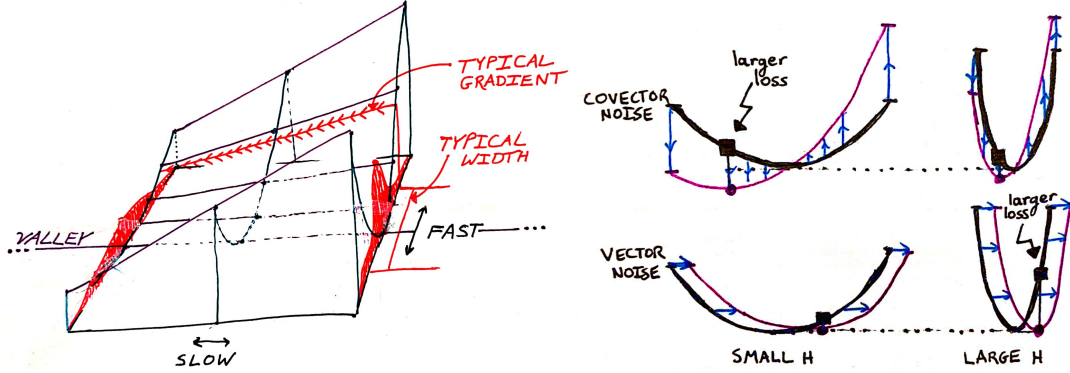
Figure 1: **Geometric intuition for curvature-noise interactions. Left**: Gradient noise pushes SGD toward flat minima (Corollary 1). The red densities show the typical θs, perturbed from the minimum due to noise C, in two cross sections of the loss valley. $J = \nabla H$ measures how curvature changes across the valley. Our theory does not assume separation between "fast" and "slow" modes, but we label them in the picture to ease comparison with Wei and Schwab (2019). Compare with Figure 4. **Right**: Both curvature and the structure of noise affect overfitting. In each of the four subplots, the ↔ axis represents weight space and the ↕ axis represents loss. ⊡⣿: *covector*-perturbed landscapes favor large Hs. ⊡⣿: *vector*-perturbed landscapes favor small Hs. SGD's implicit regularization interpolates between these rows (Corollary 2).

Physical intuition (§**??**) suggests that noise repels SGD. In particular, if two neighboring regions of weight space have high and low levels of gradient noise, respectively, then we expect the rate at which θ jumps from the former to the latter to exceed the opposite rate. There is thus a net movement toward regions of small C! This mechanism parallels the Chladni effect (Chladni, 1787) (Figure 2). [1] Our theory makes this intuition precise; the drift is in the direction of $-\nabla C$, and the effect is strongest when gradient noise is not averaged out by large batch sizes.
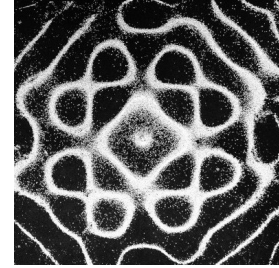

Figure 2: **Chladni plate**. Grains of sand on a vibrating plate tend toward stationary regions.

Corollary 3 (⌢) *SGD avoids high-C regions more than GD:* $l_C \triangleq \frac{N-1}{4N}\nabla^\mu C_\nu^\nu = \mathbb{E}\left[\theta_{GD} - \theta_{SGD}\right]^\mu - o(\eta^2)$. *If $\hat{l}_c$ is a smooth unbiased estimator of l_c, then GD on $l + \hat{l}_c$ has an expected test loss that agrees with SGD's to order η^2. We call this method GDC.*

An analogous form of averaging occurs over multiple epochs. For a tight comparison, we scale the learning rates appropriately so that, to leading order, few-epoch and many-epoch SGD agree. Then few and many- epoch SGD differ, to leading order, in their sensitivity to ∇C:

Corollary 4 (⌢) *SGD with $M = 1$ and $\eta = \eta_0$ avoids high-C regions more than SGD with $M = M_0$ and $\eta = \eta_0/M_0$. Precisely:* $\mathbb{E}\left[\theta_{M=M_0} - \theta_{M=1}\right]^\mu = \left(\frac{M_0-1}{4M_0}\right)N\left(\nabla^\mu C_\nu^\nu\right) + o(\eta^2)$.

[1] From Pierre Dragicevic and Yvonne Jansen's data physicalization project, Creative Commons BY-SA 3.0.

3.2. Non-Gaussian noise affects SGD but not SDE

Stochastic differential equations (SDE: see Liao et al. (2018)) are a popular theoretical approximation of 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 4), but SDE assumes uncorrelated gradient updates. Even if we restrict to single-epoch SDE, differences arise due to time discretization and non-Gaussian noise. Intuitively, SGD and SDE respond differently to changes in curvature:

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 skewed noise is $-\frac{T}{6}S_{\mu\nu\lambda}J^{\mu\nu\lambda} + o(\eta^3)$.* [1]

4. Experiments

Despite the convergence results in Theorems 1 and 2, we have no theoretical bounds for the domain and *rate* of convergence. Instead, we test our predictions by experiment. We perceive support for our theory in drastic rejections of the null hypothesis. For instance, in Figure 3 ▦, (Chaudhari and Soatto, 2018) predicts a velocity of 0 while we predict a velocity of $\eta^2/6$ Here, I bars, + signs, and shaded regions all mark 95% confidence intervals based on the standard error of the mean. §?? describes neural architectures, the definitions of artificial landscapes, sample sizes, and further plots.

4.1. Training time, epochs, and batch size; C repels SGD more than GD

We test Theorem 1's order η^3 truncation on smooth convnets for CIFAR-10 and Fashion-MNIST. Theory agrees with experiment through timescales long enough for accuracy to increase by 0.5% (Figure 3 ▦, ▦). §?? supports Corollary 4's predictions about epoch number. Figure 3 ▦ tests Corollary 3's claim that, relative to GD, high-C regions *repel* SGD. This is significant because C controls the rate at which the gen. gap (test minus train loss) grows (Corollary 2, Figure 3 ▦).

---

[1] This approximation of Example ??'s more exact expression agrees with the latter to leading order in η.
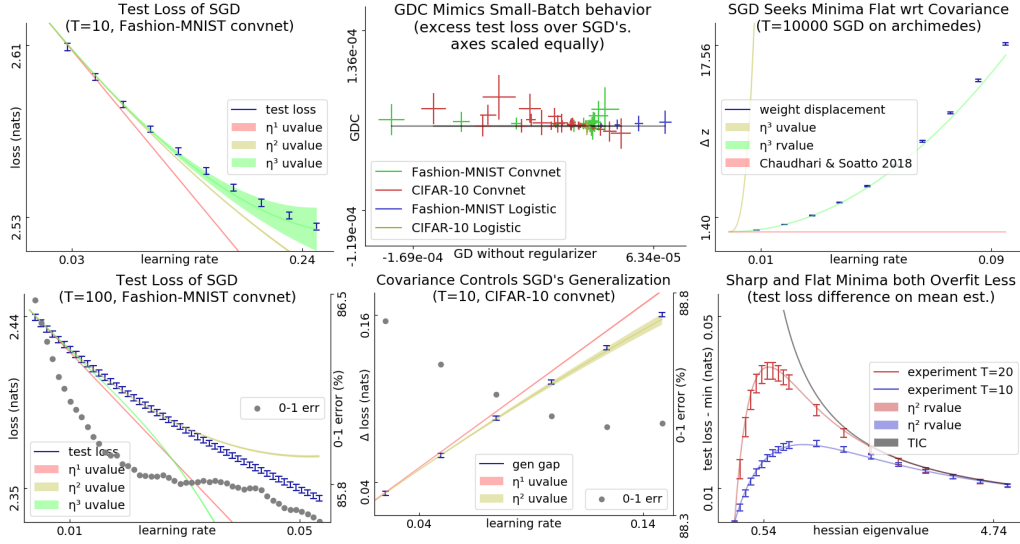
Figure 3: **Experiments on natural and artificial losses.** The label `rvalue` refers to Theorem 1's predictions, approximated as in Remark 1. Curves marked `uvalue` are polynomial approximations to Theorem 1's result (see §**??**). `uvalues` are simpler to work with but (see ▦) may be less accurate.

Left: Perturbation models SGD for small ηT. Fashion-MNIST convnet's test loss vs learning rate. In this small T setting, we choose to use our theory's simpler un-resummed values (**??**) instead of the more precise rvalues. ▦: For all init.s tested (1 shown, 11 unshown), the order 3 prediction agrees with experiment through $\eta T \approx 10^0$, corresponding to a decrease in 0-1 error of $\approx 10^{-3}$. ▦: For large ηT, our predictions break down. Here, the order-3 prediction holds until the 0-1 error improves by $5 \cdot 10^{-3}$. Beyond this, 2nd order agreement with experiment is coincidental.

Center: C controls gen. gap and distinguishes GD from SGD. With equal-scaled axes, ▦ shows that GDC matches SGD (small vertical varianec) 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, $T = 10$. ▦: 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: Predictions near minima excel for large ηT. ▦: SGD travels Archimedes' valley of global minima in the positive z direction. Note: H and C are bounded across the valley, we see drift for all small η, and we see displacement exceeding the landscape's period of 2π. So: the drift is not a pathology of well-chosen η, of divergent noise, or of ephemeral initial conditions. ▦: For Mean Estimation with fixed C and a range of Hs, initialized at the truth, the test losses after fixed-T GD are smallest for very sharp and very flat H. Near $H = 0$, our predictions improve on TIC (Dixon and Ward, 2018) and thus on AIC.

4.2. Minima that are flat *with respect to C* attract SGD

To test the claimed dependence on C, §**??** constructs a landscape, Archimedes, with non-constant C throughout its valley of global min-
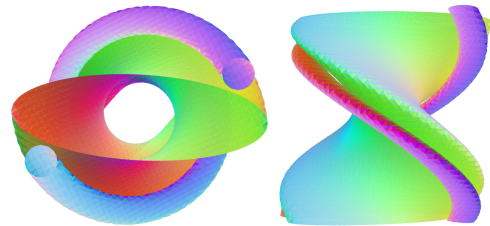

Figure 4: **Archimedes.** A **green** level surface of l twists around a valley of minima (z axis) at its center; l is large outside this surface. Due

ima. Figure 4 depicts Archimedes' chiral shape.[1] As in Archimedes' screw or Rock-Paper-Scissors, each point θ has a neighbor that, from $C(\theta)$'s perspective but not absolutely, is flatter. This permits eternal motion despite the landscape's symmetry. Indeed, Corollary 1 predicts a z-velocity of $+\eta^2/6$ per timestep, while Chaudhari and Soatto (2018)'s SDE-based analysis predicts a constant velocity of 0.[2] Our prediction agrees with experiment (Figure 3 ▦). Because SGD's motion depends smoothly on the landscape, the special case of Archimedes implies that non-conservativity is typical. One may have sought an "effective loss" $\tilde{l}$ such that, up to $\sqrt{T}$ diffusion terms, SGD on l matches ODE on $\tilde{l}$. The non-conservativity of SGD's velocity shows that no such $\tilde{l}$ exists.

4.3. 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, both phenomena occur, and noise structure determines which dominates (Corollary 2). This effect appears even in Mean Estimation (§**??**): Figure 3 ▦. To combat overfitting, we may add Corollary 2's expression for gen. gap to l. By descending on this regularized loss, we may tune smooth hyperparameters such as l_2 regularization coefficients for small datasets ($H \ll C/N$) (§**??**). Since matrix exponentiation takes time cubic in dimension, this regularizer is most useful for small models.

5. 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 avoids curvature and noise, and curvature and noise control 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 proposed a regularizer 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).

---

[1] We made these plots with the help of Paul Seeburger's online applet, CalcPlot3D.

[2] Indeed, Archimedes' velocity is η-perpendicular to the image of $(\eta C)^\mu_\nu$ in tangent space.

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.

Much as meteorologists understand how warm and cold fronts interact despite long-term forecasting's intractability, we quantify how curvature and noise contribute to 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 training loss — to account for noise.

5.1. Future work

Acknowledgments

References

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

S. Bonnabel. Sgd on riemannian manifolds. *IEEE Transactions on Automatic Control*, 2013.

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

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

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

E.F.F. Chladni. Entdeckungen über die theorie des klanges. *Leipzig*, 1787.

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

M.F. Dixon and T. Ward. Takeuchi information as a form of regularization. *Arxiv Preprint*, 2018.

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

R.P. Feynman. A space-time appxoach to quantum electrodynamics. *Physical Review*, 1949.

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.

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

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

F. Kunstner, P. Hennig, and L. Balles. Limitations of the empirical fisher approximation for natural gradient descent. *NeurIPS*, 2019.

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

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.

Y. Nesterov. Lectures on convex optimization: Minimization of smooth functions. *Springer Applied Optimization 87, Section 2.1*, 2004.

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.

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

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

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

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.

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.

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

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 sgd. *ICLR*, 2019a.

Sho Yaida. A first law of thermodynamics for sgd. *Personal Communication*, 2019b.

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

Appendix A. My Proof of Theorem 1

This is a boring technical proof.

Appendix B. My Proof of Theorem 2

This is a complete version of a proof sketched in the main text.