

A Space-Time Approach to Analyzing Stochastic Gradient Descent

Anonymous Authors¹

Abstract

We harness Feynman Diagrams to reason about the behavior of Stochastic Gradient Descent (SGD) at small learning rates η . Illustrating this technique: We construct a regularizer that causes large-batch GD to emulate small-batch SGD. We exhibit a non-conservative entropic force driving SGD. We generalize the Akaike Information Criterion (AIC) to a smooth quantity liable to descent. And we quantify the differences — due to time-discretization, inter-epoch correlations, and non-gaussian noise — between SGD and the popular approximation SDE. We empirically verify our theoretical predictions.

1. Introduction

Stochastic gradient descent (SGD) decreases an unknown objective l by performing discrete-time η -steepest* descent on noisy estimates of l . A key question is how such noise affects the final objective value. To address this question, we interpret SGD as a superposition of various processes in which weights and data interact, each process represented by a Feynman diagram[†]. When η is small, only a handful of diagrams substantively contribute, and so our formalism permits concrete computation.

In fact, unlike prior work, our formalism models discrete time and hence non-Gaussian noise. We thus obtain new results quantifying how epoch number and batch size affect test loss. We contrast SGD against popular continuous-time approximations such as ordinary or stochastic differential equations (ODE, SDE) by giving the finite- N , finite- η^{-1} corrections. Our theory of noise also recommends two novel regularizers that respectively induce GD to mimic SGD and help to tune hyperparameters such as l_2 coefficients.

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

*The “steepness” concept depends on the choice of metric for l ’s domain. We thus consider all (flat) metrics at once. Specifically, we will Taylor expand an (inverse) metric $\eta^{\mu\nu}$ around 0.

[†]Section 3 explains these words in detail.

Path integrals offer not only quantitative predictions but also an exciting new viewpoint — that of iterative optimization as a scattering process. Indeed, we import physical tools such as *crossing symmetries* (Dyson, 1949b) and *renormalization* (Gell-Mann & Goldberger, 1954) to simplify and refine our calculations. The diagrams’ combinatorics yield immediate qualitative insight as well, e.g. that to order η^3 , inter-epoch shuffling does not affect expected test loss. Overall, we present the diagram method as a flexible and elegant tool for optimization theorists. The corollaries we here prove are but advertisements for the method’s power: in the conclusion, we discuss Hessian methods and natural gradient descent as low-hanging fruit for future analyses.

1.1. Overview of Formalism

Deferring details, we survey our formalism. Say the n th datapoint participates in the t th update; we represent all such (n, t) pairs of an SGD run as the shaded cells of a grid, where datapoints index rows and times index columns. The shaded cells comprise the SGD run’s *spacetime*.

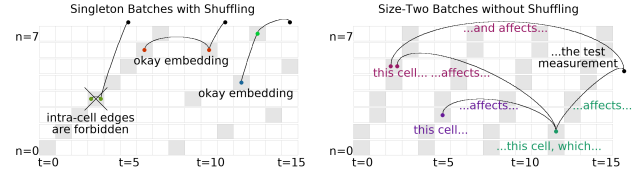





Figure 1. Two spacetimes (collections of shaded cells) for 16 steps on 8 datapoints along with some embedded Feynman diagrams. We arbitrarily choose colors to aid reference. Nodes (here black) outside the grid represent a post-training measurement of test loss; their coordinates are arbitrary. **Left:** SGD run with size-1 batches and with shuffling. The three embeddings shown are, respectively, an illegal embedding of , a legal embedding of , and a legal embedding of . **Right:** SGD run with size-2 batches and no shuffling. Depicted is an embedding of a 4-edged diagram, annotated with the data-weight interactions it represents. Section 3 quantifies the contribution of such diagrams to the test loss. The test loss is a sum of infinitely many such embeddings, each given by a different diagram embedding.

We interpret an SGD run as a superposition of many concurrent data-weight interactions. Diagrams represent such

processes (Figure 1, right). Diagrams evaluate to numbers, and an SGD run’s expected test loss is a *sum over diagrams* of those numbers, weighted by the number of ways a diagram can *embed* or “fit” into the SGD run’s spacetime.

Specifically, diagrams are graphs such as and : they are composed of thin edges and fuzzy ties, where the thin edges form a rooted tree with root drawn rightmost, and where the fuzzy ties represent a partition of nodes. Colors are arbitrarily chosen to serve as convenient labels — we might refer to a diagram’s “green nodes” — and lack mathematical meaning.

We may *embed* a diagram into spacetime by placing each node in a cell. We allow only embeddings that: *preserve left-right relationships along thin edges* and that *map two nodes into the same row exactly when they are fuzzily tied*. Figure 1, left, illustrates these rules.

Each d -edged diagram contributes to order η^d , so, for small η , the few-edged diagrams suffice to predict test losses. Our recipe for predicting test losses to within error $o(\eta^d)$ is thus: FIRST, draw all the diagrams with at most d edges. THEN, for each diagram, compute how many ways it can embed into the spacetime. FINALLY, sum the diagrams, weighed by its number of embeddings. We will formally state and prove that this correctly models SGD.

1.2. Appetizing Example: Role of Epoch Number

As an illustration, let us compare the test losses of one- and multi-epoch SGD. We scale M -epoch SGD’s η by $1/M^*$. One- and M -epoch SGD differ in their spacetimes and hence in how many times each diagram can embed. Figure 2 shows that most diagrams with at most 2 edges embed in corresponding ways into the two spacetimes. Since those embeddings’ contributions agree, we may neglect them as we compute the SGD’s difference. Only the diagram has embeddings in M -epoch SGD that do not correspond to embeddings in one-epoch SGD (Figure 2).

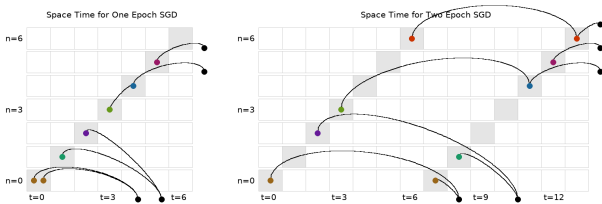


Figure 2. Examples of diagrams embedded in spacetime for one-epoch (left) and two-epoch (right) SGD. The diagrams for one and two epochs correspond by horizontal shifts *except* when same-row nodes are connected — such as in the top right diagram

*This scaling ensures that one- and M -epoch SGD agree on noiseless linear landscapes and thus a more interesting comparison.

Clearly, embeds in M -epoch spacetime in $N \cdot \binom{M}{2}$ many ways. The order η^2 discrepancy between one and multi epoch SGD is therefore[†]

$$\text{test loss difference} = \frac{N \cdot \binom{M}{2}}{M^2} \text{value} \left(\text{Diagram 3} \right) + o(\eta^2)$$

We complete the calculation by reading from the diagram’s graph structure that it evaluates to[‡]

$$\text{value} \left(\text{Diagram 3} \right) = \frac{1}{2} G^\mu \nabla_\mu C_\nu^\nu$$

We have thus computed the leading order test loss difference between one- and multi-epoch SGD.

In short: by using diagrams, we matched terms so that all but the terms canceled, thus avoiding an algebraic mess. Moreover, when we later renormalize toward the large- ηT regime (Sections 3.4, 4.4, and 4.5), diagrams will become essential. While diagrams may by direct use of our Key Lemma be sometimes avoided, Appendix B shows that to avoid them can convolute analysis. The more complicated the direct 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 $l_n : \mathcal{M} \rightarrow \mathbb{R}$. We abusively 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

[†]The factor $1/M^2$ comes from our choice to scale M -epoch’s learning rate by $1/M$ and that 2-edged diagrams are in $\Theta(\eta^2)$.

[‡]Hidden in the RHS’s notation is a factor η^2 : see Section 2.

[§]Vectors/covectors are also called column/row vectors.

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 $\theta^\mu \leftarrow \theta^\mu - \eta^{\mu\nu} \nabla_\nu \left(\frac{1}{B} \sum_{n \in \mathcal{B}} l_n(\theta) \right)$. SGD performs $N \cdot M$ many updates in total. We write l_t for the loss $\frac{1}{B} \sum_{\mathcal{B}} \dots$ over the t th 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*.

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

$$\sum_{(d_i: 0 \leq i \leq T)} (-\eta)^{\sum_i d_i} \left(\prod_{0 \leq i \leq T} \frac{(g \nabla)^{d_i}}{d_i!} \right) \Big|_{g=\nabla l_t(\theta)} (s)(\theta_0) \quad (1)$$

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

$$\begin{aligned} & -(\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla_\lambda \nabla_\mu \nabla^\nu l_b)(\nabla_\nu l_a) - (\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla_\lambda \nabla^\nu l_b)(\nabla_\mu \nabla_\nu l_a) \\ & -(\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla_\mu \nabla^\nu l_b)(\nabla_\lambda \nabla_\nu l_a) - (\nabla^\lambda l_c)(\nabla^\mu l_c)(\nabla^\nu l_b)(\nabla_\lambda \nabla_\mu \nabla_\nu l_a) \end{aligned}$$

To reduce clutter, we adapt the string notation of Feynman (1949); Penrose (1971). Then, in expectation over (l_c, l_b, l_a)

drawn i.i.d.:

$$\dots = \text{diagram 1} + \text{diagram 2} + \text{diagram 3} + \text{diagram 4} \quad (2)$$

$$= \underbrace{2 \text{diagram 1}}_{-2 \mathbb{E}[(\nabla l)(\nabla l)] \mathbb{E}[\nabla \nabla \nabla l] \mathbb{E}[\nabla l]} + \underbrace{2 \text{diagram 2}}_{-2 \mathbb{E}[(\nabla l)(\nabla l)] \mathbb{E}[\nabla \nabla l] \mathbb{E}[\nabla \nabla l]} \quad (3)$$

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 $\text{value}(D)$ of the corresponding tensor expression.

Usefully, the topology of a diagram determines its value. For instance, simplification 3 is licensed because $\text{value}(\text{diagram 1}) = \text{value}(\text{diagram 2})$. We will sometimes write a diagram D and mean $\text{value}(D)$.

Definition 2 (Fuzzy Outlines Denote the Effect of Noise).

As we study the effect of noise, we'd like to concisely notate the fuzzy ties that represent it. 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: $(\text{red green})_{\text{red green}} \triangleq \text{red green}$ and $(\text{red blue})_{\text{red blue}} \triangleq \text{red blue}$. 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 $\text{red green} \triangleq (\text{red green})_{\text{green blue}} - \text{red green}$.

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

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

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 $\text{value}(D)$ by $(-\sum_{p \in \text{parts}(D)} \text{value}(D_{rp}) - D)/N$, 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:

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

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

A P -part, d -edged diagram then contributes $\Theta((\eta N)^d N^{P-d-1})$ to the loss. For example, there are six diagrams to third order, 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. So we have 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 discretization-agnostic (hence continuous-time, noiseless) ODE approximation to SGD, while $P \leq d$ gives correction terms modeling time-discretization and hence 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 $\frac{T^3 N^{-2}}{6} \left(\text{diagram} - 3 \text{diagram} \right) = -\frac{T^3 N^{-2}}{6} \left(\left(\mathbb{E} \left[\nabla_{\mu} l_x \nabla_{\nu} l_x \nabla_{\lambda} l_x \right] - G_{\mu} G_{\nu} G_{\lambda} \right) J^{\mu\nu\lambda} - 3 C_{\mu\nu} G_{\lambda} J^{\mu\nu\lambda} \right)$. 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 $N \rightarrow \infty$ so that SDE matches SGD, optimization and generalization respectively become computationally intractable and trivial and hence less interesting.

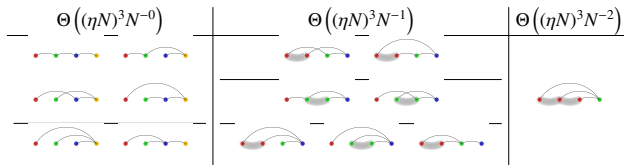


Table 1. Degree-3 scattering diagrams for $B = M = 1$ SGD's test loss. **Left:** $(d, P) = (3, 3)$. Diagrams for ODE behavior. **Center:** $(d, P) = (3, 2)$. 1st order deviation of SGD away from ODE. **Right:** $(d, P) = (3, 1)$. 2nd order deviation of SGD from ODE with appearance of non-Gaussian statistics.

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

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

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

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 $\frac{M(N-1)}{2} (\nabla_{\mu}C_{\nu}^{\mu})G^{\mu}/2$. 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.*

Corollary 5 (An Entropic Force, Unrenormalized). *When vanilla SGD is initialized at a test minimum, the weight is displaced to order η^3 by $-\frac{T(T-1)}{4} C_{\mu\nu}J_{\lambda}^{\mu\nu}G^{\lambda}$.*

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

An important idea is that of *renormalization*, i.e. the summarization of myriad small-scale interactions into an effective large-scale theory. We will renormalize — specifically, we will pair main diagrams with topologically related modifications in order to refine our d th order estimates for any fixed d . For some large- ηT limits in a positive-Hessian setting, 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 $TG_{\mu}G^{\mu}$. Though useful for small ηT , this approximation diverges as T grows. Intuitively, the consider the diagrams that are uncorrelated chains:






$\text{diagram}, \text{diagram}, \text{diagram}, \dots$. When embedded with initial and

final nodes separated by duration t , this series of diagrams sums to contributes $G(I - \eta H)^{t-1}\eta G \approx G \exp(-\eta TH)\eta G \cdot (1 + o(\eta))$. The $\exp(-\eta TH)\eta$ is thus an “effective learning rate” to be used in place of η in renormalized calculations. Summed over all embeddings, the above chains contribute $G \sum_{0 \leq t < T} (I - \eta H)^{T-t-1}\eta G$ to vanilla SGD's test loss, or:

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

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 non-root 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 $-(I - \eta H)^{\Delta t - 1} \eta$. **FILL IN** Each fuzzy tie is a fuzzy outline*.

RENORMALIZED RECIPE: In general, one sums over embeddings of irreducible diagrams, using $\text{rvalue}_f(D)$ instead of $\text{value}(D)$. In practice, we prefer to approximate sums over embeddings by integrals over times and $(I - \eta H)^t$ by $\exp(-\eta H t)$. This incurs 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*

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

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

That convergence property gives us formal license to take long-term limits of renormalized predictions. In fact, the convergence is uniform in T for sufficiently regular landscapes.

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*

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

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

Unlike prior work (Wei & Schwab, 2019), we make no assumptions of thermal equilibrium, fast-slow mode separation, or constant covariance. This added generality allows

*Actually, the simple differences denoted by fuzzy ties generalize to inclusion-exclusion (i.e. mobius-function weighted) sums for terms with more fuzzy ties than we consider in the paper’s body. We describe the general pattern in Appendix E

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

us to predict a qualitatively new dynamical phenomenon, namely that the velocity field above need not be conservative. We verify this in experiments.

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

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

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

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

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 mis-specification) near the “true” test minimum, $C = H$ is the Fisher metric, so we recover the Akaike Information Criterion (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.

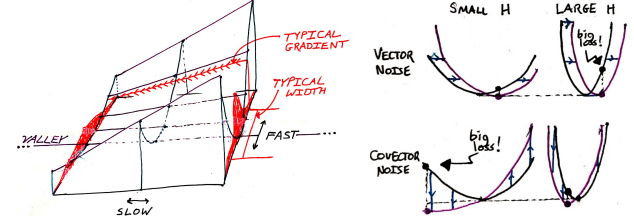


Figure 3. Left: The entropic force mechanism: gradient noise induces a flow toward minima flat with respect to the covariance. Although our analysis assumes neither thermal equilibrium nor fast-slow mode separation, we label “fast and slow directions” in this cartoon to ease comparison with Wei & Schwab (2019). In this cartoon, 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 and Applications

4.1. Vanilla SGD

FILL IN We test Proposition 1 for smooth convolutional architectures for CIFAR-10 and Fashion-MNIST by comparing measured test losses and generalization gaps with un-renormalized predictions (see Appendix G). We find good agreement between our order η^3 perturbative predictions up to $\eta T \approx 10^0$. This test verifies that our application of the diagram method hides no mistakes of proportionality of sign. See Appendix H for figures.

Likewise, Figure 4 shows that our predictions of how epochs affect overfitting are in good agreement with experiment.

4.2. Emulating Small Batches with Large Ones

Figure 4 shows that the regularizer proposed in Proposition 4 indeed enables GD to emulate SGD on a range of image-classification landscapes. For these experiments, we used a covariance estimator $\hat{C} \propto \nabla l_x(\nabla l_x - \nabla l_x)$ evaluated on two batches x, y that partition the train set.

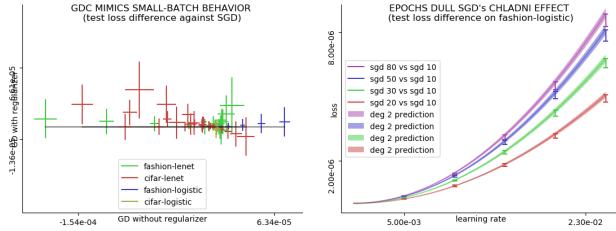


Figure 4. Left: with equal-scaled axes, this plot shows that GDC matches SGD (small vertical variation) better than GD matches SGD (large horizontal variation) in test loss, for a variety of learning rates (from 0.0025 to 0.1) and initializations (zero and multiple independent Xavier-Glorot trials) on several of image classification landscapes (logistic and convolutional CIFAR-10 and Fashion-MNIST). $T = 10$ throughout. **Right:** SGD with 2, 3, 5, 8 epochs incurs greater test loss than single-epoch SGD (difference shown in I bars) by the predicted amounts (predictions shaded) for a range of learning rates.

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 $\bullet + \frac{T}{2} \text{diagram} + \left(\frac{T}{2}\right) \text{diagram} + \frac{T}{6} \text{diagram}$. In particular, the $\left(\frac{T}{2}\right)$ differs from $T^2/2$ and hence contributes to the time-discretization error of SDE as an approximation for SDE. Moreover, the diagram contributes to non-gaussian

noise to that error. Figure 5 shows that, indeed, SDE and single-epoch SGD 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 diagram 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. By setting the covariance to lag the Hessian by a rotational phase, we construct a landscape in which this entropic force dominates. This “linear-screw” landscape is defined for three-dimensional $w \in \mathbb{R}^3$ (initialized at the origin) and one-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. This prediction, unlike the un-renormalized prediction, agrees for large ηT with experiment (Figure 5).

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}

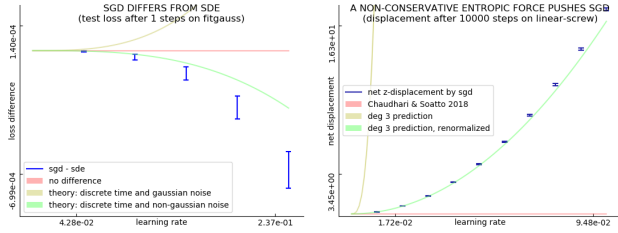


Figure 5. **Left:** SGD’s difference from SDE after $\eta T \approx 10^{-1}$ with maximal coarseness on the gaussian-fit problem. Two effects not modeled by SDE — time-discretization and non-gaussian noise oppose on this landscape but do not completely cancel. **Right:** On the linear screw landscape, 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. Hessians and covariances are uniformly bounded throughout the valley, and this effect appears at all sufficiently small η s with strength η^2 . Thus, the effect is not a pathological artifact of well-chosen learning rate or divergent covariance noise. Observe that the net displacement of $\approx 10^{1.5}$ well exceeds the z -period of 2π .

mediates this distinction.

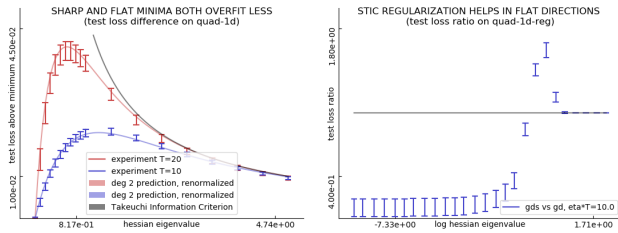


Figure 6. **Left:** For artificial quadratic landscapes with fixed covariance and a range of Hessians, initialized at the true minimum, the test losses after fixed- ηT optimization are smallest for very small and very large curvatures. This evidences our renormalized theory’s prediction that both sharp and flat minima overfit less! In particular, the Takeuchi estimate’s singularity is, as predicted, suppressed. **Right:** 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 .

Because the TIC estimates a smooth hypothesis class’s generalization gap, it is tempting to use it as an additive regularization term. However, the TIC is singular where the Hessian is (see Figure 3), and as such gives insensible results for over-parameterized models. Indeed, a prior attempt ran into numerical difficulties requiring an arbitrary cutoff (Dixon & Ward, 2018).

Fortunately, by Corollary 7, the implicit regularization of gradient descent both demands and enables a singularity-removing correction to the TIC — see Figure 6. The re-

sulting *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 the mean-estimation problem validate STIC for such model selection, especially when C/N dwarves H as in the noisy, small- N regime (see Figure 6). Because matrix exponentiation is expensive, STIC regularization without further approximations is most useful for models of small dimension on very noisy or scanty 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 making a continuous-time approximation by instead Taylor-expanding around the learning rate $\eta = 0$. In fact, we develop a diagrammatic method for evaluating each Taylor term that is inspired by the field theory methods popularized by Dyson (1949a). We then quantify the overfitting effects of batch size and epoch number, and based on this analysis, propose a regularizing term that causes large-batch GD to emulate small-batch SGD, thus establishing 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, which shows how generalization depends on the underlying metric and on the form of gradient noise, reconciles those seemingly clashing claims.

Others have imported the perturbative methods of physics to analyze descent dynamics: Dyer & Gur-Ari (2019) perturb in inverse network width, employing ’t Hooft diagrams to compute deviations of a specific class of deep architectures from Gaussian processes. Meanwhile, (Chaudhari & Soatto, 2018) and Li et al. (2017) perturb in learning rate to second order by approximating noise between updates as gaussian and uncorrelated. This approach does not generalize to higher orders, and, because correlations and heavy tails are essential obstacles to concentration of measure and hence to generalization, it does not model the generalization behavior of SGD. By contrast, we use Penrose diagrams to compute test and train losses to arbitrary order in learning rate, quantifying the effect of non-gaussian and correlated noise. Our method accounts for optimization 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’. We hence extend Roberts (2018) beyond leading order and beyond 2 time steps, allowing us to compare, for instance, the expected test losses of multi-epoch and single-epoch SGD.


6. Conclusion

We present 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* co-

variance. 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).

The diagram method invites further exploration of Lagrangian characterizations 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 $\theta \leftarrow \theta - (\eta^{-1} + \lambda \nabla \nabla l_t(\theta))^{-1} \nabla l_t(\theta)$ parameterized by small η, λ — admit diagrammatic analysis when we represent the λ term as a second type of diagram node. Though diagrams suffice for computations, it is Lagrangians that most deeply illuminate scaling and conservation laws.

Conjecture 1. *To leading order in η , the generalization gap of SGD increases with sectional curvature.*

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 Bonnabel (2013), 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.

acknowledgements commented

References

- 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

*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.

- 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 systèmes 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.
- Dyson, F. The radiation theories of tomonaga, schwinger, and feynman. *Physical Review*, 1949a.
- Dyson, F. The s matrix in quantum electrodynamics. *Physical Review*, 1949b.
- Feynman, R. A space-time approach to quantum electrodynamics. *Physical Review*, 1949.
- Fong, B. and Spivak, D. An invitation to applied category theory. *Cambridge University Press*, 2019.
- Gell-Mann, M. and Goldberger, M. Scattering of low-energy photons by particles of spin $\frac{1}{2}$. *Physical Review*, 1954.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. *ICLR*, 2019a.

Yaïda, 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.

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 three 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 char-

acter 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 a pager, 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.

THE INTERPRETATION OF A DIAGRAM

FILL IN

EXAMPLE: THE DIAGRAM 

FILL IN

EXAMPLE: THE DIAGRAM 

FILL IN

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 (single-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 dia-

grams.

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,  is shorthand for  - , 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  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  and . 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. .

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 η^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: . 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).  is shorthand for  - , 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 η^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: , , , and . 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 descendants. Thus, of the four possible diagrams (namely, , , , and ), one has no embeddings, hence has coefficient zero, hence may be ignored. Explicitly, the remaining diagrams are: , , and .

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 descendant (one-epoch). To be explicit, we write out the fuzzy outlines as the differences they represent:  - ,  - , and  - .

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 .


RENORMALIZED TECHNIQUE

Though the unrenormalized technique of Theorem 1 is simple, the renormalized technique of Theorem 2 gives conver-


gent results even for large ηT .

EXAMPLES OF SUMMING A DIAGRAM'S SPACETIME EMBEDDINGS

A.4. The Generalization Gap of SGD. Proof of Corollary 7.

The relevant irreducible diagram is . **FILL IN**

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

The relevant irreducible diagram is . **FILL IN**

A.6. SGD vs ODE. Proof of Part of Corollary 1

The rest of the proof is in Appendix B. **FILL IN**

A.7. The Effect of Batch Size and Epochs. When Does SGD Outperform GD? Proof of Proposition 2 and Corollary 3.


FILL IN

A.8. Project: Nongaussian Noise for Large Times.

We saved a result just for you, the reader! This result is absent from this paper and the literature. It might be *you* who first writes it out fully. The result is obtained 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 single-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 single-epoch, singleton-batch SGD?

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

Remark (upon finishing). Congratulations!

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

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

three 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.

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,  and . The first 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  in distinct epochs of GD's spacetime as there are in distinct epochs of SGD's spacetime. However, each same-epoch embedding of  within any one epoch of GD's spacetime corresponds by vertical shifts to an embedding of  in SGD. There are $MN \binom{N}{2}$ many such embeddings in GD's spacetime, so GD's test loss exceeds SGD's by

$$\left(\frac{\eta}{N}\right)^2 MN \binom{N}{2} \text{$$

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

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

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:

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

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

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

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

$$\begin{aligned} \mathbb{E}[(a(T))^2] &= \sum_{0 \leq t < T} \sum_{0 \leq s < T} \nabla l_{n_t} \nabla l_{n_s} \\ &= M^2 N(N-1) \mathbb{E}[\nabla l]^2 + M^2 N \mathbb{E}[(\nabla l)^2] \end{aligned}$$

Likewise,

$$\begin{aligned} \mathbb{E}[b(T)] &= \sum_{0 \leq t_0 < t_1 < T} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}} \\ &= \frac{M^2 N(N-1)}{2} \mathbb{E}[\nabla^2 l] \mathbb{E}[\nabla l] + \\ &\quad \frac{M(M-1)N}{2} \mathbb{E}[(\nabla^2 l)(\nabla l)] \end{aligned}$$

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

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

and

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

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

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

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 Nongaussian Noise at a Minimum. Proof of Part of Corollary 1.

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 nongaussian 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 nongaussian 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



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:

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

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

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

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:

$$\begin{aligned} \mathbb{E}[a_T b_T] &= - \sum_t \sum_{t_0 < t_1} \mathbb{E}[\nabla l_{n_t} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}] \\ &= - \sum_{t_0 < t_1} \sum_{t \notin \{t_0, t_1\}} \mathbb{E}[\nabla l_{n_t}] \mathbb{E}[\nabla^2 l_{n_{t_1}}] \mathbb{E}[\nabla l_{n_{t_0}}] \\ &\quad - \sum_{t_0 < t_1} \sum_{t=t_0} \mathbb{E}[\nabla l_{n_t} \nabla l_{n_{t_0}}] \mathbb{E}[\nabla^2 l_{n_{t_1}}] \\ &\quad - \sum_{t_0 < t_1} \sum_{t=t_1} \mathbb{E}[\nabla l_{n_t} \nabla^2 l_{n_{t_1}}] \mathbb{E}[\nabla l_{n_{t_0}}] \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}[a_T^3] &= - \sum_{t_a, t_b, t_c} \mathbb{E}[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}} \nabla l_{n_{t_c}}] \\ &= - \sum_{\substack{t_a, t_b, t_c \\ \text{disjoint}}} \mathbb{E}[\nabla l_{n_{t_a}}] \mathbb{E}[\nabla l_{n_{t_b}}] \mathbb{E}[\nabla l_{n_{t_c}}] \\ &\quad - 3 \sum_{t_a=t_b \neq t_c} \mathbb{E}[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}}] \mathbb{E}[\nabla l_{n_{t_c}}] \\ &\quad - \sum_{t_a=t_b=t_c} \mathbb{E}[\nabla l_{n_{t_a}} \nabla l_{n_{t_b}} \nabla l_{n_{t_c}}] \end{aligned}$$

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

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

B.3. The Effect of Inter-Epoch Shuffling. Proof of Corollary 2.

We identify the leading order effect of shuffling on test loss for pure, multi-epoch SGD. It is much harder to see by direct perturbation than by diagrams that this is an order η^4 effect, so for fairness we will assume this order it already known for both methods.

DIAGRAM METHOD

FILL IN

DIRECT PERTURBATION

We compute the displacement $\theta_T - \theta_0$ to order η^4 for vanilla SGD. Expanding $\theta_t \in \theta_0 + \eta a_t + \eta^2 b_t + \eta^3 c_t + \eta^4 d_t + o(\eta^4)$,

we find:

$$\begin{aligned}
 \theta_{t+1} &= \theta_t - \eta \nabla l_{n_t}(\theta_t) \\
 &\in \theta_0 + \eta a_t + \eta^2 b_t + \eta^3 c_t + \eta^4 d_t \\
 &\quad - \eta \left(\nabla l_{n_t} + \nabla^2 l_{n_t}(\eta a_t + \eta^2 b_t + \eta^3 c_t) \right) \\
 &\quad - \eta \left(\frac{1}{2} \nabla^3 l_{n_t}(\eta a_t + \eta^2 b_t)^2 + \frac{1}{6} \nabla^4 l_{n_t}(\eta a_t)^3 \right) + o(\eta^4) \\
 &= \theta_0 + \eta (a_t - \nabla l_{n_t}) \\
 &\quad + \eta^2 (b_t - \nabla^2 l_{n_t} a_t) \\
 &\quad + \eta^3 \left(c_t - \nabla^2 l_{n_t} b_t - \frac{1}{2} \nabla^3 l_{n_t} a_t^2 \right) \\
 &\quad + \eta^4 \left(d_t - \nabla^2 l_{n_t} c_t - \frac{1}{2} \nabla^3 l_{n_t} b_t^2 - \frac{1}{6} \nabla^4 l_{n_t} a_t^3 \right) + o(\eta^4)
 \end{aligned}$$

We thus have the recurrences $a_{t+1} = a_t - \nabla l_{n_t}$, $b_{t+1} = b_t - \nabla^2 l_{n_t} a_t$, $c_{t+1} = c_t - \nabla^2 l_{n_t} b_t - \frac{1}{2} \nabla^3 l_{n_t} a_t^2$, and $d_{t+1} = d_t - \nabla^2 l_{n_t} c_t - \frac{1}{2} \nabla^3 l_{n_t} b_t^2 - \frac{1}{6} \nabla^4 l_{n_t} a_t^3$ with solutions:

$$\begin{aligned}
 \eta a_t &= -\eta \sum_t \nabla l_{n_t} \\
 \eta^2 b_t &= +\eta^2 \sum_{t_0 < t_1} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}} \\
 \eta^3 c_t^\mu &= -\sum_{t_0 < t_1 < t_2} \nabla^\mu \nabla_\nu l_{n_{t_2}} \nabla^\nu \nabla_\sigma l_{n_{t_1}} \nabla^\sigma l_{n_{t_0}} \\
 &\quad - \frac{1}{2} \sum_{t_a, t_b < t} \nabla^\mu \nabla^\nu \nabla^\sigma l_{n_t} \nabla_\nu l_{n_a} \nabla_\sigma l_{n_b} \\
 \eta^4 d_t^\mu &= -\sum_{t_0 < t_1 < t_2} \nabla^\mu \nabla_\nu l_{n_{t_2}} \nabla^\nu \nabla_\sigma l_{n_{t_1}} \nabla^\sigma l_{n_{t_0}} \\
 &\quad - \frac{1}{2} \sum_{t_a, t_b < t} \nabla^\mu \nabla^\nu \nabla^\sigma l_{n_t} \nabla_\nu l_{n_a} \nabla_\sigma l_{n_b}
 \end{aligned}$$

We use tensor indices above because the contraction pattern would otherwise be ambiguous.

FILL IN

C. Other Perturbative Calculations

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 $\text{parts}(D)$ give the \sim -parts of D . An \mathcal{S} -map from D to $(\mathcal{G} \circ \mathcal{F})^{|\text{parts}(D)|}(\text{empty set})$ 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 $f : x \rightarrow y$, let $|\text{Aut}_f(x)|$ 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 $\pi : S \rightarrow [N]$ 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 θ s.

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 calling using the word “set” for *ordered* sequences of training points.

E. Proofs of Theorems

E.1. Regularity Hypotheses

We assume throughout this work several 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 shared (but θ -dependent) radii of convergence. *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).

PATHOLOGIES ILLUSTRATING THE HYPOTHESES’
UTILITY

FILL IN

E.2. Dyson Series for Iterative Optimizers

We first give intuition, then worry about ϵ s and δ s.

THE KEY LEMMA: PROOF IDEA

Intuitively, since ∇ Lie-generates translation, the operator $\exp(-\eta^{\mu\nu} g_\mu \nabla_\nu)$ performs translation by $-\eta g$. In particular, the case $g = \nabla l_t(\theta)$ effects a gradient step on the t th 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 ,

^{*}A metric-independent way of expressing this boundedness constraint is that the gradients all lie in some subset $S \subseteq TM$ of the tangent bundle of weight space, where, for any compact $C \subseteq M$, we have that the topological pullback — of $S \hookrightarrow TM \twoheadrightarrow M$ 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.

collecting terms of like degree while maintaining the order of operators:

$$\begin{aligned} s(\theta_T) &= \left(\prod_{0 \leq t < T} \left(\sum_{0 \leq d_t} \frac{(-\eta^{\mu\nu} g_\mu \nabla_\nu)^{d_t}}{d_t!} \right) \Big|_{g=\nabla l_t(\theta)} \right) s(\theta_0) \\ &= \sum_{0 \leq d < \infty} (-\eta)^d \sum_{\substack{(d_t: 0 \leq t < T) \\ \sum_t d_t = d}} \left(\prod_{0 \leq t < T} \frac{(g \nabla)^{d_t}}{d_t!} \Big|_{g=\nabla l_t(\theta)} \right) s(\theta_0) \end{aligned}$$

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 convenience, 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 shared radius of convergence around θ_0 , without loss also ρ . 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

$$2\eta T b < \rho \tilde{\eta} \quad (4)$$

as norms, stochastic gradient after T steps on any train set will necessarily stay within the ρ -ball. In fact, we will see that the factor of 2 ensures that SGD initialized at any point within a $\rho/2$ ball 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

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

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 coefficients of l_t, s are bounded by

some exponential decay in d , by assumption at an x -uniform rate, we have absolute convergence and may rearrange sums. We choose to group by total degree:

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

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 show that

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

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 (of course 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 messy 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, though 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 $\pi : S \rightarrow [N]$ that induces \sim , and natural d — that

$$(-\eta)^d \sum_{\substack{(d_t; 0 \leq t < T) \\ \sum_t d_t = d}} \left(\prod_{0 \leq t < T} \frac{(g \nabla)^{d_t}}{d_t!} \Big|_{g=\nabla l_t(\theta)} \right) l(\theta) = \sum_{\substack{D \in \text{im}(\mathcal{F}) \\ \text{with } d \text{ edges}}} \left(\sum_{f: D \rightarrow \mathcal{F}(S)} \frac{1}{|\text{Aut}_f(D)|} \right) \frac{\text{value}_\pi(D, f)}{B^d} \quad (6)$$

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

Remark 1 (Differentiating Products). The product rule of Leibniz easily generalizes to higher derivatives of finitary products:

$$\nabla^{|M|} \prod_{k \in K} p_k = \sum_{\nu: M \rightarrow K} \prod_{k \in K} (\nabla^{|\nu^{-1}(k)|} p_k)$$

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 \bullet . For the induction step, suppose S is a sequence of $\mathcal{M} = \min S \subseteq S$ followed by a strictly smaller

S and that the result is proven for (\tilde{d}, \tilde{S}) for every $\tilde{d} \leq d$. Let us group the terms in the left hand side of desideratum 6 by d_0 ; by applying the induction hypothesis with $\tilde{d} = d - d_0$, we find that that left hand side is:

$$\sum_{0 \leq d_0 \leq d} \sum_{\substack{\tilde{D} \in \text{im}(\mathcal{F}) \\ \text{with } d - d_0 \text{ edges}}} \frac{1}{d_0!} \sum_{\tilde{f}: \tilde{D} \rightarrow \mathcal{F}(\tilde{S})} \left(\frac{1}{|\text{Aut}_{\tilde{f}}(\tilde{D})|} \right) \cdot (-\eta)^{d_0} (g \nabla)^{d_0} \Big|_{g=\nabla l_0(\theta)} \frac{\text{value}_{\pi}(\tilde{D}, \tilde{f})}{B^{d-d_0}}$$

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 $(d - d_0 + 1)_0^{d_0}$ terms. In fact, if we expand out $g = \sum_{m \in \mathcal{M}} \nabla l_m(\theta)/B$ then there are $B^{d_0} (d - d_0 + 1)^{d_0}$ terms conveniently indexed by a pair of functions $\beta: [d_0] \rightarrow \mathcal{M}$ and $\nu: [d_0] \rightarrow \tilde{D}$. 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 $|(\beta \times \nu)^{-1}(n, \nu)|$ many additional edges from the cell of datapoint n at time 0 to the ν th node of \tilde{D} as embedded by \tilde{f} .

By the general Leibniz rule remarked on above, the sum over terms indexed by (β, ν) corresponds to a sum over embeddings f that restrict to \tilde{f} , whose terms are multiples of the corresponding and embedded 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 $f: D \rightarrow S$ are as claimed.

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

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

by the Orbit Stabilizer Theorem of elementary group theory.

It is thus enough to show that

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

We will show this by a direct bijection. First, observe that $f = \beta \sqcup \tilde{f}: [d_0] \sqcup \tilde{D} \rightarrow \mathcal{M} \sqcup \tilde{S}$. So each automorphism $\phi: D \rightarrow D$ that commutes with f induces both an automorphism $\mathcal{A} = \phi|_{\tilde{D}}: \tilde{D} \rightarrow \tilde{D}$ that commutes with \tilde{f} together with the data of a map $\mathcal{B} = \phi|_{[d_0]}: [d_0] \rightarrow [d_0]$ that both commutes with β . However, not every such pair of maps arises from a ϕ . For, in order for $\mathcal{A} \sqcup \mathcal{B}: D \rightarrow D$ 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 $|\text{Aut}_{\tilde{f}}(\tilde{D})|$ 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 $\prod_{m \in \mathcal{M}} |(\beta \times \nu)^{-1}(m, i)|!$ many choices for σ_i . The counting claim follows and with it the correctness of coefficients.

The analogous statement about generalization gaps follows similarly when we use $\sum_n l_n/N$ instead of l as the value for s . The Path Integral Theorem (Theorem 1) follows.

Remark 2 (The Case of Vanilla SGD). Because the space-time of vanilla SGD permits only embeddings with each fuzzy-connected-component in one distinct cell, the embeddings of a diagram are easily counted using factorials per Proposition 2. That proposition immediately follows from the now-proven Theorem 1.

E.4. Coefficient Convergence upon Renormalization

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 minimal element, 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 by Orbit-Stabilizer. The second we handle by Mobius Sums.

RENORMALIZATION THEOREM: MOBIUS SUM FOR GENERAL DIAGRAMS

FILL IN

RENORMALIZATION THEOREM: PROOF

We focus on test loss instead of generalization gap; the proofs are similar. The difference from the noiseless 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 discussion on Mobius Sums. Beware that the relatively noiseless embeddings may have illegal collisions of non-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 1.

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

$$\sum_{f:p \rightarrow S} \sum_{\substack{\tilde{f}:\sigma \rightarrow S \\ \exists i_*: f = \tilde{f} \circ i_*}} \frac{1}{|\text{Aut}_{\tilde{f}}(\sigma)|} = \sum_{f:p \rightarrow S} \sum_{\substack{\tilde{f}:\sigma \rightarrow S \\ \exists i_*: f = \tilde{f} \circ i_*}} \sum_{i:p \rightarrow \sigma} \frac{1}{|\text{Aut}_f(\rho)|} \quad (7)$$

Here, f is considered up to equivalence by precomposition by an automorphism of ρ ; likewise for \tilde{f} and automorphisms of σ ; i ranges through maps that induce isomorphisms of geometric realizations; and i is considered equivalent to \hat{i} when for some automorphism $\phi \in \text{Aut}_{\tilde{f}}(\sigma)$, we have $\hat{i} = i \circ \phi$. Let the set of all such i s under this equivalence relation be called X . The left hand side is the expression of Theorem 1 and the right hand side is the expression of Theorem 2; we have introduced redundant sums to structurally align the two expressions on the page.

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

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

We will prove this using Orbit-Stabilizer by presenting an action of $\text{Aut}_{\tilde{f}}(\sigma)$ on X . We simply use precomposition so that $\psi \in \text{Aut}_{\tilde{f}}(\sigma)$ sends $i \in X$ to $i \circ \psi$. Since $f \circ \psi = \tilde{f} \circ i \circ \psi$, $i \circ \psi \in X$. 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 size of i 's the stabilizer is $|\text{Aut}_{\tilde{f}}(\rho)|$. For, when $i \sim i \circ \psi$ via $\phi \in \text{Aut}_{\tilde{f}}(\sigma)$, 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 $\text{stabilizer}(i) \leftrightarrow \text{Aut}_{\tilde{f}}(\rho)$ seen to be well-defined and injective by the strictly increasing nature of structure set morphisms together with the fact that i must induce isomorphisms of geometric realizations. Conversely, every ψ that stabilizes enjoys *only* one ϕ via which $i \sim i \circ \psi$, 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 $\psi \in \text{Aut}_{\tilde{f}}(\sigma)$ such that $i_B \sim i_A \circ \psi$ by $\phi = \text{identity}$. 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 the rest of Theorem 2, we assume that H is positive. Then, for any m , the propagator $(I - \eta H)^{\otimes m^t}$ converges via an exponential decay with t to 0 (with a rate dependent on m). The Renormalization Theorem. Since up to degree d only a finite number of diagrams exist and hence only a finite

number of possible m s, 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 $\prod_{0 \leq d < D} X_d$, 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 $\mathbf{A}_{0 \leq s < 2S} x_0^{(s)} x_1^{(s)}$, where \mathbf{A} denotes averaging. In fact, the following also works:

$$S \left(\mathbf{A}_{0 \leq s < S} x_0^{(s)} \right) \left(\mathbf{A}_{0 \leq s < S} x_1^{(s)} \right) + (1 - S) \left(\mathbf{A}_{0 \leq s < S} x_0^{(s)} \right) \left(\mathbf{A}_{S \leq s < 2S} x_1^{(s)} \right) \quad (8)$$

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 $\mu : [D] \rightarrow [D]$ with $\mu(d) \leq d$ and $\mu \circ \mu = \mu$. Note that $|\mu| := |\text{im}(\mu)|$ counts μ 's parts. We then define the statistic

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

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

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

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

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

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

because

$$\begin{aligned}
 \mathbb{E}[S^D \{x\}_\mu] &= \mathbb{E} \left[\sum_{(0 \leq s_d < S) \in [S]^D} \prod_{0 \leq d < D} x_d^{(\mu(d) \cdot S + s_d)} \right] \\
 &= \sum_{\substack{(0 \leq s_d < S) \\ \in [S]^D}} \mathbb{E} \left[\prod_{0 \leq d < D} x_d^{(\min\{\bar{d} : \mu(\bar{d}) \cdot S + s_{\bar{d}} = \mu(d) \cdot S + s_d\})} \right] \\
 &= \sum_{\tau} \left| \left\{ \left(\begin{array}{c} (0 \leq s_d < S) \in [S]^D : \\ \mu(d) = \mu(\bar{d}) \Leftrightarrow \tau(d) = \tau(\bar{d}) \end{array} \right) \right\} \right| \langle x \rangle_{\tau} \\
 &= \sum_{\tau \leq \mu} \left(\prod_{0 \leq d < D} \frac{S!}{(S - |\tau(\mu^{-1}(d))|)!} \right) \langle x \rangle_{\tau}
 \end{aligned}$$

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

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

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:

$$\langle x \rangle_{\boxed{012}} = S^2 \{x\}_{\boxed{012}} - \frac{(S-1)!}{(S-3)!} \{x\}_{\boxed{01} \boxed{2}} - 3 \frac{(S-1)!}{(S-2)!} \{x\}_{\boxed{0} \boxed{12}}$$

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

G. Loss Landscapes Used for Experiments

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 and Fashion-MNIST datasets. 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.

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 scanty-data regime, e.g. 10^1 datapoints out of $\sim 10^{4.5}$ dirac masses, so overlaps are diluted.

H. Additional Figures

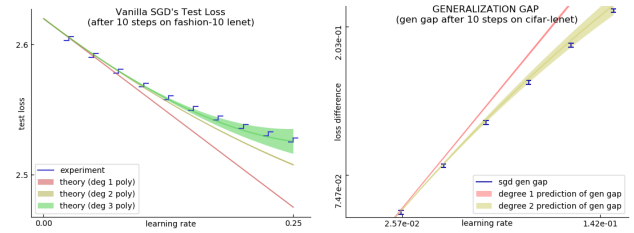


Figure 7. **Left:** Test loss vs learning rate on an image classification task. For the instance shown and all 11 other initializations unshown, the degree-3 prediction agrees with experiment through $\eta T \approx 10^0$. **Right:** Generalization gap (test minus train) vs learning rate on an image 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.

I. Glossary


As our work uses physical methods to solve problems of computer science, it may contain unfamiliar terminology. We strive in the main text to explain vital terms before their point of use; we hope this glossary complements that attempt. The entries are intuitive and, while mathematically imprecise, sufficient for daily use.

I.1. Terminology

affine manifold a shape closed under a well-defined 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 okay because the gradients will all be clockwise and counterclockwise, directions under which the circle is closed. We do not deal with projection onto feasible sets in this work. We treat weight spaces as affine manifolds in this work not for the sake of generality but for the sake of conceptual encapsulation: the results of this work are coordinate-invariant!

akaike information criterion an estimate of generalization gap: $(\text{number of parameters})/N$. 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 not (smooth or) 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 zero curl. The ODE approximation to gradient 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 product of a row and 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 an image of a cat or of a dog, but by abuse of notation also an index into the train set of the corresponding image, and by further abuse also the induced loss function of a given neural network on that corresponding image.

diagram a representation of an interaction 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.

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

entropic force a macroscopic force arising from a tendency toward disorder and marked by strong temperature (noise) dependence. For instance, rubber bands are

stretchy because of an entropic force. SGD tends toward minima flat with respect to the covariance due to a non-conservative entropic force.

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 $B \cdot M$.

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

geometric realization (of a diagram) the data of a diagram that remains when we consider chains of thin edges as equivalent to single edges. See also irreducible.

index (of a train set) a name for a datapoint that singles it out from among a train set.

index (of a stochastic process) a name for a random variable among the collection of random variables.

index (of a tensor) a name for a numpy axis. For example, a shape $a \times b \times c$ tensor will have three index positions.

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 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, considered as a function from weights θ and data x to losses $l_x(\theta)$.

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 nonzero curl. In striking contrast to the ODE approximation to gradient descent, SGD experiences a non-conservative force. See also conservative.

overfitting the process 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 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.

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

spacetime the stage on which weights and data interact. A grid-like summary of which training points participate in which gradient updates through time.

spring the familiar simple machine with energy quadratic in its degrees of freedom. 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: $C_{\mu\nu}((I - \exp(-\eta TH))^\nu)_\lambda (H^{-1})^{\lambda\mu}$.

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.

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 , though 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
	$+\eta^{\mu\nu}\eta^{\lambda\rho}(\nabla_\lambda C_{\mu\nu})G_\rho/2$	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
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(G_\sigma \nabla_\pi(C_{\mu\lambda}H_{\nu\rho})/2 - C_{\mu\lambda}G_\sigma J_{\nu\rho\pi})$	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
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\mathbb{E}[\nabla_\mu l_x \nabla_\lambda l_x \nabla_\sigma l_x] - G_\mu G_\lambda G_\sigma)J_{\nu\rho\pi}$	A datapoint thrice affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\nabla_\lambda C_{\mu\nu})H_{\rho\sigma}G_\pi/2$	A datapoint affects a later instance of itself, which affects a different point, which affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\mathbb{E}[\nabla_\nu \nabla_\lambda l_x \nabla_\rho \nabla_\sigma l_x] - H_{\nu\lambda}H_{\rho\sigma})G_\mu G_\pi$	A datapoint affects another datapoint, which in turn affects a later instance of itself, which affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\mathbb{E}[\nabla_\mu l_x \nabla_\nu \nabla_\lambda l_x \nabla_\rho \nabla_\sigma l_x] - G_\mu H_{\nu\lambda}H_{\rho\sigma})G_\pi$	A datapoint affects a later instance of itself, which affects a yet later instance of itself, which affects the test loss
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\mathbb{E}[\nabla_\mu l_x \nabla_\lambda l_x \nabla_\rho \nabla_\sigma l_x] - G_\mu G_\lambda H_{\rho\sigma})H_{\nu\pi}$	A datapoint affects the test loss twice, once directly, and once through a later instance of itself
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\mathbb{E}[\nabla_\lambda l_x \nabla_\nu \nabla_\rho \nabla_\sigma l_x] - G_\lambda J_{\nu\rho\sigma})G_\mu G_\pi$	A datapoint affects the test loss through a later encountered instance of itself, which was earlier affected by a different datapoint
	$-\eta^{\mu\nu}\eta^{\lambda\rho}\eta^{\sigma\pi}(\mathbb{E}[\nabla_\mu l_x \nabla_\lambda l_x \nabla_\nu \nabla_\rho \nabla_\sigma l_x] - G_\mu G_\lambda J_{\nu\rho\sigma})G_\pi$	A datapoint twice affects a later instance of itself that in turn affects the test loss