

# A Space-Time Approach to Analyzing Stochastic Gradient Descent

Anonymous Authors<sup>1</sup>

## Abstract

We present a diagrammatic calculus for reasoning about the behavior, at small learning rates, of SGD and its variants. We interpret the diagrams as histories of scattering events, thus offering a new physical analogy for descent. Illustrating this technique, we construct a regularizing term that causes large-batch GD to emulate small-batch SGD, present a model-selection heuristic that depends only on statistics measured before optimization, and exhibit a counter-intuitive loss landscape wherein SGD eternally cycles counterclockwise around a circle of minima.

IDEA: ASCENT?

IDEA: TEST as EXP of TRAIN?

IDEA: CHLADNI

Fashion Mnist and CIFAR 10

Correct Thm 1 to address nonconstant batch size

## 1. Introduction

Stochastic gradient descent (SGD) decreases an unknown objective  $l$  by performing discrete-time steepest descent on noisy estimates of  $l$ . A key question is how the noise affects the final objective value. We connect SGD dynamics to physical scattering theory, thus providing a quantitative and qualitative toolkit for answering this question.

Specifically, we derive a diagram-based formalism for reasoning about SGD via a path integral over possible interactions between weights and data. The formalism permits perturbative analysis, leading to predictions of learning curves for small  $\eta$ . Unlike the continuous-time limits of previous work, this framework models discrete time, and with it, the potential **non-Gaussianity** of noise. We thus obtain new results quantifying the **effect of epoch number, batch size,**

**and momentum** on SGD test loss. We also contrast SGD against popular continuous-time approximations such as ordinary or stochastic differential equations (ODE, SDE).

Path integrals offer not only quantitative predictions but also an exciting new viewpoint — that of iterative optimization as a **scattering process**. Much as individual Feynman diagrams (see [Dyson \(1949a\)](#)) depict how local particle interactions compose into global outcomes, our diagrams depict how individual SGD updates influence each other before affecting a final test loss. In fact, we import from physics tools such as **crossing symmetries** (see [Dyson \(1949b\)](#)) and **re-normalization** (see [Gell-Mann & Goldberger \(1954\)](#)) to simplify our calculations and refine our estimates. The diagrams’ combinatorial properties yield precise qualitative conclusions as well, for instance that to order  $\eta^2$ , **inter-epoch** shuffling does not affect expected test loss.

## 2. Background and Notation

### 2.1. A Smooth Stage: Tensor Conventions

We adopt summation notation for Greek, suppressing indices when clear. To expedite dimensional analysis, we regard the learning rate as an inverse metric  $\eta^{\mu\nu}$  that converts a gradient into a displacement (([Bonnabel, 2013](#))). We use  $\eta$  to raise indices; for example, with  $C$  denoting the covariance of gradients, its “trace” will be  $C_\mu^\mu = \eta^{\mu\nu} C_{\mu\nu}$ . Standard syntactic constraints make manifest which expressions transform naturally with respect to optimization dynamics.

We assume that all polynomials of the 0th and higher derivatives of the losses  $l_n$ , considered as random functions on weight space, have infinitely differentiable expectations.

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

### 2.2. Combinatorial Costumes: Structure Sets

We make use of *structure sets*, i.e. sets  $S$  equipped with a preorder  $\leq$  and an equivalence relation  $\sim$ . The morphisms of structure sets are strictly increasing maps that preserve  $\sim$  and its negation. A structure set is *pointed* if it has a unique 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  $\mathcal{F}, \mathcal{G}$ .

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

A *diagram* is a rooted tree equipped with an equivalence relation  $\sim$  on nodes. We draw the tree with thin edges, with the root at the far right, and we indicate  $\sim$  with fuzzy ties. By reading the tree as a Hasse graph, we see that each diagram  $D$  induces a structure set, by abuse of notation also named  $D$ . An  $\mathcal{S}$ -map from  $D$  to  $[P] = (\mathcal{G} \circ \mathcal{F})^P$  (empty set) is an ordering of  $D$ , where  $P$  counts  $D$ 's equivalence classes. Let  $o(D)$  count orderings of  $D$ .

Fong gives a swift introduction to these category theoretic and diagrammatic ideas (2019).

### 2.3. 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 structure set  $S$  whose preorder is a total preorder with element  $i$  inside strongly connected component  $C(i)$ : 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  $l_M = \frac{1}{M} \sum_{m \in M} l_{\pi(m)}$  is a batch average. Since the distribution of  $l_n$  is permutation invariant, the non-canonical choice of  $\pi$  does not affect the distribution of output  $\theta$ s.

Of special interest are structure sets that divide 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*.

## 3. Diagram Calculus for SGD

### 3.1. Role of Diagrams

Suppose  $s$  is smooth on weight space; e.g.  $s$  may be a test loss. We track  $s(\theta)$  as  $\theta$  evolves by SGD:

**Key Lemma 1.** *The Maclaurin series of  $s(\theta_T)$  with respect to  $\eta$  is:*

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

In averaging over training sets ( $l_t : 0 \leq t < T$ ) 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 of  $\eta$ .

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

$$\begin{aligned} & -(\nabla^4 l_c)(\nabla^\mu l_c)(\nabla_\lambda \nabla_\mu \nabla^\nu l_b)(\nabla_\nu l_a) - (\nabla^4 l_c)(\nabla^\mu l_c)(\nabla_\lambda \nabla^\nu l_b)(\nabla_\mu \nabla_\nu l_a) \\ & -(\nabla^4 l_c)(\nabla^\mu l_c)(\nabla_\mu \nabla^\nu l_b)(\nabla_\lambda \nabla_\nu l_a) - (\nabla^4 l_c)(\nabla^\mu l_c)(\nabla^\nu l_b)(\nabla_\lambda \nabla_\mu \nabla_\nu l_a) \end{aligned}$$

To reduce clutter, we adapt the string notation of Penrose (1971). Then, in expectation over  $(l_c, l_b, l_a)$  drawn i.i.d.:

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

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

Above, each node corresponds to an  $l_n$  (here, red for  $l_c$ , green for  $l_b$ , blue for  $l_a$ ), differentiated  $g$  times for a degree- $g$  node (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  $l_c$ ). The colors are redundant with the fuzzy ties. The value  $v(D)$  of a diagram  $D$  is the expected value of the corresponding tensor expression.

Crucially, for a fixed, i.i.d. distribution over  $(l_c, l_b, l_a)$ , **the topology of a diagram determines its value**. For instance,

$$\text{diagram 1} = \text{diagram 2}.$$

Thus follows the simplification of equation 3. We may convert back to explicit tensor expressions, invoking independence between untied nodes to factor the expression. However, as we will see, the diagrams offer physical intuition, streamline computations, and determine useful unbiased estimators of the statistics they represent.

We define a diagram with fuzzy outlines to be the difference between the fuzzy tied and untied versions :  $\text{diagram 1} - \text{diagram 2}$ .

The recipes for writing down test (or train) losses of SGD and its variants are straight-forward in the diagram notation because they reduce the problem of evaluating the previous dynamical expressions to the problem of counting isomorphic graphs. The more complicated the direct computation, the greater the savings of using diagrams. An appendix provides details and proofs for a variety of situations. For now, we focus on the test loss of SGD.

### 3.2. Recipe for the Test Loss of SGD

Our results all follow from this theorem and its analogues. Throughout, the  $d$ -edged diagrams give the order  $\eta^d$  terms.

**Theorem 1.** *SGD's expected test loss has a Maclaurin series given as a weighted sum of diagrams:*

$$\sum_{D \in \text{im}(\mathcal{F})} \left( \sum_{f: D \rightarrow \mathcal{F}(S)} \prod_{i \in S} \frac{|C(i)|^{f^{-1}(i)}}{|f^{-1}(i)|!} \right) v(D) \quad (4)$$

Here,  $D$  is (an isomorphism class of) a diagram of form  $\mathcal{F}(T)$  and  $f$  is a morphism in  $\mathcal{P}$ .

**Theorem 2.** SGD's expected generalization gap (test loss minus train loss) is formula 4 with  $v(D)$  replaced by  $\sum_{p \in D/\sim_D} v(D_p)/N$ . Here,  $p$  ranges through equivalence classes of  $D$ , and  $D_p$  is  $D$  with a fuzzy outline connecting  $D$ 's maximal node to  $p$ , e.g.  $(\text{---})_{p=\bullet} = \text{---}\bullet$ .

In the special case of  $B = 1, M = 1$ :

**Proposition 1.** The order  $\eta^d$  contribution to the expected test loss of one-epoch SGD with singleton batches is:

$$\frac{(-1)^d}{d!} \sum_{D \in \text{im}(\mathcal{F})} o(D) \binom{N}{P-1} \binom{d}{d_0, \dots, d_{P-1}} v(D) \quad (5)$$

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.

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,  $\eta N$  measures the **physical time** of descent, and  $1/N$  measures **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.

**Corollary 1.** For one-epoch SGD on singleton batches through fixed physical time  $T$ : the order  $N^{-1}$  deviation of SGD's test loss from ODE's is  $(T^2 N^{-1}/2)$   $\text{---}\bullet$ . The order  $N^{-2}$  deviation of SGD's test loss due to non-gaussian noise is  $-(T^3 N^{-2}/6)(\text{---}\bullet\text{---}\bullet - 3 \text{---}\bullet\text{---}\bullet)$ .

For finite  $N$ , these effects make SDE different from SGD. SDE also fails to model the correlations between updates in multiepoch SGD. On the other hand, in the  $N = \infty$  limit for which SDE matches SGD, optimization and generalization become computationally intractable and trivial, respectively.

A quick combinatorial argument shows:

**Corollary 2.** To order  $\eta^2$ , inter-epoch shuffling doesn't affect SGD's expected test loss.

Indeed, for any inter-epoch shuffling scheme:

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

$$\begin{aligned} & + MN \text{---}\bullet + MN \left( MN - \frac{1}{2} \right) \text{---}\bullet\text{---}\bullet \\ & + MN \left( \frac{M}{2} \right) \text{---}\bullet\text{---}\bullet + MN \left( \frac{M - \frac{1}{B}}{2} \right) \text{---}\bullet\text{---}\bullet \end{aligned}$$

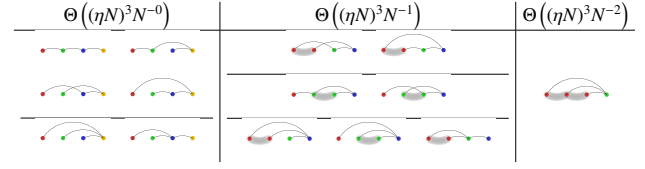


Table 1. Degree-3 scattering diagrams for  $B = M = 1$  SGD's test loss. **Left:**  $(d, P) = (3, 3)$ . Diagrams for ODE behavior. **Center:**  $(d, P) = (3, 2)$ . 1st order deviation of SGD away from ODE. **Right:**  $(d, P) = (3, 1)$ . 2nd order deviation of SGD from ODE with appearance of non-Gaussian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

**Corollary 3.** To order  $\eta^2$ , one-epoch SGD has  $\left( \frac{M-1}{M} \right) \left( \frac{B+1}{B} \right) \left( \frac{N}{2} \right)$   $\text{---}\bullet$  less test loss than  $M$ -epoch SGD with learning rate  $\eta/M$ .

Given an unbiased estimator  $\hat{C}$  of gradient covariance, we may get GD to mimic SGD:

**Corollary 4.** The expected test loss of pure SGD is, to order  $\eta^2$ , less than that of pure GD by  $\left( \frac{M(N-1)}{2} \right)$   $\text{---}\bullet$ . Moreover, GD on a modified loss  $\tilde{l}_n = l_n + \left( \frac{N-1}{4N} \right) \hat{C}_v^v(\theta)$  has an expected test loss that agrees with SGD's to second order.

### 3.3. Descent as Scattering

In sum, SGD's test loss is a weighted sum of diagrams, each  $d$ -edged diagram contributing to order  $\eta^d$ . We depict the  $\mathcal{S}$ -maps of  $f : D \rightarrow S$  Theorem 1 as an embedding of the graph  $D$  into the structure set or "spacetime"  $S$ . Thus, SGD's test loss is a sum over all embeddings in spacetime; see Figure (3.3). This loss depends on the shape of spacetime, which in encodes correlations between updates. To compute a test loss, we simply count embeddings. For instance, the order  $\eta^2$  diagrams in (3.3) all contribute the same amount to test loss, so we just need to find out how many such embedded diagrams there are. Likewise, as shown in Figure (3.3),

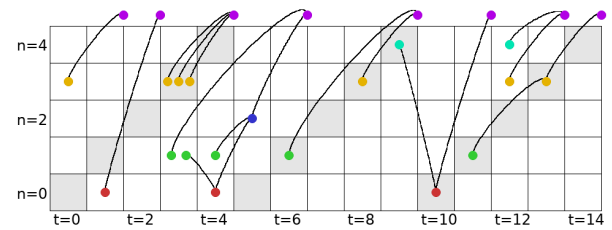


Figure 1. Some diagrams embedded in spacetime. The left four diagrams give order  $\eta^1, \eta^1, \eta^3$ , and  $\eta^5$  contributions to pure GD's test loss. The right four each contribute  $\eta^2 \mathbb{E}[\nabla^2] \mathbb{E}[\nabla]^2$ ; their equivalence demonstrates crossing symmetry. Only diagrams whose nodes fall within shaded diagonals contribute to pure SGD's test loss (with an extra factor  $N$  per edge).

the order  $\eta^2$  diagrams of pure GD and pure SGD are nearly

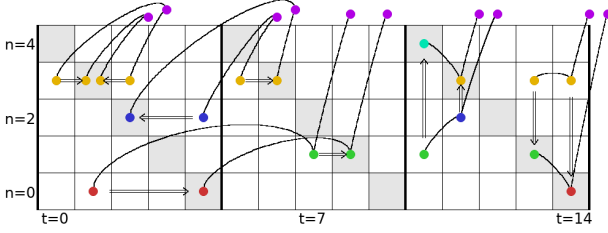


Figure 2. Comparison of pure GD's vs pure SGD's test loss. We may normalize almost every order  $\eta^2$  GD diagram to an equivalent SGD diagram by horizontal or vertical shifts (see left ten diagrams). By contrast,  $MN\binom{N}{2}$  many  $\text{---}\text{---}\text{---}$  turn into  $\text{---}\text{---}\text{---}$  (see right two diagrams). So pure GD's test loss exceeds pure SGD's test loss by  $M((N-1)/2) \text{---}\text{---}\text{---}$ .

in correspondence, except for a discrepancy that shows the two test losses differ by  $M((N-1)/2) \text{---}\text{---}\text{---}$ .

### 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 can use this two ways: **(A)** to refine our computations if we know the hessian; **(B)** to refine our computations if we know the “effective propagator”.

For example, suppose we know  $H$  exactly. Then uncorrelated chain diagrams such as  $\text{---}\text{---}\text{---}$ ,  $\text{---}\text{---}\text{---}$ ,  $\text{---}\text{---}\text{---}$ ,  $\dots$ , when embedded with initial and final nodes separated by duration  $t$ , together contribute  $G(I + \eta H)^{t-1} \eta G$ . We may thus organize diagrams together by the homeomorphism classes of their *geometric realizations*; each class yields a sum. For example, the above chains contribute the following to test loss for vanilla SGD:

$$G \sum_{0 \leq t < T} (I + \eta H)^{T-t-1} \eta G = G_\mu (\Delta_T)^{\mu\nu} \eta G_\nu \quad (6)$$

where  $(\Delta_T)^\mu_\nu = ((I + \eta H)^T - I)/(\eta H)$  is a “propagator”. Likewise, for V structures, there is an approximate contribution

$$\frac{1}{2} G_\mu (\Delta_T)^{\mu\nu} G_\rho (\Delta_T)^{\rho\sigma} H_{\nu\sigma} \quad (7)$$

This is approximate because it does not account for correlations. For tree structions such as  $\text{---}\text{---}\text{---}$ , we have:

$$\sum_t \frac{1}{2} G_\mu (\Delta_t)^{\mu\nu} G_\rho (\Delta_t)^{\rho\sigma} J_{\nu\sigma\lambda} ((I + \eta H)^{T-t-1} \eta)^{\lambda\pi} G_\pi \quad (8)$$

Here, we integrate over internal — i.e. non-leaf non-root — nodes.

Now, suppose we know

## 4. Consequences and Applications

### 4.1. Vanilla SGD

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. As a simplest example where the loss landscape is not a Gaussian Process, consider fitting a centered normal  $\mathcal{N}(0, \sigma^2)$  to some one-dimensional data. We parameterize the landscape by  $h = \log(\sigma^2)$ . The gradient at sample  $x$  and weight  $\sigma$  is then  $g_x(h) = (1 - x^2 \exp(-h))/2$ . If  $x \sim \mathcal{N}(0, 1)$  is standard normal, then  $g_x(h)$  will be affinely related to a chi-squared, and in particular non-gaussian. At  $h = 1$ , the expected gradient vanishes, and the test loss only involves diagrams with no singleton leaves; to third order, it is  $\text{---}\text{---}\text{---} + \frac{N}{2} \text{---}\text{---}\text{---} + \binom{N}{2} \text{---}\text{---}\text{---} + \frac{N}{6} \text{---}\text{---}\text{---}$

### 4.2. Emulating Small Batches with Large Ones

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

small T: eta curve	test loss decrease near minimum	batch matching over one init	batch matching over multiple inits
small T gen gap: actual vs predicted	nongaussian example	scan over betas	long time comparison

Figure 3. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

Figure 4. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

...sunt in culpa qui officia deserunt mollit anim id est laborum.

### 4.3. Analyzing Second Order Methods

We demonstrate how our approach extends to more sophisticated optimizers by analyzing momentum and a hessian-based method.

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. momentum

Now consider a hessian-based update parameterized by a scalar  $\lambda$ :

$$\theta \leftarrow \theta - (\eta^{-1} + \lambda \nabla \nabla l_t(\theta))^{-1} \nabla l_t(\theta)$$

...sunt in culpa qui officia deserunt mollit anim id est laborum.

invhess

#### 4.5. Myopic Model Selection

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

Figure 5. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

#### 4.4. Epochs and Overfitting

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

rankings: actual vs  
predicted

architecture vs opti-  
mization ease

Figure 7. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

...sunt in culpa qui officia deserunt mollit anim id est laborum.

multiepoch vs sgd  
limit

multiepoch vs gd  
limit

#### 4.6. Comparison to Continuous Time

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

Figure 6. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...



...sunt in culpa qui officia deserunt mollit anim id est laborum. Also, sgd interepoch correlations

distinguishing landscape

ode vs sde vs sgd performance on landscape

Figure 8. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

#### 4.7. Thermodynamic Engine

We clarify Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. We constructed a counter-intuitive loss landscape wherein, for arbitrarily small learning rates, SGD cycles counterclockwise around a circle of minima. Our mechanism differs from that discovered by Chaudhari & Soatto (2018) discuss the thermodynamic significance of both

loss landscape: mean and covariance

net theta vs time: ours, chaudhari, naive

Figure 9. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

## 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 networks trained via SGD 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 (2019), 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). Using this technique, we 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 pre-

cise version of the Covariance-BatchSize-Generalization relationship conjectured in Jastrzębski et al. (2018).

While we make rigorous, architecture-agnostic predictions of learning curves, these predictions become vacuous for large  $\eta$ . In particular, while our work does not assume convexity of the loss landscape, it also is blind to large- $\eta T$  convergence of SGD. Other discrete-time dynamical analyses allow large  $\eta$  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), citetwa18), 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 makes explicit the dependence of generalization on the underlying metric and on the form of gradient noise, reconciles those latter, 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 non-infinitely-wide deep learning 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 of 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. 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 presented a novel diagrammatic tool for analyzing gradient descent. We introduced a novel regularizing term, thus showing that **large-batch GD can be made to emulate small-batch SGD** and completing a project suggested by Roberts (2018). This is significant because, while small batch sizes can lead to better generalization (Bottou, 1991), modern infrastructure increasingly rewards large batch sizes (Goyal et al., 2018). We showed also that in multi-epoch SGD, inter-epoch shuffling induces only a 3rd order effect on test loss. Intuitively, we proved that **the hessian matters asymptotically more than shuffling order**.

The diagram method is also a rich source of intuitions and physical analogies. For example, it offers a clearer understanding of the empirically verified limit cycles found in

Chaudhari. As our physical analogy emphasizes the underlying metric, it reconciles competing views of whether sharp or flat minima generalize. Further exploration of this bridge to particle physics, especially within the framework of renormalization theory, pose a promising direction for future research.

## VARIANCES

### 6.1. Acknowledgements

We thank Dan A. Roberts and Sho Yaida for patient introductions to their work and for precisely posing several of the questions we answer here. We feel deeply grateful to Sho Yaida and Josh B. Tenenbaum for their compassionate guidance. We appreciate the generosity of Andrzej Banburski and Wenli Zhao in offering crucial feedback on writing.

## References

- Bartlett, P., Foster, D., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 2017.
- Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 2013.
- Bottou, L. Stochastic gradient learning in neural networks. *Neuro-Nimes*, 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.
- 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.
- 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., N., B., Fischer, A., Y., B., and A., S. 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.

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.

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.

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

Wu, L., C., M., 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*, 2019.

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

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

## A. Derivation of Diagram Rules

### 6.2. Dyson Series for Iterative Optimizers

If a density  $\rho$  governs a point  $\theta$  in weight space, then after a sequence of updates  $\theta \mapsto \theta - \eta^{\mu\nu} \nabla_{\mu} l(\theta)$  on losses ( $l_t : 0 \leq t < T$ ), the following density (up to an error term whose Maclaurin series vanishes; all perturbative results will implicitly carry such terms) will govern the new point:

$$\exp\left(+\eta^{\mu\nu} \nabla_{\mu} l_{T-1}(\theta) \nabla_{\nu}\right) \cdots \exp\left(+\eta^{\mu\nu} \nabla_{\mu} l_0(\theta) \nabla_{\nu}\right) \rho \quad (9)$$

or  $\prod \exp(+\eta \nabla / \nabla) \rho$  for short. The exponent above is a linear operator that acts on a space of sufficiently smooth maps; in particular, the  $\nabla_{\nu}$  does not act on the  $\nabla_{\mu} l(\theta)$  with which it pairs. Integrating by parts, we write the expectation over initial values after  $T$  steps of a function  $s$  of weight space (e.g.  $s$  may be test loss) as:

$$\int_{\theta} \rho(\theta) \left( \prod_{0 \leq t \leq T} \exp\left(-\eta^{\mu\nu} \nabla_{\mu} l(\theta) \nabla_{\nu}\right) s \right) (\theta) \quad (10)$$

Since the exponentials above 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  $\eta$ . Setting the initialization  $\rho(\theta) = \delta(\theta - \theta_0)$  to be deterministic, and labeling as  $\theta_t$  the weight after  $t$  steps, we find:

$$s(\theta_T) = \sum_{0 \leq d < \infty} (-\eta)^d \sum_{\substack{(d_t: 0 \leq t < T) \\ \sum_t d_t = d}} \left( \prod_{0 \leq t < T} \frac{(\nabla l_t(\theta) \nabla)^{d_t}}{d_t!} \right) s(\theta_0) \quad (11)$$

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

## B. Tutorial on Diagram Rules

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.

## C. Derivations of Perturbative Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit...

...sunt in culpa qui officia deserunt mollit anim id est laborum.







## D. 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 and the new formalism on some sample problems.

### Effect of Batch Size

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  $\eta^2$ .

#### 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

$$\eta^2 \frac{MN\binom{N}{2}}{N^2} (\text{Diagram 1} - \text{Diagram 2}) = \eta^2 \frac{M(N-1)}{2} \text{Diagram 3}.$$

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

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

See Figure FILL IN for a visualization.

#### DIRECT PERTURBATION

We compute the displacement  $\theta_T - \theta_0$  to order  $\eta^2$  for pure SGD and separately for pure GD. Writing  $\theta_t \in \theta_0 + \eta a(t) + \eta^2 b(t) + o(\eta^2)$ , 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  $a(t) = -\sum_{0 \leq t' < t} \nabla l_{n_{t'}}$  and  $b(t) = \sum_{0 \leq t_0 < t_1 < t} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}$ . Now, with  $l$  the test landscape, we have

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

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,

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

Likewise, 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

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

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

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


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

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

## Effect of Nongaussian Noise at a Minimum

We consider pure SGD initialized at a local minimum of the test loss. One expects  $\theta$  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  $\eta^3$  because at lower orders, gaussian and nongaussian landscapes will by definition match.

### DIAGRAM METHOD

Because  $\mathbb{E}[\nabla l]$  vanishes at the initialization, all diagrams with a degree-one vertex that is a singleton vanish. Because we work at order  $\eta^3$ , we consider 3-edged diagrams. Finally, because all expectations and covariances of gradients match 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

#### Generalization Gap vs Curvature and Covariance

We relate the generalization gap of vanilla SGD to the ambient curvature and gradient noise. Prior work has alternately claimed that sharp minima generalize better (after all,  $l^2$  regularization increases curvature) and that flat minima generalize better (after all, small displacements have smaller costs near flat minima). We resolve this seeming conflict.

### DIAGRAM METHOD

The generalization gap of vanilla SGD is

$$\begin{aligned}&\frac{1}{N} (\eta N \text{  - \eta^2 \binom{N}{2} (\text{  +  +  )) + o(\eta^2) \\ &= \eta \text{  - \eta^2 \frac{N-1}{2} (3 \text{  +  ) \\ &\quad - \eta^2 \frac{1}{2} (\text{  -  ) + o(\eta^2)\end{aligned}$$

### DIRECT PERTURBATION

#### Sample Complexity

### DIAGRAM METHOD

### DIRECT PERTURBATION