A Space-Time Approach to Analyzing Stochastic Gradient Descent

Samuel C. Tenka [1]

Abstract

We harness Feynman Diagrams to reason about the behavior of Stochastic Gradient Descent (SGD) at small learning rates η. We empirically verify our theoretical predictions on artificial datasets and convnets for CIFAR-10 and Fashion-MNIST.

1. Introduction

Stochastic gradient descent (SGD) decreases an unknown objective l via T steps of discrete-time η-steepest[*] descent on noisy estimates of l. Practitioners benefit from the intuition that SGD descends on l itself (Bottou, 1991). We present a novel framework for refining this intuition to account for noise. For instance, we show that SGD moves to disalign the hessian H from the current covariance C of gradients, and that this force can dominate after SGD has reached a valley of train minima. We prove our predictions for small η, and our experiments show that even a single evaluation of our force laws at a weight θ suffices to predict how SGD drifts from θ through macroscopic timescales, i.e. ηT large enough to improve accuracy by 1 percentage point.

Departing from prior work, we model discrete time and hence non-Gaussian noise. Indeed, we give the finite-T, finite-η^{-1} corrections to continuous-time approximations such as ordinary and stochastic differential equations (ODE, SDE). We thus obtain new results quantifying how epoch number and batch size affect test loss. Our theory of noise recommends two novel regularizers that respectively induce GD to mimic SGD and help to tune hyperparameters such as l_2 coefficients. We verify our predictions on landscapes including CIFAR-10 convnets (details in Appendix G).

The theoretical formalism underlying our results is especially novel: we interpret SGD as a superposition of several

---

[1] Computer Science and AI Lab, Massachusetts Institute of Technology, Cambridge, MA, USA . Correspondence to: Samuel C. Tenka < `colimit.edu`; insert @-sign appropriately >.

[*] To define "steepness" requires a metric on l's domain. We will consider all flat metrics by Taylor expanding an (inverse) metric $\eta^{\mu\nu}$ around 0.

weight-data interactions, each represented by a Feynman-style diagram (§3 explains these words). This viewpoint offers not only quantitative predictions but also qualitative insight, e.g. that to order η^3, inter-epoch shuffling does not affect expected test loss. We believe that our diagram method is an elegant and general tool for studying stochastic optimization for T large and ηT small. Looking toward future work, our conclusion discusses Hessian methods and natural GD as low-hanging fruit.

1.1. Overview of Approach

Consider an SGD run on N train points for T steps, starting at a weight θ_0. Our method expresses the expectation (over randomly sampled train sets) of quantities such as the final weight (or test or train loss) as a sum of diagrams, where each diagram evaluates to a statistic of the loss landscape at initialization. Diagrams with e edges contribute only $O(\eta^e)$ to the quantities of interest, so for small η we sum only the few-edged diagrams and incur an $o(\eta^e)$ error term.

For example, the diagram •⌒• evaluates to the dot product ηGG, where $G = \nabla l$ is the gradient of the expected loss l, evaluated at θ_0. Likewise, •⌒•⌒• $= \eta^2 GHG$, where $H = \nabla\nabla l$ is the hessian. The rule is that each degree-d node evaluates to the dth derivative of l evaluated at θ_0, and the edges indicate the order in which those derivatives are multiplied. Further examples include • $= l(\theta_0)$, a 0th derivative, and

•⌒⌢⌒• $= \eta^3 GGJG$, where $J = \nabla\nabla\nabla l$ is l's third derivative.[†]

A diagram tells us about the loss landscape but not about SGD parameters such as T or inter-epoch shuffling. We summarize those parameters as set of pairs (n, t), one for each participation of the nth datapoint in the tth update. Full-batch GD will have NT many pairs, for instance, while singleton-batch SGD will have T many pairs.

[†] We color nodes for convenient reference (e.g. to a diagram's "green nodes"). As mere labels, colors lack mathematical meaning.
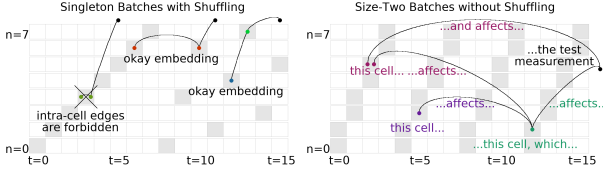
Figure 1. **Diagrams in spacetime dictate SGD's behavior.**

Each of a diagram's nodes abstractly represents an event at such a pair, and we may "concretize" a diagram by assigning to each node a specific pair (n, t). Reading a concrete diagram from left to right, an edge from (n, t) to (n', t') represents how the nth train point's participation in the tth timestep affects the n'th train point's participation in the t'th timestep. Thus, we permit only concretizations whose edges have $t < t'$. The rightmost node represents measurement at test time, so we do not assign a pair (n, t) to it.

Theorem (Main Theorem, Informal). *Fix SGD parameters, namely N, T, a batch size B, and a deterministic routine to sample the tth batch from a train set. Sum the diagrams with at most d edges, where a diagram with e edges and c many concretizations is weighted by $c/(-B)^e$. This sum agrees with SGD's expected final test loss to order $o(\eta^d)$.*

Example 1. What is SGD's expected final test loss to order η^1? There are only two diagrams with ≤ 1 edges: $\bullet = l(\theta_0)$ and $\bullet\!\!-\!\!\bullet = \eta GG$. For SGD with batchsize 1, $\bullet\!\!-\!\!\bullet$ has T many concretizations, since its rightmost node must represent the test measurement and its other node can represent any of T many (n, t) pairs. By the Theorem, the answer is $l(\theta_0) - T \cdot \eta GG + o(\eta^1)$.

Example 1 is well-known (e.g. Nesterov (2004)). Indeed, it quantifies the intuition that, in each of T steps, SGD moves the weight by ηG and hence decreases the loss by ηGG. The expression is exact for a noiseless linear landscape, but, because it fails to model how gradients depend on the current weight (curvature) or on the current train point (noise), it is typically an approximation. Our formalism corrects the expression by introducing new diagrams and hence new terms modeling curvature and noise.

Like Example 1, our predictions depend only on loss statistics near θ_0 and hence break down for ηT large enough that the weight moves far from initialization. Our theory is thus most useful in small-movement contexts. For instance, we analyze SGD overfitting near a minimum (Corollary 7). Moreover, by setting T large enough that SGD dynamics senses curvature and noise yet small enough that our predictions are valid, we analyze how curvature and noise — and not just gradients — repel or attract the evolving weight (Corollary 6). Theorem 2 establishes that this range of medium Ts is non-empty.

1.2. A First Look at Curvature and Noise

Intuitively, our diagrams' edges depict higher derivatives and hence the loss landscape's curvature. Meanwhile, we depict correlations and hence noise by a new structure: fuzzy "ties". For example, $\bullet\!\cdot\!\!-\!\!\bullet$ and $\bullet\!\!-\!\!\bullet$ are two valid and distinct diagrams. Fuzzy ties determine which derivatives occur within the same expectation, so we have

$$\bullet\!\cdot\!\!-\!\!\bullet \triangleq \eta^2 \mathbb{E}[\nabla l_n] \mathbb{E}[\nabla \nabla l_n] \mathbb{E}[\nabla l_n] = \eta^2 GHG = \eta^2 \frac{\nabla(GG)}{2} G$$

and, writing C for the covariance of gradients,

$$\bullet\!\!-\!\!\bullet \triangleq \eta^2 \mathbb{E}[\nabla l_n \nabla \nabla l_n] \mathbb{E}[\nabla l_n] = \eta^2 \frac{\nabla(GG + C)}{2} G$$

The general rule is that nodes in the same fuzzy-tie connected component occur in the same expectation brackets. Since fuzzy ties depict correlations, we demand that each concretization of a diagram sends any two fuzzily-tied nodes to pairs $(n, t), (n, t')$ that share a train point index n.

Example 2. When $N = T$, then singleton-batch SGD permits permits no concretizations of $\bullet\!\!-\!\!\bullet$, since the edge constraint $t < t'$ conflicts with the tie constraint $n = n'$ when, as in this case, the permitted (n, t) pairs comprise a bijection between ns and ts.

Example 3. By constrast, when $N = T$, full-batch GD permits $N\binom{N}{2}$ many concretizations of $\bullet\!\!-\!\!\bullet$, since all NT possible (n, t) pairs occur. Those concretizations $(n, t), (n, t')$ have as close analogues the concretizations $(n, t), (n', t')$ in Example 2's setting of the tie-less diagram $\bullet\!\cdot\!\!-\!\!\bullet$.

Comparing the two examples above reveals a difference between batchsize-1 and batchsize-N descent for $N = T$: by the Main Theorem, the latter incurs an additional test loss

$$\frac{c}{(-B)^e} \left(\bullet\!\!-\!\!\bullet - \bullet\!\cdot\!\!-\!\!\bullet \right) = \overset{algebra}{\cdots} = \frac{\eta^2(N-1)}{4} G\nabla C$$

It turns out that $\bullet\!\!-\!\!\bullet$ is the only 2-edged diagram whose concretizations in SGD and GD differ. Thus, this test loss difference between SGD and GD is correct to order η^2.
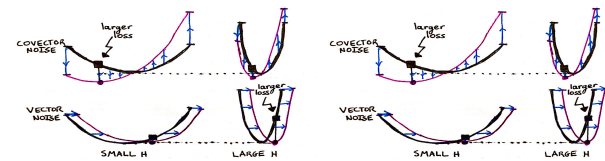

Figure 2. Hi

The above generalizes Roberts (2018)'s $T = 2$ result, proved without diagrams, to arbitrary T. In principle, one could avoid diagrams completely by direct use of our Key Lemma.

However, as demonstrated by side-by-side comparison in the Appendix, the concept of counting concretizations of diagrams helps to streamline calculation, yielding arguments 4 times shorter and arguably more insightful conceptual than direct perturbation. The more complicated the computation, the greater the savings of using diagrams.

2. Background and Notation

2.1. The Loss Landscape

We henceforth fix a loss landscape on a weight space $\mathcal{M}$, i.e. a mean-l distribution over smooth functions $l_n : \mathcal{M} \to \mathbb{R}$. We refer both to n and to l_n as *datapoints*.

More precisely, we fix a real-valued stochastic process indexed by the points of the *weight space* $\mathcal{M}$, an affine manifold. The process furnishes a *train sequence* $(l_n : 0 \leq n < N)$ of i.i.d. samples, each an unbiased estimate of l. We henceforth assume the regularity conditions listed in Appendix E, for instance that l, l_n are analytic and that all moments exist.

For example, these conditions admit tanh networks with cross entropy loss on bounded data — and with arbitrary weight sharing, skip connections, soft attention, dropout, batch-normalization with disjoint batches, and weight decay.

2.2. Tensor Conventions

We use $G_\mu, H_{\mu\nu}, J_{\mu\nu\lambda}$ for the first, second, and third derivatives of l and $C_{\mu\nu}$ for the covariance of gradients. By convention, our repeated Greek indices are implicitly summed: if A_μ, B^μ are the coefficients of a covector A and a vector B[*], 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 vector displacement (Bonnabel, 2013). We use η to *raise* indices. In $H^\mu_\lambda \triangleq \eta^{\mu\nu} H_{\nu\lambda}$, for instance, 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 with respect to optimization dynamics.

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 updates on the train sequence $(l_n : 0 \leq n < N)$. For each of $M \cdot B$ *epochs*, SGD partitions the N datapoints into a length-N/B sequence of size-B *batches*. For each batch $\mathcal{B}$, SGD updates $\theta^\mu \leftsquigarrow \theta^\mu - \eta^{\mu\nu} \nabla_\nu \left(\frac{1}{B} \sum_{n \in \mathcal{B}} l_n(\theta) \right)$. SGD performs $N \cdot M$ many updates in total. We write l_t for the loss $\frac{1}{B} \sum_{\mathcal{B}} \cdots$ over the tth batch. Together with an inter-epoch shuffling pattern, N, B, M determine the SGD algorithm.

The cases $B = 1$ and $B = N$ we call *pure SGD* and *pure GD*. The $M = 1$ case of pure SGD we call *vanilla SGD*.

2.4. Diagrams and Embeddings

3. Diagram Calculus for SGD

3.1. How Diagrams Arise

Suppose s is analytic on weight space, for example $s = l$. We track $s(\theta)$ as SGD updates θ:

Key Lemma. *For all T: for η sufficiently small, $s(\theta_T)$ is*

$$\sum_{(d_t : 0 \leq t < T)} (-\eta)^{\sum_t d_t} \left(\prod_{0 \leq t < T} \frac{(g\nabla)^{d_t}}{d_t!} \Bigg|_{g = \nabla l_t(\theta)} \right) (s)(\theta_0) \quad (1)$$

Moreover, the expectation symbol (over train sets) commutes with the sum over ds.

In averaging over train sets we may factor the expectation of the above product according to independence relations between the l_t. We view various training procedures (e.g. pure GD, pure SGD) as *prescribing different independence relations* that lead to different factorizations and hence to potentially different generalization behavior at each order.

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

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

To reduce clutter, we draw contractions as thin edges and datapoint equalities as fuzzy ties[†]. Then, in expectation over (l_c, l_b, l_a) drawn i.i.d.:

$$\cdots = \text{⌢⌢} + \text{⌢⌢} + \text{⌢⌢} + \text{⌢⌢} \quad (2)$$

$$= \quad 2\,\text{⌢⌢} \quad + \quad 2\,\text{⌢⌢} \quad (3)$$

$$\underbrace{}_{-2\ \mathbb{E}[(\nabla l)(\nabla l)]\ \mathbb{E}[\nabla\nabla\nabla l]\ \mathbb{E}[\nabla l]} \qquad \underbrace{}_{-2\ \mathbb{E}[(\nabla l)(\nabla l)]\ \mathbb{E}[\nabla\nabla l]\ \mathbb{E}[\nabla\nabla l]}$$

Each degree-g node corresponds to an l_n (here, l_c, l_b, l_a) differentiated g times (for instance, l_b is differentiated thrice in the first diagram and twice in the second). Thin *edges* mark contractions by $-\eta$. Fuzzy *ties* denote correlations by connecting identical loss functions (here, l_c with l_c), and are visually emphasized by otherwise meaningless colors.

Definition 1. A diagram D evaluates to the expected value value(D) of the corresponding tensor expression.

Usefully, *the topology of a diagram determines its value*. For instance, simplification 3 is licensed because value$\left(\text{⌢⌢} \right)$ = value$\left(\text{⌢⌢} \right)$. We will sometimes write a diagram D and mean value(D).

Definition 2 (Fuzzy Outlines Denote Noise's Net Effect). So, if D is a diagram and $p, \tilde{p}$ are two parts of D's partition, let $D_{p\tilde{p}}$ be the diagram with $p, \tilde{p}$ joined into a single part. So: $(\overset{\frown}{\cdots})_{\text{red green}} \triangleq \bullet\!\!-\!\!\bullet$ and $(\overset{\frown}{\cdots})_{\text{red blue}} \triangleq \overset{\frown}{\bullet\!\!-\!\!\bullet}$. Differences such as $D_{p\tilde{p}} - D$ will often concern us, so we define a diagram with fuzzy *outlines* to represent the difference between the fuzzy tied and untied versions so that $\overset{\frown}{\bullet\!\!-\!\!\bullet} \triangleq (\overset{\frown}{\cdots})_{\text{green blue}} - \overset{\frown}{\cdots\cdot} = \bullet\!\!-\!\!\bullet - \overset{\frown}{\cdots\cdot}$.

3.2. Recipe for SGD's Test Loss and Generalization

Our main tool, proved in Appendix E, follows.

Theorem 1 (Test Loss as a Path Integral). *For all T: for η sufficiently small, SGD's expected test loss is*

$$\sum_{D\in\text{im}(\mathcal{F})} \left(\sum_{f:D\to\mathcal{F}(S)} \frac{1}{|\text{Aut}_f(D)|} \right) \frac{\text{value}(D)}{B^{|\text{edges}(D)|}}$$

Here, D is a diagram whose root r participates in no fuzzy edge, f represents an embedding of D into spacetime, and $|\text{Aut}_f(D)|$ counts the graph-automorphisms of D that preserve f's assignment of nodes to cells. Replacing each value(D) by $\left(-\sum_{p\in\text{parts}(D)} \text{value}(D_{rp} - D)/N \right)$, we obtain the expected generalization gap (test minus train loss).

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

$$\frac{(-1)^d}{d!} \sum_{D\in\text{im}(\mathcal{F})} |\text{ords}(D)| \binom{N}{P-1} \binom{d}{d_0,\cdots,d_{P-1}} \text{value}(D)$$

where D ranges over d-edged diagrams whose equivalence classes have sizes $d_p : 0 \le p \le P$, with $d_P = 1$ and, without loss, are each antichains. The modification to compute the generalization gap is the same as in Theorem 1.

Remark 1 (Vanilla SGD's Deviation from ODE). A P-part, d-edged diagram contributes $\Theta\left((\eta N)^d N^{P-d-1}\right)$ to the loss. For example, there are 6 diagrams to order 3, and they have $(4+2) + (2+2+3) + (1)$ many orderings (see Table 1). Intuitively, ηN measures the physical time of descent and $1/N$ measures the coarseness of time discretization. Thus arises a double-series in $(\eta N)^d N^{P-d-1}$, where d counts thin edges and $d+1-P$ counts fuzzy ties; the $P = d+1$ terms correspond to a noiseless, discretization-agnostic (hence ODE) approximation to SGD, while $P \le d$ gives correction terms modeling time-discretization and noise.

3.3. Consequences of the Recipe

For accessibility, we write our results with Greek indices instead of diagrams. However, as Appendices A, B, and C show, diagrams are crucial for proving — and for identifying the data-weight processes that govern — these results.

Corollary 1 (SGD Differs from ODE and SDE). *The test loss of vanilla SGD deviates at order N^{-1} from ODE by $\frac{T^2 N^{-1}}{2} C_{\mu\nu} H^{\mu\nu}$. Its order N^{-2} deviation due to non-Gaussian noise is $\frac{T^3 N^{-2}}{6} \left(\overset{\frown}{\bullet\!\!-\!\!\bullet} - 3\overset{\frown}{\cdots\cdot} \right) = -\frac{T^3 N^{-2}}{6} \left(\left(\mathbb{E}\left[\nabla_\mu l_x \nabla_\nu l_x \nabla_\lambda l_x \right] - G_\mu G_\nu G_\lambda \right) J^{\mu\nu\lambda} - 3 C_{\mu\nu} G_\lambda J^{\mu\nu\lambda} \right)$. These effects contribute to SGD's difference from SDE.*

For finite N, these effects separate SDE from SGD. SDE also fails to model multi-epoch SGD's inter-update correlations. Conversely, as $N \to \infty$ so that SDE matches SGD, optimization and generalization respectively become computationally intractable and trivial and hence less interesting.

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

Corollary 2 (Shuffling Barely Matters). *To order η^3, inter-epoch shuffling doesn't affect SGD's expected test loss.*

Indeed, for any inter-epoch shuffling scheme:

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

$$l - MN G_\mu G^\mu + MN\left(MN - \frac{1}{2} \right) G_\mu H^\mu_\nu G^\nu$$

$$+ MN\left(\frac{M}{2} \right) C_{\mu\nu} H^{\mu\nu} + MN\left(\frac{M - \frac{1}{B}}{2} \right) \left(\nabla_\mu C^\nu_\nu \right) G^\mu / 2$$

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

Upon analyzing $\overset{\frown}{\bullet\!\!-\!\!\bullet}$: given a smooth unbiased estimator $\hat{C}$ of gradient covariance, we may cause GD to mimic SGD:

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

Corollary 5 (An Entropic Force, Unrenormalized). *When vanilla SGD is initialized at a test minimum, the weight is displaced to order η^3 by $-\binom{T}{2} C_{\mu\nu} J^{\mu\nu\lambda}$.*

Corollary 5 was first proven by Yaida (2019b); our methods reproduce it elegantly. We will renormalize it to Corollary

6, which applies for large times, hence demonstrating a persistent and non-conservative entropic force.

3.4. Effective Theories

The mathematics of *renormalization* summarizes myriad small-scale interactions into an effective large-scale theory. We will renormalize by pairing diagrams of interest with topologically related diagrams in order to refine our dth order estimates for any fixed d. In some large-ηT settings, the unrenormalized theory does not converge while the renormalized theory does. Thus, renormalization will help us reason about large-time behavior.

As a motivating example, consider that, to order η^1, SGD reduces test loss by $TG_\mu G^\mu$ (c.f. Nesterov (2004)). Though useful for small ηT, this approximation diverges as T grows. Intuitively, the this approximation suffers because it does not account for evolving gradients, i.e. for the Hessian. Let us thus consider the diagrams that are uncorrelated chains: •⌐, •⌐•, •⌐⌐•, ⋯. This infinite sequence epitomizes the effect of the Hessian (degree 2 nodes). When embedded with initial and final nodes separated by duration Δt, this series of diagrams sums to $-G(I - \eta H)^{\Delta t - 1}\eta G \approx -G \exp(-\eta \Delta t H)\eta G \cdot (1 + o(\eta))$. The $\exp(-\eta \Delta t H)\eta$ is thus an "effective learning rate" to be used in place of η in renormalized calculations. Summed over all embeddings, the above chains contribute $G \sum_{0 \le t < T}(I - \eta H)^{T - t - 1}\eta G$ to vanilla SGD's test loss, or:

$$\approx -G \left(\int_t \exp(-\eta(T - t)H) \right)\eta G = -G \left(\frac{I - \exp(-\eta T H)}{H} \right)G$$

In contrast to the unrenormalized prediction $TG_\mu G^\mu$, the above converges in the large-T limit (when H is positive).

So the chains form a class of diagrams usefully analyzed together. We may likewise organize other diagrams. We will represent each class by its minimal element:

Definition 3. A diagram is *irreducible* when it lacks nonroot un-fuzzily-tied degree-two nodes (such as •⌐⌐•'s green node). So •⌐ and ⌐• but neither •⌐• nor •⌐• are irreducible.

Definition 4 (Embedding-Sensitive Values). Let $\text{rvalue}_f(D)$ be the expected value of D's corresponding tensor expression, where instead of using $-\eta$ to contract two tensors embedded to times $t, t + \Delta t$, we use $-(I - \eta H)^{\Delta t - 1}\eta$. Actually, as we study the effect of noise, it will be most convenient to let rvalues represent a *difference* from the noiseless case. For the small diagrams we consider, we obtain rvalues by replacing fuzzy ties by fuzzy outlines; larger diagrams present complications addressed in Appendix E.4.

RENORMALIZED RECIPE: In general, one sums over embeddings of irreducible diagrams, using $\text{rvalue}_f(D)$ instead

of value(D). In practice, we prefer to approximate sums over embeddings by integrals over times and $(I - \eta H)^t$ by $\exp(-\eta H t)$, hence incurring a term-by-term multiplicative error of $1 + o(\eta)$ that preserves leading order results. Diagrams thus induce easily evaluated integrals of exponentials.

Theorem 2 (Renormalization Gives Large-T Limits). *For any T: for η sufficiently small, SGD's expected test loss exceeds the noiseless case by*

$$\sum_{\substack{D \in \text{im}(\mathcal{F}) \\ \text{irreducible}}} \left(\sum_{f: D \to \mathcal{F}(S)} \frac{1}{|\text{Aut}_f(D)|} \right) \frac{\text{rvalue}_f(D)}{B^{|\text{edges}(D)|}}$$

In contrast to Theorem 1: when H is positive, the dth order truncation converges as T diverges.

That convergence property formally licenses long-term limits of renormalized predictions. In fact, the convergence is uniform in T for sufficiently regular landscapes.
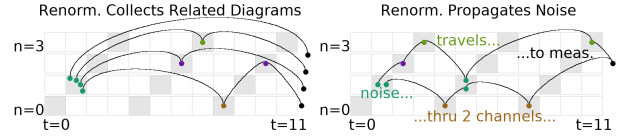

Figure 3. **Renormalization summarizes Hessians' long-term effect. Left**: Some contributions to rvalue$_f$(•⌐•) for an example embedding f. Theorem 2 shows how (geometric series) and why (improved convergence) we consider these diagrams as a group. **Right**: A contribution to rvalue$_f$(⌐⌐•). The leftmost teal nodes depict a disturbance by noise (multi-node cells or rows correspond higher moments and hence to noise). The noise travels through two channels (top, bottom) before interacting with the test Hessian (degree-two black node). Such embeddings give the small-η, large-T interaction of noise and curvature. **Both**: In violation of our rule for valid embeddings, the nodes inserted into the irreducible diagrams may occupy the same row despite lacking fuzzy ties. Overcounting accompanies our $(I - \eta H)^{\Delta t - 1}$ prescription; it is countered as described Appendix E.4 and implicit in Theorem 2.

Integrating rvalue$_f$(⌐⌐•) over embeddings f, we see:

Corollary 6 (A Nonconservative Entropic Force). *Initialized at a test minimum, vanilla SGD's weight moves to order η^2 with a long-time-averaged[*] expected velocity of*

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

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

Unlike Wei & Schwab (2019), we make no assumptions of thermal equilibrium, fast-slow mode separation, or constant covariance. This generality reveals a novel, empirically

---

[*] That is, T so large that $C \exp(-\eta K T)$ is negligible. Appendix C gives a similar expression for general T.

verified dynamical phenomenon, namely that the velocity field above need not be conservative.

Integrating rvalue$_f$(⬭) and rvalue$_f$(⬤) yields:

Corollary 7 (Flat, Sharp Minima Overfit Less). *Initialized at a test minimum, pure GD's test loss is to order η*

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

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

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

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

Corollary 7'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, in the classical setting of maximum likelihood (ML) estimation (with no model misspecification) near the "true" test 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 makes no model well-specification assumptions.
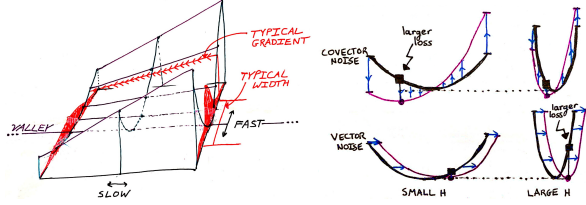

Figure 4. **Renormalization reveals novel phenomena with physical interpretations. Left**: The entropic force mechanism: gradient noise induces a flow toward minima flat *with respect to to the covariance*. Although our analysis assumes neither thermal equilibrium nor fast-slow mode separation, we label "fast and slow directions" in this figure to ease comparison with Wei & Schwab (2019). In this figure, red densities denote the spread predicted by a renormalized $C^{\mu\nu}$, and the spatial variation of curvature corresponds to $J_{\mu\nu\lambda}$. **Right**: Noise structure determines how curvature affects overfitting. To use a physical metaphor, for a fixed displacement scale, stiffer springs incur greater energy gains. But for a fixed force scale, limper springs incur greater energy gains. Geometrically, for (empirical risk minimization on) a vector-perturbed landscape, small Hessians are favored (top row), while for covector-perturbed, landscape large Hessians are favored (bottom row). Corollary 7 shows how the implicit regularization of fixed-ηT descent mediates between the two rows.

4. 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 6, (Chaudhari & 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 G for experimental procedure including architectures and sample size.

4.1. Basic Predictions

We test Proposition 1 by direct measurement of smooth convnets on CIFAR-10 and Fashion-MNIST. Our order η^3 perturbative predictions up to $\eta T \approx 10^0$ (Figure 5, left). Likewise, Corollary 3 predictions about epochs agree with experiment (Appendix H). These tests verify that the Proposition hides no mistakes of proportionality or sign.

4.2. Emulating Small Batches with Large Ones

Figure 5 (right) shows that the regularizer proposed in Corollary 4 indeed enables GD to emulate SGD on a range of image-classification landscapes. For 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 architectures composed of linear transforms and coordinate-wise nonlinearities represented by elementary functions, we may compute $\nabla \hat{C}$ with the same memory and time as the usual gradient ∇l_t, up to a multiplicative constant.
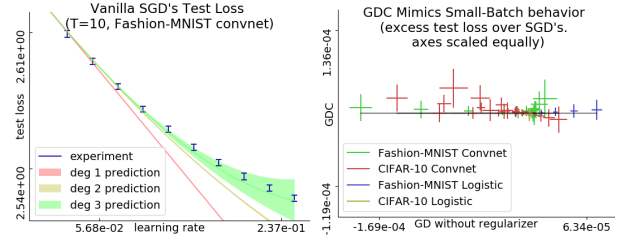

Figure 5. **Perturbation models SGD and GD. Left**: Test loss vs learning rate on a Fashion-MNIST convnet. For the instance shown and all 11 other initializations unshown, the degree-3 prediction agrees with experiment through $\eta T \approx 10^0$. **Right**: with equal-scaled axes, this plot shows that GDC matches SGD (small vertical variation) better than GD matches SGD (large horizontal variation) in test loss, for a variety of learning rates (from 0.0025 to 0.1) and initializations (zero and multiple independent Xavier-Glorot trials) on several of image classification landscapes (logistic and convolutional CIFAR-10 and Fashion-MNIST). $T = 10$ throughout.

4.3. Comparison to Continuous Time

Consider fitting a centered normal $\mathcal{N}(0, \sigma^2)$ to some centered standard normal data. 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 σ 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. At $h = 0$, the expected gradient vanishes, and the test loss of vanilla SGD only involves diagrams with no singleton leaves; to third order, it is $\cdot + \frac{T}{2} \overset{\frown}{\smile} + \binom{T}{2} \overset{\frown}{\smile} + \frac{T}{6} \overset{\frown}{\smile}$ In particular, the $\binom{T}{2}$ differs from $T^2/2$ and hence contributes to the time-discretization error of SDE as an approximation for SDE. Moreover, non-Gaussian noise contributes via $\overset{\frown}{\smile}$ to that error. Appendix H shows that SDE and one-epoch SGD indeed differ. For multi-epoch SGD, the effect of overfitting to finite training data further separates SDE and SGD.

4.4. A Nonconservative Entropic Force

To test Corollary 6'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 & 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.

Intuitively, the presence of the term $\overset{\frown}{\smile}$ in our test loss expansion indicates that *SGD descends on a covariance-smoothed landscape*. 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 $S(z)(w) = \cos(z - \pi/4)w_x + \sin(z - \pi/4)w_y$. We see that 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 renormalized theory predicts a z-speed of $\eta^2/6$ per timestep. Upon renormalization, our predictions thus agree for large ηT with experiment (Figure 6, left).

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

4.5. 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 7 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 4, right).
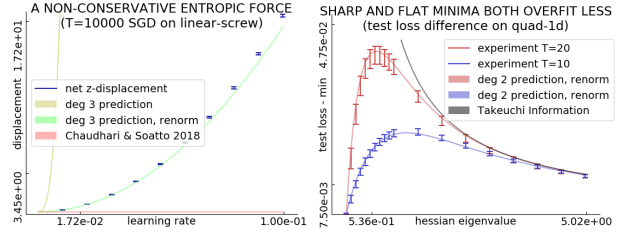

Figure 6. **Renormalized predictions excel even for large ηT. Left**: On Linear Screw, the persistent entropic force pushes the weight through a valley of global minima not at a $T^{1/2}$ diffusive rate but at a directional T^1 rate. Since Hessians and covariances are bounded throughout the valley and 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 , it gives insensible results for over-parameterized models. Indeed, Dixon & Ward (2018) reports numerical difficulties requiring an arbitrary cutoff.

Fortunately, by Corollary 7, the implicit regularization of gradient descent both demands and enables a singularity-removing correction to the TIC (Figure 6, right). The resulting *Stabilized TIC* (STIC) uses the metric implicit in gradient descent to threshold flat from sharp minima. 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 G, plot in Appendix H) recommend STIC for model selection when C/N dwarves H as in the noisy, small-N regime.[*] Because diagonalization typically takes time cubic in dimension, STIC regularization without further approximations is

---

[*]In fact, this analysis generalizes James-Stein style shrinkage estimators by agreeing in the quadratic Gaussian setting.

most useful for small models on noisy and limited data.

5. Related Work

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

Several lines of work quantify the overfitting of SGD-trained networks (Neyshabur et al., 2017a). For instance, Bartlett et al. (2017) controls the Rademacher complexity of deep hypothesis classes, leading to generalization bounds that are optimizer-agnostic. However, since SGD-trained networks generalize despite their seeming ability to shatter large sets (Zhang et al., 2017), one infers that generalization arises from the aptness to data of not only architecture but also optimization (Neyshabur et al., 2017b). Others have focused on the implicit regularization of SGD itself, for instance by modeling descent via stochastic differential equations (SDEs) (e.g. Chaudhari & Soatto (2018)). However, per Yaida (2019a), 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 Jastrzębski et al. (2018).

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 (Liao et al., 2018), by exhibiting generalization of local minima *flat* with respect to the standard metric (see Hoffer et al. (2017), Keskar et al. (2017), Wang et al. (2018)), or by exhibiting generalization of local minima *sharp* with respect to the standard metric (see Stein (1956), Dinh et al. (2017), Wu et al. (2018)). Our work reconciles those seemingly clashing claims.

Others have perturbatively analyzed descent: Dyer & Gur-Ari (2019) perturb in inverse network width, employing Feynman-'t Hooft diagrams to correct the Gaussian Process approximation for a specific class of deep networks. Mean-

while, (Chaudhari & Soatto, 2018) and Li et al. (2017) perturb in learning rate to second order by approximating noise between updates as Gaussian and uncorrelated. That approach does not generalize to higher orders, and, in neglecting correlations and heavy tails, ignores the generalization behavior of SGD. 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)[*], it applies to inexact-likelihood landscapes such as VAEs'.

6. Conclusion

We presented an elegant diagram-based framework for understanding small-ηT SGD. In fact, our Renormalization Theorem licences predictions for large ηT, beyond the reach of direct perturbation. The theory makes novel predictions:

STORY OF ⌣: 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 and to tune hyperparameters by gradient descent.

STORY OF ⌣: Refining Wei & Schwab (2019) to nonconstant, nonisotropic covariance, we find that SGD descends on a loss landscape smoothed by the *current* covariance. As the covariance evolves, the smoothing mask and thus the effective landscape evolves. This dynamics is generically nonconservative. In contrast to Chaudhari & Soatto (2018)'s SDE approximation, SGD does not generically converge to a limit cycle.

STORY OF ⌣: As conjectured by Roberts (2018), 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).

6.1. Limitations

As expected for truncated Taylor series, our perturbative predictions of final test loss based on initial data dis-

---

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

agree with experiment for large ηT. Moreover, our predictions often involve many-indexed tensors (such as J or $(I \otimes I)/(\eta H \otimes I + I \otimes \eta H)$) inconvenient to compute in real-world settings.

That said, we have demonstrated that our predictions agree with experiment for ηT large enough that continuous time approximations fail. That is, we predict force laws different from those of continuous-time approximations. For example, we supplement the intuition that "SGD performs steepest descent on the train landscape" by showing that, even at a train loss minimum, SGD drifts in a specific direction so as to dis-align H from the current C. To the practitioner, our contributions may thus play the same role as that intuition of steepest descent.

6.2. Questions

The diagram method opens the door to exploration of Lagrangian formalisms and curved backgrounds[*]:

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 $\theta \leftsquigarrow \theta - (\eta^{-1} + \lambda \nabla \nabla l_t(\theta))^{-1} \nabla l_t(\theta)$ 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 computations, 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 spaces[†]. 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 & Kiela (2017)'s *hyperbolic embeddings*. We are optimistic our formalism may resolve conjectures such as above.

6.3. Acknowledgements

References

Absil, P.-A., Mahony, R., and Sepulchre, R. Optimization algorithms on matrix manifolds, chapter 4. *Princeton University Press*, 2007.

Amari, S.-I. Natural gradient works efficiently. *Neural Computation*, 1998.

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

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

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

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

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

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

Dixon, M. and Ward, T. Takeuchi's information criteria as a form of regularization. *Arxiv Preprint*, 2018.

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

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

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

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

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

---

[*] Landau and Lifshitz introduce these concepts with grace and brevity (1960; 1951).

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

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

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

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

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

Krizhevsky, A. Learning multiple layers of features from tiny images. *UToronto Thesis*, 2009.

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

Landau, L. and Lifshitz, E. The classical theory of fields. *Addison-Wesley*, 1951.

Landau, L. and Lifshitz, E. Mechanics. *Pergamon Press*, 1960.

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

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

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

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

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

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

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. *ICML*, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

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

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

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

Rota, G.-C. Theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheoriei und Verwandte Gebiete*, 1964.

Roux, N., Bengio, Y., and Fitzgibbon, A. Improving first and second-order methods by modeling uncertainty. *Book Chapter: Optimization for Machine Learning, Chapter 15*, 2012.

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

Wang, H., Keskar, N., Xiong, C., and Socher, R. Identifying generalization properties in neural networks. *Arxiv Preprint*, 2018.

Wei, M. and Schwab, D. How noise affects the hessian spectrum in overparameterized neural networks. *Arxiv Preprint*, 2019.

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

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

Xiao, H., Rasul, L., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *Arxiv Preprint*, 2017.

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

Yaida, S. A first law of thermodynamics for stochastic gradient descent. *Personal Communication*, 2019b.

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

Zhang, H., Reddi, S., and Sra, S. Fast stochastic optimization on riemannian manifolds. *NeurIPS*, 2016.

Invitation to Appendices

The first three appendices deal with CALCULATION and contain the proofs — some heavily annotated for pedagogical purposes — of all the corollaries in the text. APPENDIX A guides the reader through a hands-on tutorial of the diagram method, including renormalization. It concludes with a simple project for the reader, the solution of which is a novel result easily obtained by the diagram method but not present in the literature. APPENDIX B demonstrates how and why the diagram method excels over direct application of the Key Lemma, even if we desire only unrenormalized results. APPENDIX C provides the remaining computations promised.

The next two appendices offer the foundational PROOFS that license the computations and corollaries treated in the preceding appendices. APPENDIX D sets up the objects of study with full mathematical precision. APPENDIX E, after cataloging the regularity conditions required of loss landscapes (and implicitly assumed as a hypothesis of all results herein), proves Theorems 1 and 2, along the way proving the Key Lemma.

The next three appendices further describe EXPERIMENTS, methods, and results. APPENDIX F explains how we obtained loss landscape statistics such as expected entropic force for real datasets. APPENDIX G details our training, testing, and architecture for the datasets (e.g. CIFAR-10) that we used in experiments. APPENDIX H exhibits additional plots that could (but fail to) falsify our theory.

The final appendix is for convenient REFERENCE. APPENDIX I glosses physical and mathematical terminology and tabulates the values and interpretations for some common diagrams.

A. Tutorial on Diagram Rules

After reviewing the diagram method's recipe step by step, we will work through sample problems. As in the main text, we relegate precise combinatorial definitions to Section D, preferring in this section to appeal to intuition and to hands-on examples.

Recall the computational recipe. To compute a quantity of interest, list all the relevant diagrams, then evaluate each diagram over all of its embeddings into spacetime. After reviewing the intuitive interpretation and combinatorial character of diagrams, we illustrate each step of this process. We finish with example computations and an easy project.

A.1. Anatomy of a Diagram

THE DATA OF A DIAGRAM

A diagram has thin edges and fuzzy ties. The thin edges must form a rooted tree. When drawing diagrams on paper,

our convention is to record the root by drawing it to the right of all other nodes. The fuzzy edges are just a mechanism to represent a partition of the diagram's nodes. For a diagram with $d + 1$ nodes and p parts, one can specify the parts using $d + 1 - p$ many fuzzy ties and no fewer. Intuitively: thin edges represent differentiation and fuzzy ties represent noise.

The following two diagrams are equivalent, not only in that they evaluate to the same numbers but in that their combinatorial data is the same: , . Only a redundancy of our ink description makes them look different. The following two diagrams are inequivalent but evaluate to the same numbers in an un-renormalized setting: , . We recognize them as inequivalent when we observe that their roots (here blue) play different roles. The diagrams thus depict different data-weight processes. In a renormalized setting, they will evaluate to different numbers.

We define a diagram with fuzzy *outlines* to represent the difference between the fuzzy tied and untied versions so that $\triangleq$ − .

EVALUATING A DIAGRAM

The thin edges represent contraction by $-\eta$, the fuzzy ties represent correlations, and each degree-g node represents a dth derivative. Intuitively, thin edges represent differentiation and fuzzy ties represent noise.

Question. *What does* *evaluate to?*

Answer. Each top, upside-down U represents an $-\eta^{\mu\nu}$, and the two red fuzzily-tied sprouts together represent a $C_{\mu\nu} + G_\mu G_\nu$. The green vertex has three edges so it is the third derivative $J_{\mu\nu\eta}$. The blue vertex, having one edge, is just a first derivative, G_μ.

We read the thin edges carefully to see how to compose these tensors together: So: $-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(C_{\mu\lambda} + G_\mu G_\lambda)J_{\nu\rho\sigma}G_\pi$. This happens to give the entropic force near a test minimum.

Question. *What does* *evaluate to?*

Answer. The top, upside-down U represents an $-\eta^{\mu\nu}$, and the bottom two fuzzily-outlined sprouts represent a $C_{\mu\nu}$. So: $-\eta^{\mu\nu}C_{\mu\nu}$. This is (minus) the trace of the covariance.

A.2. Listing Relevant Diagrams

REVIEW OF RULES

A diagram with d thin edges evaluates to an order-η^d expression. Thus, if we wish the order η^d contribution to a quantity, we know immediately that we will only need to consider diagrams with d thin edges.

Depending on what we want to compute, we can rule out

more diagrams. The type of quantity (test loss / test loss minus train loss), the optimization schedule (one-epoch / effect of batch size comparison between two optimizers), and special posited structure (initialize at minimum / no noise / Gaussian noise, quadratic loss) can all rule out diagrams.

Indeed, Theorem 1 says that the *test loss* involves only diagrams whose root (a.k.a. rightmost node) participates in no fuzzy ties. And it says that the *generalization gap* (ie. test minus train loss) involves only diagrams whose root participates in a fuzzy outline with one other node.

Remark (Role of Fuzzy Outlines)**.** As originally defined, diagrams have thin edges and fuzzy ties but not fuzzy outlines. But because taking train losses minus test losses is such a common operation, we have introduced the fuzzy outline notation to mean a difference of diagrams, one diagram with the fuzzy outline replaced by a fuzzy tie, and the other diagram with the fuzzy outline deleted. For example, [diagram] is shorthand for [diagram] − [diagram], which is a (summand in the) train loss minus (a corresponding summand in the) test loss.

Moreover, because diagrams must ultimately embed into spacetime in order to contribute, we might rule out ill-fitting diagrams if we know about our specific spacetime. As a simplest example, one-epoch SGD has only one spacetime cell per row and thus does not permit diagrams such as [diagram] to be embedded because such embeddings would have intra-cell edges. Thus, for one-epoch SGD, we may ignore diagrams whose thin-edge trees have a fuzzily tied ancestor-descendant pair. As another example of this type of simplification: the spacetimes of any two (N, B, M)-style SGD runs, with same N, M for fair comparison but different B, correspond in many-to-one fashion so that no diagram embedded with each node in a different epoch contributes to the difference between GD and SGD.

Finally, knowledge of loss landscape structure can further rule out diagrams. As a simplest example, it is often interesting to initialize at a test minimum to then measure subsequent overfitting. At a test minimum, the expected gradient G vanishes, and hence every diagram with a factor of G vanishes. These are the diagrams with a leaf node that is not fuzzily tied. So, at a test minimum, we may ignore diagrams such as [diagram] and [diagram]. As a further example of this type of simplification: in the noiseless case where covariances are 0, we may ignore every diagram with size-2 fuzzy-outline-cliques, e.g. [diagram].

EXAMPLES OF LISTING RELEVANT DIAGRAMS

Question. *Which diagrams contribute to general SGD's generalization gap (i.e. test minus train loss) at order η^1?*

**Answer.** The d-edged diagrams contribute to order eta^d,

so we wish to enumerate the diagrams with one edge. Per Theorem 1, the generalization loss is a sum over diagrams whose root (a.k.a. rightmost node) participates in one fuzzy outline. So we seek 1-edged diagrams whose rightmost node is fuzzily outlinedly tied to one other node. It turns out there is just one such diagram: [diagram]. Indeed, there is only one possible rooted tree of thin edges. And since there are only two nodes, the rightmost node's fuzzy-outline partner is determined.

Remark (Fuzzy Outlines as Shorthand)**.** [diagram] is shorthand for [diagram] − [diagram], which is a (summand in the) train loss minus (a corresponding summand in the) test loss.

**Question.** *Which diagrams contribute to general SGD's test loss at order η^2?*

Answer. The d-edged diagrams contribute to order eta^d, so we wish to enumerate the diagrams with two edges. Per Theorem 1, the test loss is a sum over diagrams whose root (a.k.a. rightmost node) participates in no fuzzy ties. So we seek 2-edged diagrams whose rightmost node isn't fuzzily tied. It turns out there are four such diagrams: [diagram], [diagram], [diagram], and [diagram]. Indeed, two permissible rooted trees multiply with two permissible fuzzy tie patterns to yield four diagrams.

Question. *Which diagrams contribute to one-epoch SGD's test loss at order η^2?*

Answer. We seek 2-edged diagrams whose rightmost node isn't fuzzily tied. As we saw before, there are four such diagrams. However, consideration of one-epoch, singleton-batch SGD's spacetime shows us that no node in the thin-edge tree may be fuzzily tied to any of its descendents. Thus, of the four possible diagrams (namely, [diagram]), one has no embeddings, hence has coefficient zero, hence may be ignored. Explictly, the remaining diagrams are: [diagram], [diagram], and [diagram].

Question. *Which diagrams contribute to one-epoch SGD's generalization gap at order η^2?*

Answer. We seek 2-edged diagrams (order η^2) whose rightmost node has a fuzzy outline connecting to one other node (generalization gap). We ignore trees whose fuzzy-outline-erased versions involve fuzzy ties between any pair of ancestor and descendent (one-epoch). To be explicit, we write out the fuzzy outlines as the differences they represent: [diagram] − [diagram], [diagram] − [diagram], and [diagram] − [diagram].

A.3. Summing a Diagram's Spacetime Embeddings

UNRENORMALIZED TECHNIQUE

The unrenormalized technique of Theorem 1 is simpler than the renormalized technique of Theorem 2, but it only has guarantees for small ηT.

We may *embed* a diagram into spacetime by placing each *non-root* node in a shaded cell. We allow only embeddings that obey both:

thin edge condition have a node c in a column strictly prior to a node p if c's path to the root encounters the root.

fuzzy tie condition have two nodes in the same row if and only if they are fuzzily tied.

The un-renormalized contribution of a diagram D to a loss is a sum over all of its embeddings — each inversely weighted by the number of ways to permute the nodes of the diagram such that nodes stay within the original cells assigned by the embedding and such that each pair of nodes connected by thin edges map by the permutation to a pair of nodes connected by thin edges — of

un-renormalized value : D), that is, the expectation over the loss landscape's randomness of the the $-\eta$-contracted tensor expression to which the diagram corresponds.

RENORMALIZED TECHNIQUE

Though the unrenormalized technique of Theorem 1 is simple, the renormalized technique of Theorem 2 gives convergent results even for large ηT.

We may *embed* a diagram into spacetime by placing each *non-root* node in a shaded cell. We allow only embeddings that obey both:

thin edge condition have a node c in a column strictly prior to a node p if c's path to the root encounters the root.

fuzzy tie condition have two nodes in the same row if and only if they are fuzzily tied.

In the renormalized theory, we only consider irreducible diagrams, that is, diagrams with the property that every thin-edge-degree two node participates in some fuzzy edges.

The renormalized contribution of a diagram D to a loss is a sum over all of its embeddings f — each inversely weighted by the number of ways to permute the nodes of the diagram such that nodes stay within the original cells assigned by the embedding and such that each pair of nodes connected by thin edges map by the permutation to a pair of nodes connected by thin edges, of:

renormalized value rvalue$_f(D)$, that is, the expectation over the loss landscape's randomness of the the

$-\exp(-\delta T\eta H)\eta$-contracted tensor expression to which the diagram corresponds. (minus the noiseless baseline as motivated in Definition 4 and described in Appendix E.4; this helps us calculate the net effect of noise by calculating the loss with noise minus the loss without noise).

EXAMPLES OF SUMMING A DIAGRAM'S SPACETIME EMBEDDINGS

Consider Figure 7.
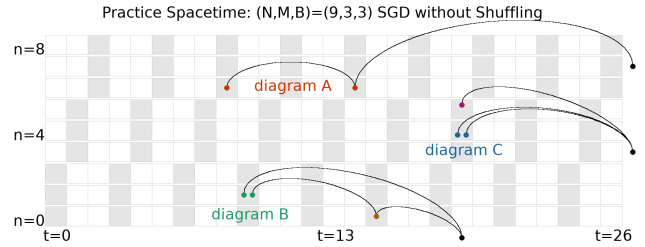
Figure 7. Sample embeddings — one for each of , ,

and — in the spacetime of $(9, 3, 3)$ SGD.

Question. *How many automorphisms does the depicted embedding of diagram A enjoy?*

Answer. One (the identity), since each node is in a different cell. This induces a coefficient $1/1$.

Question. *How many ways does diagram A embed into the spacetime?*

Answer. As many as there are pairs of distinct same-row cells. There are $N = 9$ possible rows and $\binom{BM}{2} = \binom{9}{2}$ possible pairs of epochs. So: $N\binom{BM}{2} = 9\binom{9}{2}$.

Question. *How many automorphisms does the depicted embedding of diagram B enjoy? This induces a coefficient $1/1$.*

Answer. One (the identity), since the two cell-sharing nodes are distinguished by their thin-edge structure and hence may not be interchanged.

Question. *How many ways does diagram B embed into the spacetime?*

Answer. As many as there are pairs of different-column cells (for the connected teal and bronze nodes), times the number of epochs (for the other teal node). So: $B^2\binom{MN}{2}MB$.

Question. *How many automorphisms does the depicted embedding of diagram C enjoy? This induces a coefficient $1/2$.*

Answer. Two: we may switch or not switch the two cell-sharing nodes.

Question. *What is the total coefficient on diagram C after summing over its embeddings?*

Answer. The blue nodes occupy one of N rows and one of $(MB)^2/2$ sets of columns, and the magenta node occupies any one of the $MB(N-1)$ cells in the remaining row. So: $(N(MB)^2/2) \cdot (MB(N-1))$.

Remark. Most of these embeddings of diagram C will have automorphism group of size 1, because the same-row nodes will nevertheless by in different cells. Those will be weighted with coefficient $1/1$. The rest will be weighted with the $1/2$ we found above. This permits us to "count" a noninteger number such as $N(MB)^2/2$ of embeddings: the "count" is shorthand for this weighted sum.

To illustrate renormalization, we prove Corollary 7:

A.4. Example: Overfitting and Generalization of GD. Proof of Corollary 7.

OVERFITTING

We work in the setting of Corollary 7 and aim to prove its first part. We apply Theorem 2.

The relevant irreducible diagram is ⬳ (which equals

⬳ because we are at a test minimum). This diagram has one embedding for each pair of same-row shaded cells, potentially identical, in spacetime; for GD, the spacetime has every cell shaded, so each *non-decreasing* pair of durations in $[0, T]^2$ is represented; the symmetry factor for the case where the cells is identical is $1/2$, so we lose no precision by interpreting a automorphism-weighted sum over the *non-decreasing* pairs as half of a sum over all pairs. Each of these may embed into N many rows, hence the factor below of N. The two integration variables (say, $t, \tilde{t}$) separate, and we have:

$$\frac{N}{B^{\text{degree}}} \frac{C_{\mu\nu}}{2} \int_t (\exp(-t\eta H))^\mu_\lambda \int_{\tilde{t}} (\exp(-\tilde{t}\eta H))^\nu_\rho \, \eta^{\lambda\sigma} \eta^{\rho\pi} H_{\sigma\pi}$$

Since for GD we have $N = B$ and we are working to degree 2, the prefactor is $1/N$. Since $\int_t \exp(at) = (I - \exp(-aT))/a$, the desired result follows.

GENERALIZATION

We work in the setting of Corollary 7 and aim to prove its second part. We apply the generalization gap modification described in Theorem 1 to Theorem 2's result about test losses; this is licensed, as seen by inspecting their proofs.

The relevant irreducible diagram is ⬳. This diagram has one embedding for each shaded cell of spacetime; for GD, the spacetime has every cell shaded, so each duration from 0 to T is represented. So the generalization gap is, to leading order,

$$+\frac{C_{\mu\nu}}{N} \int_t (\exp(-t\eta H))^\mu_\lambda \, \eta^{\lambda\nu}$$

Here, the minus sign from the gen-gap modification canceled with the minus sign from the odd power of $-\eta$. Integration finishes the proof.

A.5. Project: Non-Gaussian Noise for Large Times.

One may obtain a novel and unpublished result by answering the following questions.

Remark (Gaussian Third Moments)**.** Recall that Gaussian processes fit arbitrary first and second moments, but that their third moments are determined by

$$\langle x^3 \rangle = 3 \langle x^2 \rangle \langle x \rangle - 2 \langle x \rangle^3$$

Question. *Which diagrams that contribute to one-epoch SGD's test loss at order η^3 detect non-Gaussian noise?*

Question. *What is the renormalized contribution of* ⬳

as embedded in the spacetime of one-epoch, singleton-batch SGD?

Question. *What is the leading order contribution of non-Gaussian noise to vanilla SGD's test loss, for large ηT?*

B. Diagram Rules vs Direct Perturbation

Diagram methods from Stueckelberg to Peierls have flourished in physics because they enable swift computations and offer immediate intuition that would otherwise require laborious algebraic manipulation. We demonstrate how our diagram formalism likewise streamlines analysis of descent by comparing direct perturbation[*] to the new formalism on two sample problems.

Aiming for a conservative comparison of derivation ergonomics, we lean toward explicit routine when using diagrams and allow ourselves to use clever and lucky simplifications when doing direct perturbation. For example, while solving the first sample problem by direct perturbation, we structure the SGD and GD computations so that the coefficients (that in both the SGD and GD cases are) called $a(T)$ manifestly agree in their first and second moments. This allows us to save some lines of argument.

Despite these efforts, the diagram method yields arguments about *four times shorter* — and strikingly more conceptual — than direct perturbation yields. These examples specifically suggest that: diagrams obviate the need for meticulous index-tracking, from the start focus one's attention on non-cancelling terms by making visually obvious which terms will eventually cancel, and allow immediate exploitation of a setting's special posited structure, for instance that we are initialized at a test minimum or that the batch size is 1. We regard these examples as evidence that diagrams offer a practical tool for the theorist.

---

[*] By "direct perturbation", we mean direct application of our Key Lemma.

We make no attempt to compare the renormalized version of our formalism to direct perturbation because the algebraic manipulations involved for the latter are too complicated to carry out.

B.1. Effect of Batch Size. Proof of Corollary 4.

We compare the test losses of pure SGD and pure GD. Because pure SGD and pure GD differ in how samples are correlated, their test loss difference involves a covariance and hence occurs at order η^2.

DIAGRAM METHOD

Since SGD and GD agree on noiseless landscapes, we consider only diagrams with fuzzy ties. Since we are working to second order, we consider only two-edged diagrams. There are only two such diagrams, ⟨diagram⟩ and ⟨diagram⟩. The first diagram, ⟨diagram⟩, embeds in GD's space time in N^2 as many ways as it embeds in SGD's spacetime, due to horizontal shifts. Likewise, there are N^2 times as many embeddings of ⟨diagram⟩ in distinct epochs of GD's spacetime as there are in distinct epochs of SGD's spacetime. However, each same-epoch embedding of ⟨diagram⟩ within any one epoch of GD's spacetime corresponds by vertical shifts to an embedding of ⟨diagram⟩ in SGD. There are $MN\binom{N}{2}$ many such embeddings in GD's spacetime, so GD's test loss exceeds SGD's by $\frac{MN\binom{N}{2}}{N^2}$ ⟨diagram⟩. Reading the diagram's value from its graph structure, we unpack that expression as:

$$\eta^2 \frac{M(N-1)}{4} G\nabla C$$

DIRECT PERTURBATION

We compute the displacement $\theta_T - \theta_0$ to order η^2 for pure SGD and separately for pure GD. Expanding $\theta_t \in \theta_0 + \eta a(t) + \eta^2 b(t) + o(\eta^2)$, we find:

$$\theta_{t+1} = \theta_t - \eta\nabla l_{n_t}(\theta_t)$$
$$\in \theta_0 + \eta a(t) + \eta^2 b(t) - \eta(\nabla l_{n_t} + \eta\nabla^2 l_{n_t}a(t)) + o(\eta^2)$$
$$= \theta_0 + \eta(a(t) - \nabla l_{n_t}) + \eta^2(b(t) - \nabla^2 l_{n_t}a(t)) + o(\eta^2)$$

To save space, we write l_{n_t} for $l_{n_t}(\theta_0)$. It's enough to solve the recurrence $a(t+1) = a(t) - \nabla l_{n_t}$ and $b(t+1) = b(t) - \nabla^2 l_{n_t}a(t)$. Since $a(0), b(0)$ vanish, we have $a(t) = -\sum_{0 \le t < T} \nabla l_{n_t}$ and

$b(t) = \sum_{0 \le t_0 < t_1 < T} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}$. We now expand l:

$$l(\theta_T) \in l + (\nabla l)(\eta a(T) + \eta^2 b(T))$$
$$+ \frac{1}{2}(\nabla^2 l)(\eta a(T) + \eta^2 b(T))^2 + o(\eta^2)$$
$$= l + \eta((\nabla l)a(T)) + \eta^2((\nabla l)b(T) + \frac{1}{2}(\nabla^2 l)a(T)^2) + o(\eta^2)$$

Then $\mathbb{E}[a(T)] = -MN(\nabla l)$ and, since the N many singleton batches in each of M many epochs are pairwise independent,

$$\mathbb{E}\left[(a(T))^2\right] = \sum_{0 \le t < T} \sum_{0 \le s < T} \nabla l_{n_t} \nabla l_{n_s}$$
$$= M^2 N(N-1)\mathbb{E}[\nabla l]^2 + M^2 N \mathbb{E}\left[(\nabla l)^2\right]$$

Likewise,

$$\mathbb{E}[b(T)] = \sum_{0 \le t_0 < t_1 < T} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}$$
$$= \frac{M^2 N(N-1)}{2}\mathbb{E}\left[\nabla^2 l\right]\mathbb{E}[\nabla l] +$$
$$\frac{M(M-1)N}{2}\mathbb{E}\left[(\nabla^2 l)(\nabla l)\right]$$

Similarly, for pure GD, we may demand that a, b obey recurrence relations $a(t+1) = a(t) - \sum_n \nabla l_n/N$ and $b(t+1) = b(t) - \sum_n \nabla^2 l_n a(t)/N$, meaning that $a(t) = -t\sum_n \nabla l_n/N$ and $b(t) = \binom{t}{2}\sum_{n_0}\sum_{n_1}\nabla^2 l_{n_0}\nabla l_{n_1}/N^2$. So $\mathbb{E}[a(T)] = -MN(\nabla l)$ and

$$\mathbb{E}\left[(a(T))^2\right] = M^2 \sum_{n_0}\sum_{n_1} \nabla l_{n_0}\nabla l_{n_1}$$
$$= M^2 N(N-1)\mathbb{E}[\nabla l]^2 + M^2 N\mathbb{E}\left[(\nabla l)^2\right]$$

and

$$\mathbb{E}[b(T)] = \binom{MN}{2}\frac{1}{N^2}\sum_{n_0}\sum_{n_1}\nabla^2 l_{n_0}\nabla l_{n_1}$$
$$= \frac{M(MN-1)(N-1)}{2}\mathbb{E}\left[\nabla^2 l\right]\mathbb{E}[\nabla l] +$$
$$\frac{M(MN-1)}{2}\mathbb{E}\left[(\nabla^2 l)(\nabla l)\right]$$

We see that the expectations for a and a^2 agree between pure SGD and pure GD. So only b contributes. We conclude that pure GD's test loss exceeds pure SGD's by

$$\eta^2\left(\frac{M(MN-1)(N-1)}{2} - \frac{M^2 N(N-1)}{2}\right)\mathbb{E}\left[\nabla^2 l\right]\mathbb{E}[\nabla l]^2$$
$$+ \eta^2\left(\frac{M(MN-1)N}{2} - \frac{M(M-1)N}{2}\right)\mathbb{E}\left[(\nabla^2 l)(\nabla l)\right]\mathbb{E}[\nabla l]$$
$$= \eta^2\frac{M(N-1)}{2}\mathbb{E}[\nabla l]\left(\mathbb{E}\left[(\nabla^2 l)(\nabla l)\right] - \mathbb{E}\left[\nabla^2 l\right]\mathbb{E}[\nabla l]\right)$$

Since $(\nabla^2 l)(\nabla l) = \nabla((\nabla l)^2)/2$, we can summarize this difference as

$$\eta^2 \frac{M(N-1)}{4} G\nabla C$$

B.2. Effect of Non-Gaussian Noise at a Minimum. Proof of Corollary 1's Second Part.

The rest of the proof is in Appendix A.

We consider vanilla SGD initialized at a local minimum of the test loss. One expects θ to diffuse around that minimum according to gradient noise. We compute the effect on test loss of non-Gaussian diffusion. Specifically, we compare SGD test loss on the loss landscape to SGD test loss on a different loss landscape defined as a Gaussian process whose every covariance agrees with the original landscape's. We work to order η^3 because at lower orders, the Gaussian landscapes will by construction match their non-Gaussian counterparts.

DIAGRAM METHOD

Because $\mathbb{E}[\nabla l]$ vanishes at initialization, all diagrams with a degree-one vertex that is a singleton vanish. Because we work at order η^3, we consider 3-edged diagrams. Finally, because all first and second moments match between the two landscapes, we consider only diagrams with at least one partition of size at least 3. The only such test diagram is

⬤⬤⬤. This embeds in T ways (one for each spacetime cell of vanilla SGD) and has symmetry factor $1/3!$ for a total of

$$\frac{T\eta^3}{6} \mathbb{E}[\nabla^3 l] \mathbb{E}\left[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}} \nabla l_{n_{t_c}}\right]$$

DIRECT PERTURBATION

We compute the displacement $\theta_T - \theta_0$ to order η^3 for vanilla SGD. Expanding $\theta_t \in \theta_0 + \eta a_t + \eta^2 b_t + \eta^3 c_t + o(\eta^3)$, we find:

$$\theta_{t+1} = \theta_t - \eta \nabla l_{n_t}(\theta_t)$$
$$\in \theta_0 + \eta a_t + \eta^2 b_t + \eta^3 c_t$$
$$- \eta\left(\nabla l_{n_t} + \nabla^2 l_{n_t}(\eta a_t + \eta^2 b_t) + \frac{1}{2}\nabla^3 l_{n_t}(\eta a_t)^2\right) + o(\eta^3)$$
$$= \theta_0 + \eta\left(a_t - \nabla l_{n_t}\right)$$
$$+ \eta^2\left(b_t - \nabla^2 l_{n_t} a_t\right)$$
$$+ \eta^3\left(c_t - \nabla^2 l_{n_t} b_t - \frac{1}{2}\nabla^3 l_{n_t} a_t^2\right) + o(\eta)$$

We thus have the recurrences $a_{t+1} = a_t - \nabla l_{n_t}$, $b_{t+1} = b_t -$

$\nabla^2 l_{n_t} a_t$, and $c_{t+1} = c_t - \nabla^2 l_{n_t} b_t - \frac{1}{2}\nabla^3 l_{n_t} a_t^2$ with solutions: $a_t = -\sum_t \nabla l_{n_t}$ and $\eta^2 b_t = +\eta^2 \sum_{t_0 < t_1} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}$. We do not compute c_t because we will soon see that it will be multiplied by 0.

To third order, the test loss of SGD is

$$l(\theta_T) \in l(\theta_0) + (\nabla l)(\eta a_T + \eta^2 b_T + \eta^3 c_T)$$
$$+ \frac{\nabla^2 l}{2}(\eta a_T + \eta^2 b_T)^2$$
$$+ \frac{\nabla^3 l}{6}(\eta a_T)^3 + o(\eta)^3$$
$$= l(\theta_0) + \eta\left((\nabla l)a_T\right)$$
$$+ \eta^2\left((\nabla l)b_T + \frac{\nabla^2 l}{2}a_T^2\right)$$
$$+ \eta^3\left((\nabla l)c_T + (\nabla^2 l)a_T b_T + \frac{\nabla^3 l}{6}a_T^3\right) + o(\eta)^3$$

Because $\mathbb{E}[\nabla l]$ vanishes at initialization, we neglect the (∇l) terms. The remaining η^3 terms involve $a_T b_T$, and a_T^3. So let us compute their expectations:

$$\mathbb{E}[a_T b_T] = -\sum_t \sum_{t_0 < t_1} \mathbb{E}\left[\nabla l_{n_t} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}\right]$$
$$= -\sum_{t_0 < t_1} \sum_{t \notin \{t_0, t_1\}} \mathbb{E}[\nabla l_{n_t}] \mathbb{E}\left[\nabla^2 l_{n_{t_1}}\right] \mathbb{E}[\nabla l_{n_{t_0}}]$$
$$- \sum_{t_0 < t_1} \sum_{t = t_0} \mathbb{E}\left[\nabla l_{n_t} \nabla l_{n_{t_0}}\right] \mathbb{E}\left[\nabla^2 l_{n_{t_1}}\right]$$
$$- \sum_{t_0 < t_1} \sum_{t = t_1} \mathbb{E}\left[\nabla l_{n_t} \nabla^2 l_{n_{t_1}}\right] \mathbb{E}[\nabla l_{n_{t_0}}]$$

Since $\mathbb{E}[\nabla l]$ divides $\mathbb{E}[a_T b_T]$, the latter vanishes.

$$\mathbb{E}\left[a_T^3\right] = -\sum_{t_a, t_b, t_c} \mathbb{E}\left[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}} \nabla l_{n_{t_c}}\right]$$
$$= -\sum_{\substack{t_a, t_b, t_c \\ \text{disjoint}}} \mathbb{E}\left[\nabla l_{n_{t_a}}\right] \mathbb{E}\left[\nabla l_{n_{t_b}}\right] \mathbb{E}\left[\nabla l_{n_{t_c}}\right]$$
$$- 3\sum_{t_a = t_b \neq t_c} \mathbb{E}\left[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}}\right] \mathbb{E}\left[\nabla l_{n_{t_c}}\right]$$
$$- \sum_{t_a = t_b = t_c} \mathbb{E}\left[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}} \nabla l_{n_{t_c}}\right]$$

As we initialize at a test minimum, only the last line remains, at it has T identical summands. When we plug into the expression for SGD test loss, we get

$$\frac{T\eta^3}{6} \mathbb{E}[\nabla^3 l] \mathbb{E}\left[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}} \nabla l_{n_{t_c}}\right]$$

C. Other Perturbative Calculations

C.1. SGD vs ODE. Proof of Corollary 1's First Part

The rest of the proof is in Appendix B.

We work in the setting of Corollary 1. This corollary's first part follows immediately from Remark 1 in the case that $d = 2$, $P = 2$, and $(\eta N)^d$ is considered fixed while N^{P-d-1} is considered changing.

C.2. The Effect of Batch Size and Epochs: When Does SGD Outperform GD? Proof of Proposition 2 and Corollaries 3, 4.

To prove Proposition 2, we simply count the embeddings of the diagrams, noting that the automorphism groups are all of size 1 or 2. Since we use fuzzy outlines instead of fuzzy ties, we allow untied nodes to occupy the same row, since the excess will be canceled out by the term subtract in the definition of fuzzy outlines.

diagram	embed.s w/ $\left|\text{Aut}_f\right| = 1$	embed.s w/ $\left|\text{Aut}_f\right| = 2$
·	1	0
·—·	MNB	0
·—·—·	$\binom{MNB}{2}$	0
⌒·—·	$N\binom{MB}{2}$	0
·—⌒·	$\binom{MNB}{2}$	0
⌒⌒·	$N\binom{MB}{2}$	MNB

The two mentioned corollaries follow from plugging in appropriate values of M, N, B.

C.3. Inter-Epoch Shuffling Doesn't Matter Much For Small ηT Proof of Corollary 2.

We work in the setting of Corollary 2.

For two shuffling patterns (i.e. spacetimes) and fixed N, M, B, the embeddings (of diagrams through order 3) into one spacetime are in bijective and value-preserving correspondence with the embeddings into the other spacetime.

Indeed, if all three non-root nodes are embedded into one epoch, or conversely if they are embedded into three distinct epochs, then this follows from the exchangeability of the distribution over train sets.

The remaining case is that the non-root nodes are embedded into two epochs, one epoch carrying one node (say A) and the other carrying two (say B, C). We may move A to any cell of its epoch without violating the thin-edge-condition; this produces a collection of several diagram embeddings with shared tree structure and potentially different fuzzy tie structure. Crucially, these diagrams are all order at most 3 and as a collection of values are invariant to the move of interchanging B's row with C's row. Therefore: the order-3-and-lower diagrams partition into collections each amenable

to the exchangeability argument of the previous paragraph. QED.

C.4. The 3rd Order Curl: Which Minima Does SGD Prefer? Proof of Corollaries 5 and 6.

UN-RENORMALIZED

We work in the setting of Corollary 5.

Because we seek to compute a net displacement instead of a test loss, we consider diagrams with their root amputated away: the free thin edge that used to connect to the root now represents the uncontracted tensor index that gives the displacement. To third order and at a test minimum, the only non-vanishing such diagram is ⌒·—· (which equals

⌒·—· because we are at a test minimum) — with the blue root amputated. It evaluates to $-C_\mu \nu J^{\mu\nu\lambda}$, and it embeds in $\binom{T}{2}$ ways into vanilla SGD's spacetime: one for each increasing pair of cells into which to place the red nodes and the green node, respectively. The result follows.

RENORMALIZED

We work in the setting of Corollary 6.

The relevant irreducible diagram is ⌒·—· (amputated as in the previous subsubsection). An embedding of this diagram into vanilla SGD's spacetime has two relevant durations, t from red to green and $\tilde{t}$ from green to blue, such that $t + \tilde{t} \leq T$ The automorphism group of every embedding has size 2: identity or switch the red nodes. So the answer is:

$$C_{\mu\nu} J_\sigma^{\rho\lambda} \left(\int_{t+\tilde{t} \leq T} (\exp(-t\eta H)\eta)^{\mu\rho} (\exp(-t\eta H)\eta)^{\nu\lambda} (\exp(-\tilde{t}\eta H)\eta)^{\sigma\pi} \right)$$

Again, standard integration gives the desired result. QED.

D. Mathematical Background

This Appendix provides the mathematical context for our proofs. For the practitioner of the diagram method, such background is likely unnecessary. We list the regularity conditions assumed of the loss landscape not here but in Appendix E.

D.1. The Combinatorial Costumes: Structure Sets

The main text and Appendix A give practical perspective on spacetimes and diagrams. To define those concepts with mathematical precision, however, we employ the language of category theory. What linear algebra does to clarify and systematize an otherwise unwieldy world of coordinate transformations, category theory does for combinatorial structures and notions of sameness. We recommend Fong

& Spivak (2019) for a practical introduction to applied category theory.

We define both diagrams and spacetimes in terms of *structure sets*, i.e. sets S equipped with a preorder $\leq$ and an equivalence relation $\sim$. There need not be any relationship between $\leq$ and $\sim$. Morphisms between structure sets are strictly increasing maps that preserve $\sim$ and its negation. A structure set is *pointed* when it has a unique $\leq$-maximum element and this element forms a singleton $\sim$-class. The categories $\mathcal{S}$ of structure sets and $\mathcal{P}$ of pointed structure sets enjoy a free-forgetful adjunction $\mathcal{F}, \mathcal{G}$.

When $\leq$ is a total preorder, we say that S is a *spacetime*. When $\leq$ has arbitrary joins and its geometric realization is a tree, we say that S is a *diagram*.

Let parts(D) give the $\sim$-parts of D. An $\mathcal{S}$-map from D to $(\mathcal{G} \circ \mathcal{F})^{|\text{parts}(D)|}$(empty set) is an *ordering* of D. Let $|\text{edges}(D)|$ and $|\text{ords}(D)|$ count edges and orderings of D. In any category, and with any morphism $f : x \to y$, let $\left|\text{Aut}_f(x)\right|$ count the automorphisms of x that commute with x.

D.2. The Parameterized *Personae*: Forms of SGD

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

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

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

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

We follow convention in using the word "set" for ordered sequences of training points.

E. Proofs of Theorems

E.1. Regularity Hypotheses

We assume throughout this work the following regularity properties of the loss landscape. *Existence of Taylor Moments* — we assume that each finite collection of polynomials of the 0th and higher derivatives of the l_x, all evaluated at any point θ, may be considered together as a random variable insofar as they are equipped with a probability measure

upon of the standard Borel algebra. *Analyticity Uniform in Randomness* — we moreover assume that the functions $\theta \mapsto l_x(\theta)$, as well as the expectations of polynomials of their 0th and higher derivatives, exist and are analytic with radii of convergence bounded from 0 (by a potentially θ-dependent function). *Boundedness of Gradients* — we also assume that the gradients $\nabla l_x(\theta)$, considered as random covectors, are bounded by some continuous function of θ.[*][†]

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

E.2. Dyson Series for Iterative Optimizers

We first give intuition, then worry about ϵs and δs.

THE KEY LEMMA: PROOF IDEA

Intuitively, since ∇ Lie-generates translation, the operator $\exp\left(-\eta^{\mu\nu}g_\mu\nabla_\nu\right)$ performs translatation by $-\eta g$. In particular, the case $g = \nabla l_t(\theta)$ effects a gradient step on the tth batch. A product of such exponential operators will give the loss after a sequence of updates $\theta \mapsto \theta - \eta^{\mu\nu}\nabla_\mu l(\theta)$ on losses $(l_t : 0 \leq t < T)$. Because the operators might not commute, we may not compose the product of exponentials into an exponential of a sum. We instead compute an expansion in powers of η, collecting terms of like degree while maintaining the order of operators:

$$s(\theta_T) = \left(\prod_{0 \leq t < T}\left(\sum_{0 \leq d_t}\frac{(-\eta^{\mu\nu}g_\mu\nabla_\nu)^{d_t}}{d_t!}\bigg|_{g=\nabla l_t(\theta)}\right)s\right)(\theta_0)$$

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

We finish by taking expectations.

THE KEY LEMMA: PROOF

We work in a neighborhood of the initialization so that the tangent space of weight space is a trivial bundle. For con-

---

[*] A metric-independent way of expressing this boundedness constraint is that the gradients all lie in some subset $\mathcal{S} \subseteq TM$ of the tangent bundle of weight space, where, for any compact $C \subseteq M$, we have that the topological pullback — of $\mathcal{S} \hookrightarrow TM \twoheadrightarrow M$ and $C \hookrightarrow M$ — is compact. We hope that the results of this paper expose how important the choice of metric can be and hence underscore the value of determining whether a concept is metric-independent.

[†] Some of our experiments involve Gaussian noise, which is not bounded and hence violates one of our hypotheses. For experimental purposes, however, Gaussians are effectively bounded, on the one hand in the sense that with high probability no standard normal sample encountered on Gigahertz hardware within the age of the universe will much exceed $\sqrt{2\log(10^{30})} \approx 12$, and on the other hand in the sense that our predictions vary smoothly with the first few moments of this distribution, so that a ± 12-clipped Gaussian will yield almost the same predictions.

venience, we fix a flat coordinate system, and with it the induced flat, non-degenerate inverse metric $\tilde{\eta}$; the benefit is that we may compare our varying η against one fixed $\tilde{\eta}$. Henceforth, a "ball" unless otherwise specified will mean a ball with respect to $\tilde{\eta}$ around the initialization θ_0. Since s is analytic, its Taylor series converges to s within some positive radius ρ ball. By assumption, every l_t is also analytic with radius of convergence around θ_0 at least some $\rho > 0$. Since gradients are x-uniformly bounded by a continuous function of θ, and since in finite dimensions the closed ρ-ball is compact, we have a strict gradient bound b uniform in both x and θ on gradient norms within that closed ball. When

$$2\eta T b < \rho \tilde{\eta} \tag{4}$$

as norms, SGD after T steps on any train set will necessarily stay within the ρ-ball.* We note that the above condition on η is weak enough to permit all η within some open neighborhood of $\eta = 0$.

Condition 4 together with analyticity of s then implies that $(\exp(-\eta g\nabla)s)(\theta) = s(\theta - \eta g)$ when θ lies in the $\tilde{\eta}$ ball (of radius ρ) and its η-distance from that $\tilde{\eta}$ ball's boundary exceeds b, and that both sides are analytic in η, θ on the same domain — and *a fortiori* when θ lies in the ball of radius $\rho(1 - 1/(2T))$. Likewise, a routine induction through T gives the value of s (after doing T gradient steps from an initialization θ) as

$$\left(\prod_{0 \le t < T} \exp(-\eta g\nabla)\Big|_{g=\nabla l_t(\theta)} \right)(s)(\theta)$$

for any θ in the $\rho(1 - T/(2T))$-ball (that is, the $\rho/2$-ball), and that both sides are analytic in η, θ on that same domain. Note that in each exponential, the ∇_ν does not act on the $\nabla_\mu l(\theta)$ with which It pairs.

Now we use the standard expansion of exp. Because (by analyticity) the order d coeffients of l_t, s are bounded by some exponential decay in d that has by assumption an x-uniform rate, we have absolute convergence and may rearrange sums. We choose to group by total degree:

$$\cdots = \sum_{0 \le d < \infty} (-\eta)^d \sum_{\substack{(d_t : 0 \le t < T) \\ \sum_t d_t = d}} \left(\prod_{0 \le t < T} \frac{(g\nabla)^{d_t}}{d_t!} \Big|_{g=\nabla l_t(\theta)} \right) s(\theta) \quad (5)$$

The first part of the Key Lemma is proved. It remains to show that expectations over train sets commute with the above summation.

We will apply Fubini's Theorem. To do so, it suffices to

* In fact, the factor of 2 helps ensure that SGD initialized at any point within a $\rho/2$ ball will necessarily stay within the ρ-ball.

show that

$$|c_d((l_t : 0 \le t < T))| \triangleq \left| \sum_{\substack{(d_t : 0 \le t < T) \\ \sum_t d_t = d}} \left(\prod_{0 \le t < T} \frac{(g\nabla)^{d_t}}{d_t!} \Big|_{g=\nabla l_t(\theta)} \right) s(\theta) \right|$$

has an expectation that decays exponentially with d. The symbol c_d we introduce purely for convenience; that its value depends on the train set we emphasize using function application notation. Crucially, no matter the train set, we have shown that the expansion 5 (that features c_d appear as coefficients) converges to an analytic function for all η bounded as in condition 4. The uniformity of this demanded bound on η implies by the standard relation between radii of convergence and decay of coefficients that $|c_d|$ decays exponentially in d at a rate uniform over train sets. If the expectation of $|c_d|$ exists at all, then, it will likewise decay at that same shared rate.

But $|c_d|$ indeed has an expectation, for it is a bounded continuous function of a (finite-dimensional) space of T-tuples (each of whose entries can specify the first d derivatives of an l_t) and because the latter space enjoys a joint distribution (over the standard Borel algebra). So Fubini's Theorem applies. The Key Lemma follows.

E.3. Terms and Diagram Embeddings Correspond

Path Integral Theorem: Proof Idea

We now seek to describe the terms that appear in the Key Lemma. Theorem 1 does so by matching each term to an embedding of a diagram in spacetime, so that the infinite sum becomes a sum over all diagram spacetime configurations. The main idea is that the combinatorics of diagrams parallels the combinatorics of repeated applications of the product rule for derivatives applied to the expression in the Key Lemma. Balancing against this combinatorial explosion are factorial-style denominators, again from the Key Lemma, that we summarize in terms of the sizes of automorphism groups.

Consolation

The following proof is messy. It compresses into a reusable package the intricacies of direct perturbation (see Appendix B for samples of uncompressed computations), and as such equates two conceptually clean sides via a jungle of canceling sums and factorials.

How can we increase our confidence in the correctness of a theorem so unappetizingly proved? We regard three pieces of evidence as supplementing this proof: *Aesthetic* evidence — the Theorem assumes a form familiar to mathematicians and physicists: it is a sum over combinatorial objects weighted inversely by the order of their respective

automorphism groups. *Comparative* evidence — the Theorem's predictions agree with direct perturbation in the cases we report in Appendix B. *Empirical* evidence — the Theorem, while compactly stated, precisely predicts the existence and intensity of the phenomena we report in the main body up to third order.

Path Integral Theorem: Proof

We first prove the statement about test losses. Due to the analyticity property established in our proof of the Key Lemma, it suffices to show agreement at each degree d and train set individually. That is, it suffices to show — for each train set $(l_n : 0 \leq n < N)$, spacetime S, function $\pi : S \to [N]$ that induces $\sim$, and natural d — that

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

$$\sum_{\substack{D \in \text{im}(\mathcal{F}) \\ \text{with } d \text{ edges}}} \left(\sum_{f : D \to \mathcal{F}(S)} \frac{1}{\left| \text{Aut}_f(D) \right|} \right) \frac{\text{value}_\pi(D, f)}{B^d} \qquad (6)$$

Here, value_π is the value of a diagram embedding before taking expectations over train sets. We have for all f that $\mathbb{E}\left[\text{value}_\pi(D, f) \right] = \text{value}(D)$. Observe that both sides of 6 are finitary sums.

Remark 2 (Differentiating Products)**.** The product rule of Leibniz easily generalizes to higher derivatives of finitary products:

$$\nabla^{|M|} \prod_{k \in K} p_k = \sum_{v : M \to K} \prod_{k \in K} \left(\nabla^{|v^{-1}(k)|} p_k \right)$$

The above has $|K|^{|M|}$ many term indexed by functions to K from M.

We proceed by joint induction on d and S. The base cases wherein S is empty or $d = 0$ both follow immediately from the Key Lemma, for then the only embedding is the unique embedding of the one-node diagram $\cdot$. For the induction step, suppose S is a sequence of $\mathcal{M} = \min S \subseteq S$ followed by a strictly smaller $\tilde{S}$ and that the result is proven for $(\tilde{d}, \tilde{S})$ for every $\tilde{d} \leq d$. Let us group by d_0 the terms on the left hand side of desideratum 6. Applying the induction hypothesis with $\tilde{d} = d - d_0$, we find that that left hand side is:

$$\sum_{0 \leq d_0 \leq d} \sum_{\substack{\tilde{D} \in \text{im}(\mathcal{F}) \\ \text{with } d - d_0 \text{ edges}}} \frac{1}{d_0!} \sum_{\tilde{f} : \tilde{D} \to \mathcal{F}(\tilde{S})} \left(\frac{1}{\left| \text{Aut}_{\tilde{f}}(\tilde{D}) \right|} \right) \cdot$$

$$(-\eta)^{d_0} \left. (g\nabla)^{d_0} \right|_{g = \nabla l_0(\theta)} \frac{\text{value}_\pi(\tilde{D}, \tilde{f})}{B^{d - d_0}}$$

Since $\text{value}_\pi(\tilde{D}, \tilde{f})$ is a multilinear product of $d - d_0 + 1$ many tensors, the product rule for derivatives tells us that

$(g\nabla)^{d_0}$ acts on $\text{value}_\pi(\tilde{D}, \tilde{f})$ to produce $(d - d_0 + 1)^{d_0}$ terms. In fact, $g = \sum_{m \in \mathcal{M}} \nabla l_m(\theta)/B$ expands to $B^{d_0}(d - d_0 + 1)^{d_0}$ terms, each conveniently indexed by a pair of functions $\beta : [d_0] \to \mathcal{M}$ and $\nu : [d_0] \to \tilde{D}$. The (β, ν)-term corresponds to an embedding f of a larger diagram D in the sense that it contributes $\text{value}_\pi(D, f)/B^{d_0}$ to the sum. Here, (f, D) is $(\tilde{f}, \tilde{D})$ with $\left| (\beta \times \nu)^{-1}(n, v) \right|$ many additional edges from the cell of datapoint n at time 0 to the vth node of $\tilde{D}$ as embedded by $\tilde{f}$.

By the Leibniz rule of Remark , this (β, ν)-indexed sum by corresponds to a sum over embeddings f that restrict to $\tilde{f}$, whose terms are multiples of the value of the corresponding embedding of D. Together with the sum over $\tilde{f}$, this gives a sum over all embeddings f. So we now only need to check that the coefficients for each $f : D \to S$ are as claimed.

We note that the (β, ν) diagram (and its value) agrees with the $(\beta \circ \sigma, \nu \circ \sigma)$ diagram (and its value) for any permutation σ of $[d_0]$. The corresponding orbit has size

$$\frac{d_0!}{\prod_{(m,i) \in \mathcal{M} \times \tilde{D}} \left| (\beta \times \nu)^{-1}(m, i) \right|!}$$

by the Orbit Stabilizer Theorem of elementary group theory.

It is thus enough to show that

$$\left| \text{Aut}_f(D) \right| = \left| \text{Aut}_{\tilde{f}}(D) \right| \prod_{(m,i) \in \mathcal{M} \times \tilde{D}} \left| (\beta \times \nu)^{-1}(m, i) \right|!$$

We will show this by a direct bijection. First, observe that $f = \beta \sqcup \tilde{f} : [d_0] \sqcup \tilde{D} \to \mathcal{M} \sqcup \tilde{S}$. So each automorphism $\phi : D \to D$ that commutes with f induces both an automorphism $\mathcal{A} = \phi|_{\tilde{D}} : \tilde{D} \to \tilde{D}$ that commutes with $\tilde{f}$ together with the data of a map $\mathcal{B} = \phi_{[d_0]} : [d_0] \to [d_0]$ that both commutes with β. However, not every such pair of maps arises from a ϕ. For, in order for $\mathcal{A} \sqcup \mathcal{B} : D \to D$ to be an automorphism, it must respect the order structure of D. In particular, if $x \leq_D y$ with $x \in [d_0]$ and $y \in \tilde{D}$, then we need

$$\mathcal{B}(x) \leq_D \mathcal{A}(y)$$

as well. The pairs $(\mathcal{A}, \mathcal{B})$ that thusly preserve order are in bijection with the $\phi \in \text{Aut}_f(D)$. There are $\left| \text{Aut}_{\tilde{f}}(\tilde{D}) \right|$ many $\mathcal{A}$. For each $\mathcal{A}$, there are as many $\mathcal{B}$ as there are sequences $(\sigma_i : i \in \tilde{D})$ of permutations on $\{j \in [d_0] : j \leq_D i\} \subseteq [d_0]$ that commute with $\mathcal{B}$. These permutations may be chosen independently; there are $\prod_{m \in \mathcal{M}} \left| (\beta \times \nu)^{-1}(m, i) \right|!$ many choices for σ_i. Claim **??** follows, and with it the correctness of coefficients.

The argument for generalization gaps parallels the above when we use $l - \sum_n l_n/N$ instead of l as the value for s. The Path Integral Theorem (Theorem 1) is proved.

Remark 3 (The Case of Vanilla SGD)**.** The spacetime of vanilla SGD permits all and only those embeddings that

assign to each part of a diagram's partition a distinct cell. Such embeddings factor through a diagram ordering and are thus easily counted using factorials per Proposition 1. That proposition immediately follows from the now-proven Theorem 1.

E.4. Coefficient Convergence upon Renormalization

RENORMALIZATION: rvalue$_f(D)$'s GENERAL RECIPE

As mentioned in Definition 4, because we wish to study the difference between noisy and non-noisy optimization, we define our rvalues as differences of multiple diagrams, all sharing one thin-edge structure but some more fuzzily tied than others. In the cases encountered in the text, these differences are simply those captured by our fuzzy tie notation, e.g. ⬤⌢⬤·⬤ ≜ ⬤⌢⬤·⬤ − ·⌢·⬤·. However, because the renormalized picture's every irreducible diagram implicitly represents a whole family of topologically related diagrams, to use this difference as our definition for larger diagrams leads to combinatorial overcounting.

For example, let D be an irreducible diagram (e.g., ⬤⬤⬤·). Removing all fuzzy ties, we obtain another diagram $D_\flat$ (in the continuing example, ·⬤·⬤·), in turn is in the same class as some potentially smaller irreducible diagram $D_\star$ (·⬤·). The overcounting problem arises because some terms in $D_\star$'s renormalized value overlap with some terms in D's renormalized value, even though they are distinct irreducible diagrams in correspondingly distinct classes.

We counter this overcounting using the standard technique of Möbius inversion (Rota, 1964). The relevant partial ordered set is the set of diagrams in the interval $[D_\flat, D]$, where in this discussion we consider diagrams with isomorphic thin edge structures as ordered by refinement of partitions. We proceed to define the relevant Möbius function. If rvalue$_f^\star$ gives the renormalized values as in 4 before replacing fuzzy ties by fuzzy outlines, we let for any $D_a \leq D_\flat$:

$$\text{rvalue}_f^\sharp(D_a, D_b) \triangleq \text{rvalue}^\star(D_b) - \sum_{D_a \leq D_c < D_b} \text{rvalue}_f^\sharp(D_b, D_c)$$

We then set rvalue$_f(D) \triangleq \text{rvalue}_f^\sharp(D_\flat, D)$.

For instance,

$$\text{rvalue}_f(\text{⬤⬤⬤·}) \triangleq \text{rvalue}_f^\star(\text{⬤⬤⬤·})$$

$$- \left(\text{rvalue}_f^\star(\text{⬤⬤·⬤·}) - \text{·⬤·⬤·}\right)$$

$$- \left(\text{rvalue}_f^\star(\text{⬤·⬤·}) - \text{·⬤·⬤·}\right)$$

$$- \left(\text{rvalue}_f^\star(\text{·⬤⬤·⬤·}) - \text{·⬤·⬤·}\right)$$

$$- \text{rvalue}_f^\star(\text{·⬤·⬤·})$$

RENORMALIZATION THEOREM: PROOF IDEA

The diagrams summed in Theorem 2 may be grouped by their geometric realizations. Each nonempty class of diagrams with a given geometric realization has a unique element with minimally many edges, and in this way all and only irreducible diagrams arise.

We encounter two complications: on one hand, that the sizes of automorphism groups might not be uniform among the class of diagrams with a given geometric realization. On the other hand, that the embeddings of a specific member of that class might be hard to count. The first we handle using Orbit-Stabilizer. The second we address as described by via Möbius sums.

RENORMALIZATION THEOREM: PROOF

We focus on test loss instead of generalization gap; the proofs are similar. The difference in loss from the noise-less case is given by all the diagram embeddings with at least one fuzzy tie, where the fuzzy tie pattern is actually replaced by a difference between noisy and noiseless cases as prescribed by the preceding discussion on Möbius Sums. Beware that even relatively noiseless embeddings may have illegal collisions of non-fuzzily-tied nodes within a single spacetime (data) row. Throughout the rest of this proof, we permit such illegal embeddings of the fuzz-less diagrams that arise from the aforementioned decomposition.

Because the Taylor series for analytic functions converge absolutely in the interior of the disk of convergence, the rearrangement of terms corresponding to a grouping by geometric realizations preserves the convergence result of Theorem 1.

Let us then focus on those diagrams σ with a given geometric realization represented by an irreducible diagram ρ. By Theorem 1, it suffices to show that

$$\sum_{f:\rho \to S} \sum_{\substack{\tilde{f}:\sigma \to S \\ \exists i_\star: f = \tilde{f} \circ i_\star}} \frac{1}{|\text{Aut}_{\tilde{f}}(\sigma)|} = \sum_{f:\rho \to S} \sum_{\substack{\tilde{f}:\sigma \to S \\ \exists i_\star: f = \tilde{f} \circ i_\star}} \sum_{\substack{i:\rho \to \sigma \\ f = \tilde{f} \circ i}} \frac{1}{|\text{Aut}_f(\rho)|} \tag{7}$$

Here, f is considered up to an equivalence defined by pre-composition with an automorphism of ρ. We likewise consider $\tilde{f}$ up to automorphisms of σ. And above, i ranges through maps that induce isomorphisms of geometric realizations, where i is considered equivalent to $\hat{i}$ when for some automorphism $\phi \in \text{Aut}_{\tilde{f}}(\sigma)$, we have $\hat{i} = i \circ \phi$. Name as X the set of all such is under this equivalence relation.

In equation 7, we have introduced redundant sums to structurally align the two expressions on the page; besides this rewriting, we see that equation 7's left hand side matches Theorem 1 resulting formula and tgat its right hand side is the desired formula of Theorem 2.

To prove equation 7, it suffices to show (for any $f, \tilde{f}, i$ as above) that

$$\left|\text{Aut}_f(\rho)\right| = \left|\text{Aut}_{\tilde{f}}(\sigma)\right| \cdot |X|$$

We will prove this using the Orbit Stabilizer Theorem by presenting an action of $\text{Aut}_f(\rho)$ on X. We simply use pre-composition so that $\psi \in \text{Aut}_f(\rho)$ sends $i \in X$ to $i \circ \psi$. Since $f \circ \psi = f$, $i \circ \psi \in X$. Moreover, the action is well-defined, because if $i \sim \hat{i}$ by ϕ, then $i \circ \psi \sim \hat{i} \circ \psi$ also by ϕ.

The stabilizer of i has size $\left|\text{Aut}_{\tilde{f}}(\rho)\right|$. For, when $i \sim i \circ \psi$ via $\phi \in \text{Aut}_{\tilde{f}}(\rho)$, we have $i \circ \psi = \phi \circ i$. This relation in fact induces a bijective correspondence: *every* ϕ induces a ψ via $\psi = i^{-1} \circ \phi \circ i$, so we have a map stabilizer$(i) \longleftrightarrow \text{Aut}_{\tilde{f}}(\rho)$ seen to be well-defined and injective because structure set morphisms are by definition strictly increasing and because is must induce isomorphisms of geometric realizations. Conversely, every ψ that stabilizes enjoys *only* one ϕ via which $i \sim i \circ \phi$, again by the same (isomorphism and strict increase) properties. So the stabilizer has the claimed size.

Meanwhile, the orbit is all of $|X|$. Indeed, suppose $i_A, i_B \in X$. We will present $\psi \in \text{Aut}_f(\rho)$ such that $i_B \sim i_A \circ \psi$ by $\phi = $ identity. We simply define $\psi = i_A^{-1} \circ i_B$, well-defined by the aforementioned (isomorphisms and strict increase) properties. It is then routine to verify that $f \circ \psi = \tilde{f} \circ i_A \circ i_A^{-1} \circ i_B = \tilde{f} \circ i_B = f$. So the orbit has the claimed size, and by the Orbit Stabilizer Theorem, the coefficients in the expansions of Theorems 2 and 1 match.

To prove Theorem 2's remaining convergence-strengthening result, we assume that H is positive. Then, for any m, the propagator $(I - \eta H)^{\otimes mt}$ converges via an exponential decay with t to 0 (with a rate dependent on m). Since up to degree d only a finite number of diagrams exist and hence only a finite number of possible ms, the exponential rates are bounded away from 0. Moreover, for any fixed t_{big}, the number of diagrams — involving no exponent t exceeding t_{big} — is eventually constant as T grows. Meanwhile, the number involving at least one exponent t exceeding that threshold grows polynomially in T (with degree d). The exponential decay of each term overwhelms the polynomial growth in the number of terms, and the convergence statement of

Theorem 2 follows.'

F. Bessel Factors for Estimating Multipoint Correlators from Data

Given samples from a joint probability space $\prod_{0 \le d < D} X_d$, we seek unbiased estimates of multipoint correlators (i.e. products of expectations of products) such as $\langle x_0 x_1 x_2 \rangle \langle x_3 \rangle$. For example, say $D = 2$ and from $2S$ samples we'd like to estimate $\langle x_0 x_1 \rangle$. Most simply, we could use $\underset{0 \le s < 2S}{\mathsf{A}} x_0^{(s)} x_1^{(s)}$, where A denotes averaging. In fact, the following also works:

$$S \left(\underset{0 \le s < S}{\mathsf{A}} x_0^{(s)} \right)\left(\underset{0 \le s < S}{\mathsf{A}} x_1^{(s)} \right) + (1 - S) \left(\underset{0 \le s < S}{\mathsf{A}} x_0^{(s)} \right)\left(\underset{S \le s < 2S}{\mathsf{A}} x_1^{(s)} \right) \tag{8}$$

When multiplication is expensive (e.g. when each $x_d^{(s)}$ is a tensor and multiplication is tensor contraction), we prefer the latter, since it uses $O(1)$ rather than $O(S)$ multiplications. This in turn allows more efficient use of large-batch computations on GPUs. We now generalize this estimator to higher-point correlators (and $D \cdot S$ samples).

For uniform notation, we assume without loss that each of the D factors appears exactly once in the multipoint expression of interest; such expressions then correspond to partitions on D elements, which we represent as maps $\mu : [D] \to [D]$ with $\mu(d) \le d$ and $\mu \circ \mu = \mu$. Note that $|\mu| := |im(\mu)|$ counts μ's parts. We then define the statistic

$$\{x\}_\mu := \prod_{0 \le d < D} \underset{0 \le s < S}{\mathsf{A}} x_d^{(\mu(d) \cdot S + s)}$$

and the correlator $\langle x \rangle_\mu$ we define to be the expectation of $\{x\}_\mu$ when $S = 1$. In this notation, 8 says:

$$\langle x \rangle_{\boxed{0}\,\boxed{1}} = \mathbb{E}\left[S \cdot \{x\}_{\boxed{01}} + (1 - S) \cdot \{x\}_{\boxed{0}\,\boxed{1}} \right]$$

Here, the boxes indicate partitions of $[D] = [2] = \{0, 1\}$. Now, for general μ, we have:

$$\mathbb{E}\left[S^D \{x\}_\mu \right] = \sum_{\tau \le \mu} \left(\prod_{0 \le d < D} \frac{S!}{\left(S - \left|\tau(\mu^{-1}(d))\right| \right)!} \right) \langle x \rangle_\tau \tag{9}$$

where '$\tau \le \mu$' ranges through partitions *finer* than μ, i.e. maps τ through which μ factors. In smaller steps, 9 holds

because

$$\mathbb{E}\left[S^D\{x\}_\mu\right] = \mathbb{E}\left[\sum_{(0 \le s_d < S) \in [S]^D} \prod_{0 \le d < D} x_d^{(\mu(d) \cdot S + s_d)}\right]$$

$$= \sum_{\substack{(0 \le s_d < S) \\ \in [S]^D}} \mathbb{E}\left[\prod_{0 \le d < D} x_d^{(\min\{\tilde{d} \,:\, \mu(\tilde{d}) \cdot S + s_{\tilde{d}} = \mu(d) \cdot S + s_d\})}\right]$$

$$= \sum_\tau \left|\left\{\begin{matrix}(0 \le s_d < S) \in [S]^D : \\ \left(\begin{smallmatrix}\mu(d) = \mu(\tilde{d}) \\ \wedge \; s_d = s_{\tilde{d}}\end{smallmatrix}\right) \Leftrightarrow \tau(d) = \tau(\tilde{d})\end{matrix}\right\}\right| \langle x \rangle_\tau$$

$$= \sum_{\tau \le \mu} \left(\prod_{0 \le d < D} \frac{S!}{\left(S - \left|\tau(\mu^{-1}(d))\right|\right)!}\right) \langle x \rangle_\tau$$

Solving 9 for $\langle x \rangle_\mu$, we find:

$$\boxed{\langle x \rangle_\mu = \frac{S^D}{S^{|\mu|}} \mathbb{E}\left[\{x\}_\mu\right] - \sum_{\tau < \mu} \left(\prod_{d \in im(\mu)} \frac{(S-1)!}{\left(S - \left|\tau(\mu^{-1}(d))\right|\right)!}\right) \langle x \rangle_\tau}$$

This expresses $\langle x \rangle_\mu$ in terms of the batch-friendly estimator $\{x\}_\mu$ as well as correlators $\langle x \rangle_\tau$ for τ *strictly* finer than μ. We may thus (use dynamic programming to) obtain unbiased estimators $\langle x \rangle_\mu$ for all partitions μ. Symmetries of the joint distribution and of the multilinear multiplication may further streamline estimation by turning a sum over τ into a multiplication by a combinatorial factor. For example, with complete symmetry:

$$\langle x \rangle_{\boxed{012}} = S^2 \{x\}_{\boxed{012}} - \frac{(S-1)!}{(S-3)!} \{x\}_{\boxed{0}\,\boxed{1}\,\boxed{2}} - 3\frac{(S-1)!}{(S-2)!} \{x\}_{\boxed{0}\,\boxed{12}}$$

We use such expressions throughout our experiments to estimate the (expected) values of diagrams.

G. Experiment Details: Data, Models, Train/Test Parameters

G.1. Software and Execution

All code and data-wrangling scripts can be found on github.com/???????/perturb. This link will be made available after the period of double-blind review.

Our code uses PyTorch 0.4.0 (Paszke et al., 2019) on Python 3.6.7; there are no other substantive dependencies. The code's randomness is thoroughly parameterized by random seeds and hence reproducible.

We ran experiments on a Lenovo laptop and on our institution's clusters; we consumed about 100 GPU-hours.

G.2. Artificial Loss Landscapes

Gaussian Fit

The "*Gaussian fit*" landscape is a distribution over functions $l_x : \mathbb{R}^1 \to \mathbb{R}$ on 1-dimensional weight space, indexed by standard-normally distributed 1-dimensional datapoints x and defined by the expression:

$$l_x(h) \triangleq \frac{1}{2}\left(h + x^2 \exp(-h)\right)$$

To measure overfitting, we initialize at the true test minimum $h = 0$.

Linear Screw

The "*linear screw*" landscape is a distribution over functions $l_x : \mathbb{R}^3 \to \mathbb{R}$ on 3-dimensional weight space, indexed by standard-normally distributed 1-dimensional datapoints x and defined by the expression:

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

Here, $H(z)(w, w) = w_x^2 + w_y^2 + (\cos(z)w_x + \sin(z)w_y)^2$ and $S(z)(w) = \cos(z - \pi/4)w_x + \sin(z - \pi/4)w_y$. We consider initializing at $x = y = z = 0$, which lies within a valley of global minima defined by $x = y = 0$. We note that the landscape has a three-dimensional continuous screw symmetry.

Mean Estimation

The "*mean estimation*" family of landscapes has as elements distributions over functions $l_x : \mathbb{R}^1 \to \mathbb{R}$ on 13-dimensional weight space, indexed by standard-normally distributed 1-dimensional datapoints x and defined by the expression:

$$l_x(w) \triangleq \frac{1}{2}Hw^2 + xSw$$

Here, H, S are positive reals parameterizing the family; they give the hessian and (square root of) gradient covariance, respectively.

For our hyperparameter-selection experiment, we introduce an l_2 term λ as follows:

$$l_x(w, \lambda) \triangleq \frac{1}{2}(H + \lambda)w^2 + xSw$$

Here, we constrain $\lambda \ge 0$ during optimization using projections; we found similar results when parameterizing $\lambda = \exp(h)$, which obviates the need for projection but necessitates a non-canonical choice of initialization. We initialize $\lambda = 0$.

G.3. Loss Landscapes for Image Classification: CIFAR-10 and Fashion-MNIST

Architectures

In addition to the clarifyingly artificial loss landscapes (Gaussian Fit, Linear Screw, and Mean Estimation) described in the main text, we tested our predictions on logistic

linear regression and simple convolutional networks (2 convolutional weight layers each with kernel 5, stride 2, and 10 channels, followed by two dense weight layers with hidden dimension 10) for the CIFAR-10 (Krizhevsky, 2009) and Fashion-MNIST datasets (Xiao et al., 2017). The convolutional architectures used tanh activations and Gaussian Xavier initialization. We parameterized the model so that the Gaussian-Xavier initialization of the linear maps in each layer differentially pulls back to standard normal initializations of the parameters.

DATASETS

For these non-artificial landscapes, we regard the finite amount of available data as the true (sum of diracs) distribution from which we sample test and train sets in i.i.d. manner (and hence "with replacement"). We do this to gain practical access to a ground truth against which we may compare our predictions. One might object that this sampling procedure would cause test and train sets to overlap, hence biasing test loss measurements. In fact, test and train sets overlap only in reference, not in sense: the situation is analogous to a text prediction task in which two training points culled from different corpora happen to record the same sequence of words, say, "Thank you!". In any case, all of our experiments focus on the limited-data regime, e.g. 10^1 datapoints out of $\sim 10^{4.5}$ dirac masses, so overlaps are diluted.

G.4. Measurement Process

DIAGRAM EVALUATION ON REAL LANDSCAPES

We implemented the formulae of Appendix F in order to estimate diagram values from real data measured at initialization from batch averages of products of derivatives.

DESCENT SIMULATIONS

We recorded test and train losses for each of the trials below. To improve our estimation of average differences, when we compared two optimizers, we gave them the same random seed (and hence the same train sets).

We ran $2 \cdot 10^5$ trials of Gaussian Fit with SDE and SGD, initialized at the test minimum with $T = 1$ and η ranging from $5 \cdot 10^{-2}$ to $2.5 \cdot 10^{-1}$. We ran $5 \cdot 10^1$ trials of Linear Screw with SGD with $T = 10^4$ and η ranging from 10^{-2} to 10^{-1}. We ran 10^3 trials of Mean Estimation with GD and STIC with $T = 10^2$, H ranging from 10^{-4} to $4 \cdot 10^0$, a covariance of gradients of 10^2, and the true mean 0 or 10 units away from initialization.

We ran $5 \cdot 10^4$ trials of the CIFAR-10 convnet on each of 6 Glorot-Xavier initializations we fixed once and for all through these experiments for the optimizers SGD, GD, and GDC, with $T = 10$ and η between 10^{-3} and $2.5 \cdot 10^{-2}$.

We did likewise for the linear logistic model on the one initialization of 0.

We ran $4 \cdot 10^4$ trials of the Fashion-MNIST convnet on each of 6 Glorot-Xavier initializations we fixed once and for all through these experiments for the optimizers SGD, GD, and GDC with $T = 10$ and η between 10^{-3} and $2.5 \cdot 10^{-2}$. We did likewise for the linear logistic model on the one initialization of 0.

G.5. Optimizers

We approximated SDE by refining time discretization by a factor of 16, scaling learning rate down by a factor of 16, and introducing additional noise in the shape of the covariance in proportion as prescribed by the Wiener process scaling.

Our GDC regularizer we implemented as described in the text, to work on real data. For our tests of the STIC regularizer, we exploited the low-dimensional special structure of our artificial landscape in order to avoid diagonalizing to perform the matrix exponentiation: precisely, we used that, even on training landscapes, the covariance of gradients would be degenerate in all but one direction, and so we need only exponentiate a scalar.
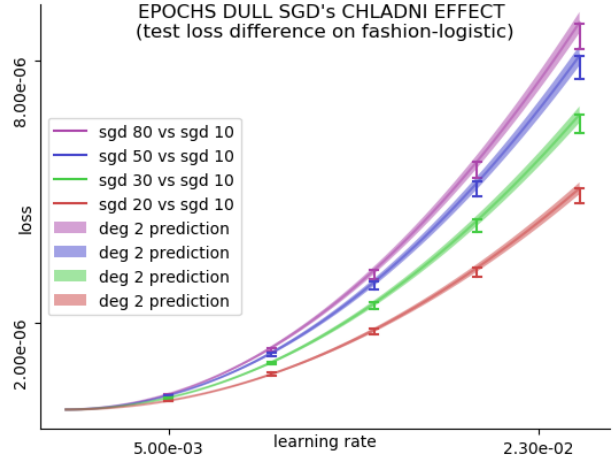
H. Additional Figures

Figure 8. SGD with $2, 3, 5, 8$ epochs incurs greater test loss than one-epoch SGD (difference shown in I bars) by the predicted amounts (predictions shaded) for a range of learning rates.

I. Glossary

As our work uses physical methods to solve problems of computing science, it may employ terminology unfamiliar to the reader. We strive in the main text to explain vital terms before their point of use, and we hope this glossary complements that attempt. The entries provide intuition
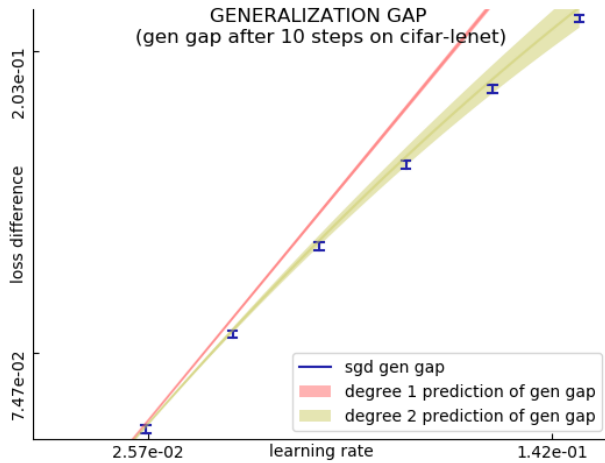
Figure 9. Generalization gap (test minus train) vs learning rate on an image classification task. For the instance shown and all 11 other initializations unshown, the degree-2 prediction agrees with experiment through $\eta T \approx 10^0$. Throughout, measurements are in blue and theory is in other colors. Vertical spans indicate 95% confidence intervals for the mean.

only; for formal statements, we direct the reader to our formal definitions as well as the cited reference works.

I.1. Terminology

affine manifold a shape equipped with and closed under a notion of displacement. For example, a plane or a circle but not a disc. The problem with a disc is that gradients might point toward the boundary and an SGD update might overshoot and fall off. By contrast, a circle is permitted because each nonzero gradient will point clockwise or counterclockwise, directions under which the circle is closed.

AIC see akaike information criterion.

akaike information criterion an estimate of generalization gap equal to (number of parameters)/N. Discrete-valued for fixed N, hence not liable to gradient descent.

analytic (of a function) locally equal to a convergent Taylor series. Most smooth functions one encounters in daily life are analytic. The ReLU function is neither smooth nor analytic.

automorphism a structure preserving map from an object to itself that has a structure preserving inverse. The automorphisms of sets are permutations; the automorphisms of graphs are conceptually analogous. For example, ⬡ has two automorphisms: the identity map and the map that switches the two red nodes.

conservative (of a covector field) with path-independent line integrals. The ODE approximation to gradient
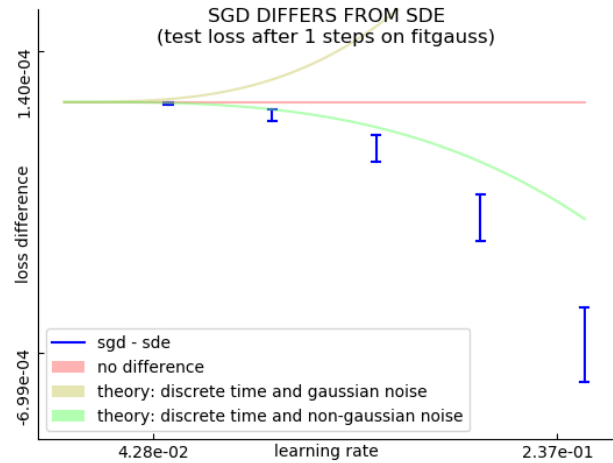

Figure 10. SGD's difference from SDE after $\eta T \approx 10^{-1}$ with maximal coarseness on the Gaussian-fit problem. Two effects not modeled by SDE — time-discretization and non-Gaussian noise oppose on this landscape but do not completely cancel. Our theory approximates the above curve with a correct sign and order of magnitude; we expect that the fourth order corrections would improve it further.

descent always yields a conservative covector field of gradients. See also nonconservative.

contraction (of two tensors) the numpy operation of multiplying along a pair of axes, then summing. The action of a row vector on a column vector gives the simplest example of contraction. For more complicated tensors, one performs that simplest product along the specified pair of axes while maintaining all other axes in an SQL-style join operation.

covector a numpy array of shape $1 \times p$. For example, a gradient. Compare to vector.

crossing symmetry a numerical relationship between diagrams with related shapes.

datapoint for us, a heavily overloaded term: an image of a cat or of a dog **OR** an index into the train set of the corresponding image **OR** the induced loss function of a given neural network on that corresponding image.

diagram a representation of an interaction process between weights and data as a rooted tree equipped with a partition of nodes. Drawing conventions: thin edges represent the tree, and fuzzy ties indicate the partition. The root is specified from among the nodes by placing it rightmost on the page.

edge see thin edge.

embedding (of a diagram into a spacetime) an assignment of diagram nodes into spacetime cells such that two nodes occupy the same row precisely when they are
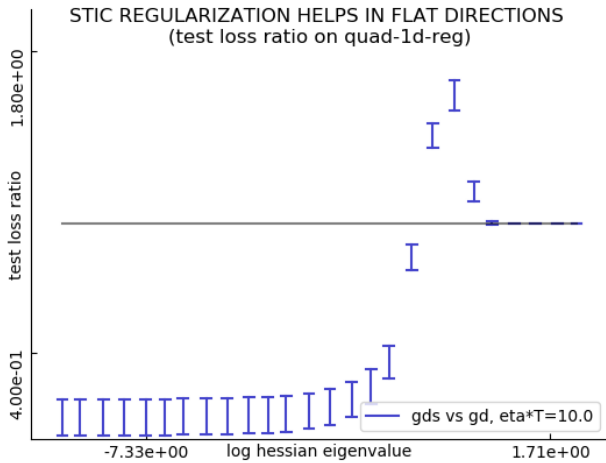
Figure 11. Blue intervals below the black bar correspond to STIC regularization outperforming gradient descent. For artificial quadratic landscapes with fixed covariance and a range of Hessians, initialized a fixed distance *away* from the true minimum, joint descent on an l_2 penalty coefficient λ by means of STIC improves on plain gradient descent for most Hessians. That there is at all a discrepancy from theory is possible because λ is not perfectly tuned according to STIC but instead descended on for finite ηT.

fuzzily tied and such that strict leafward-rootward relations along thin edges are reflected as strict past-future relations in the spacetime.

entropic force a macroscopic force arising from a tendency toward disorder and marked by strong dependence on temperature or noise scale. For instance, rubber bands are stretchy due to an entropic force arising from the tendency toward disorder of their constituent polymers. We showed that, due to a non-conservative entropic force, SGD tends toward minima that are flat with respect to the covariance.

epoch a temporal interval (within an SGD run) within which each datapoint participates in exactly one update. Our notation represents the number of epochs as $B \cdot M$.

fuzzy outline a convenient notation to express a common pattern of differences between a more fuzzily-tied diagram and a less fuzzily-tied diagram. Measures the net difference between more noisy and less noisy landscapes, hence extracts the effect due to and only to noise. Compare to fuzzy tie.

fuzzy tie a graphical depiction of correlations and hence noise in a loss landscape. At a technical level, a generator of a diagram's partition of nodes. For example, ⌒• has 1 fuzzy tie. See diagram. Compare to thin edge. Compare to fuzzy outline.

GD a special case of SGD in which batches have maximal size. Unlike generic SGD, GD operates deterministically once a train set is specified.

GDC a stochastic optimizer defined as GD descending on the original objective plus a loss term that makes GD mimic SGD. See Corollary 4

generalization gap for a model trained on some train set, the test loss minus the train loss. Compare to overfitting.

geometric realization (of a diagram) the data of a diagram that remains when we consider each chain of thin edges (such that each interior node is a singleton with respect to the partition) as equivalent to a single edge. See also irreducible.

inverse metric a notion of size for row vectors, in the sense of kernel methods. For instance, a covariance of weight displacements. Compare to metric.

irreducible (of a diagram) the property of being minimal among all diagrams with the same geometric realization. Concretely, the property that every thin-edge-degree two node participates in some fuzzy edges. See also geometric realization.

landscape a neural network together with a dataset, considered as a distribution, indexed by datapoints x, over functions from weights θ losses to $l_x(\theta)$.

landscape, convnet a neural network with three weight layers as described in Appendix G. Used for classifying CIFAR-10 or Fashion-MNIST images.

landscape, logistic logistic linear regression with bias as described in Appendix G. Used for classifying CIFAR-10 or Fashion-MNIST images.

landscape, gaussian fit maximum-likelihood fitting of a univariate gaussian density to samples drawn from a gaussian. Discussed in Section 4.3.

landscape, linear screw Constructed in Section 4.4.

landscape, mean estimation Quadratic-mean landscape with linear gaussian noise. Described in Appendix G.

metric a notion of size for column vectors, in the sense of kernel methods. For example, a covariance of gradients. Compare to inverse metric.

non-conservative (of a covector field) with path-dependent line integrals. In striking contrast to the ODE approximation to gradient descent, SGD experiences a non-conservative force. See also conservative.

ODE a continuous-time approximation to SGD, here construed as the large-batch, large-time, small-learning rate limit of SGD, where the rates of the three limits are synchronized so as to yield a deterministic algorithm that matches single-step SGD in the first moment of the final distribution of weights to first order in physical time ηT. Compare to SDE.

outline see fuzzy outline.

overfitting the act of responding to noise as if it is signal. In this work, we quantify overfitting by initializing a weight at a test minimum, then training. The greater the net gain in test loss, the more we regard the optimization process as having overfitted. Compare to generalization gap.

perturbation the technique of analyzing a complicated system by decomposing it into a simple system plus a small complication, then applying Taylor's theorem to extract the effects of that complication.

raising indices the algebraic step of contracting an inverse metric with a covector to produce a vector. To use an example from smooth convex optimization in the quadratic case, the dualization step of identifying a point's position from the objective's slope at that point.

renormalization the theory-building process of summarizing myriad small-scale interactions into a more easily manipulated large-scale interaction, often by appeal to scaling and symmetry. Here and historically, intertwined with resolving issues of convergence.

renormalized value (of a diagram embedding) used when applying Theorem 2 to obtain more complicated but precise expressions than Theorem 1 provides. See value.

scattering process the brief interaction of otherwise self-contained and easily-understood objects. For example, billiard balls scatter off of each other. We show how to interpret weights as scattering off of data. Often treated perturbatively.

SDE a continuous-time approximation to SGD, here construed as the large-batch, large-time, small-learning rate limit of SGD, where the rates of the three limits are synchronized so as to match single-step SGD in the first two moments of the final distribution of weights to first order in physical time ηT. Compare to ODE.

spacetime the stage on which weights and data interact. A grid-like summary that answers the question, *Which training points participate in which gradient updates at which times?*

spring the familiar simple machine with energy quadratic in its degrees of freedom, useful as an intuitive model for some of the results in this work. The sequential composition of two stiff springs yields a limper spring. We thus see that the potential energy of a static spring bearing a weight scales *inversely* with the spring's stiffness. This story formally parallels the Takeuchi prediction that flat minima generalize divergently badly.

stabilized takeuchi information criterion an estimate of generalization gap defined to be $C_{\mu\nu}(I - \exp(-\eta T H))^{\nu}_{\lambda}(H^{-1})^{\lambda\mu}$. Compare to TIC and AIC.

STIC see stabilized takeuchi information criterion.

stochastic process a collection of related random variables. For example, the validation losses of two neural networks with different and frozen weights forms a collection of 2 random variables, where the randomness is over a shared validation set. See also index (of a stochastic process).

takeuchi information criterion an estimate of generalization gap: $C_{\mu\nu}(H^{-1})^{\mu\nu}/N$. Diverges for small H, hence not suited to gradient descent.

tensor a numpy array of potentially long shape, e.g. shape $a \times b \times c \times d$. For example, the collection of 3rd derivatives of a multivariate function comprise a shape-$p \times p \times p$ tensor.

thin edge a graphical depiction of tensor contractions and hence a gradient operation on a loss landscape. At a technical level, a generator of a diagram's partial order. For example, has 3 thin edges. See diagram. Compare to fuzzy tie.

TIC see takeuchi information criterion.

tie see fuzzy tie.

value (of a diagram) the numeric contribution of the weight-data interaction depicted by the diagram to a test loss or generalization gap. Computed algorithmically as described in the text.

vector a numpy array of shape $p \times 1$. For example, a weight displacement. Compare to covector.

I.2. Diagrams for Computing Test Losses

We present all 3rd order diagrams relevant to test loss computations. Actually, the rows are indexed by topological families of diagrams. For example, the diagrams ·⌢· and ⌢·· , through distinct as diagrams, are topologically equivalent. They thus have the same unrenormalized value (an example of crossing symmetry!), and for brevity we treat them in the same row, labeled arbitrarily with one of them (here ⌢··). The interpretation of a diagram as a weight-data interaction process depends on the exact diagram, not just its topological family. So the interpretation row should be regarded as providing examples instead of being a complete enumeration.

Diagram	Unrenormalized Value	Interpretation
·	$+l$	Trivial process: no data-weight interaction
	$-\eta^{\mu\nu}G_\mu G_\nu$	A datapoint directly affects the test loss
	$+\eta^{\mu\nu}\eta^{\lambda\rho}\left(\nabla_\lambda C_{\mu\nu}\right)G_\rho/2$	A datapoint affects the test loss through a later encountered instance of itself
	$+\eta^{\mu\nu}\eta^{\lambda\rho}G_\mu G_\lambda H_{\nu\rho}$	Two different datapoints both affect the test loss
	$+\eta^{\mu\nu}\eta^{\lambda\rho}C_{\mu\lambda}H_{\nu\rho}$	A datapoint twice affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}G_\mu H_{\nu\lambda}H_{\rho\sigma}G_\pi$	A datapoint affects a different datapoint that affects yet another datapoint that affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}G_\mu G_\lambda G_\sigma J_{\nu\rho\pi}$	Three different datapoints affect the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}G_\mu H_{\nu\lambda}G_\sigma H_{\rho\pi}$	A datapoint twice affects the loss, once directly and once through a later-encountered and different datapoint
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(G_\sigma\nabla_\pi\left(C_{\mu\lambda}H_{\nu\rho}\right)/2 - C_{\mu\lambda}G_\sigma J_{\nu\rho\pi}\right)$	A datapoint affects a different point that itself twice affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}C_{\mu\lambda}G_\sigma J_{\nu\rho\pi}$	A datapoint twice affects the test loss while a different datapoint affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\mathbb{E}\left[\nabla_\mu l_x\nabla_\lambda l_x\nabla_\sigma l_x\right] - G_\mu G_\lambda G_\sigma\right)J_{\nu\rho\pi}$	A datapoint thrice affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\nabla_\lambda C_{\mu\nu}\right)H_{\rho\sigma}G_\pi/2$	A datapoint affects a later instance of itself, which affects a different point, which affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\mathbb{E}\left[\nabla_\nu\nabla_\lambda l_x\nabla_\rho\nabla_\sigma l_x\right] - H_{\nu\lambda}H_{\rho\sigma}\right)G_\mu G_\pi$	A datapoint affects another datapoint, which in turn affects a later instance of itself, which affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\mathbb{E}\left[\nabla_\mu l_x\nabla_\nu\nabla_\lambda l_x\nabla_\rho\nabla_\sigma l_x\right] - G_\mu H_{\nu\lambda}H_{\rho\sigma}\right)G_\pi$	A datapoint affects a later instance of itself, which affects a yet later instance of itself, which affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\mathbb{E}\left[\nabla_\mu l_x\nabla_\lambda l_x\nabla_\rho\nabla_\sigma l_x\right] - G_\mu G_\lambda H_{\rho\sigma}\right)H_{\nu\pi}$	A datapoint affects the test loss twice, once directly, and once through a later instance of itself
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\mathbb{E}\left[\nabla_\lambda l_x\nabla_\nu\nabla_\rho\nabla_\sigma l_x\right] - G_\lambda J_{\nu\rho\sigma}\right)G_\mu G_\pi$	A datapoint affects the test loss through a later encountered instance of itself, which was earlier affected by a different datapoint
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}\left(\mathbb{E}\left[\nabla_\mu l_x\nabla_\lambda l_x\nabla_\nu\nabla_\rho\nabla_\sigma l_x\right] - G_\mu G_\lambda J_{\nu\rho\sigma}\right)G_\pi$	A datapoint twice affects a later instance of itself that in turn affects the test loss