

A Space-Time Approach to Analyzing Stochastic Gradient Descent

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

May 31, 2020

Abstract

We analyze of 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 interpretation of SGD as a superposition of concurrent interactions between weights and data, each represented by a diagram analogous to those of ?? In the conclusion, we discuss this bridge to physics — and its relation to Hessian methods and natural GD — as topics for future research.

1.1 Example of diagram-based reasoning


Our theory analyzes SGD in terms of combinatorial objects we call *diagrams*. Deferring details, we illustrate how our theory yields non-trivial results via short arguments. First, we list how components of diagrams encode statistics of the loss $l_x(\theta)$ at weight θ and datapoint x :

$$\begin{aligned}
 G &\triangleq \mathbb{E}_x [\nabla l_x(\theta)] \triangleq \text{red node} \\
 H &\triangleq \mathbb{E}_x [\nabla \nabla l_x(\theta)] \triangleq \text{red node with 2 edges} & C &\triangleq \mathbb{E}_x [(\nabla l_x(\theta) - G)^2] \triangleq \text{red node with 2 edges, fuzzy outline} \\
 J &\triangleq \mathbb{E}_x [\nabla \nabla \nabla l_x(\theta)] \triangleq \text{red node with 3 edges} & S &\triangleq \mathbb{E}_x [(\nabla l_x(\theta) - G)^3] \triangleq \text{red node with 3 edges, fuzzy outline}
 \end{aligned}$$

Table 1: **Notation.** Throughout, G, H, J denote the 1st, 2nd, and 3rd derivatives of the loss function. We write C, S for the 2nd and 3rd cumulants of the gradient distribution. We differentiate w.r.t. the weight θ and we take expectations w.r.t. the datapoints x . Note: the tensors J, S have three indices. Each $\nabla^d l_x$ corresponds to a node with d thin edges emanating, and fuzzy outlines connect nodes that occur within the same expectation.

We may connect Table 1’s diagrams together to obtain *complete diagrams* without loose ends. For example, we may connect two copies of $G = \text{red node}$ with one copy of $H = \text{red node with 2 edges}$ to obtain .¹ If we run SGD for T gradient steps with learning rate η starting at θ_0 , then by Taylor expansion we may express the expected test loss at the final weight θ_T in terms of the statistics in Table evaluated at the initialization θ_0 . Diagrams organize the computation of this Taylor series.


Main Idea (Informal). There is a method to assign to any complete diagram a number that depends on η, T . SGD’s expected test loss is a sum, over all complete diagrams, of these numbers. We incur only an $o(\eta^d)$ error if we consider only diagrams with at most d edges.

Example 1 (How does non-Gaussian noise affect test loss?). Assume² θ_0 minimizes the test loss and that we run SGD for 1 epoch with batch size 1. The skew S is 0 for Gaussians, and we seek the effect of non-zero S . To compute the leading-order effect of S on test loss, we identify the fewest-edged complete diagrams containing $S = \text{red node with 3 edges, fuzzy outline}$. In this case, there is one such diagram: . Then, working in a basis that diagonalizes ηH , we obtain the leading-order effect of S on test loss (with error $o(\eta^3)$):

$$-\frac{\eta^3}{6} \sum_{\mu\nu\lambda} S_{\mu\nu\lambda} \frac{1 - \exp(-T\eta(H_{\mu\mu} + H_{\nu\nu} + H_{\lambda\lambda}))}{\eta(H_{\mu\mu} + H_{\nu\nu} + H_{\lambda\lambda})} J_{\mu\nu\lambda}$$

¹ We color nodes for convenient reference (e.g. to a diagram’s “green nodes”). As mere labels, colors lack mathematical meaning.

² for simplicity. Our theory is not limited to this setting.

Remark 1. The S , the three H 's, and the J above respectively correspond to 's group of red nodes, three thin edges, and green node. Each diagram encodes many Taylor terms, and the fact that we may evaluate each diagram as a whole is an advantage of our calculational framework. Intuitively, each diagram gives the net effect of a certain combination of gradients (G), noise (C, S, \dots) and curvature (H, J, \dots). After developing our theory more precisely, we will return to these intuitive interpretations.

2 Background and Notation

2.1 Loss landscape

We henceforth fix a space \mathcal{M} of weights on which a loss function $l : \mathcal{M} \rightarrow \mathbb{R}$ is defined. SGD operates on unbiased estimates of l drawn from some fixed probability distribution \mathcal{D} . We thus denote by $(l_n : 0 \leq n < N)$ an i.i.d. training sequence of such estimates. We will refer both to n and to l_n as *training points*. We likewise write l_x for a sample from \mathcal{D} independent from the training sequence. We assume the regularity conditions listed in Appendix **FILL IN**, e.g. that l, l_x are analytic and that all moments exist.

E.g.: our theory applies to tanh networks with cross entropy loss on bounded data — and with arbitrary weight sharing, skip connections, soft attention, dropout, and weight decay.

2.2 Tensor conventions

Adopting Einstein's convention, we implicitly sum repeated Greek indices: if A_μ, B^μ are the coefficients of a covector A and a vector B ¹, indexed by basis elements 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 a gradient covector into a displacement vector [?], and we use η to raise indices: e.g., in $H^\mu_\lambda \triangleq \eta^{\mu\nu} H_{\nu\lambda}$, η raises one of $H_{\mu\nu}$'s indices. Another example is $C^\mu_\mu \triangleq \sum_{\mu\nu} \eta^{\mu\nu} \cdot C_{\nu\mu}$. Standard syntactic constraints make manifest which expressions transform naturally.

We say two expressions *agree to order* η^d when their difference, divided by some homogeneous degree- d polynomial of η , tends to 0 as η shrinks. Their difference is then $\in o(\eta^d)$.

2.3 SGD terminology

SGD decreases the objective l via T steps of discrete-time η -steepest² 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 $M = E/B = T/N$. Concretely, SGD performs $T = NM$ updates of the form:

$$\theta^\mu \leftarrow \theta^\mu - \eta^{\mu\nu} \nabla_\nu \left(\frac{1}{B} \sum_{n \in \mathcal{B}_t} l_n(\theta) \right)$$


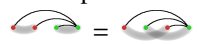
¹ Vectors/covectors are also called column/row vectors.

² To define “steepest” requires a metric on l 's domain. We regard $\eta^{\mu\nu}$ as an (inverse) metric.

We write l_t for the loss $\frac{1}{B} \sum_{\mathcal{B}_t} l_n$ on the t th batch. **no VANILLA!**

2.4 Diagrams and embeddings

Though a rough, intuitive understanding of concepts such as *diagram* suffices for absorbing this paper’s main results, the following definitions may help the reader who wishes to follow our mathematics closely.

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.

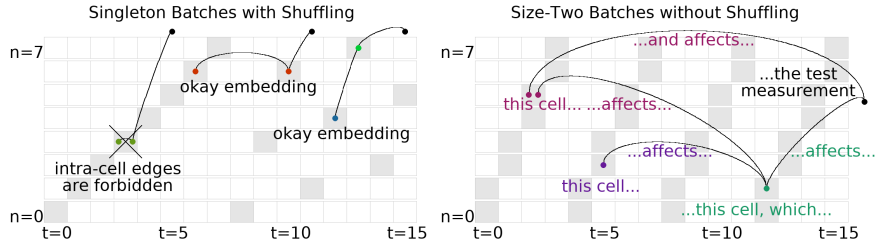





Figure 1: **Diagrams in Spacetime Depict SGD’s Subprocesses.** Two spacetimes with $N = 8, T = 16$. **Left:** Batchsize $B = 1$ with inter-epoch shuffling. Embeddings, legal and illegal, of , , and . **Right:** Batchsize $B = 2$ without inter-epoch shuffling. Interpretation of an order η^4 diagram embedding.

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 1 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 η . We denote D 's un-resummed value by $\text{value}(D)$, 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 $\text{value}(D)$ 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.:

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

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 ?):

$$\begin{aligned} \text{Diagram with 2 fuzzy outlines} &\triangleq \text{Diagram with 2 ties} - \text{Diagram with 1 tie} - \text{Diagram with 0 ties} \\ &\triangleq \text{Diagram with 2 ties} - \text{Diagram with 1 tie} - \text{Diagram with 0 ties} + 2 \cdot \text{Diagram with 1 tie} \end{aligned}$$

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 , .

3 Diagram Calculus for SGD

3.1 Recipe for SGD's expected test loss

Our Main Theorem expresses SGD's test loss as a sum over diagram embeddings. Recalling that a diagram with d edges is $O(\eta^d)$, we may read this Theorem as a Taylor series in the learning rate. In practice, we truncate the series to small d , thus focusing on the few-edged diagrams.

Theorem 1 (Test Loss as a Path Integral). *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 diagrams drawn with fuzzy outlines instead of ties, 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 the 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 at a Local Minimum). *When SGD is initialized at a local minimum of test loss, and when $\nabla \nabla l_x$ is bounded below by some positive form that doesn't depend on x , then the d th-order truncation of Theorem 1 converges as T diverges.*

Remark 2 (Approximation by integrals). In practice, we approximate sums over embeddings by integrals over times and $(I - \eta H)^t$ by $\exp(-\eta H t)$. This incurs a multiplicative error of $1 + o(\eta)$ that preserves leading order results. So diagrams induce easily evaluated integrals of exponentials.

Remark 3 (Using Un-resummed Values). $\text{value}(D)$ is simpler to work with than $\text{rvalue}_f(D)$. Theorem 1 remains true if we replace each $\text{rvalue}_f(D)$ by $\text{value}(D)$, so long as we drop the constraint that D be irreducible and we use diagrams drawn with fuzzy ties instead of fuzzy outlines. However, Theorem 2's convergence guarantee no longer applies, and empirically we find deteriorated predictions.

Remark 4 (Variants). The above gives SGD's expected test loss. What if we seek train instead of test losses? Or net weight displacements instead of losses? Or variances instead of expectations? Theorem 1 and Remark 3 have simple analogues for each of these ²³ possibilities, which we discuss in the appendix.

3.2 Single-Epoch, Singleton-Batch SGD

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*

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

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.

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, ηT 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 2.










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

Table 2: **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.

4 Insights from the Formalism





4.1 SGD descends on a C -smoothed landscape

Integrating $\text{rvalue}_f(\text{diagram})$ over embeddings f , we see:


Corollary 1 (Minima flat w.r.t. C attract SGD). *Initialized at a test minimum, vanilla SGD's weight moves to order η^2 with a long-time-averaged¹ expected velocity of*

$$v^\pi = C_{\mu\nu} (F^{-1})^{\mu\nu}_{\rho\lambda} J^\rho_\sigma \left(\frac{I - \exp(-T\eta H)}{T\eta H} \eta \right)^{\sigma\pi}$$

per timestep. Here, $F = \eta H \otimes I + I \otimes \eta H$, a 4-valent tensor.

The intuition behind the Corollary is that the diagram  contains a subdiagram  = CH ; by a routine check, this subdiagram is the leading-order loss increase when we convolve the landscape with a C -shaped Gaussian. Since  connects the subdiagram to the test measurement via 1 edge, it couples  to the linear part of the

¹ That is, T so large that $C \exp(-\eta KT)$ is negligible. Appendix ?? gives a similar expression for general T .

test loss and hence represents a displacement of weights away from high CH . In short,  reveals that *SGD descends on a covariance-smoothed landscape*. See Figure 2 (right).

An un-resummed version of this result was first reported by ?; however, for fixed T , the un-resummed result scales with η^3 while Corollary 1 scales with η^2 . The discrepancy occurs, intuitively, because the re-summed analysis accounts for the accumulation of noise from many updates, hence amplifying the contribution of C . Our experiments verify our scaling law.

Unlike Wei and Schwab [2019], we make no assumptions of thermal equilibrium, fast-slow mode separation, or constant covariance. This generality reveals a novel dynamical phenomenon, namely that the velocity field above need not be conservative (see Section 5.4)

4.2 Curvature controls overfitting

Integrating $\text{rvalue}_f(\text{img alt="A small diagram showing a red arrow pointing right and a green arrow pointing left, with a blue arrow pointing right above them." data-bbox="345 395 385 410"/>) and $\text{rvalue}_f(\text{img alt="A small diagram showing a red arrow pointing right and a green arrow pointing left, with a blue arrow pointing right above them." data-bbox="455 395 495 410"/>) yields:$$

Corollary 2 (Flat, Sharp Minima Overfit Less). *Initialized at a test minimum, pure GD's test loss is to order η*

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

above the minimum. This vanishes when H does. Likewise, pure GD's generalization gap is to order η :


$$\frac{1}{N} C_{\mu\nu} (I - \exp(-\eta TH))^\nu_\lambda (H^{-1})^{\lambda\mu}$$

In contrast to the later-mentioned Takeuchi estimate, this does not diverge as H shrinks.

Corollary 2's generalization gap converges after large T to $C_{\mu\nu}(H^{-1})^{\mu\nu}/N$, also known as Takeuchi's Information Criterion (TIC). In turn, $C = H$ is the Fisher metric in the classical setting of maximum likelihood (ML) estimation (in well-specified models) near the "true" test minimum, 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 makes no model well-specification assumptions.

4.3 Effects of epochs and of batch size

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

Analyzing , we find that we may cause GD to mimic SGD using any smooth unbiased estimator \hat{C} of C :

Corollary 4 (Batch Size). *The expected test loss of pure SGD is, to order η^2 , less than that of pure GD by $\frac{M(N-1)}{2} (\nabla_\mu C_\nu) G^\mu/2$. Moreover, 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.*

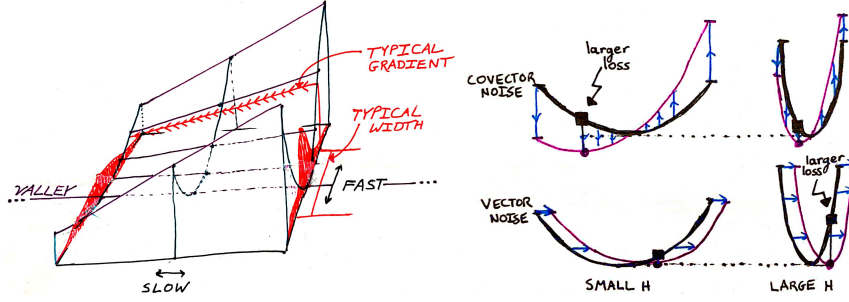


Figure 2: **Re-summation reveals novel phenomena.** **Left:** The entropic force mechanism: gradient noise induces a flow toward minima *with respect to the covariance*. Though our analysis assumes neither thermal equilibrium nor fast-slow mode separation, we label “fast and slow directions” to ease 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.

4.4 Non-Gaussian noise affects SGD but not SDE

Stochastic Differential Equations (SDE: see ?) 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, more interestingly, due to 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 leading order 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.

5 Experiments

We focus on experiments whose rejection of the null hypothesis (and hence support of our theory) is so drastic as to be visually obvious. For example, in Figure 5, [Chaud-

¹ This expression differs from the more exact expression of Example 1 because here we use Remark 3’s substitution. One may check that the two expressions agree to leading order.

hari and Soatto, 2018] predicts a velocity of 0 while we predict a velocity of $\eta^2/6$. Throughout, I bars and + marks denote a 95% confidence interval based on the standard error of the mean, in the vertical or vertical-and-horizontal directions, respectively. See Appendix ?? for experimental procedure including architectures and sample size.

5.1 Basic predictions

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

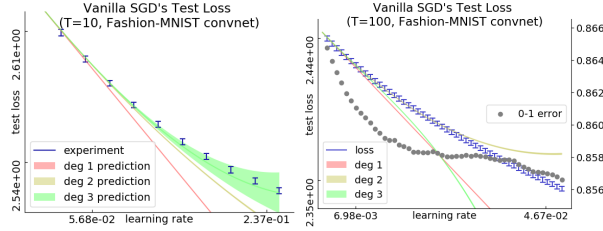


Figure 3: **Perturbation models SGD for small ηT .** Test loss vs learning rate on a Fashion-MNIST convnet, with un-re-summed predictions. **Left:** For the instance shown and all 11 other initializations unshown, our degree-3 prediction agrees with experiment through $\eta T \approx 10^0$, which corresponds to a decrease in 0-1 error of $\approx 10^{-3}$. **Right:** For larger ηT , our predictions can break down. Here, the order-3 prediction holds until the 0-1 error improves by $5 \cdot 10^{-3}$. Beyond this, close agreement with experiment is coincidental.

5.2 Emulating small batches with large ones

By Corollary 4, SGD avoids high- C regions more than GD. We artificially correct GD accordingly, yielding an optimizer, GDC, that indeed behaves like SGD on a range of landscapes (Figure 4 (left)). It may be important to emulate SGD's avoidance of high- C regions because we C controls the rate at which each new update increases the generalization gap¹ (Figure 4 (right)).

The connection between generalization and covariance was first established by Roberts [2018] in the case $T = 2$ and to order η^2 . In fact, that work conjectures the possibility of emulating GD with SGD. This sub-section extends that work by generalizing to arbitrary T and arbitrary orders η^d , and by concretely defining GDC.

In these experiments, we used a covariance estimator $\hat{C} \propto \nabla l_x(\nabla l_x - \nabla l_y)$ evaluated on two batches x, y that evenly partition the train set. For typical architectures, we may compute $\nabla \hat{C}$ with the same memory and time as the usual gradient ∇l_t , up to a multiplicative constant.

¹ Reminder: for us, generalization gap is test minus train loss.

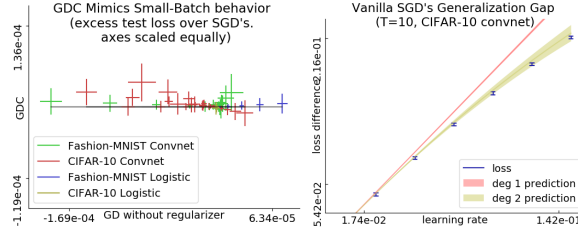


Figure 4: **C controls generalization and distinguishes GD from SGD.** **Left:** With equal-scaled axes, this plot shows that GDC matches SGD (small vertical variation) better than GD matches SGD (large horizontal variation) in test loss, for a variety of learning rates ($\approx 10^{-3} - 10^{-1}$) and initializations (zero and several Xavier-Glorot trials) on logistic and architectures for image classification. Here, $T = 10$. **Right:** CIFAR-10 generalization gaps. For the instance shown and all 11 other initializations unshown, the degree-2 prediction agrees with experiment through $\eta T \approx 5 \cdot 10^{-1}$.

5.3 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.

5.4 Nonconservative entropic force

To test Corollary 1's predicted force, we construct a counter-intuitive loss landscape wherein, for arbitrarily small learning rates, SGD steadily increases the weight's z component despite 0 test gradient in that direction. Our mechanism differs from that discovered by Chaudhari and Soatto [2018]. Specifically, because in this landscape the force is η -perpendicular to the image of ηC , that work predicts an entropic force of 0. This disagreement in predictions is possible because our analysis does not make any assumptions of equilibrium, conservatism, or continuous time.

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-dimensional $w \in \mathbb{R}^3$ (initialized to 0) and 1-dimensional $x \sim \mathcal{N}(0, 1)$:

$$l_x(w) \triangleq \frac{1}{2} H(z)(w, w) + x \cdot S(z)(w)$$

Here, $H(z)(w, w) = w_x^2 + w_y^2 + (\cos(z)w_x + \sin(z)w_y)^2$ and $S(z)(w) = \cos(z - \pi/4)w_x + \sin(z - \pi/4)w_y$. There is a valley of global minima defined by $x = y = 0$. If SGD

is initialized there, then to leading order in η and for large T , the re-summed theory predicts a z -speed of $\eta^2/6$ per timestep. Our re-summed predictions agree with experiment for ηT so large that the weight moves about 5 times the landscape’s natural length scale of 2π (Figure 5, left).

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.

5.5 Sharp and flat minima both overfit less

Prior work has varyingly found that *sharp* minima overfit less (after all, l^2 regularization increases curvature) or that *flat* minima overfit less (after all, flat minima are more robust to small displacements in weight space). Corollary 2 reconciles these competing intuitions by showing how the relationship of generalization and curvature depends on the learning task’s noise structure and how the metric η^{-1} mediates this distinction (Figure 2, right).

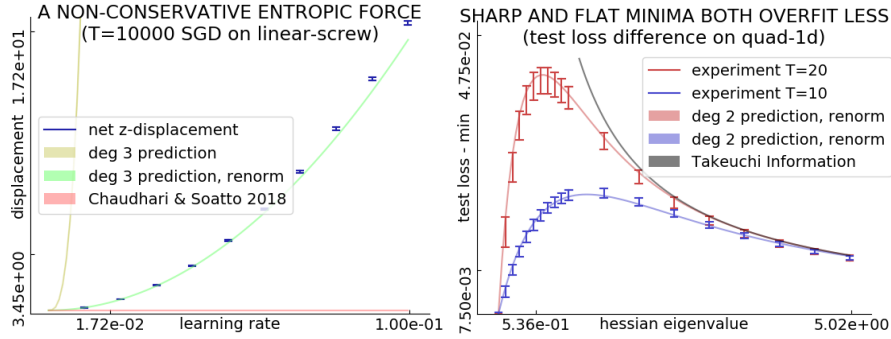


Figure 5: Re-summed predictions excel even for large ηT for SGD near minima. **Left:** On Linear Screw, the persistent entropic force pushes the weight through a valley of global minima not at a $T^{1/2}$ diffusive rate but at a directional T^1 rate. Since Hessians and covariances are bounded throughout the valley and the effect appears for all sufficiently small η , the effect is not a pathological artifact of well-chosen learning rate or divergent covariance noise. The net displacement of $\approx 10^{1.5}$ well exceeds the z -period of 2π . **Right:** For Mean Estimation with fixed covariance and a range of Hessians, initialized at the true minimum, the test losses after fixed- ηT optimization are smallest for very small and very large curvatures. This evidences our prediction that both sharp and flat minima overfit less and that TIC’s singularity is suppressed.

Because the TIC estimates a smooth hypothesis class’s generalization gap, it is tempting to use it as an additive regularization term. However, since the TIC is singular where the Hessian is singular, it gives insensible results for over-parameterized models. Indeed, ? report numerical difficulties requiring an arbitrary cutoff.

Fortunately, by Corollary 2, the implicit regularization of gradient descent both demands and enables a singularity-removing correction to the TIC (Figure 5, right).

The resulting *Stabilized TIC* (STIC) uses the metric η^{-1} implicit in gradient descent to threshold flat from sharp minima¹. It thus offers a principled method for optimizer-aware model selection easily compatible with automatic differentiation systems. By descending on STIC, we may tune smooth hyperparameters such as l_2 coefficients. Experiments on an artificial Mean Estimation problem (task in Appendix ??, plot in Appendix ??) recommend STIC for model selection when H is negligible compared to C/N as in the noisy, small- N regime. Because diagonalization typically takes time cubic in dimension, exact STIC regularization is most useful for small models on noisy and limited data.

6 Related Work

It was ? who, in uniting gradient descent [?] with stochastic approximation [?], invented SGD. Since the development of back-propagation for efficient differentiation [?], SGD has been used to train connectionist models including neural networks [Bottou, 1991], in recent years to remarkable success [?].

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

Following Roberts [2018], we avoid continuous-time approximations and Taylor-expand around $\eta = 0$. We hence extend that work beyond leading order and beyond 2 time steps, allowing us to compare, for instance, the expected test losses of multi-epoch and one-epoch SGD. We also quantify the overfitting effects of batch size, whence we propose a regularizer that causes large-batch GD to emulate small-batch SGD. In doing so, we establish a precise version of the relationship — between covariance, batch size, and generalization — conjectured by ?.

While we make rigorous, architecture-agnostic predictions of learning curves, these predictions become vacuous for large η . Other discrete-time dynamical analyses allow large η by treating deep generalization phenomenologically, whether by fitting to an empirically-determined correlate of Rademacher bounds [?], by exhibiting generalization of local minima *flat* with respect to the standard metric (see ?, Keskar et al. [2017], Wang et al. [2018]), or by exhibiting generalization of local minima *sharp* with respect to the standard metric (see ?, Dinh et al. [2017], Wu et al. [2018]). Our work reconciles those seemingly clashing claims.


Others have perturbatively analyzed descent: ? perturb in inverse network width, employing Feynman-’t Hooft diagrams to correct the Gaussian Process approximation


¹ The notion of H ’s width depends on a choice of metric. Prior work chooses this metric arbitrarily. We show that choosing η^{-1} is a natural choice because it leads to a prediction of the gen. gap.


for a specific class of deep networks. Meanwhile, Chaudhari and Soatto [2018] and ? perturb in learning rate to second order by approximating noise between updates as Gaussian and uncorrelated. In neglecting correlations and heavy tails, that work neither extends to higher orders nor describes SGD’s generalization behavior. By contrast, we use Feynman-Penrose diagrams to compute test and train losses to arbitrary order in learning rate. Our method accounts for non-Gaussian and correlated noise and applies to *any* sufficiently smooth architecture. For example, since our work does not rely on information-geometric relationships between C and H [Amari, 1998]¹, it applies to inexact-likelihood landscapes such as VAEs’.

7 Conclusion

TODO We present a diagram-based method for studying stochastic optimization on short timescales. Theorem 1 justifies long-time predictions of SGD’s dynamics near minima. Our theory answers the following questions.

Which Minima Overfit Less? By analyzing , we find that flat and sharp minima both overfit less than minima of curvature comparable to $(\eta T)^{-1}$. Flat minima are robust to vector-valued noise, sharp minima are robust to covector-valued noise, and medium minima attain the worst of both worlds. We thus reconcile prior intuitions that sharp [Keskar et al., 2017, Wang et al., 2018] or flat [Dinh et al., 2017, Wu et al., 2018] minima overfit worse. These considerations lead us to a smooth generalization of AIC enabling hyperparameter tuning by gradient descent.

Which Minima Does SGD Prefer? Analyzing , we refine Wei and Schwab [2019] to nonconstant, nonisotropic covariance to reveal that SGD descends on a loss landscape smoothed by the *current* covariance C . In particular, SGD moves toward regions flat with respect to C . As C evolves, the smoothing mask and thus the effective landscape evolves. These dynamics are generically nonconservative. In contrast to Chaudhari and Soatto [2018]’s SDE approximation, SGD does not generically converge to a limit cycle.

Can GD Emulate SGD? By analyzing , we prove the conjecture of Roberts [2018], that large-batch GD can be made to emulate small-batch SGD. We show how to do this by adding a multiple of an unbiased covariance estimator to the descent objective. This emulation 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].

7.1 Consequences

Our analysis of which minima (among a valley of minima) SGD prefers — and our characterization of when SGD overfits less in certain minima — together offer insight into SGD’s success in training over-parameterized models.

Our results may also help to analyze fine-tuning procedures such as the meta-learning of MAML Finn et al. [2017]. Indeed, those methods seek models initialized near minima

¹ Disagreement of C and H is typical in modern learning [??].

and tunable to new data through a small number of updates, a setting matched to our theory’s assumptions.

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.

Yet, even short-time predictions show how curvature and noise — and not just averaged gradients — repel or attract SGD’s current weight. For example, we proved that SGD in a valley moves toward regions flat with respect to the current covariance C . Much as meteorologists understand how warm and cold fronts interact despite the intractability of long-term weather forecasting, we quantify the counter-intuitive dynamics governing SGD’s short-time behavior.¹ Our results enhance the intuitions of practitioners — e.g. that "SGD descends on the train loss" — by summarizing the effect of noise in closed-form dynamical laws valid in each short-term interval of SGD’s trajectory.

7.2 Questions

The diagram method opens the door to exploration of Lagrangian formalisms and curved backgrounds²:

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 η, λ — admit diagrammatic analysis when we represent the λ 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 η , 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 spaces³. 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.

¹ Because our analysis holds for any initialization, one may imagine SGD’s coarse-grained trajectory as an integral curve of the vector field given by our theory.

² Landau and Lifshitz [1960, 1951] introduce these concepts.

³ One may represent the affine connection as a node, thus giving rise to non-tensorial and hence gauge-dependent diagrams.

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 4.3 shows how to modify the loss landscape so that large-batch GD enjoys the stochastic regularizing properties of small-batch SGD, dually, 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 7 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 4.2 and 5.5 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 privacy, adversaries, pedestrian safety, and bias reflected in training data, it becomes increasingly important to model the fact that training sets and test sets may differ. By quantifying effect of sample noise in learning, our work contributes to this goal.

Acknowledgements

We feel deep gratitude to Sho Yaida, 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 Andrzej 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.
- A.-L. Cauchy. Méthode générale pour la résolution des systèmes d’équations simultanées. *Comptes rendus de l’Académie des Sciences*, 1847.
- 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.
- M.F. Dixon and T. Ward. Takeuchi’s information criteria 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.
- 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.
- S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in sgd. *Arxiv Preprint*, 2018.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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 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.