A Perturbative Analysis of Stochastic Gradient Descent

Samuel C. Tenka

MIT, CSAIL

C O L I@M I T . e d u

Abstract

We quantify how gradient noise shapes the dynamics of stochastic gradient descent (SGD) by Taylor-expanding with respect to the learning rate. We present a new diagram-based technique that permits re-summation of that series to convergent results on short timescales or near local minima. We physically interpret the resulting terms as entropic corrections to deterministic descent. Our theory predicts that gradient noise biases SGD's trajectory toward low-curvature, low-noise regions of the loss landscape. We contrast our theory against popular approximations of SGD such as SDE and experimentally verify our theory's predictions.

Introduction

Gradient estimates, measured on minibatches and thus noisy, form the primary learning signal in deep learning. While users of deep learning benefit from the intuition that *stochastic gradient descent* (SGD) approximates deterministic descent (GD) (Bottou 1991; LeCun, Bengio, and Hinton 2015), SGD's gradient noise in practice alters training dynamics and testing losses (Goyal et al. 2018; Wu et al. 2020). We analyze these dynamics for SGD on short timescales or near minima. We apply this theory to find that **gradient noise biases learning** toward low-curvature, low-noise weights.

Specifically, we study the expectation over training sets of the testing loss by Taylor expanding that loss in the learning rate η. By induction on T, the loss changes by $-TG\eta G + o(\eta)$ after T training steps, where $G = \mathbb{E}_x[\nabla l_x(\theta_0)]$ is the expectation over testing samples x of the gradient at initialization. This estimate is exact when $\nabla l_x(\theta)$ depends on neither x nor θ, i.e., for deterministic linear loss landscapes. We compute how noise and curvature correct this estimate.

A Taylor series analysis of SGD presents three challenges. *First*, the terms explode in variety. Even the main correction to $-TG\eta G$ represents the diverse ways that some past update may affect a future weight θ_t and thus a future update involving $\nabla l_{x_t}(\theta_t)$. That is, updates **interact**. *Second*, SGD's gradient noise is correlated between timesteps: the same training sample reappears in each epoch. Such **finite-sample** effects complicate the simple induction that led to $-TG\eta G$. *Third*, the series' dth order truncation **diverges** as

T grows. Indeed, on landscapes such as least squares linear regression, the loss grows exponentially with time for $\eta < 0$. So on no neighborhood of $\eta = 0$ does any Taylor truncation suffer an error uniform in T.

We address the three challenges by using diagrams such as ⌢, ⌢⌢, ⌢⌢ to organize and evaluate many terms at once, including correlation effects.[1] We identify topologically related higher order diagrams that balance out a given diagram's divergence. In thus modifying plain truncation, we tame the large-T divergence. These '**re-summed**' expressions' errors vanish uniformly in T for quadratic landscapes (with non-Gaussian gradient noise) and are provably finite and empirically small for convolutional landscapes.

Roughly, the analytical procedure we develop is this. We interpret each diagram as depicting a class of interactions or "**histories**" between updates. E.g. ⌢⌢ depicts an update's (red's) double effect on a future update (green) that in turn affects the testing loss (blue). The rightmost or rootnode always "occurs" post-training. Up to $o(\eta^d)$ error, the testing loss is a sum over all histories of all diagrams with $\leq d$ edges of certain diagram-dependent tensor expressions.
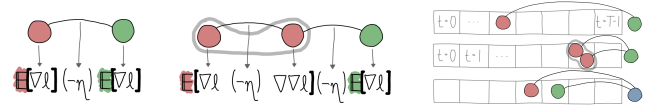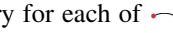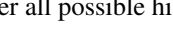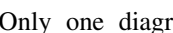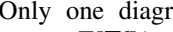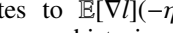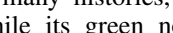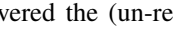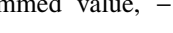

Figure 1: Diagrams evaluate to tensor expressions (■■□): a degree-d node gives a dth derivative; an edge gives a $(-\eta)$; expectation brackets enclose each fuzzy group. Each diagram depicts a class of histories; □□■ shows an example history for each of ⌢, ⌢⌢, ⌢⌢. Summing diagram values over all possible histories gives the expected testing loss.

Only one diagram, ⌢, has $d = 1$ edges. ⌢ evaluates to $\mathbb{E}[\nabla l](-\eta)\mathbb{E}[\nabla l] = -G\eta G$. And ⌢ describes T many histories, since its red node can be any update while its green node represents testing. We've thus recovered the (un-resummed) value $-TG\eta G$ above. The re-summed value, $-TG\frac{1-\exp(-T\eta H)}{H}G$, includes contributions from ⌢, ⌢⌢, ⌢⌢⌢, ⋯.

Paradigmatic example

We illustrate our theory by deriving a "non-conservative entropic force" in SGD. We avoid the next section's generality.

Notation and assumptions, I

We formalize the loss — suffered by a fixed architecture on a random datapoint — as a distribution $\mathcal{D}$ over functions from a space $\mathcal{M}$ of weights. The *testing loss* $l : \mathcal{M} \to \mathbb{R}$ is $\mathcal{D}$'s mean. We write $\theta \in \mathcal{M}$, $l_x \sim \mathcal{D}$ for generic elements. We consider training sequences $(l_n : 0 \le n < N) \sim \mathcal{D}^N$. We call n and l_n *training points*. Each initialization $\theta_0 \in \mathcal{M}$ then induces via SGD a distribution over trajectories $(\theta_t : 0 \le t \le T)$. Specifically, SGD runs T steps of η-steepest descent:

$$\theta_{t+1}^\mu := \theta_t^\mu - \sum_\nu \eta^{\mu\nu}\nabla_\nu l_{n_t}(\theta_t)$$

Each sequence $(n_t : kN \le t < kN+N)$ is a permutation of $(n : 0 \le n < N)$. Our Greek indices name components of θ, η, ∇ w.r.t. a fixed basis. We view η as a bilinear form so that the only type-correct expressions have geometric significance.

We heavily use the **gradient** $G_\mu = \mathbb{E}[\nabla_\mu l_x(\theta)]$, **hessian** $H_{\mu\nu} = \mathbb{E}[\nabla_\mu \nabla_\nu l_x(\theta)]$, **jerk** $J_{\mu\nu\xi} = \mathbb{E}[\nabla_\mu \nabla_\nu \nabla_\xi l_x(\theta)]$, **covariance** $C_{\mu\nu} = \mathbb{E}[(\nabla l_x(\theta)-G)^{\otimes 2}]_{\mu\nu}$, and **skew** $S_{\mu\nu\xi} = \mathbb{E}[(\nabla l_x(\theta)-G)^{\otimes 3}]_{\mu\nu\xi}$, typically evaluated at initialization ($\theta = \theta_0$). So G, H, J, C, S respectively have $1, 2, 3, 2, 3$ many axes, each of which transforms under change-of-basis like a covector.[1]

To illustrate our notation we quote Nesterov (2004), §2.1:

Prop 0. $G = \nabla l(\theta_0)$ *controls the loss to leading order. Precisely,* $\mathbb{E}[l(\theta_T) - l(\theta_0)]^\mu = -T \sum_{\mu\nu} G_\mu \eta^{\mu\nu} G_\nu + o(\eta^1)$.

Our work identifies how noise and curvature correct Prop 0.

SGD descends on a C-smoothed landscape

Let's study the displacement $\mathbb{E}[\theta_T - \theta_0]$ of one-epoch, batch-size-one SGD initialized at a testing loss minimum.

We sketched above how a sum over diagrams gives $\mathbb{E}[l(\theta_T)]$ for the testing loss l. We may likewise compute $\mathbb{E}[s(\theta_T)]$ for other $s : \mathcal{M} \to \mathbb{R}$; we need only replace the factor $\mathbb{E}[\nabla^k l(\theta_0)]$ corresponding to each diagram's root by the factor $\mathbb{E}[\nabla^k s(\theta_0)]$. Specifically, a linear $s(\theta) = w \cdot \theta$ reveals the displacement's w-component. Since s's higher derivatives vanish, any diagram whose root has degree > 1 will evaluate to zero. So only diagrams with degree-1 root are relevant to computing displacements (e.g. ⌢, ⌢⌢, ⌢⌢).

Because in this example we initialize at a minimum ($G = 0$), we may ignore yet more diagrams: those with a factor of $\mathbb{E}[\nabla l]$, i.e., those with a degree-1 non-root node that's not fuzzily grouped. So we rule out ⌢, ⌢⌢, ···. Only one fewest-edged diagram remains: ⌢⌢, which evaluates to $\mathbb{E}[\nabla l\nabla l](-\eta)^2\mathbb{E}[\nabla\nabla\nabla l](-\eta)\mathbb{E}[\nabla s] = -\eta^3(C\nabla H)\cdot(\nabla s)$. So to leading order the displacement is some time-dependent number of histories times $-\eta^3(C\nabla H)$.

Figure 2: **Left**: Gradient noise pushes SGD toward minima flat w.r.t. C. A 2D loss near a valley of minima. Red densities show typical θs, perturbed by noise (C), in two cross sections of the valley. The hessian changes across the valley: $J \ne 0$. **Right**: HELIX is defined on an $\mathcal{M} = \mathbb{R}^3$ extending into the page. A helical level surface S (orange-green) of l winds around a 1D valley of minima orthogonal to the page. $l \gg 1$ outside S. Gradient noise is parallel to the page and to the line between outer tubes. Thanks to CalcPlot3D.

Since $-\eta^3 C\nabla H$ points toward small H, **SGD moves toward flat minima** (Figure 2). This effect scales with C. Resummation more precisely quantifies this 'entropic force':[23]

Corollary 1. *Start SGD at a minimum of l with $H > 0$, $N = T$. Use an eigenbasis of $K = \eta H$. For any T and with $\mathcal{P}_T(s) = (1 - \exp(-Ts))/s$, the expected displacement is*

$$-\sum_{\substack{\mu\nu \\ \acute{o}\pi\rho}} C_{\mu\acute{o}}\mathcal{P}_T(K_{\mu\mu} + K_{\acute{o}\acute{o}})\eta^{\mu\nu}\eta^{\acute{o}\pi}J_{\nu\pi\xi}\mathcal{P}_T(K_{\xi\xi})\eta^{\xi\rho}/2 + o(\eta^3)$$

Consider a 1D valley of near-minima wherein ηH has spectrum $\lambda_0 \ll 1/T \ll \lambda_1 \le \cdots$ for eigenvectors v_i. Let's ignore diffusion along the valley: $(\eta C)v_0 = 0$. Then $\mathcal{P}_T(\lambda_0) \approx T$ and every $C_{ij}\mathcal{P}_T(\lambda_i + \lambda_j)$ is $O(T^0)$. So $\mathbb{E}[\theta_T - \theta_0]$ scales linearly with T. We thus expect SGD to move with velocity $\approx -\eta^2 C\nabla H/2$ per timestep toward flat minima. Observe that $\nabla(\mathfrak{G}_C \star l) = \nabla l + C\nabla H/2 + o(C)$, where $\mathfrak{G}_C\star$ denotes convolution with a *fixed* centered C-shaped Gaussian; we conclude with the intuition that *SGD descends on a C-smoothed landscape that changes as C does.*

To test Corollary 1's C-dependence, we construct a landscape, HELIX, on whose valley of global minima C varies (Figure 3). As in Rock-Paper-Scissors, each θ has a neighbor that is more attractive (flatter) w.r.t. $C(\theta)$. This induces eternal motion into the page despite HELIX's discrete translation symmetry. Corollary 1 predicts a velocity of $+\eta^2/6$ per timestep, while (Chaudhari and Soatto 2018)'s SDE-based analysis predicts a constant velocity of 0.[4] Wrapping HELIX in a loop makes SGD circulate. This is possible because $C\nabla H$, unlike $\nabla(CH)$, is not a total derivative. In avoiding (Wei and Schwab 2019)'s constant-C assumption, we find that SGD's velocity field can have curl.

---

[1]Gradients (covectors, dollars-per-mile) and displacements (vectors, miles) have geometrically distinct types. We respect this distinction throughout; that's why our η is a tensor. (Misner, Thorne, and Wheeler 1973) (§2.5) visualizes this geometry; (McCullagh 1987) (§1.4) discusses vectors vs covectors in statistics.
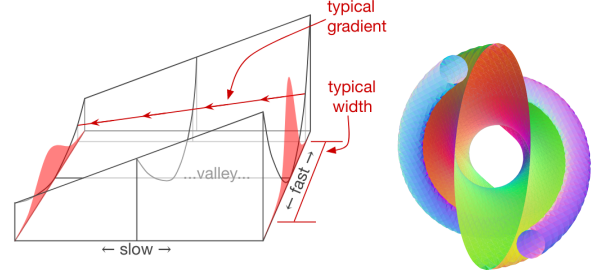
[2]Thermal systems tend toward disorder as if pushed by an 'entropic force'. So arises the tension of rubber bands: their polymers can wreathe in many ways but be straight in only one. Such 'forces' characteristically scale with temperature (the noise intensity C).

[3]Our result ($T \gg 1$) is $\Theta(\eta^2)$; (Yaida 2019a)'s ($T = 2$) is $\Theta(\eta^3)$. We integrate noise over time, amplifying C's effect.

[4]For, HELIX' velocity is η-perpendicular to the image of $(\eta C)_\nu^\mu$.
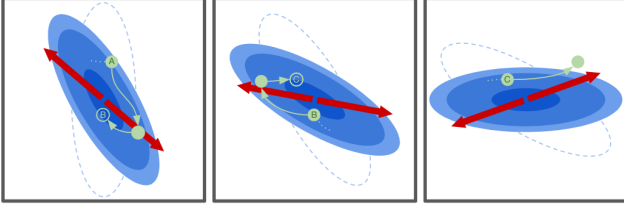
Figure 3: Green: SGD's trajectory over cross sections of HE-LIX descend progressively into the page. Blue: l's contours; l's valley intersects each pane's center. Dotted curves help compare adjacent panes. Red: C's major axis. Gradient noise kicks θ from A; θ then falls (■□□) to B in □■□. At C, noise kicks θ uphill (□□■); θ thus never settles and the descent persists.

Perturbative theory of SGD

We introduce the Taylor series approach to studying SGD dynamics and then the diagram notation for the terms that thus appear. After showing how to evaluate diagrams to numbers, we state our main result: that diagram-based analysis is correct.

Notation and assumptions, II

REGULARITY CONDITIONS — We assume the following throughout. *Derivative bounds*: there are compact sets (K_k : $k \geq 0$) so that $\nabla^n l_x(\theta) \in K_k$ for all θ, l_x. Here $\nabla^k l_x(\theta)$ is a kth derivative, a k-axis tensor. **Analytic moments**: any polynomial p of l_x and its higher derivatives induces a random variable so that $\mathbb{E}[p] : \mathcal{M} \to \mathbb{R}$ (exists and) is analytic in θ; moreover, $\mathbb{E}[p]$'s radii of convergence are strictly bounded from 0, even as θ varies. Consequently, $\nabla \mathbb{E}[p] = \mathbb{E}[\nabla p]$.

BATCHES AND EPOCHS — Generalizing the previous section, our theory describes SGD with any number $\mathbf{N}$ of training points, $\mathbf{T}$ of updates, and $\mathbf{B}$ of points per batch. SGD runs T updates (hence $\mathbf{E} = TB/N$ epochs or $\mathbf{M} = T/N$ updates per training point) of the form

$$\theta_{t+1}^\mu := \theta_t^\mu - \sum_\nu \eta^{\mu\nu} \nabla_\nu \sum_{n \in \mathcal{B}_t} l_n(\theta)/B$$

where in each epoch we sample the tth batch $\mathcal{B}_t$ without replacement from the training sequence.

COMPARING TENSORS — A (potentially tensor) quantity q *vanishes to order η^d* when for some homogeneous degree-d polynomial p $\lim_{\eta \to 0} q/p(\eta) = 0$; we then say $q \in o(\eta^d)$. We write $A \preceq B$ ($\prec$) for symmetric bilinear forms A, B when $A(v, v) \leq B(v, v)$ ($<$) for all $v \neq 0$.

CONTINUOUS TIME — Ordinary and stochastic differential equations (ODE, SDE) are popular models of SGD (Liao et al. 2018; Barrett and Dherin 2021). They correspond to continuous-time limits of large-training-set SGD with independent noise ($E = B = 1$, $N = T = kT_0$, $\eta = \eta_0/k$, $k \to \infty$) ODE descends on a noiseless version of the landscape $l_x(\theta)$, while SDE descends on a version whose gradient noise is independent gaussian of shape $C(\theta) = \mathbb{E}[\nabla l_x \nabla l_x] - \mathbb{E}[\nabla l_x]\mathbb{E}[\nabla l_x]$, scaled as in a Wiener process:

$$l_x^{\text{ODE}}(\theta) - l(\theta) = 0 \qquad l_x^{\text{SDE}}(\theta) - l(\theta) \sim \mathcal{N}(0, k \cdot \theta^T C(\theta)\theta)$$

In this paper we shall take k to be large but finite and express results about ODE, SDE with error terms such as $o(1/k)$. We emphasize that the constants in these little-os are permitted to depend on the loss landscape and optimization parameters such as η, T. It is physical intuition and experiment that determine when such error terms are negligible.

Taylor series: method and challenges

We'll prove Prop 0 (c.f. (Nesterov 2004; Roberts 2018)).

Proof. By gradient bounds: $\theta_T - \theta_0$ is $O(\eta^1)$. We **claim** that $(\theta_T - \theta_0)^\mu = -T \sum_\nu \eta^{\mu\nu} G_\nu + o(\eta^1)$. The claim holds when $T = 0$. Say the claim holds for $\tilde{T}$-step SGD with $T = \tilde{T} + 1$. The displacement $(\theta_T - \theta_{\tilde{T}})^\mu$ evaluates to:

$$-\sum_\nu \eta^{\mu\nu} \nabla_\nu l_{n_{\tilde{T}}}(\theta_{\tilde{T}})$$
$$= -\sum_\nu \eta^{\mu\nu} \nabla_\nu \left(l_{n_{\tilde{T}}}(\theta_0) + \boxed{\sum_\xi \nabla_\xi l_{n_{\tilde{T}}}(\theta_0)(\theta_{\tilde{T}} - \theta_0)^\xi} + o(\theta_{\tilde{T}} - \theta_0) \right)$$
$$= -\sum_\nu \eta^{\mu\nu} \nabla_\nu \left(l_{n_{\tilde{T}}}(\theta_0) + \nabla l_{n_{\tilde{T}}}(\theta_0) \cdot O(\eta^1) + o(O(\eta^1)) \right)$$
$$= \boxed{-\sum_\nu \eta^{\mu\nu} \nabla_\nu l_{\tilde{T}}(\theta_0)} + o(\eta^1)$$

Applying the induction hypothesis proves the claim. We plug the claim into l's Taylor series:

$$\mathbb{E}[l(\theta_T) - l(\theta_0)] = \sum_\mu \nabla_\mu l(\theta_0) \boxed{\mathbb{E}[\theta_T - \theta_0]^\mu} + \mathbb{E}[o(\theta_T - \theta_0)]$$
$$= \sum_\mu \nabla_\mu l(\theta_0)(-T\eta G + o(\eta^1)) + o(O(\eta^1))$$
$$= \boxed{-\sum_{\mu\nu} T G_\mu \eta^{\mu\nu} G_\nu} + o(\eta^1)$$

Indeed, due to the assumption of analytic moments, the above expectations of $o(\eta^1)$ terms are still $o(\eta^1)$. □

Diagrams in brief That development begins with the observation that each η 'connects' two ∇ operators as indices prescribe. So we draw ηs as edges, $\nabla^k l$s as nodes, and each summand of the following form (evaluated at $\theta = \theta_0$)

$$\sum_{\text{all Greek indices}} \left(\prod_{j \in J} \eta^{\mu_j \nu_j} \right) \left(\prod_{i \in I} \left(\prod_{k \in K_i} \nabla_{\xi_{i,k}} \right) l_{x_i} \right) \left(\prod_{k \in K_\star} \nabla_{\xi_{\star,k}} \right) l$$

as an undirected graph with edges indexed by $j \in J$, nodes indexed by $i \in I \sqcup \{\star\}$, and an edge j incident to a node i when $\{\xi_{i,k} : k \in K_i\}$ meets $\{\mu_j, \nu_j\}$. Per FACTORING, we also equip these graphs with a partition of nodes to account for correlation structure. Such diagrams offer the same advantages over direct notation that decimal numerals offer over unary numerals. More importantly, their topology has dynamical significance and leads to a physical interpretation of SGD.

Diagrams overcome Taylor Series challenges

Definition 1. A **diagram** is a rooted tree equipped with an equivalence relation on (i.e. a partition of) its non-root nodes. We orient the tree left-to-right so that children precede parents; the root is thus rightmost. We draw the partition structure with fuzzy outlines. ◇

Valid diagrams include ⌢, ⌢⌢, ⌢⌢⌢, ⌢⌢⌢, ⌢⌢⌢, ⌢⌢ but not ⌣, ⌢, ⌢, ⌢, ⌢, ⌢ .

Since a diagram is just a rooted tree and partition, ⌢⌢ = ⌢⌢ = ⌢⌢ are the same diagrams.

Definition 2. A diagram D's *un-resummed value* (**uvalue**) is a product with one factor of l_x's dth derivative for each degree-d node, grouped under cumulant symbols $\mathbb{C}$ (think: expectation symbols $\mathbb{E}$)[1] per D's fuzzy groups, and tensor-contracted via a factor $\eta^{\mu\nu}$ for each edge. ◇

E.g. uvalue(⌢⌢) is:

$$\sum_{\mu\nu\xi\acute{o}\pi\rho} \eta^{\mu\xi}\eta^{\nu\acute{o}}\eta^{\pi\rho} \; \mathbb{C}[(\nabla_\mu l_x)\cdot(\nabla_\nu l_x)] \; \mathbb{C}[(\nabla_\xi \nabla_{\acute{o}} \nabla_\pi l_x)] \; \mathbb{C}[(\nabla_\rho l_x)]$$

There are dozens of small diagrams. In many analyses, only a few diagrams are relevant. Examples: for fixed T, **to order d we may neglect diagrams with more than d edges**; if $E = B = 1$ (§), each diagram with an ancestor-descendant pair in the same part contributes zero to $\mathbb{E}[l(\theta_T)]$; for θ_0 a minimum of l, all diagrams vanish that contain a leaf node not fuzzily grouped.

SGD as a Sum over Histories Having expressed terms in $\mathbb{E}[l(\theta_T)]$'s Taylor expansion as uvalues of diagrams, we seek the coefficient for each uvalue. Intuitively, a diagram represents a process (as in Figure 9) and a diagram's contribution scales with the number of ways that process may occur. Specifically, the Key Lemma in § establishes that a diagram's coefficient in the expansion is the number of its *histories* (weighted by symmetry factors to counter over-counting), where we define *histories* below with respect to given SGD hyperparameters N, E, B.[2]

Definition 3. An **history** of a diagram is an assignment of non-root nodes to *(n,t)* pairs such that: the nth training point participates in the tth batch; parents' ts strictly exceed their children's ts; and any two nodes' ns are equal if the nodes are in the same part of the partition. ◇

E.g. ⌢ has just one non-root node. It has as many histories as there are (n, t) cells where n participates in the tth update, i.e., $B \cdot T$ histories. Since ⌢ is the only 1-edged diagram, it gives the full η^1 contribution to final testing loss.

Diagrams streamline analysis of SGD because it is in practice straightforward to count a diagram's histories. Also, the topology of diagrams has dynamical significance: the $t^d T^{-p}$-th order correction[3] to the ODE approximation of SGD is given by diagrams with d edges and $d + 1 - p$ many fuzzy groups (counting each node not fuzzily grouped as its

---

[1]Inconsequential technicality: uvalues are products of *cumulants* such as C, not of un-centered moments such as $GG + C$. The symbol $\mathbb{C}[a]$ gives a's mean; $\mathbb{C}[a \cdot b]$, a, b's covariance; we center higher cumulants $\mathbb{C}[\prod_i a_i]$ with respect to lower cumulants.

[2]and w.r.t. a deterministic selector of the tth batch. One may take expectations over such algorithms — §.

[3]We compare ODE integrated to time t to T steps of SGD with $\eta = \eta_\star t/T$ and $E = B = 1$, and we assume $p \neq 0$.
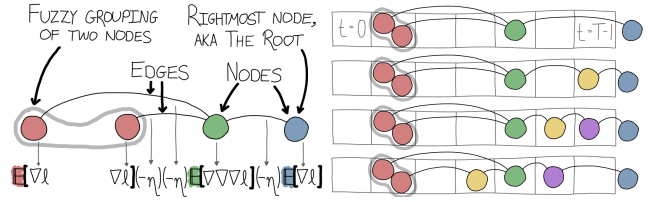

Figure 4: ■: a diagram consists of nodes, edges, and fuzzy groupings that dictate the diagram's corresponding tensor expression. □ shows one of a diagram's histories. Re-summation balances that history with the of topologically related histories of related diagrams such as □.

own group). Likewise, if we seek to isolate the effect, say, of C or H or S or J, we may consider only those diagrams that contain the corresponding subgraph.

Re-summation cures large-T divergences So far, diagrams have been a convenient but dispensible book-keeping tool; so far, §'s polynomial divergence remains in (3). We now show how diagrams enable us to tame this divergence.

Let us collect similar diagrams, where our notion of 'similar' permits chains to grow or shrink (see Definition 4). We obtain lists (each conveniently represented by its smallest member) such as ⌢⌢, ⌢⌢, ⌢⌢, ⋯.We will express in closed form the total contribution to (3) of all diagrams in such a list. The idea is that the uvalues of chains are powers of hessians — e.g. uvalue(⌢⌢⌢) = $(-\eta)^4 GH^3 G$ — so we may sum over chain lengths via geometric series.

We define an embedded diagram's *resummed value* or *rvalue* as we defined the uvalue, except that we use $(I - \eta H)^{\Delta t-1}\eta$ instead of η to contract a pair of tensors embedded Δt timesteps apart. For example, take Figure 7's history of ⌢ (topmost of four). The associated uvalue is $\sum_{\mu\nu} G_\mu \eta^{\mu\nu} G_\nu$: a G for each degree-one node and an η for each edge. By contrast, the associated rvalue is $\sum_{\mu\nu} G_\mu ((I - \eta H)^{11-1}\eta)^{\mu\nu} G_\nu$ since the edge spans 11 timesteps. Distributing out this expression reveals uvalues for histories of ⌢, etc.

Definition 4. A **link** is a degree-2 non-root node that is not fuzzily grouped. E.g. ⌢⌢ has one link (green). To *reduce* at a link, we replace the link by a black edge connecting the link's two neighbors. E.g. ⌢⌢ ⤳ ⌢⌢. Reduction generates an equivalence relation on diagrams. Each equivalence class contains exactly one **linkless** diagram. ◇

Main result

Theorem 1. $\forall T : \exists \eta_0 > 0 : \forall 0 \leq \eta < \eta_0 :$ *the final testing loss is a sum over* linkless *diagrams:*

$$\mathbb{E}[l(\theta_T)] = \sum_{\substack{D \; a \; linkless \\ diagram}} \sum_{\substack{f \; an \; embed- \\ -ding \; of \; D}} \frac{1}{|\mathrm{Aut}_f(D)|} \frac{\mathrm{rvalue}_f(D)}{(-B)^{|\mathrm{edges}(D)|}}$$

Here, $|\mathrm{Aut}_f(D)|$ *counts the graph automorphisms of D that preserve f. (Typically* $|\mathrm{Aut}_f(D)| = 1$.)

diagram	#embed.s	interpretation
	T	naïve descent
	$\binom{T+1}{2}$	θ-dependent loss
	T	gradient noise
	0	correl'd batches

Remark 1. A diagram with d edges scales as $O(\eta^d)$, so the Theorem expresses a series in η. In practice, we truncate to small d (thus focusing on few-edged diagrams) and we replace sums over histories by integrals over t; $(I - \eta H)^t$, by $\exp(-\eta H t)$, thus reducing to a routine integration of exponentials at the cost an error factor $1 + o(\eta)$. $\diamond$

Remark 2. Theorem 1 gives us the expectation of the testing loss. Straightforward variations of the theorem permit us to compute variances instead of expectations, training statistics instead of testing statistics, and weight displacements instead of losses. See §appendix:solve-variants. $\diamond$

Theorem 2. *For constant-$M = P$ or constant-$N = P$ SGD:* $\forall \theta_\star, P : (G(\theta) = 0 \wedge H(\theta) > 0) \implies \exists U \ni \theta_\star \text{ open} : \forall \theta_0 \in U : \text{Theorem 1's dth-order truncation converges as } T \to \infty.$

Remark 3. Thm 2 claims only that a limit $\lim_{T\to\infty} L_d(T, \eta)$ of the dth-order truncation $L_d(T, \eta)$ exists. It does not compare $\lim_{d\to\infty} \lim_{T\to\infty} L_d(T, \eta)$ with $\lim_{T\to\infty} \lim_{d\to\infty} L_d(T, \eta)$ (however, we note that we have not in practice observed pathologies of non-commuting limits). Our theory suggests but does not guarantee that when d, T are large (and finite), then computation of $L_d(T, \eta)$ by Taylor methods gives insight into SGD's behavior. It is by empirical tests and physical intuition that we decide whether in a given situation our theory's little-o error terms may be ignored.

Example Computation A We improve on Prop 0 for SGD with $E = B = 1$. The one 1-edged diagram ( ) embeds in T ways (one for each timestep) and contributes (let $K^\mu_\nu = \sum_\xi \eta^{\mu\xi} H_{\xi\nu}$):

$$\sum_{0 \le t < T} \sum_{\mu\nu} G_\mu \left[(I - K)^{T-t-1}\eta\right]^{\mu\nu} G_\nu = \sum_{\mu\nu} G_\mu \left[\frac{I - K^T}{I - K}\eta\right]^{\mu\nu} G_\nu$$

to the loss. This is the re-summed Prop.

Example Computation B There are three 2-edged linkless diagrams (see table for intuitive interpretations). The two diagrams involving higher cumulants (i.e., gray outlines) give the leading order effect of gradient noise. Since $E = 1$, has no histories; so only contributes. Up to a $1 + o(\eta)$ factor, its rvalues sum to (a sum over all indices of):

$$\int_t C_{\mu\nu} \left[\exp\left(-(T-t)(K \otimes I + I \otimes K)\right)\right]^{\mu\nu}_{\xi\acute{o}} \eta^{\xi\pi} \eta^{\acute{o}\rho} H_{\pi\rho}$$

$$= C_{\mu\nu} \left[\frac{I - \exp\left(-T(K \otimes I + I \otimes K)\right)}{(K \otimes I + I \otimes K)}\right]^{\mu\nu}_{\xi\acute{o}} \eta^{\xi\pi} \eta^{\acute{o}\rho} H_{\pi\rho}$$

We used Remark 1 to approximate $(I - K)^\mu_\xi (I - K)^\nu_{\acute{o}}$ by $\exp(-K \otimes I - I \otimes K)^{\mu\nu}_{\xi\acute{o}}$. The above is the leading contribution

of the gradient covariance C to SGD's final testing loss. We have derived an $E = B = 1$ variant of Corollary 6's $E = T$, $B = N$ result (§).

On Factoring

Consequences of the theory

By Cor.s 2 and 7, gradient noise repels SGD. By Cor. 5, SGD senses changes in H more than SDE; in fact, (Cor. 1) SGD seeks small-H weights. Cor. 6 relates C and H to overfitting.

Gradient noise repels SGD

Physical intuition suggests that noise repels SGD: if two neighboring regions of weight space have high and low levels of gradient noise, respectively, then the rate at which θ jumps from the former to the latter exceedsthe opposite rate. There is thus a net movement toward regions of small C.[1] Our theory makes this precise; θ drifts in the direction $-\nabla C$, and the effect is weaker when gradient noise is averaged out by large batch sizes:

Corollary 2 (Computed from). *SGD with $E = B = 1$ avoids high-C regions more than GD:* $\mathbb{E}[\theta_{GD} - \theta_{SGD}]^\rho = T \cdot \frac{N-1}{4N} \sum_{\mu\nu\xi} \eta^{\mu\rho} \eta^{\nu\xi} \nabla_\mu C_{\nu\xi} + o(\eta^2)$.

(Roberts 2019) obtained a version of this Corollary with a nearly equal error of $O(\eta^2/N) \vee o(\eta^2)$. The Corollary's proof implies that if $\hat{l}_c$ is a smooth unbiased estimator of $\frac{N-1}{4N} C^\nu_\nu$, then GD on $l + \hat{l}_c$ has an expected testing loss that agrees with SGD's to order η^2. We call this method **GDC**.

An analogous form of averaging occurs over multiple epochs.

Time discretization penalizes sloped regions

The following corollary recovers (Barrett and Dherin 2021)'s main dynamical result:

Corollary 3 ( ). *On a noiseless landscape, SGD prefers small-G^2 regions more than ODE:* $\mathbb{E}[\theta_{SGD} - \theta_{ODE}]^\rho = -\frac{T}{4} \sum_{\mu\nu} \eta^{\xi\rho} \eta^{\mu\nu} \nabla_\xi (G_\mu G_\nu) + o(\eta^2) + o(1/k)$.

Due to time discretization, in the presence of curvature SGD's response to gradient noise 'overshoots' more than SDE. The following corollary makes this precise and separates SDE from SGD, even on landscapes obeying SDE's assumption of gaussian noise:

Corollary 4 ( ). *The covariance of gradient noise contributes* $\frac{T}{2} \sum_{\mu\nu\xi\acute{o}} C_{\mu\nu} \eta^{\mu\xi} \eta^{\nu\acute{o}} H_{\xi\acute{o}} + o(\eta^2) + o(1/k)$ *to $E = B = 1$ SGD's final testing loss excess over SDE's.*

Jerk distinguishes SDE and SGD

SDE differs from SGD in ways beyond time-discretization effects. For instance, the inter-epoch noise correlations in multi-epoch SGD measurably affect SGD's final testing

---

[1]This is the same mechanism by which sand on a vibrating plate accumulates in quiet regions (Chladni 1787). We thus dub the SGD phenomenon the Chladni drift.
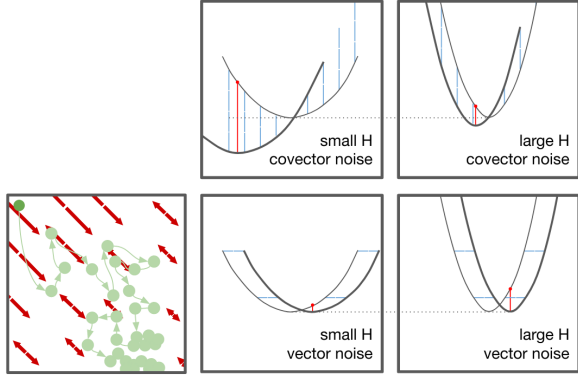
Figure 5: Left. Chladni drift on $\mathcal{M} = \mathbb{R}^2$. Red bi-arrows depict $C(\theta)$'s major axis. SGD updates (green) tend toward small C. Both curvature and noise affect overfitting. In each pane, the $\leftrightarrow$ axis represents weight space and the $\updownarrow$ axis represents loss. Noise (blue) transforms the testing loss (thin curve) into the observed loss (thick curve). Red dots mark the testing loss at the arg-min of the observed loss. □▦: *covector*-perturbed landscapes favor large Hs. □▦: *vector*-perturbed landscapes favor small Hs.

loss (Corollary 7), but SDE assumes uncorrelated gradient updates. Even if we restrict to single-epoch SGD, non-Gaussian noise lead SGD and SDE to respond differently to changes in curvature:

Corollary 5 (⌒). *For $E = B = 1$ SGD and up to error $o(\eta^3) + o(1/k)$: the skewness of gradient noise contributes*

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

to the excess final testing loss over SDE (in an eigenbasis of ηH).

Both flat and sharp minima overfit less

Intuitively, sharp minima are robust to slight changes in the average *gradient* and flat minima are robust to slight *displacements* in weight space (Figure 5 □▦). However, as SGD by definition equates displacements with gradients, it may be unclear how to reason about overfitting in the presence of curvature. Our theory accounts for the implicit regularization of fixed-T descent and shows that both effects play a role. In fact, by routine calculus on Corollary 6, overfitting is maximized for medium minima with curvature $H \sim (\eta T)^{-1}$.

Corollary 6 (from ⌒, ⌒). *Initialize GD at a non-degenerate test minimum $\theta_\star$. The overfitting (testing loss minus $l(\theta_\star)$) and generalization gap (testing minus training loss) due to training are:*

$$\sum_{\mu\nu\rho\lambda} (C/(2NH))_{\mu\nu}^{\rho\lambda} \left((I - \exp(-\eta TH))^{\otimes 2} \right)_{\rho\lambda}^{\mu\nu} + o(\eta^2)$$

and

$$\sum_{\mu\nu\rho\lambda} (C/(2NH))_{\mu\nu}^{\mu\lambda} \ (I - \exp(-\eta TH))_{\lambda}^{\nu} + o(\eta)$$

The generalization gap tends to $C_{\mu\nu}(H^{-1})^{\mu\nu}/N$ as $T \to \infty$. For maximum likelihood (ML) estimation in well-specified models near the "true" minimum, $C = H$ is the Fisher metric, so we recover the AIC: (model dimension)/N. Unlike AIC, our more general expression is descendably smooth, may be used with MAP or ELBO tasks instead of just ML, and does not assume a well-specified model.

Experiments

Our theory does not control our Taylor series' rates of convergence. We thus test our theory by experiment. We perceive support for our theory in drastic rejections of the null hypothesis. For instance, in Figure 3, (Chaudhari and Soatto 2018) predict a velocity of 0 while we predict a velocity of $\eta^2/6$. Likewise, published intuitions (§) suggest that Figure 6 ▯▯▯■ overfitting (test loss minus test minimum) is monotonic in a landscape's hessian, whereas we do not. Here, I bars, + signs, and shaded regions all mark 95% confidence intervals based on the standard error of the mean. § describes neural architectures, artificial landscapes, sample sizes, and further plots.

Discussion

Related work

Kiefer and Wolfowitz (1952) united gradient descent (Cauchy 1847) with stochastic approximation (Robbins and Monro 1951) to invent SGD. Since the development of backpropagation (Werbos 1974), SGD has been used to train connectionist models, e.g. neural networks (Bottou 1991), recently to remarkable success (LeCun, Bengio, and Hinton 2015). Several research programs treat overfitting of SGD-trained networks. (Neyshabur et al. 2017a). Bartlett, Foster, and Telgarsky (2017) controls the Rademacher complexity of deep hypothesis classes, leading to optimizer-agnostic generalization bounds. Yet SGD-trained networks generalize despite their ability to shatter large sets (Zhang et al. 2017), so generalization must arise from not only architecture but also optimization(Neyshabur et al. 2017b).

Some analyses of implicit regularization use a Langevin or SDE approximation (e.g. Chaudhari and Soatto (2018); Zhu et al. (2019)), but, per Yaida (2019b), such continuous-time or uncorrelated-noise analyses treat SGD noise incorrectly. We avoid these pitfalls by Taylor expanding around $\eta = 0$ as in Roberts (2018). Unlike that work, we generalize beyond order η^1 and $T = 2$. Our interpretation of the resulting terms offers a new qualitative picture of SGD as a superposition of simpler information-flow processes. Other research focuses on *double descent* and suggests that some highly overparameterized models share implicit regularization properties with linear least-squares models (Belkin et al. 2019), for example by bounding log-determinants (and hence the effective dimensions) of feature matrices and weight spaces (Mei and Montanari 2020).[1] Our work reveals new dynamics toward and within valleys of minima, dynamics that may also reduce the effective dimension of model

---

[1](Mei and Montanari 2020)'s eq. 75 bounds a log-determinant defined in eq. 61 of a transformed feature matrix. C.f. to linear Representer Theorems (Mohri, Rostamizadeh, and Talwalkar 2018).
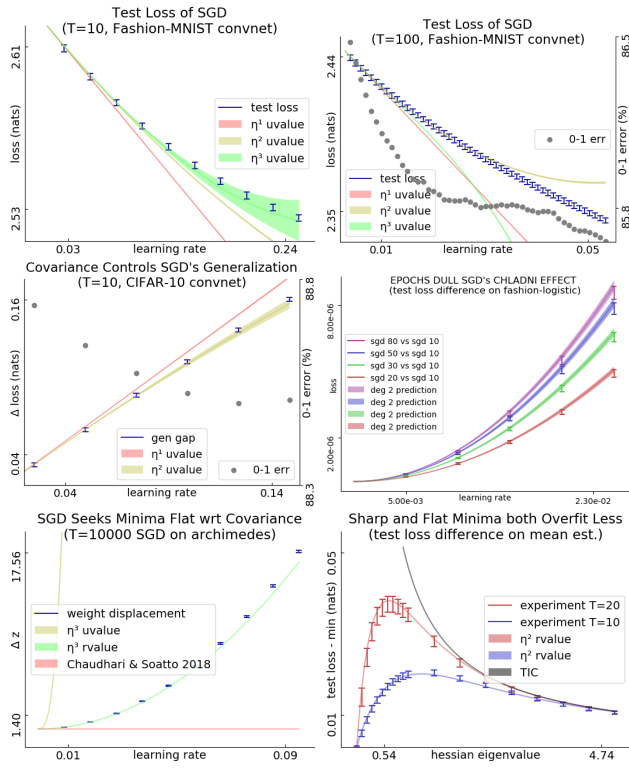
Figure 6: ■□□□: Fashion-MNIST convnet's testing loss vs learning rate. For all initializations tested (1 shown, 11 unshown), the order 3 prediction agrees with experiment through $\eta T \approx 10^0$, corresponding to a decrease in 0-1 error of $\approx 10^{-3}$. □■□□: Fashion-MNIST convnet's testing loss. For large ηT, our predictions break down. Here, the order 3 prediction holds until the 0-1 error improves by $5 \cdot 10^{-3}$. Beyond this, 2nd order agreement with experiment is coincidental. □□■□: CIFAR-10 convnet generalization gaps. For all initializations tested (1 shown, 11 unshown), the degree-2 prediction agrees with experiment through $\eta T \approx 5 \cdot 10^{-1}$. □□□■: For Mean Estimation with fixed C and a range of Hs, initialized at the truth, the testing losses after fixed-T GD are smallest for very sharp and very flat H. Near $H = 0$, our predictions improve on AIC,TIC (Dixon and Ward 2018). **Right**: Predictions near minima excel for large ηT.

space. However, our focus on the structure of gradient noise may be overspecific, since recent work finds that GD and SGD may both converge to the same global minima (Zou et al. 2020) or that noise covariance but not higher moments are relevant to regularization (Wu et al. 2020).

Our predictions are vacuous for large η. Other work treats large-η learning phenomenologically, whether by finding empirical correlates of the generalization gap (Liao et al. 2018), by showing that *flat* minima generalize (Hoffer, Hubara, and Soudry 2017; Keskar et al. 2017; Wang et al. 2018), or by showing that *sharp* minima generalize (Stein 1956; Dinh et al. 2017; Wu, Ma, and E 2018). SGD's implicit regularization mediates between these seemingly clashing intuitions (§**??**).

Prior work analyzes SGD perturbatively: (Dyer and Gur-Ari 2019) perturb in inverse network width, using 't Hooft diagrams to correct the Gaussian Process approximation for specific nets. Perturbing to order η^2, (Chaudhari and Soatto 2018) and (Li, Tai, and E 2017) assume uncorrelated Gaussian noise while (Barrett and Dherin 2021) compares GD to ODE. By contrast, we use Penrose diagrams (Penrose 1971) to compute testing losses and to compare to ODE and SDE to *arbitrary order* in η. We allow correlated, non-Gaussian noise and thus *any* smooth architecture.

Conclusion

This paper presents a new physics-inspired perspective on SGD. We use diagrams to study stochastic optimization on short timescales or near minima. Generalizing (Liao et al. 2018; Wei and Schwab 2019; Zhu et al. 2019; Barrett and Dherin 2021), we model correlated, non-gaussian, non-isotropic, non-constant gradient noise and find qualitative differences in dynamics. For example, we construct a non-pathological loss landscape on which SGD's trajectory *ascends*. We verify our theory on convolutional CIFAR-10 and Fashion-MNIST landscapes. Corollaries 1 and 6 together show that SGD avoids curvature and noise, which to leading order control generalization.

Our theory offers a new physics-inspired perspective of SGD as a superposition of concurrent information-flow processes. We interpret the resulting terms as describing processes by which data influence weights. Notating processes with such diagrams, we show how to compute the effect of each process and that summing the finitely many processes with d or fewer edges suffices to answer dynamical questions to error $o(\eta^d)$. We thus factor the analysis of SGD into the analyses of individual processes, a technique that may power future theoretical inquiries.

Since our predictions depend only on loss data near initialization, they break down after the weight moves far from initialization. Our theory thus best applies to small-movement contexts, whether for long times (large ηT) near an isolated minimum or for short times (small ηT) in general. Thus, the theory might aid future analysis of fine-tuners such as (Finn, Abbeel, and Levine 2017)'s MAML.

Much as meteorologists understand how warm and cold fronts interact despite long-term forecasting's intractability, we quantify how curvature and noise contribute to counterintuitive dynamics governing each short-term interval of SGD's trajectory. Equipped with our theory, users of deep learning may refine intuitions — e.g. that SGD descends on the training loss — to account for noise.
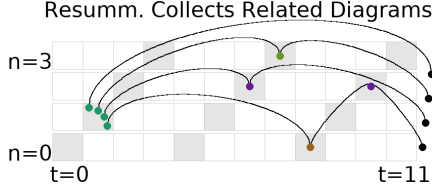
Figure 7: **Resummation propagates information damped by curvature**. Each resummed value (here, for ⌢) represents many un-resummed values, four shown here, each modulated by the Hessian (∨) in a different way.

$$
\begin{aligned}
G_\mu &= \mathbb{E}[\nabla_\mu l_x(\theta)] &\longleftrightarrow \\
H_{\mu\nu} &= \mathbb{E}[\nabla_\mu \nabla_\nu l_x(\theta)] &\longleftrightarrow \\
J_{\mu\nu\xi} &= \mathbb{E}[\nabla_\mu \nabla_\nu \nabla_\xi l_x(\theta)] &\longleftrightarrow \\
C_{\mu\nu} &= \mathbb{E}[(\nabla l_x(\theta) - G)^{\otimes 2}]_{\mu\nu} &\longleftrightarrow \\
S_{\mu\nu\xi} &= \mathbb{E}[(\nabla l_x(\theta) - G)^{\otimes 3}]_{\mu\nu\xi} &\longleftrightarrow
\end{aligned}
$$

Figure 8: **Named tensors**, typically evaluated at initialization ($\theta = \theta_0$). Def. **??** explains how diagrams depict tensors.

For a tight comparison, we scale the learning rates appropriately so that, to leading order, few-epoch and many-epoch SGD agree. Then few-epoch and many-epoch SGD differ, to leading order, in their sensitivity to ∇C:

Corollary 7 (⌢). *SGD with $E = B = 1$, $\eta = \eta_0$ avoids high-C regions more than SGD with $E = E_0$, $B = 1$, $\eta = \eta_0/E_0$. Precisely:* $\mathbb{E}[\theta_{E=E_0} - \theta_{E=1}]^\mu = \left(\frac{E_0 - 1}{4E_0}\right) N\eta^{\mu\rho}\eta^{\nu\xi}\nabla_\mu C_{\nu\xi} + o(\eta^2)$.

In sum, high-C regions repel small-(E, B) SGD more than large-(E, B) SGD. We thus extend the $T = 2$ result of (Roberts 2018) and resolve some questions posed therein.

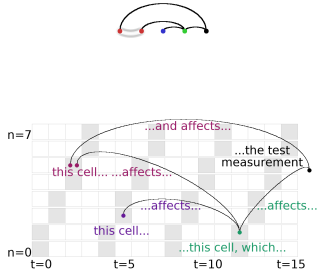Figure 9 shows an instance of the process[1]

Figure 9: **A sub-process of SGD**. Timesteps index columns; training data index rows. The 5th datum participates in the 2nd SGD update. This $(n = 5, t = 2)$ event affects the testing loss both directly and via the $(1, 12)$ event, which is itself modulated by the $(2, 5)$ event.

Taylor series: method and challenges

Let's study $\mathbb{E}[\theta_T]$, $\mathbb{E}[l(\theta_T)]$. To warm up, we'll prove Prop 0 (c.f. (Nesterov 2004; Roberts 2018)).

---

[1]Throughout, colors help us refer to parts of diagrams; colors lack mathematical meaning.

Proof. By gradient bounds: $\theta_T - \theta_0$ is $O(\eta^1)$. We **claim** that $(\theta_T - \theta_0)^\mu = -T\sum_\nu \eta^{\mu\nu}G_\nu + o(\eta^1)$. The claim holds when $T = 0$. Say the claim holds for $\tilde{T}$-step SGD with $T = \tilde{T} + 1$. The displacement $(\theta_T - \theta_{\tilde{T}})^\mu$ evaluates to:

$$
\begin{aligned}
&- \sum_\nu \eta^{\mu\nu}\nabla_\nu l_{n_{\tilde{T}}}(\theta_{\tilde{T}}) \\
=& -\sum_\nu \eta^{\mu\nu}\nabla_\nu \left(l_{n_{\tilde{T}}}(\theta_0) + \boxed{\sum_\xi \nabla_\xi l_{n_{\tilde{T}}}(\theta_0)(\theta_{\tilde{T}} - \theta_0)^\xi} + o(\theta_{\tilde{T}} - \theta_0)\right) \\
=& -\sum_\nu \eta^{\mu\nu}\nabla_\nu \left(l_{n_{\tilde{T}}}(\theta_0) + \nabla l_{n_{\tilde{T}}}(\theta_0) \cdot O(\eta^1) + o(O(\eta^1))\right) \\
=& \boxed{-\sum_\nu \eta^{\mu\nu}\nabla_\nu l_{\tilde{T}}(\theta_0)} + o(\eta^1)
\end{aligned}
$$

Applying the induction hypothesis proves the claim. We plug the claim into l's Taylor series:

$$
\begin{aligned}
\mathbb{E}[l(\theta_T) - l(\theta_0)] &= \sum_\mu \nabla_\mu l(\theta_0) \boxed{\mathbb{E}[\theta_T - \theta_0]^\mu} + \mathbb{E}[o(\theta_T - \theta_0)] \\
&= \sum_\mu \nabla_\mu l(\theta_0)(-T\eta G + o(\eta^1)) + o(O(\eta^1)) \\
&= \boxed{-\sum_{\mu\nu} TG_\mu \eta^{\mu\nu}G_\nu} + o(\eta^1)
\end{aligned}
$$

Indeed, due to the assumption of analytic moments, the above expectations of $o(\eta^1)$ terms are still $o(\eta^1)$. □

What happens when we keep higher order terms? MULTIPLE MOMENTS — We used above that, to order η^1, $\mathbb{E}[l(\theta_T)]$ depends on the training data only through the first moment $\boxed{\mathbb{E}[\theta_T - \theta_0]}$. But to compute $\mathbb{E}[l(\theta_T)]$ to higher order, we'd also need the kth moments $M_k^{\mu_0\mu_1\cdots} = \mathbb{E}\left[\prod_i (\theta_T - \theta_0)^{\mu_i}\right]$. We may achieve this by inductively proving multiple **claim**s, one for each moment.

TUPLES OF TIMES — Complications arise even as we compute M_1 to order η^2. We may not neglect the gradient correction $\nabla(\boxed{\nabla l_{n_{\tilde{T}}}(\theta_0) \cdot (\theta_{\tilde{T}} - \theta_0)})$ at the $\tilde{T}$th induction step. As the displacement $\theta_{\tilde{T}} - \theta_0$ contains (to order η^1) $\tilde{T}$ terms, so will the correction. Totalling the correction over time thus yields $\sum_{0 \leq \tilde{T} < T} \tilde{T} = \binom{T}{2}$ summands, each (e.g. $\nabla\nabla l_5 \nabla l_2$) involving a *pair* of times. Order-d corrections represent the joint influence of d-tuples of times. Prop 0's result $\sum_{\tilde{T}} \left(-\eta \nabla l_{\tilde{T}}(\theta_0)\right)$ is degree 1 in T; but the order-d displacement is a degree d polynomial — very divergent — in T.

FACTORING'S FAILURE — To obtain $\boxed{-TG\eta G}$, we multiplied l's derivatives by the expectations of such summands. In contrast to Prop 0, these expectations, even those of a fixed degree in η, now vary in form due to noise: some (e.g. $\nabla\nabla l_5 \nabla l_2$) have statistically independent factors that permit expectations to factor; others (e.g. $\nabla\nabla l_5 \nabla l_5$) do not. This is how ∇l_x's higher cumulants (such as the covariance and skew of the gradient distribution) appear in our analysis.

DIVERSE DERIVATIVES — At order η^3, a hessian correction $\nabla((\theta_{\tilde{T}} - \theta_0) \cdot \nabla\nabla l_{n_{\tilde{T}}}(\theta_0) \cdot (\theta_{\tilde{T}} - \theta_0)/2)$ augments the gradient correction. Then M_1's order-η^3 summands vary in form, even when all expectations factor (as happens on noiseless landscapes). For instance, the hessian and gradient corrections respectively induce order-η^3 summands of $\mathbb{E}[l(\theta_T)]$ such as

$$
\sum_{\substack{\mu\nu\xi \\ \acute{o}\pi\rho}} \eta^{\mu\acute{o}}\eta^{\nu\pi}\eta^{\xi\rho}(\nabla_\mu l_x)(\nabla_{\acute{o}}\nabla_\nu l_y)(\nabla_\pi \nabla_\xi l_z)(\nabla_\rho l)
$$

And M_2, M_3's terms are yet more diverse. In short, a Taylor expansion even to low degrees yields a combinatorial explosion of terms. Our paper develops tools to organize and interpret these terms.

Since ↶ has only one non-rightmost node, it has as many histories as there are timesteps. Since two nodes have degree 1, each history's contribution $G(-\eta(I - \eta H)^{\Delta-1})G$ involves two 1st derivatives: G. We use geometric series to sum from $\Delta = 1$ through $\Delta = T$ to find the testing loss change up to $o(\eta^1)$ error: $-G((I - (I - \eta H)^T)/H)G$. This is a more precise form of the above $-TG\eta G$ result.

The displacement $\mathbb{E}[\theta_T - \theta_0]$ contains many order-η^3 summands, including those of the form

$$\Delta_{xyz}^{\xi} \propto - \sum_{\substack{\mu\nu \\ \acute{o}\pi\rho}} \eta^{\mu\acute{o}} \eta^{\nu\pi} \eta^{\xi\rho} \mathbb{E}\left[(\nabla_\mu l_x)(\nabla_\nu l_y)(\nabla_{\acute{o}}\nabla_\pi\nabla_\rho l_z)\right]$$

where $0 \leq x, y, z < N$ label datapoints. Let $\Delta_\circ = \mathbb{E}[\Delta_{xxz} - \Delta_{xyz}]$ for x, y, z distinct. l_x, l_y, l_z are i.i.d., so: $\Delta_\circ^{\xi} \propto -\sum_{...} \eta^{\mu\acute{o}} \eta^{\nu\pi} \eta^{\xi\rho} C_{\mu\nu} J_{\acute{o}\pi\rho}$ or, schematically, $\boxed{\Delta_\circ^{\xi} \propto -\eta^3 C \nabla H}$. Here, $C_{\mu\nu} = \mathbb{E}_x[\nabla_\mu l_x \nabla_\nu l_x] - G_\mu G_\nu$ is the covariance of gradients, $H_{\pi\rho} = \nabla_\pi\nabla_\rho l$ is l's hessian, and $J_{\acute{o}\pi\rho} = \nabla_{\acute{o}} H_{\pi\rho}$ is l's 'jerk'.[1]

References

Absil, P.-A.; Mahony, R.; and Sepulchre, R. 2007. Optimization Algorithms on Matrix Manifolds, Chapter 4. *Princeton University Press* .

Amari, S.-I. 1998. Natural Gradient Works Efficiently. *Neural Computation* .

Barrett, D.; and Dherin, B. 2021. Implicit Gradient Regularization. *ICLR* .

Bartlett, P.; Foster, D.; and Telgarsky, M. 2017. Spectrally-Normalized Margin Bounds for Neural Networks. *NeurIPS* .

Belkin, M.; Hsu, D.; Ma, S.; and Mandal, S. 2019. Reconciling Modern Machine Learning Practice and the Bias-Variance Trade-off. *PNAS* .

Bottou, L. 1991. Stochastic Gradient Learning in Neural Networks. *Neuro-Nîmes* .

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

[1] 'Jerk' names 3rd derivatives in dynamical systems: ISO 2041.

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

Chladni, E. 1787. Entdeckungen über die Theorie des Klanges. *Leipzig* .

Comon, P. 2014. An Introduction to Tensors. *IEEE Signal Processing Magazine* .

Dinh, L.; Pascanu, R.; Bengio, S.; and Bengio, Y. 2017. Sharp Minima Can Generalize For Deep Nets. *ICLR* .

Dixon, M.; and Ward, T. 2018. Takeuchi Information as a form of Regularization. *Arxiv Preprint* .

Dyer, E.; and Gur-Ari, G. 2019. Asymptotics of Wide Networks from Feynman Diagrams. *ICML Workshop* .

Dyson, F. 1949. The Radiation Theories of Tomonaga, Schwinger, and Feynman. *Physical Review* .

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *ICML* .

Gauss, C. 1823. Theoria Combinationis Obsevationum Erroribus Minimis Obnoxiae, section 39. *Proceedings of the Royal Society of Gottingen* .

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2018. Accurate, Large Minibatch SGD. *Data @ Scale* .

Hoffer, E.; Hubara, I.; and Soudry, D. 2017. Train Longer, Generalize Better. *NeurIPS* .

Keskar, N.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ICLR* .

Kiefer, J.; and Wolfowitz, J. 1952. Stochastic Estimation of the Maximum of a Regression Function. *Annals of Mathematical Statistics* .

Kolář, I.; Michor, P.; and Slovák, J. 1993. Natural Operations in Differential Geometry. *Springer* .

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *UToronto Thesis* .

Landau, L.; and Lifshitz, E. 1951. The Classical Theory of Fields. *Addison-Wesley* .

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

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

Li, Q.; Tai, C.; and E, W. 2017. Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms I. *PMLR* .

Liao, Q.; Miranda, B.; Banburski, A.; Hidary, J.; and Poggio, T. 2018. A Surprising Linear Relationship Predicts Test Performance in Deep Networks. *Center for Brains, Minds, and Machines Memo 91* .

McCullagh, P. 1987. Tensor Methods in Statistics. *Chemical Rubber Company Press* .

Mei, S.; and Montanari, A. 2020. The Generalization Error of Random Features Regression. *Arxiv Preprint* .

Misner, C.; Thorne, K.; and Wheeler, J. 1973. Gravitation. *W.H. Freeman and Company* .

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. Foundations of Machine Learning, Section 6.3.2. *MIT Press* .

Nesterov, Y. 2004. Lectures on Convex Optimization: Minimization of Smooth Functions. *Springer Applied Optimization 87, Section 2.1* .

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017a. Exploring Generalization in Deep Learning. *NeurIPS* .

Neyshabur, B.; Tomioka, R.; Salakhutdinov, R.; and Srebro, N. 2017b. Geometry of Optimization and Implicit Regularization in Deep Learning. *Chapter 4 from Intel CRI-CI: Why and When Deep Learning Works Compendium* .

Nickel, M.; and Kiela, D. 2017. Poincaré Embeddings for Learning Hierarchical Representations. *ICML* .

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *NeurIPS* .

Penrose, R. 1971. Applications of Negative Dimensional Tensors. *Combinatorial Mathematics and its Applications* .

Robbins, H.; and Monro, S. 1951. A Stochastic Approximation Method. *Pages 400-407 of The Annals of Mathematical Statistics.* .

Roberts, D. 2018. SGD Implicitly Regularizes Generalization Error. *NeurIPS: Integration of Deep Learning Theories Workshop* .

Roberts, D. 2019. SGD. *Personal communication* .

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

Sidiropoulos, N.; Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E.; and Faloutsos, C. 2017. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing* .

Stein, C. 1956. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Berkeley Symposium on Mathematical Probability* .

Wang, H.; Keskar, N.; Xiong, C.; and Socher, R. 2018. Identifying Generalization Properties in Neural Networks. *Arxiv Preprint* .

Wei, M.; and Schwab, D. 2019. How Noise Affects the Hessian Spectrum in Overparameterized Neural Networks. *Arxiv Preprint* .

Werbos, P. 1974. Beyond Regression: New Tools for Prediction and Analysis. *Harvard Thesis* .

Wu, J.; Hu, W.; Xiong, H.; Huan, J.; Braverman, V.; and Zhu, Z. 2020. On the Noisy Gradient Descent that Generalizes as SGD. *ICML* .

Wu, L.; Ma, C.; and E, W. 2018. How SGD Selects the Global Minima in Over-Parameterized Learning. *NeurIPS* .

Xiao, H.; Rasul, L.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *Arxiv Preprint* .

Yaida, S. 2019a. A First Law of Thermodynamics for SGD. *Personal Communication* .

Yaida, S. 2019b. Fluctuation-Dissipation Relations for SGD. *ICLR* .

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

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

Zhu, Z.; Wu, J.; Yu, B.; and Ma, J. 2019. The Anisotropic Noise in Stochastic Gradient Descent. *ICML* .

Zou, D.; Cao, Y.; Zhou, D.; and Gu, Q. 2020. Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks. *MLJ* .

Organization of the appendices

The following three appendices serve three respective functions:
- to explain how to calculate using diagrams;
- to prove our results (and pose a conjecture);
- to specify our experimental methods and results.

In more detail, we organize the appendices as follows.

Tutorial: how to use diagrams

This paper presents a new technique for calculating the expected learning curves of SGD in terms of statistics of the loss landscape near initialization. Here, we explain this technique. There are are **four steps** to computing the expected testing loss, or other quantities of interest, after a specific number of gradient updates:
- **Specify, as a grid**, the batch size, training set size, and number of epochs.
- **Draw histories**, of diagrams into the grid, as needed for the desired precision.
- **Evaluate each diagram history**, whether exactly (via rvalues) or roughly (via uvalues).
- **Sum the histories' values** to obtain the quantity of interest as a function of η.

After presenting two example calculations that follow these four steps, we detail each step individually. Though we focus on the computation of expected testing losses, we describe how the four steps may give us other quantities of interest: variances instead of expectations, training statistics instead of testing statistics, or weight displacements instead of losses.

Two example calculations

We illustrate the four step procedure above by using it to answer the following two questions.

Our first example calculation reproduces Prop 0. In other words, it answers the question:

Question 1 (Leading order effect of gradients)**.** *What's the leading order loss decrease $\mathbb{E}[l(\theta_T) - l(\theta_0)]$? We seek an answer expressed in terms of the landscape statistics at initialization: $G, H, C, \cdots$. We expect only G to be relevant.*

Our second example is (an illustrative case of) Corollary 7.

Question 2 (Leading order effect of epochs)**.** *How does multi-epoch SGD differ from single-epoch SGD? Specifically, what is the difference between the final testing losses of the following two versions of SGD?*
- *SGD over $T = M_0 \times N$ time steps, learning rate η_0/M, and batch size $B = 1$*
- *SGD over $T = N$ time steps, learning rate η_0, and batch size $B = 1$*

We seek an answer expressed in terms of the landscape statistics at initialization: $G, H, C, \cdots$.

To make our discussion concrete, we will set $M_0 = 2$; our analysis generalizes directly to larger M_0.

We scaled the above two versions of SGD deliberately, to create an interesting comparison. Specifically, on a noiseless linear landscape $l_x = l \in (\mathbb{R}^n)^*$, the versions attain equal testing losses, namely $l(\theta_0) - T l_\mu \eta^{\mu\nu}$. So Question 2's answer will be second-order (or higher-order) in η.

Computations: Grids   We begin by asking a question about the final testing loss of some form of SGD. We specify the batch size, training set size, and number of epochs of the setting under analysis by drawing an appropriate grid. That is, we
- **draw an** $N \times T$ **grid** and
- shade its cells, shading the (n, t)th cell **when the tth batch includes the nth data point**.

Thus, each column contains B (batch size) many shaded cells and each row contains E (epoch number) many shaded cells.

EFFECT OF GRADIENTS (QUESTION 1)

Question 1 does not specify a batch size, epoch number, or training set size and so does not specify a grid. In fact, we wish to answer the Question for any choice of those hyperparameters. E.g. we'll answer the Question for SGD with hyperparameters $B, E, N = 2, 4, 8$:

Size-Two Batches without Shuffling
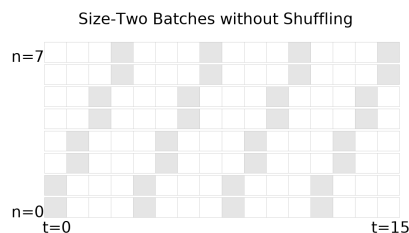
A grid for SGD with batch size $B = 2$ run for $E = 4$ epochs on $N = 8$ training points for a total of $T = 16$ timesteps.

EFFECT OF EPOCHS (QUESTION 2)

Two grids are relevant to Question 2: one for multi-epoch sgd and another for single-epoch SGD. See below.

Two Epoch SGD
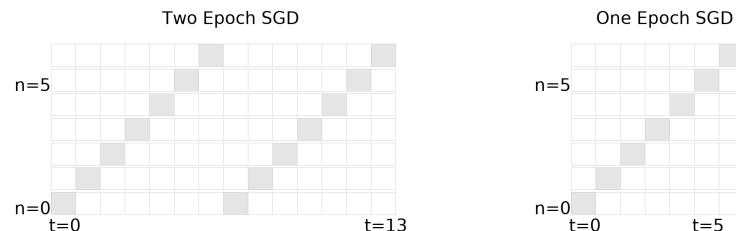
One Epoch SGD

Grids for single-epoch and multi-epoch SGD. Both grids depict $N = 7$ training points and batch size $B = 1$. **Left**: SGD with $M = 2$ update per training sample for a total of $T = MN = 2N$ many updates. **Right**: SGD with $M = 1$ update per training sample for a total of $T = MN = N$ many updates.

Computations: Embeddings of diagrams into grids   Say we permit order-d errors. We draw all relevant diagrams with d or fewer edges and then characterize the histories of those diagrams in §'s grid. An *history* of a diagram D in a grid is an assignment of D's non-root nodes to shaded cells (n, t) obeying the following criteria:

- **time-ordering condition**: the times t strictly increase along each path from leaf to root; and
- **correlation condition**: if two nodes are in the same part of D's partition, then they are assigned to the same datapoint n.

We draw histories by placing nodes in their assigned shaded (n, t) cells; we draw the root nodes outside the grids (at arbitrary positions).

EFFECT OF GRADIENTS (QUESTION 1)

We seek an order 1 result and thus consider one-edged diagrams; there is only one: ⌒. We now describe the histories of this diagram:

Size-Two Batches without Shuffling
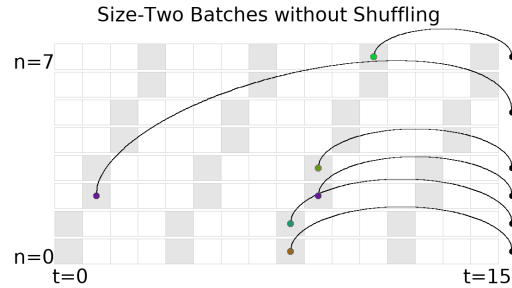
⌒ **has one history for each shaded cell**. An history of a non-root node at cell (n, t) represents the influence of the datapoint n on the testing loss due to the tth update. To first order (i.e. one-edged diagrams), the influences of different timesteps do not interact. The combinatorics of histories is thus straightforward.

EFFECT OF EPOCHS (QUESTION 2)

We seek an order 2 result and thus consider two-edged diagrams; there are four: ⌒, ⌒, ⌒, and ⌒. The figure below shows some histories of order-1 and order-2 diagrams (i.e. one-edged and two-edged diagrams) into the grid relevant to Question 2. Specifically, from top to bottom in each grid, the five diagrams embedded are ⌒ (or ⌒), ⌒, ⌒, ⌒, and ⌒ (or ⌒). The diagram ⌒ may be embedded wherever the diagram ⌒ may be embedded, but not vice versa. Likewise for ⌒ and ⌒.

Spacetime for Two-Epoch SGD One-Epoch SGD
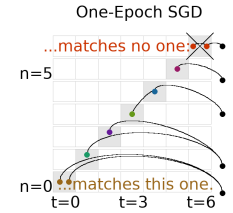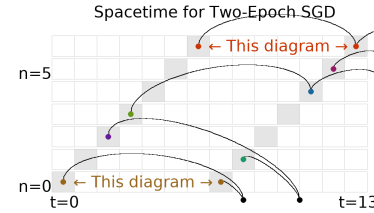
Here, ⌒ **embeds into the multi-epoch but not single-epoch grid.**

Left: ⌒ embeds into the multi-epoch grid. **Right**: ⌒ cannot embed into the single-epoch grid: the correlation condition forces both red nodes into the same row and thus the same cell; the time-ordering condition forces the red nodes into distinct columns and thus distinct cells.

Computations: Evaluating each diagram history   We evaluate §'s diagrams. We choose here to compute uvalues (apt for fixed T), not rvalues. These rules translate diagrams to numbers:
   - **Node rule**: Replace each degree d node by $\nabla^d l_x$.
   - **Outline rule**: surround the nodes in each part of the partition by a "cumulant bracket". If a part contains one node x, the cumulant bracket is the expectation: $\mathbb{E}[x]$. If the part contains two nodes x, y, the cumulant bracket is the covariance: $\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$.[1]
   - **Edge rule**: insert a $\eta^{\mu\nu}$ for each edge. The indices μ, ν should match the corresponding indices of the two nodes incident to the edge.

EFFECT OF GRADIENTS (QUESTION 1)	EFFECT OF EPOCHS (QUESTION 2)

EFFECT OF GRADIENTS (QUESTION 1)

In § we determined the histories of ⌢. Now we evaluate uvalue(⌢). The node rule suggests that we begin with

$$\nabla_\mu l_x \nabla_\nu l_x$$

We have two factors, each with one derivative because the diagram has two nodes, each of degree one. Note that the number of indices (here, two) is the total degree over all nodes and thus also twice the number (here, one) of edges. The outline rule transforms this to

$$\mathbb{E}\left[\nabla_\mu l_x\right]\mathbb{E}\left[\nabla_\nu l_x\right] = G_\mu G_\nu$$

since all parts in the diagram's partition have size one. The edge rule inserts a factor $\eta^{\mu\nu}$ to yield:

$$\text{uvalue}(⌢) = G_\mu G_\nu \eta^{\mu\nu} = G_\mu G^\mu$$

EFFECT OF EPOCHS (QUESTION 2)

In § we saw that ⌢ embeds similarly into multi-epoch and single-epoch grids: its multi-epoch histories correspond by a $M_0^2 : 1$ map to its single-epoch histories. Since we scaled the learning rate of the two SGD versions by a factor of M_0, and since ⌢ (being two-edged) scales as η^2, *the total* uvalue *of its multi-epoch histories will match the total* uvalue *of its single-epoch histories.* So we need not compute ⌢'s contribution.

We see that this cancellation happens for all of the order-2 diagrams *except* for ⌢. Therefore, we must only compute uvalue(⌢).

The node rule suggests that we begin with $\nabla_\mu l_x \nabla_\nu \nabla_\lambda l_x \nabla_\rho l_x$. The outline rule transforms this to

$$\left(\mathbb{E}\left[\nabla_\mu l_x \nabla_\nu \nabla_\lambda l_x\right] - \mathbb{E}\left[\nabla_\mu l_x\right]\mathbb{E}\left[\nabla_\nu \nabla_\lambda l_x\right]\right)\mathbb{E}\left[\nabla_\rho l_x\right] = (\nabla_\nu C_{\mu\lambda}/2)G_\rho$$

The edge rule inserts a factor $\eta^{\mu\lambda}\eta^{\nu\rho}$ to yield:

$$\text{uvalue}(⌢) = (\nabla_\nu C_{\mu\lambda}/2)G_\rho \eta^{\mu\lambda}\eta^{\nu\rho} = G^\nu \nabla_\nu C^\mu_\mu/2$$

---

[1]The general pattern is that the cumulant bracket $\mathbb{C}[\prod_{i\in I} x_i]$ of a product indexed by I is (here, P ranges over partitions of I with at least two parts; $I = \sqcup_{p\in P} p$):

$$\mathbb{C}[\prod_{i\in I} x_i] = \mathbb{E}[\prod_I x_i] - \sum_{\text{partition } P} \prod_{p\in P} \mathbb{C}[\prod_{i\in p} x_i]$$

Computations: Summing the histories' values   Our Key Lemma('s restatement) says that to compute a testing loss,

- we sum §'s uvalues, each weighted by the number of ways its diagram embeds in the grid,
- where histories with s many symmetries count only $1/s$ much toward the total number of histories.

A symmetry of an history f of a diagram D, i.e. an element of $\text{Aut}_f(D)$, is defined to be a relabeling of D's nodes that simultaneously preserves D's rooted tree structure, D's partition structure, and f's assignment of nodes to (n, t) cells of the grid. This is a strong constraint, so there will typically be no symmetries except for the identity, meaning that $s = 1$.

EFFECT OF GRADIENTS (QUESTION 1)

Referring again to §, we see that $D = \frown$ has TB many histories (B many histories per column for T many columns). Since D has no non-trivial automorphisms (i.e. no non-trivial relabeling of nodes that preserves the root, the graph structure, and the equivalence relation on non-root nodes), D has no non-trivial automorphisms that preserve any given history. Thus $|\text{Aut}_f(D)| = 1$ for each history of D. We conclude that the Restated Key Lemma's expression (3)

$$\sum_{\substack{D \text{ a} \\ \text{diagram}}} \sum_{\substack{f \text{ an embed-} \\ \text{-ding of } D}} \frac{(-B)^{-|\text{edges}(D)|}}{|\text{Aut}_f(D)|} \text{uvalue}(D)$$

has as its contribution from $D = \frown$ the value

$$(\text{\#of histories } f) \cdot \frac{(-B)^{-1}}{1} G_\mu G^\mu = TB \cdot (-G_\mu G^\mu / B)$$

Prop 0's expression $-TG_\mu G^\mu$ follows.

EFFECT OF EPOCHS (QUESTION 2)

Referring again to §, we see that $\frown$ has $\binom{M_0}{2} N$ many histories into the multi-epoch grid (one history per pair of distinct epochs, per row) — and no histories into the single-epoch grid. Moreover, each history of $\frown$ has $|\text{Aut}_f(D)| = 1$. We conclude that the testing loss of $M = M_0$ SGD exceeds the testing loss of $M = 1$ SGD by this much:

$$\binom{M_0}{2} N \cdot \frac{(-1)^2}{1} \cdot (\nabla_\nu C_{\mu\lambda}/2) G^\rho \eta^{\mu\lambda} \eta^{\nu\rho} + o(\eta^2)$$

Since Question 2 defines $\eta^2 = \eta_0^2 / M_0^2$, we can rewrite our answer as:

$$l(\theta_{M=M_0, \eta=\eta_0/M_0}) - l(\theta_{M=1, \eta=\eta_0}) = \frac{M_0 - 1}{4M_0} N \cdot G^\nu (\nabla_\nu C_\mu^\mu) + o(\eta_0^2)$$

where we use η_0 to raise indices. This completes the example problem.

How to identify the relevant grid

Diagrams tell us about the loss landscape but not about SGD's batch size, number of epochs, and training set size. We encode this SGD data as a set of pairs (n, t), where we have one pair for each participation of the nth datapoint in the tth update. For instance, full-batch GD has NT many pairs, and singeleton-batch SGD has T many pairs. We will draw these (n, t) pairs as shaded cells in an $N \times T$ grid; we will call the shaded grid the SGD's **grid**. See Figure 10.
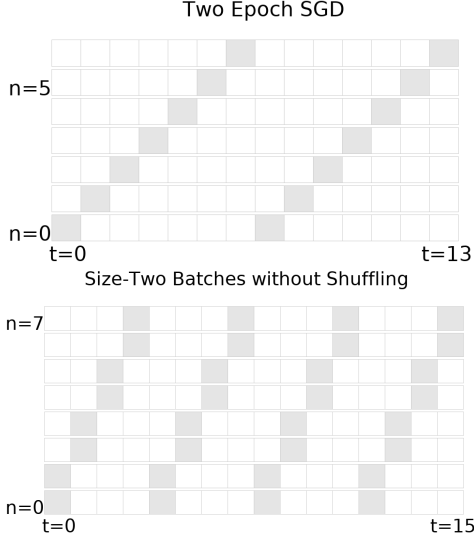
Figure 10: **The grids of two SGD variants.** Shaded cells show (n, t) pairs (see text). **Left**: Two epoch SGD with batch size one. **Right**: Four epoch SGD with batch size two.

Note: A grid may also depict the inter-epoch permuting of training sets due to which the bth batch in one epoch differs from the bth

batch in a different epoch. For instance, see the grid to the right. Since each grid commits to a concrete sequence of training set permutations, we may analyze SGD with randomized permutations by taking expectations over multiple grids. However, the corollaries in this text are invariant to inter-epoch training set permutations, so we will not focus on this point.

How to identify the relevant diagram histories

We explain below what the words mean in this green summary box.

A *diagram* is a finite rooted tree equipped with a partition of its nodes, such that the root node occupies a part of size 1. For example, there are four diagrams with two edges: ⌢, ⌢, ⌢, and ⌢. As always, we specify a diagram's root by drawing it rightmost.

A diagram is *linkless* when each of its degree-2 nodes is in a part of size one. Intuitively, this rules out multi-edge chains unadorned by fuzzy ties. Thus, only the first diagram in the list ⌢, ⌢, ⌢, ⋯ is linkless. Only the first diagram in the list ⌢, ⌢, ⋯ is linkless. Only the first diagram in the list ⌢, ⌢, ⋯ is linkless.

An *history* of a diagram D into a grid is an assignment of D's non-root nodes to shaded cells (n, t) that obeys the following two criteria:
- **time-ordering condition**: the times t strictly increase along each path from leaf to root; and
- **correlation condition**: if two nodes are in the same part of D's partition, then they are assigned to the same datapoint n.

We may conveniently draw histories by placing nodes in the shaded cells to which they are assigned. Then, the time-ordering condition forbids (among other things) intra-cell edges and the correlation condition demands that fuzzily tied nodes are in the same row. See Figure 11.
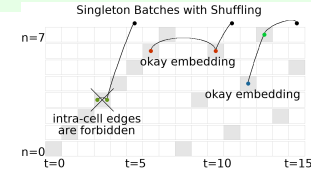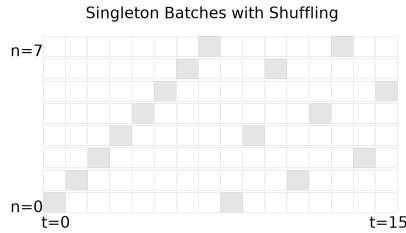
Figure 11: Embeddings, legal and illegal. **Left**: illegal history of ⌢, since the time-ordering condition is not obeyed. For the same reason, not a legal history of ⌢.
Middle: an history of ⌢. Also an history of ⌢, since the correlation condition is obeyed. **Right**: a legal history of ⌢. Not an history of ⌢, since the correlation condition is not obeyed.

In principle, the relevant diagrams for a calculation with error $o(\eta^d)$ are the diagrams with at most d edges. For d greater than 2, there will be many such diagrams. However, in practice we gain insight even from considering one diagram at a time:

Remark. *In this paper's corollaries, we seek to extract the specific effect of a specific landscape or optimization feature such as skewed noise (Prop ??) or multiple epochs (§). In these cases, it is usually the case that most diagrams are irrelevant. For example, because a diagram evaluates to a product of its components, the only way the skewness of gradient noise can appear in our calculations is through diagrams such as* ⌢ *that have a part of size* 3. *Likewise, in § we argued by considering which histories that the only diagram relevant to Question 2 is* ⌢. ◇

Here are some further examples. Table 1 shows the 6 diagrams that may embed into the grid of $E = B = 1$. It shows each diagram in multiple ways to underscore that diagrams are purely topological and to suggest the ways in which these diagrams may embed into a grid.

$\Theta\left((\eta N)^3 N^{-0}\right)$	$\Theta\left((\eta N)^3 N^{-1}\right)$

Table 1: **Multiple ways to draw the** 6 **distinct degree-**3 **diagrams for** $B = E = 1$ **SGD's testing loss.** Because the grid of $B = E = 1$ SGD has only one cell per row and one cell per column, the only diagrams that have a non-zero number of histories are the diagrams such that each ancestor-descendant pair in the rooted tree occupie two different parts of the partition. We show $(4 + 2) + (2 + 2 + 3) + (1)$ ways to draw the 6 diagrams. In fact, these drawings show all of the time-orderings of the diagrams' nodes that are consistent with the time-ordering condition. **Organization**: We organize the diagrams into columns by the number of parts in their partitions. Because partitions (fuzzy outlines) indicate correlations between nodes (i.e. noise), diagrams with fuzzy outlines show deviations of SGD away from deterministic ODE. The big-Θ notation that heads the columns gives the asymptotics of the sum-over-histories of each diagram's uvalues (for N large and η small even relative to $1/N$). **Left**: Diagrams for ODE behavior. **Center**: 1st order deviation of SGD away from ODE. **Right**: 2nd order deviation of SGD from ODE with appearance of non-Gaussian statistics.

How to evaluate each history

We will discuss how to compute both rvalues and uvalues. Both are ways of turning a diagram history into a number. The paper body mainly mentions rvalues. uvalues are simpler to calculate, since they depend only on a diagram's topology, not on the way it is embedded. Physical intution suggests that rvalues are more accurate; in particular, when we initialize near a non-degenerate local minimum, rvalues do not diverge to $\pm\infty$ as $T \to \infty$.

We will explain the following green summary box.

TURN AN HISTORY OF A DIAGRAM into its uvalue or rvalue by applying the fo[...]
• **Node rule**: Replace each degree d node by $\nabla^d l_x$.
• **Outline rule**: surround the nodes in each part of the partition by a[...] the cumulant bracket is the expectation: $\mathbb{E}[x]$. If the part contains two[...] $\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$. (The general pattern is that the cumulant bracket $\mathbb{C}[\prod_{i \in I} x_i]$ o[...] I with at least two parts and $I = \sqcup_{p \in P} p$): $\mathbb{C}[\prod_{i \in I} x_i] = \mathbb{E}[\prod x_i] - \sum_{\text{partition } P} \prod_{p \in P}$[...] fine partitions, i.e., those partitions whose parts have size one.)
• If we wish to compute a uvalue, then we apply the **Edge rule for uval**[...] match the corresponding indices of the two nodes incident to the edge.
• If we wish to compute an rvalue, then we apply the **Edge rule for rvalu**[...] insert a factor of $K^{|t'-t|-1}\eta$, where $K \triangleq (I - \eta H)$. Here, we consider the ro[...]

Un-resumed values: uvalue(D) Each part in a diagram's partition looks like one of the following fragments (or one of the infinitely many analogous fragments):

$$G \triangleq \mathbb{E}_x[\nabla l_x(\theta)] \triangleq$$

$$H \triangleq \mathbb{E}_x[\nabla\nabla l_x(\theta)] \triangleq$$

$$J \triangleq \mathbb{E}_x[\nabla\nabla\nabla l_x(\theta)] \triangleq$$

$$\mathbb{E}_x[(\nabla l_x(\theta) - G)(\nabla\nabla l_x(\theta) - H)] \triangleq$$

$$\mathbb{E}_x[(\nabla\nabla l_x(\theta) - H)(\nabla\nabla l_x(\theta) - H)] \triangleq$$

$$\mathbb{E}_x[(\nabla l_x(\theta) - G)(\nabla\nabla\nabla l_x(\theta) - J)] \triangleq$$

$$C \triangleq \mathbb{E}_x\left[(\nabla l_x(\theta) - G)^2\right]$$

$$S \triangleq \mathbb{E}_x\left[(\nabla l_x(\theta) - G)^3\right]$$

$$\mathbb{E}_x\left[(\nabla l_x(\theta) - G)^4\right] - 3C^2$$

$$\mathbb{E}_x\left[(\nabla l_x(\theta) - G)^5\right] - 10CS$$

The above examples illustrate the **Node rule**: each degree d node evaluates to $\nabla^d l_x$.

Fuzzy outlines dictate how to collect the $\nabla^d l_x$s into expectation brackets. For example, we could collect the nodes within each part (of the partition) into a pair of expectation brackets $\mathbb{E}_x[\cdot]$ — call the result the **moment value**. However, this would yield (un-centered) moments such as $\mathbb{E}_x\left[(\nabla l_x(\theta))^2\right]$ instead of cumulants such as $C = \mathbb{E}_x\left[(\nabla l_x(\theta) - G)^2\right]$. For technical reasons (§?? and §), cumulants will be easier to work with than moments, so we will choose to define the values of diagrams slightly differently as follows.[1]

Outline rule: surround the nodes in each part of the partition by a "cumulant bracket". The cumulant bracket $\mathbb{C}[\prod_{i \in I} x_i]$ of a product indexed by I is (here, P ranges over partitions of I with at least two parts; $I = \sqcup_{p \in P} p$):

$$\mathbb{C}[\prod_{i \in I} x_i] = \mathbb{E}[\prod_{i} x_i] - \sum_{\text{partition } P} \prod_{p \in P} \mathbb{C}[\prod_{i \in p} x_i]$$

Thus, a cumulant bracket of a diagram is the moment bracket of that diagram minus other terms. Those other terms are obtained by considering diagrams with the same graph structure but strictly more parts in their partition. The recursive definition of $\mathbb{C}$ grounds out because the maximal number of parts in a partition of a finite set is finite.

---

[1]This is just the standard Möbius recursion for defining cumulants (see (Rota 1964)).

For example, if a part contains one node x, the cumulant bracket is the expectation: $\mathbb{E}[x]$. If the part contains two nodes x, y, the cumulant bracket is the covariance: $\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$. If a part contains three nodes x, y, z, then the cumulant bracket is

$$\mathbb{E}[xyz] - \mathbb{E}[x]\mathbb{E}[yz] - \mathbb{E}[y]\mathbb{E}[xz] - \mathbb{E}[z]\mathbb{E}[yz] + 2\mathbb{E}[x]\mathbb{E}[y]\mathbb{E}[z]$$

We visualize the above in the following example:

Example 1. For example, if we denote moment values by solid gray fuzzy ties (instead of fuzzy outlines), then:

We will use the concept of "moment values" again in §**??**. ◇

Finally, we come to edges. **Edge rule**: insert a factor of $\eta^{\mu\nu}$ for each edge. The indices μ, ν should match the corresponding indices of the two nodes incident to the edge.

Example 2 (Un-resummed value)**.** Remember that $\overset{\smile}{} =$ $C_{\mu\nu}$ and $\vee = H_{\lambda\rho}$, so that $\overset{\smile}{}\vee = C_{\mu\nu}H_{\lambda\rho}$. Then

$$\text{uvalue}(\overset{\frown}{}) = C_{\mu\nu}H_{\lambda\rho}\eta^{\mu\lambda}\eta^{\nu\rho}$$

Here, $\overset{\frown}{}$ has two edges, which correspond in this example to the tensor contractions via $\eta^{\mu\lambda}$ and via $\eta^{\nu\rho}$, respectively. ◇

Resummed values: $\text{rvalue}_f(D)$ The only difference between rvalues and uvalues is in their rule for evaluating edges.

Edge rule: if an edge's endpoints are embedded to times t, t', insert a factor of $K^{|t'-t|-1}\eta$, where $K \triangleq (I - \eta H)$. Here, we consider the root node as embedded to the time T.

Example 3 (Re-summed value)**.** Recall as in Example 2 that $\overset{\smile}{} = C_{\mu\nu}$ and $\vee = H_{\lambda\rho}$, so that $\overset{\smile}{}\vee = C_{\mu\nu}H_{\lambda\rho}$. Then if f is an history of $\overset{\frown}{}$ that sends the diagram's red part to a time t (and its green root to T), we have:

$$\text{rvalue}_f(\overset{\frown}{}) = C_{\mu\nu}H_{\lambda\rho}\left(K^{T-t-1}\eta\right)^{\mu\lambda}\left(K^{T-t-1}\eta\right)^{\nu\rho}$$

Here, $\overset{\frown}{}$ has two edges, which correspond in this example to the tensor contractions via $(K^{\cdots}\eta)^{\mu\lambda}$ and via $(K^{\cdots}\eta)^{\nu\rho}$, respectively. ◇

Overall In sum, we evaluate an history of a diagram by using the **node**, **outline**, and **edge** rules to build an expression of $\nabla^d l_x$s, $\mathbb{E}_x$s and ηs. The difference between uvalues and rvalues lies only in their edge rule.

How to sum the histories' values

We give examples of automorphism groups and we illustrate the integration mentioned in this green summary box:

WE OBTAIN OVERALL final testing loss expressions by adding together automorphism-group sizes as in 1.

If we are using rvalues instead of uvalues, we approximate sums over his exp$(-\eta Ht)$, and we apply:rule

$$\int_{0 \le u < T} du \exp(-uA) = (I - \exp$$

When written in an eigenbasis of ηH, this A's coefficients are sums of on edge involved in the relevant degrees of freedom over which we integrate

The Restated Key Lemma and Theorem 1 together say

Theorem. *For any T: for η small enough, SGD has expected testing loss*

$$\sum_{\substack{D \text{ a linkless} \\ \text{diagram}}} \sum_{\substack{f \text{ an embed-} \\ \text{-ding of } D}} \frac{(-B)^{-|\text{edges}(D)|}}{\left|\text{Aut}_f(D)\right|} \text{rvalue}_f(D)$$

which is the same as

$$\sum_{\substack{D \text{ a} \\ \text{diagram}}} \sum_{\substack{f \text{ an embed-} \\ \text{-ding of } D}} \frac{(-B)^{-|\text{edges}(D)|}}{\left|\text{Aut}_f(D)\right|} \text{uvalue}(D)$$

Here, B is the batch size.

How do we evaluate the above sum? Summing uvalues reduces to counting histories, which in all the applications reported in this text is a routine combinatorial exercise. However, when summing rvalues, it is often convenient to replace a sum over histories by an integral over times, and the power $(I - \eta H)^{\Delta t - 1}$ by the exponential $\exp{-\Delta t\eta H}$. This incurs a term-by-term $1 + o(\eta)$ error factor, meaning that it preserves leading order results.

Example 4. Let us return to $D = \overset{\frown}{}$, embedded, say, in the grid of one-epoch one-sample-per-batch SGD. From Example 3, we know that we want to sum the following value over all histories f, i.e. over all $0 \le t < T$ to which the red part of the diagram's partition may be assigned:

$$\text{rvalue}_f(\overset{\frown}{}) = C_{\mu\nu}\left(K^{T-t-1}\eta\right)^{\mu\lambda}\left(K^{T-t-1}\eta\right)^{\nu\rho}H_{\lambda\rho}$$

Each history has a factor $(-B)^{-|\text{edges}(D)|}/\left|\text{Aut}_f(D)\right| = (-B)^{-2}/2$; we will multiply in this factor at the end so we now focus on the $\sum_f$. So, using the aforementioned approximation, we seek to evaluate

$$\int_{0 \le t < T} dt\, C_{\mu\nu}\left(\exp\left(-(T-t)\eta H\right)\eta\right)^{\mu\lambda}\left(\exp\left(-(T-t)\eta H\right)\eta\right)^{\nu\rho}H_{\lambda\rho} =$$

$$C_{\mu\nu}\left(\int_{0 \le t < T} dt\, \exp\left(-(T-t)((\eta H)\otimes I + I \otimes (\eta H))\right)^{\mu\nu}_{\pi\sigma}\right)\eta^{\pi\lambda}\eta^{\sigma\rho}H_{\lambda\rho}$$

We know from linear algebra and calculus that $\int_{0 \le u < T} du \exp(-uA) = (I - \exp(-TA))/A$ (when A is a non-singular linear endomorphism). Applying this rule for

$u = T - t$ and $A = (\eta H) \otimes I + I \otimes (\eta H)$, we evaluate the integral as:

$$\cdots = C_{\mu\nu} \left(\frac{I - \exp\left(-T((\eta H) \otimes I + I \otimes (\eta H))\right)}{(\eta H) \otimes I + I \otimes (\eta H)} \right)^{\mu\nu}_{\pi\sigma} \eta^{\pi\lambda}\eta^{\sigma\rho} H_{\lambda\rho}$$

This is perhaps easier to write in an eigenbasis of ηH:

$$\cdots = \sum_{\mu\nu} C_{\mu\nu} \frac{1 - \exp\left(-T((\eta H)^\mu_\mu + (\eta H)^\nu_\nu)\right)}{(\eta H)^\mu_\mu + (\eta H)^\nu_\nu} (\eta H \eta)^{\mu\nu}$$

Multiplying this expression by the aforementioned $(-B)^{-2}/2$ gives the contribution of ⌢ to SGD's test loss. ◇

An **automorphism of D that preserves an history f of** D is a relabeling of D's nodes that preserves the root, the graph structure, the equivalence relation on non-root nodes, and that at the same time respects f: f must send a node to the same (n, t) pair to which the node's relabeling is sent. The sizes of automorphism groups appear in the denominators of our main results. They are usually of size one (that is, they usually contain only the the identity relabeling).

Example 5 (Automorphisms). We take as examples the diagram histories of §?? (figure reproduced here). All the histories in the **left figure** have $|\mathrm{Aut}_f(D)| = 1$. For instance, the bottom-most history (of ⌢) has a non-trivial relabeling (namely, swap the two non-root nodes) that obeys all of the automorphism conditions *except* the "respects f" condition. Indeed, the two non-root nodes are assigned to different cells in the grid, so we may not swap them without violating the "respects f" condition.

By contrast, one of the histories (among the valid histories) shown in the **right figure** has a non-trivial automorphism group. This is the bottom-most history (of ⌢, again). Observe that swapping that diagram's two non-root nodes preserves the root, preserves the graph structure, and preserves the equivalence relation on non-root nodes. Moreover, such swapping respects f, since the two swapped nodes embed into the same cell. Thus, in this case, $|\mathrm{Aut}_f(D)| = 2$. ◇

Spacetime for Two-Epoch SGD

One-Epoch SGD
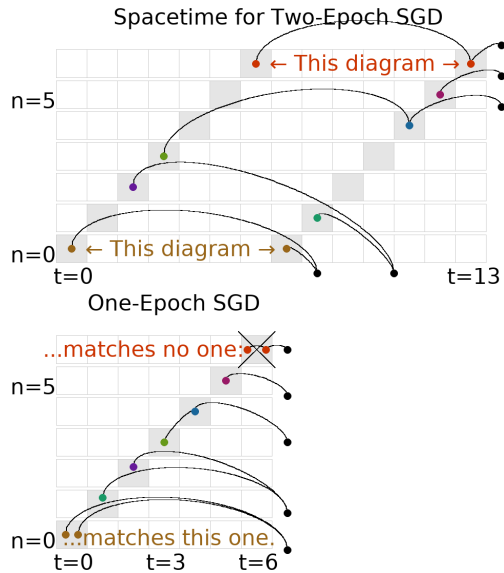
How to solve variant problems

In §, we briefly discuss second-order methods and natural gradient descent. Here, we briefly discuss modifications. We omit proofs, which would closely follow §'s proof of the expectation-of-test-loss case.

Variance (instead of expectation) To compute variances instead of expectations (with respect to the noise in the training set), one considers generalized diagrams that have "two roots" instead of one. More precisely, to compute, say, the un-centered second moment of testing loss, one uses diagrams whose edge structures are not rooted trees but instead forests consisting of two rooted trees. We require that the set of roots (now a set of size two instead of size one) is a part of the diagram's partition. We draw the two roots rightmost.

For example, the generalized diagrams ⌢ or ⌢⌢ may appear in this computation.

Measuring on the training (instead of test) set To compute the training loss, we compute with all the same diagrams as the testing loss, and we also allow all the additional generalized diagrams that violate the constraint that a diagram's root should be in a part of size one. Therefore, to compute the generalization gap (i.e. testing loss minus training loss), we sum over all the diagrams that expressly violate this constraint (and then, since gen. gp is test minus train instead of train minus test, we multiply the whole answer by -1). For example, the generalized diagrams ⌢ or ⌢⌢ may appear in this computation.

Weight displacement (instead of loss) To compute displacements instead of losses, one considers generalized diagrams that have a "loose end" instead of a root. For example, the generalized diagrams ⌄ or ⌢ may appear in this computation.

Do diagrams streamline computation?

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[1] to the new formalism on two sample problems.

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

---

[1]By "direct perturbation", we mean direct application of our Key Lemma (§).

manifestly agree in their first and second moments. This allows us to save some lines.

Despite these efforts, the diagram method yields arguments about *four times shorter* — and strikingly more conceptual — than direct perturbation yields. (We make no attempt to compare the re-summed version of our formalism to direct perturbation because the algebraic manipulations involved for the latter are too complicated to carry out.) 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 now compare Diagram Rules vs Direct Perturbation .

Effect of batch size   We compare the testing losses of pure SGD and pure GD. Because pure SGD and pure GD differ in how samples are correlated, their testing 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 histories of ⌢ in distinct epochs of GD's spacetime as there are in distinct epochs of SGD's spacetime. However, each same-epoch history of ⌣ within any one epoch of GD's spacetime corresponds by vertical shifts to an history of ⌣ in SGD. There are $MN\binom{N}{2}$ many such histories in GD's spacetime, so GD's testing loss exceeds SGD's by $\frac{MN\binom{N}{2}}{N^2}$ ⌣. Reading the diagram's value from its graph structure, we unpack that expression as:

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

DIRECT PERTURBATION —

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

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

To save space, we write l_{n_t} for $l_{n_t}(\theta_0)$. It's enough to solve the recurrence $a(t+1) = a(t) - \nabla l_{n_t}$ and $b(t+1) = b(t) - \nabla^2 l_{n_t} a(t)$. Since $a(0), b(0)$ vanish, we have $a(t) = -\sum_{0 \le t < T} \nabla l_{n_t}$ and $b(t) = \sum_{0 \le t_0 < t_1 < T} \nabla^2 l_{n_{t_1}} \nabla l_{n_{t_0}}$. We now expand l:

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

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

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

Likewise,

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

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

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

and

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

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

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

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

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

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

DIAGRAM METHOD —

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

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

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

DIRECT PERTURBATION —

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

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

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 testing loss of SGD is

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

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

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

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

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

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

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

Mathematics of the theory

Assumptions and Definitions

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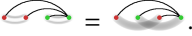
Existence of Taylor Moments — we assume that each finite collection of polynomials of the 0th and higher derivatives of the l_x, all evaluated at any point θ, may be considered together as a random variable insofar as they are equipped with a probability measure upon the standard Borel algebra.

Analyticity Uniform in Randomness — we assume that the functions $\theta \mapsto l_x(\theta)$ — and the expectations of polynomials of their 0th and higher derivatives — exist and are analytic with radii of convergence bounded from 0 (by a potentially θ-dependent function). So expectations and derivatives commute.

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.[1] 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.

Now we turn to definitions.

Definition 5 (Diagrams)**.** *A diagram is a finite rooted tree equipped with a partition of non-root nodes. We draw the tree using thin "edges". By convention, we draw each node to the right of its children; the root is thus always rightmost. We draw the partition by connecting the nodes within each part via fuzzy "ties". For example, has 2 parts. We insist on using as few fuzzy ties as possible so that, if d counts edges and c counts ties, then $d + 1 - c$ counts parts of the partition. There may be multiple ways to draw a single diagram, e.g. = .*

Definition 6 (Embedding a Diagram into a Grid)**.** *An history of a diagram into a grid is an assignment of that diagram's non-root nodes to pairs (n, t) such that each node occurs at a time t' strictly after each of its children and such that two nodes occupy the same row n if they inhabit the same part of D's partition. In situations of interest, the permissible pairs (n, t) are those pairs such that the nth trainingpoint participates in the tth SGD update. We prefer to draw all such pairs by shading the corresponding cells in a square grid whose rows are indexed by training points n and whose columns aare indexed by timesteps t.*

We define $\mathrm{uvalue}(D)$ and $\mathrm{rvalue}_f(D)$ as in §.

A key lemma à la Dyson

Suppose s is an analytic function defined on the space of weights. The following Lemma, reminiscent of (Dyson 1949), helps us track $s(\theta)$ as SGD updates θ:

---

[1]Some of our experiments involve Gaussian noise, which is not bounded and so violates the hypothesis. In practice, Gaussians are effectively bounded in 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.

Key Lemma. *For all T: for η sufficiently small, $s(\theta_T)$ is a sum over tuples of natural numbers:*

$$s(\theta_T) = \sum_{(d_t : 0 \le t < T) \in \mathbb{N}^T} (-\eta)^{\sum_t d_t} \left(\prod_{0 \le t < T} \left(\frac{(g\nabla)^{d_t}}{d_t!} \Big|_{g = \sum_{n \in \mathcal{B}_t} \nabla l_n(\theta)/B} \right) \right)(s)(\theta_0) \tag{1}$$

Moreover, the expectation symbol (over training sets) commutes with the outer sum.

Here, we consider each $(g\nabla)^{d_t}$ as a higher order function that takes in a function f defined on weight space and outputs a function equal to the d_tth derivative of f, times g^{d_t}. The above product then indicates composition of $(g\nabla)^{d_t}$'s across the different t's. In total, that product takes the function s as input and outputs a function equal to some polynomial of s's derivatives.

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

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

as norms, SGD after T steps on any train set will necessarily stay within the ρ-ball.[2] We note that the above condition on η is weak enough to permit all η within some open neighborhood of $\eta = 0$.

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

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

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

Now we use the standard expansion of exp. Because (by analyticity) the order d coeffients of l_t, s are bounded by

---

[2]The 2 ensures that SGD initialized at any point within a $\rho/2$ ball will necessarily stay within the ρ-ball.

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

$$\cdots = \sum_{0 \leq d < \infty} (-\eta)^d \sum_{\substack{(d_t : 0 \leq t < T) \\ \sum_t d_t = d}} \left(\prod_{0 \leq t < T} \left. \frac{(g\nabla)^{d_t}}{d_t!} \right|_{g = \nabla l_t(\theta)} \right) s(\theta) \quad (3)$$

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

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

has an expectation that decays exponentially with d. The symbol c_d we introduce purely for convenience; that its value depends on the train set we emphasize using function application notation. Crucially, no matter the train set, we have shown that the expansion 3 (that features c_d appear as coefficients) converges to an analytic function for all η bounded as in condition 2. 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.

Finally, $|c_d|$ indeed has a well-defined expected value, for $|c_d|$ 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. So Fubini's Theorem applies. The Key Lemma follows. □

From Dyson to diagrams

We now describe the terms that appear in the Key Lemma. The following result looks like Theorem 1, except it has $\text{uvalue}(D)$ instead of $\text{uvalue}_f(D)$, and the sum is over all diagrams, not just linkless ones. In fact, we will use Theorem 3 to prove Theorem 1.

Theorem 3 (Test Loss as a Path Integral). *For all T: for η sufficiently small, SGD's expected test loss is*

$$\sum_D \sum_{\text{histories } f} \frac{1}{|\text{Aut}_f(D)|} \frac{\text{uvalue}(D)}{(-B)^{|\text{edges}(D)|}}$$

Here, D is a diagram whose root r does not participate in any fuzzy edge, f is an history of D into a grid, and $\left| \text{Aut}_f(D) \right|$ counts the graph-automorphisms of D that preserve f's assignment of nodes to cells. If we replace D by $\left(-\sum_{p \in \text{parts}(D)} (D_{rp} - D)/N \right)$, where r is D's root, we obtain the expected generalization gap (testing minus training loss).

Theorem 3 describe the terms that appear in the Key Lemma by matching each term to an history of a diagram

in a grid, so that the infinite sum becomes a sum over all diagram grid 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.

Proof of Theorem 3. We first prove the statement about testing 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 $(l_n : 0 \leq n < N)$, grid S, function $\pi : S \to [N]$ that induces $\sim$, and natural d — that

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

$$\sum_{\substack{D \in \text{im}(\mathcal{F}) \\ \text{with } d \text{ edges}}} \left(\sum_{f : D \to \mathcal{F}(S)} \frac{1}{|\text{Aut}_f(D)|} \right) \frac{\text{uvalue}_\pi(D, f)}{B^d} \quad (4)$$

Here, uvalue_π is the value of a diagram history before taking expectations over train sets. We have for all f that $\mathbb{E}\left[\text{uvalue}_\pi(D, f)\right] = \text{uvalue}(D)$. Observe that both sides of 4 are finitary sums.

Remark 4 (Differentiating Products). *The product rule of Leibniz easily generalizes to higher derivatives of finitary products:*

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

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

We proceed by joint induction on d and S. The base cases wherein S is empty or $d = 0$ both follow immediately from the Key Lemma, for then the only history is the unique history of the one-node diagram $\cdot$. For the induction step, suppose S is a sequence of $\mathcal{M} = \min S \subseteq S$ followed by a strictly smaller S and that the result is proven for $(\tilde{d}, \tilde{S})$ for every $\tilde{d} \leq d$. Let us group by d_0 the terms on the left hand side of desideratum 4. Applying the induction hypothesis with $\tilde{d} = d - d_0$, we find that that left hand side is:

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

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

Since $\text{uvalue}_\pi(\tilde{D}, \tilde{f})$ 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 $\text{uvalue}_\pi(\tilde{D}, \tilde{f})$ to produce $(d - d_0 + 1)^{d_0}$ terms. In fact, $g = \sum_{m \in \mathcal{M}} \nabla l_m(\theta)/B$ expands to $B^{d_0}(d - d_0 + 1)^{d_0}$ terms, each conveniently indexed by a pair of functions

$\beta : [d_0] \to \mathcal{M}$ and $\nu : [d_0] \to \tilde{D}$. The (β, ν)-term corresponds to an history f of a larger diagram D in the sense that it contributes $\mathrm{uvalue}_\pi(D, f)/B^{d_0}$ to the sum. Here, (f, D) is $(\tilde{f}, \tilde{D})$ with $\left|(\beta \times \nu)^{-1}(n, v)\right|$ many additional edges from the cell of datapoint n at time 0 to the vth node of $\tilde{D}$ as embedded by $\tilde{f}$.

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

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

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

by the Orbit Stabilizer Theorem of elementary group theory.

It is thus enough to show that

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

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

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

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

The argument for generalization gaps parallels the above when we use $l - \sum_n l_n/N$ instead of l as the value for s. The orem 3 is proved. $\square$

Remark 5 (The Case of $E = B = 1$ SGD). *The grid of $E = B = 1$ SGD permits all and only those histories that assign to each part of a diagram's partition a distinct cell. Such histories factor through a diagram ordering and are thus easily counted using factorials per Prop 1. That prop immediately follows from the now-proven Theorem 3.*

Prop 1. *The order η^d contribution to the expected testing loss of one-epoch SGD with singleton batches is:*

$$\frac{(-1)^d}{d!} \sum_D \left|\mathrm{ords}(D)\right| \binom{N}{P-1} \binom{d}{d_0, \cdots, d_{P-1}} \mathrm{uvalue}(D)$$

where D ranges over d-edged diagrams. Here, D's parts have sizes $d_p : 0 \leq p \leq P$, and $\left|\mathrm{ords}(D)\right|$ counts the total orderings of D s.t. children precede parents and parts are contiguous.

Proof of Theorem 1

(We say an history is **strict** if it assigns to each part a different datapoint n. Then, by Möbius inversion ((Rota 1964)), a sum over strict histories of moment values (§) matches a sum over all histories of uvalues.)

The diagrams summed in Theorem 1 and 2 may be grouped by their geometric realizations. Each nonempty class of diagrams with a given geometric realization has a unique element with minimally many edges, and in this way all and only linkless 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 histories of a specific member of that class might be hard to count. The first we handle using Orbit-Stabilizer. The second we address as described by §**??** via Möbius sums.

Proof of Theorem 1. We apply Möbius inversion (§**??**) to Theorem 3 (§).

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

Because the Taylor series for analytic functions converge absolutely in the interior of the disk of convergence, the rearrangement of terms corresponding to a grouping by geometric realizations preserves the convergence result of Theorem 3.

Let us then focus on those diagrams σ with a given geometric realization represented by an linkless diagram ρ. By Theorem 3, it suffices to show that

$$\sum_{f:\rho\to S} \sum_{\substack{\tilde{f}:\sigma\to S \\ \exists i_\star : f = \tilde{f} \circ i_\star}} \frac{1}{\left|\mathrm{Aut}_{\tilde{f}}(\sigma)\right|} = \sum_{f:\rho\to S} \sum_{\substack{\tilde{f}:\sigma\to S \\ \exists i_\star : f = \tilde{f} \circ i_\star}} \sum_{\substack{i:\rho\to\sigma \\ f = \tilde{f} \circ i}} \frac{1}{\left|\mathrm{Aut}_f(\rho)\right|} \tag{5}$$

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

In equation 5, we have introduced redundant sums to structurally align the two expressions on the page; besides this rewriting, we see that equation 5's left hand side matches Theorem 3 resulting formula and tgat its right hand side is the desired formula of Theorem 1.

To prove equation 5, it suffices to show (for any $f, \tilde{f}, i$ as above) that

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

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

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

Meanwhile, the orbit is all of $|X|$. Indeed, suppose $i_A, i_B \in X$. We will present $\psi \in \text{Aut}_f(\rho)$ such that $i_B \sim i_A \circ \psi$ by $\phi = $ identity. We simply define $\psi = i_A^{-1} \circ i_B$, well-defined by the aforementioned (isomorphisms and strict increase) properties. It is then routine to verify that $f \circ \psi = \tilde{f} \circ i_A \circ i_A^{-1} \circ i_B = \tilde{f} \circ i_B = f$. So the orbit has the claimed size, and by the Orbit Stabilizer Theorem, the coefficients in the expansions of Theorems 1 and 3 match. □

Proof of Theorem 2

Proof of Theorem 2. Since we assumed hessians are positive: for any m, the propagator $K^t = ((I - \eta H)^{\otimes m})^t$ exponentially decays to 0 (at a rate dependent on m). Since up to degree d only a finite number of diagrams exist and hence only a finite number of possible ms, the exponential rates are bounded away from 0. Moreover, for any fixed t_{big}, the number of diagrams — involving no exponent t exceeding t_{big} — is eventually constant as T grows. Meanwhile, the number involving at least one exponent t exceeding that threshold grows polynomially in T (with degree d). The exponential decay of each term overwhelms the polynomial growth in the number of terms, and Theorem's first part follows. □

Proofs of corollaries

Corollary 1

Proof. The relevant linkless diagram is colorred (amputated as in the previous subsubsection). An history of this diagram into $E = B = 1$ SGD's grid is determined by two durations — t from red to green and $\tilde{t}$ from green to blue — obeying $t + \tilde{t} \leq T$. The automorphism group of each history has size 2: identity or switch the red nodes. So the answer is:

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

Standard calculus then gives the desired result. □

Corollary 6's first part

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

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

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

Corollary 6's second part We apply the generalization gap modification (described in §) to Theorem 1's result about testing losses.

Proof. The relevant linkless diagram is . This diagram has one history for each shaded cell of grid; for GD, the grid has every cell shaded, so each duration from 0 to T is represented. So the generalization gap is, to leading order,

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

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

Corollaries 7 and 2 Corollary 7 and Corollary 2 follow from plugging appropriate values of M, N, B into the following prop.

Prop 2. *To order η^2, the testing loss of SGD — on N samples for $T = MN$ timesteps with batch size B dividing N and with any shuffling scheme — has expectation*

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

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

of Prop 2. To prove Prop 2, we simply count the histories of the diagrams, noting that the automorphism groups are all of size 1 or 2. See Table 2. □

Corollary 5 The corollary's first part follows immediately from Prop 2.

Proof of second part. 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 part (in their partition) of size at

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

Table 2: Terms used in proof of Prop 2

least 3. The only such test diagram is ⌒ . This embeds in T ways (one for each grid cell — recall that $E = B = 1$) and has symmetry factor 1/3! for a total of

$$\frac{T\eta^3}{6}\mathbb{E}\left[\nabla^3 l\right]\mathbb{E}\left[\nabla l_{n_{t_a}}\nabla l_{n_{t_b}}\nabla l_{n_{t_c}}\right] = \frac{T\eta^3}{6}S_{\mu\nu\sigma}J^{\mu\nu\sigma}$$

This is the un-resummed expression. To obtain the re-summed expression, we replace $\eta^{\mu\nu}$ with $(I - \eta H)^{\Delta t - 1}\eta)^{\mu\nu}$. The histories range over T many times uniformly spaced in $[0, T]$. So we may integrate (let's name our variable of integration τ, and let's have it represent $\tau = \Delta t - 1 = T - t - 1$):

$$\int_{0 \le \tau < T}(\exp(-\tau\eta H)\eta)^{\mu\nu}(\exp(-\tau\eta H)\eta)^{\pi\sigma}(\exp(-\tau\eta H)\eta)^{\lambda\rho}S_{\mu\pi\lambda}J_{\nu\sigma\rho}$$

Observe that

$$(\exp(-\tau\eta H))^{\nu}_{\mu}(\exp(-\tau\eta H))^{\sigma}_{\pi}(\exp(-\tau\eta H))^{\rho}_{\lambda} = (\exp(-\tau Y))^{\sigma\lambda\rho}_{\mu\nu\pi}$$

where $Y = \eta H \otimes I \otimes I + I \otimes \eta H \otimes I + I \otimes I \otimes \eta H$ is a six-index tensor. We finish by recalling that

$$\int_t \exp(At) = \frac{\exp(At)}{A}\big|_t$$

$\square$

Future topics

Our diagrams invite exploration of Lagrangian formalisms and curved backgrounds:[1]

Question 3. *Does some least-action principle govern SGD; if not, what is an essential obstacle to this characterization?*

Lagrange's least-action formalism intimately intertwines with the diagrams of physics. Together, they afford a modular framework for introducing new interactions as new terms or diagram nodes. In fact, we find that some *higher-order* methods — such as the Hessian-based update $\theta \leftsquigarrow \theta - (\eta^{-1} + \lambda\nabla\nabla l_t(\theta))^{-1}\nabla l_t(\theta)$ parameterized by small η, λ — admit diagrammatic analysis when we represent the λ term as a second type of diagram node. Though diagrams suffice for computation, it is Lagrangians that most deeply illuminate scaling and conservation laws.

Our work assumes a flat metric $\eta^{\mu\nu}$, but it might generalize to weight spaces curved in the sense of Riemann.[2]

---

[1](Landau and Lifshitz 1960, 1951) review these concepts.

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

Such curvature finds concrete application in the *learning on manifolds* paradigm of (Absil, Mahony, and Sepulchre 2007; Zhang, Reddi, and Sra 2016), notably specialized to (Amari 1998)'s *natural gradient descent* and (Nickel and Kiela 2017)'s *hyperbolic histories*. While that work focuses on *optimization* on curved weight spaces, in machine learning we also wish to analyze *generalization*. Starting with the intuition that "smaller" hypothesis classes generalize better and that curvature controls the volume of small neighborhoods, we conjecture that sectional curvature regularizes learning:

Conjecture 1 (Sectional curvature regularizes). *If $\eta(\tau)$ is a Riemann metric on weight space, smoothly parameterized by τ, and if the sectional curvature through every 2-form at θ_0 increases as τ grows, then the gen. gap attained by fixed-T SGD with learning rate $c\eta(\tau)$ (when initialized from θ_0) decreases as τ grows, for all sufficiently small $c > 0$.*

We are optimistic our formalism may resolve conjectures such as above.

Experimental methods

What artificial landscapes did we use?

We define three artificial landscapes, called Gauss, Helix, and Mean Estimation.

Gauss   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). More explicitly, the Gauss landscape is a probability distribution $\mathcal{D}$ over functions $l_x : \mathbb{R}^1 \to \mathbb{R}$ on 1-dimensional weight space, indexed by standard-normally distributed 1-dimensional datapoints x and defined by the expression:

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

The gradient at sample x and weight σ is then $g_x(h) = (1 - x^2 \exp(-h))/2$. Since $x \sim \mathcal{N}(0, 1)$, the gradient $g_x(h)$ will be affinely related to a chi-squared, and in particular non-Gaussian.

To measure overfitting, we initialize at the true test minimum $h = 0$, then train and see how much the testing loss increases. At $h = 0$, the expected gradient vanishes, and the testing loss of SGD involves only diagrams that have no leaves of size one.

Helix   The Helix landscape has chirality, much like Archimedes' screw. Specifically, the Helix landscape has weights $\theta = (u, v, z) \in \mathbb{R}^3$, data points $x \sim \mathcal{N}(0, 1)$, and loss:

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

Here,

$$H(\theta) = u^2 + v^2 + (\cos(z)u + \sin(z)v)^2$$

is quadratic in u, v, and

$$S(\theta) = \cos(z - \pi/4)u + \sin(z - \pi/4)v$$

is linear in u, v. Also, since $x \sim \mathcal{N}(0, 1)$, the $x \cdot S(\theta)$ term has expectation 0. In fact, the landscape has a three-dimensional continuous screw symmetry consisting of translation along z and simulateous rotation in the $u - v$ plane. Our experiments are initialized at $u = v = z = 0$, which lies within a valley of global minima defined by $u = v = 0$.

The paper body showed that SGD travels in Helix' $+z$ direction. By topologically quotienting the weight space, say by identifying points related by a translation by $\Delta z = 200\pi$, we may turn the line-shaped valley into a circle-shaped valley. Then SGD eternally travels, say, counterclockwise. Alternatively, one may preserve the homotopy type of the underlying weight space by Nash-embedding a flat solid torus

$$[-10^1, +10^1] \times [-10^1, +10^1] \times [-10^3, +10^3]/((x, y, -10^3) \sim (x, y, +10^3))$$

in a higher-dimensional Euclidean space and extending Helix from that torus to the ambient space.

Slightly modifying Helix by adding a linear term $\alpha \cdot z$ to l for $\eta\alpha^2 \ll \eta^2/6$ leads SGD to perpetually ascend.

Mean Estimation   The Mean Estimation family of landscapes has 1 dimensional weights θ and 1-dimensional datapoints x. It is defined by the expression:

$$l_x(\theta) \triangleq \frac{1}{2}H\theta^2 + xS\theta$$

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

For our hyperparameter-selection experiment (Figure 13 □□■) we introduce an l_2 regularization term as follows:

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

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

What image-classification landscapes did we use?

Architectures   In addition to the artificial loss landscapes Gauss, Helix, and Mean Estimation, we tested our predictions on logistic linear regression and simple convolutional networks (2 convolutional weight layers each with kernel 5, stride 2, and 10 channels, followed by two dense weight layers with hidden dimension 10) for the CIFAR-10 (Krizhevsky 2009) and Fashion-MNIST datasets (Xiao, Rasul, and Vollgraf 2017). The convolutional architectures used tanh activations and Gaussian Xavier initialization. To set a standard distance scale on weight space, 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.

Some of our experiments involve Gaussian noise, which is not bounded and so violates the our assumptions. In practice, Gaussians are effectively bounded in 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. Even more experiments permit arbitrarily large losses and thus also violate our boundedness assumptions; since in practice SGD with small learning rates does not explore regions of very-large loss, we consider this violation negligible.

Datasets   For image classification landscapes, we regard the finite amount of available data as the true (sum of diracs) distribution $\mathcal{D}$ from which we sample testing and training 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 testing and training sets to overlap, hence biasing testing loss measurements. In fact, testing and training sets overlap only in reference, not in sense: the situation is analogous to a text prediction task in which two training points culled from different corpora happen to record the same sequence of words, say, "Thank you!". In any case, all of our experiments focus on the limited-data regime, e.g. 10^1 datapoints out of $\sim 10^{4.5}$ dirac masses, so overlaps are rare.

Measurement process

Diagram evaluation on real landscapes We implemented the formulae of § in order to estimate diagram values from real data measured at initialization from batch averages of products of derivatives.

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

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

We ran $5 \cdot 10^4$ trials of the CIFAR-10 convnet on each of 6 Glorot-Xavier initializations we fixed once and for all through these experiments for the optimizers SGD, GD, and GDC, with $T = 10$ and η between 10^{-3} and $2.5 \cdot 10^{-2}$. We did likewise for the linear logistic model on the one initialization of 0.

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

Implementing optimizers

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

Our GDC regularizer was implemented using the unbiased estimator

$$\hat{C} \triangleq (l_x - l_y)_\mu l_{xv}/2$$

For our tests of regularization based on Corollary 6, we exploited the low-dimensional special structure of the artificial landscape in order to avoid diagonalizing to perform the matrix exponentiation: precisely, we used that, even on training landscapes, the covariance of gradients would be degenerate in all but one direction, and so we need only exponentiate a scalar.

Software frameworks and hardware

All code and data-wrangling scripts can be found on github.com/???????/perturb. This link will be made available after the period of double-blind review. Our code uses PyTorch 0.4.0 (Paszke et al. 2019) on Python 3.6.7; there are no other substantive dependencies. The code's randomness is parameterized by random seeds and hence reproducible. We ran experiments on a Lenovo laptop and on our institution's clusters; we consumed about 100 GPU-hours.

Unbiased estimators of landscape statistics

We use the following method — familiar to some but apparently nowhere described in writing — for obtaining unbiased estimates for various statistics of the loss landscape. The method is merely an elaboration of Bessel's factor (Gauss 1823). For completeness, we explain it here.

Given samples from a joint probability space $\prod_{0 \le d < D} X_d$, we seek unbiased estimates of *multipoint correlators* (i.e. products of expectations of products) such as $\langle x_0 x_1 x_2 \rangle \langle x_3 \rangle$. Here, angle brackets denote expectations over the population. 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 $\mathsf{A}_{0 \le s < 2S} \, x_0^{(s)} x_1^{(s)}$, where A denotes averaging over the sample. In fact, the following also works:

$$S \left(\underset{0 \le s < S}{\mathsf{A}} x_0^{(s)} \right)\left(\underset{0 \le s < S}{\mathsf{A}} x_1^{(s)} \right) + (1 - S)\left(\underset{0 \le s < S}{\mathsf{A}} x_0^{(s)} \right)\left(\underset{S \le s < 2S}{\mathsf{A}} x_1^{(s)} \right) \tag{6}$$

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 batch computations on GPUs. We now generalize this estimator to higher-point correlators (and $D \cdot S$ samples).

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

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

and the correlator $\langle x \rangle_\mu$ we define to be the expectation of $\{x\}_\mu$ when $S = 1$. In this notation, 6 says:

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

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

$$\mathbb{E}\left[S^D \{x\}_\mu \right] = \sum_{\tau \le \mu} \left(\prod_{0 \le d < D} \frac{S!}{\left(S - \left| \tau(\mu^{-1}(d)) \right| \right)!} \right) \langle x \rangle_\tau \tag{7}$$

where '$\tau \le \mu$' ranges through partitions *finer* than μ, i.e. maps τ through which μ factors. In smaller steps, 7 holds because

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

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

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

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

Solving 7 for $\langle x \rangle_\mu$, we find:

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

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

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

Additional figures

In the rightmost figure, we add Corollary 6's generalization gap estimate to l. By descending on this regularized loss, we may tune smooth hyperparameters such as l_2 regularization coefficients for small datasets ($H \ll C/N$) (§). Since matrix exponentiation takes time cubic in dimension, this regularizer is most useful for small models.

Review of Tensors

The linear algebra of tensors plays a role in our analysis similar to the role of types in C. One eschew types entirely, but to do so would hinder interpretation of the raw bitstrings we manipulate and would obscure which operations (e.g. dereferencing of a pointer) are natural and which are not (e.g. dereferencing of a float). The language of tensors supplies our analysis with abstractions and operations appropriate to derivatives and moments in high dimensional loss landscapes.

For example, we recognize $C^{-1} + H$ as ill-formed, even though both C^{-1} and H are square grids of numbers of the same shape. This is because C^{-1} and H are tensors of types $\genfrac{}{}{0pt}{}{2}{0}$ and $\genfrac{}{}{0pt}{}{0}{2}$, respectively.

So here's a brief refresher for tensors. We recommend these sources for more details: (Sidiropoulos et al. 2017) §1 for motivational background; (Misner, Thorne, and Wheeler 1973) §2.5 for helpful pictures; (Comon 2014) §2 for examples of how new tensors arise from old; (Kolář, Michor, and Slovák 1993) §7 and §14 for precise formalism; and (McCullagh 1987) §1.4 for statistics-relevant examples.[12]

Vectors versus covectors

Imagine an air-conditioned hallway's temperature gradient (0.1°K/meter) and length (20 meters). When we switch units from m to cm, the temperature gradient numerically *decreases* (to 0.001) but the length numerically *increases* (to 2000). So temperature gradients and spatial displacements are instances of distinct geometric types. A displacement inhabits a vector space V of primary interest; a gradient inhabits V's *dual space* $V^\star$, defined as the set of linear maps from V to $\mathbb{R}$. When V is clear from context (e.g. throughout our paper V consists of tangent vectors on $\mathcal{M}$), we call elements of V *vectors* and elements of $V^\star$ *covectors*.[34]

Imagine a smooth real-valued function $f : V \to \mathbb{R}$. Then we can specify the first derivative $\nabla f(x)$ (say at $x = 0$) by specifying f's p many partials with respect to a basis of V. Imagine a (say, compactly supported) probability measure μ

---

[1]Some of these sources contrast the physicists' and statisticians' uses of tensors and of linear algebra overall. Physicists often use vector spaces whose elements represent changes through physical space and time; statisticians often use vector spaces whose elements represent mixtures of experimental subjects. Our paper doesn't commit to or rely on any domain interpretation of our loss landscape and its associated vector spaces, so to us these distinctions are irrelevant.

[2]Some of these sources refer to *symmetric tensors* and *antisymmetric tensors* (a.k.a. *alternating forms*). Our paper does not use these concepts, so we invite the reader to black-box adjectives such as 'symmetric' when consulting those sources. A symmetric tensor (etc) is just a special case of tensors as we define in this appendix.

[3]This duality is most apparent when we recall that elements in V correspond bijectively with linear maps from $\mathbb{R}$ to V. Then a *vector* (a.k.a. *column* a.k.a. *displacement* a.k.a. *primal* vector) is a linear map of type $\mathbb{R} \to V$ and a *covector* (a.k.a. *row* a.k.a. *gradient* a.k.a. *dual* vector) is a linear map of type $V \to \mathbb{R}$.

[4]Here's why the word *dual* is apt. We may map $V \to (V^\star)^\star$ by evaluation: $v \mapsto (f \mapsto f(v))$. In finite dimensions, each $\omega \in (V^\star)^\star$ arises this way. So we identify $(V^\star)^\star = V$. So a space's double dual is the space itself.

on V. We can specify the first moment $\mathbb{E}_\mu[v]$ of μ by specifying $\mathbb{E}_\mu[v]$'s p many projections with respect to a basis of V.

So both $\nabla f(0)$ and $\mathbb{E}_\mu[v]$ are objects of dimension p. But $\mathbb{E}_\mu[v] \in V$ is a vector while $\nabla f(0) \in V^\star$ is a covector: the two differ in their geometric properties. There is a canonical way to push μ forward along a linear map $\phi : V \to W$ and thereby to send $\mathbb{E}_\mu[v]$ to $\mathbb{E}_\mu[\phi(v)]$. And there is a canonical way to pull f back along a linear map $\psi : U \to V$ and thereby to send $\nabla f(0)$ to $\nabla(f \circ \psi)(0)$. But there is no natural non-zero way to pull $\mu, \mathbb{E}_\mu[v]$ backward or to push $f, \nabla f(0)$ forward (try it!).

What is a tensor?

We've seen that the first derivative ∇f of a smooth function $f : V \to \mathbb{R}$ (evaluated, say, at $x \in V$) inhabits $V^\star$, not V. How about higher derivatives? To answer this question, we introduce **tensor**s: multi-axis grids of numbers viewed as linear algebraic objects.

Take the third derivative $\nabla\nabla\nabla f$ evaluated at $x \in V$. We can describe $\nabla\nabla\nabla f(x)$ completely by specifying the p^3 many partial derivatives with respect to a basis $\mathcal{E} = (e^i : 0 \le i < p)$ of V:

$$\left[D^{\mathcal{E}}_{i,j,k} f\right](x) \triangleq \frac{\partial}{\partial t}\frac{\partial}{\partial t'}\frac{\partial}{\partial t''} f(x + t \cdot e^i + t' \cdot e^j + t'' \cdot e^k)\Big|_{t=t'=t''=0}$$

The p^3 many components $\left[D^{\mathcal{E}}_{i,j,k} f\right](x)$ enjoy key algebraic properties such as symmetry:

$$\left[D^{\mathcal{E}}_{i,j,k} f\right](x) = \left[D^{\mathcal{E}}_{j,i,k} f\right](x) = \left[D^{\mathcal{E}}_{i,k,j} f\right](x)$$

and linearity (which we see from the chain rule for derivatives):

$$\left[D^{\mathcal{F}}_{\hat{i},\hat{j},\hat{k}} f\right](x) = \sum_{\hat{i}}\sum_{\hat{j}}\sum_{\hat{k}} C^i_{\hat{i}} C^j_{\hat{j}} C^k_{\hat{k}} \left[D^{\mathcal{E}}_{i,j,k} f\right](x)$$

Here, $\mathcal{F}$ is a basis expressible in terms of $\mathcal{E}$ via $f^i = \sum_{\hat{i}} C^i_{\hat{i}} e^{\hat{i}}$.

By showing how $\nabla\nabla\nabla f$'s components change coherently as we change basis, the linearity rule helps us interpret $\nabla\nabla\nabla f$ as a geometric object with magnitude and directional information rather than as a mere grid of numbers.

We capture the essence of such linearity rules of by defining what it means for a vector space P to be a tensor product of given vector spaces V, W; a 'tensor' is then just an element of P. The DEFINING QUALITY of P is that **a linear map from P to X is 'as good as' a bilinear map from V, W to X, for any vector space X.**[1] We write $V \otimes W$ for the two-axis tensor P. More generally, we write $U \otimes V \otimes \cdots \otimes W$ for

---

[1]Precisely, a tensor product of V, W is a vector space P equipped with a natural family of vector-space isomorphisms

$$\iota_X : \mathrm{Bilin}(V, W; X) \xrightarrow{\sim} \mathrm{Lin}(P; X)$$

from the vector space of bilinear maps to the vector space of linear maps. The word 'natural' just means that for any linear map $\phi : X \to Y$ and any $b \in \mathrm{Bilin}(V, W; X)$:

$$\phi \circ \iota_X(b) = \iota_Y(\phi \circ b)$$

Naturality captures the intuition that ι just 're-packages' data without 'altering' its content. For instance, if $\kappa : X \to X$ is a non-trivial

---

an m-axis tensor P, the maps from which correspond to the m-linear maps from $U, V, \cdots, W$.

We may now interpret the third derivative $\nabla\nabla\nabla f(x)$ as a linear-algebraic object:

$$\nabla\nabla\nabla f(x) \in (V \otimes V \otimes V)^\star$$

The defining quality of $V \otimes V \otimes V = P$ says that such an element is specified the moment we specify a trilinear map from V, V, V to $\mathbb{R}$. This trilinear map is simply

$$(u, v, w) \mapsto \frac{\partial}{\partial t}\frac{\partial}{\partial t'}\frac{\partial}{\partial t''} f(x + t \cdot u + t' \cdot v + t'' \cdot w)$$

So $\nabla\nabla\nabla f(x)$ is a tensor. We now explain what we mean when we say that it is a tensor of type $^0_3$.

Tensors of type ^u_d

We've seen two ways to make new finite-dimensional vector spaces from old: we can dualize (to get $V^\star$ from V) and we can tensor (to get $U \otimes V \otimes \cdots \otimes W$ from $U, V, \cdots W$).

No matter how we compose these operations, we'll get a space of the standard form

$$(U \otimes V \otimes \cdots W) \otimes (X^\star \otimes Y^\star \otimes \cdots Z^\star)$$

In other words, we have a tensor product of **u** many non-dualized vector spaces and **d** many dualized vector spaces. We won't provide a proof.[2]

An example is that $((U^\star \otimes V)^\star \otimes W^\star)^\star$, though not of the above form, is naturally isomorphic to $U^\star \otimes V \otimes W$, which *is* of the above form. That example illustrates a general algorithm: the factors in an un-standardized expression correspond with the factors in a standardized expression; whether a standardized factor is dualized or not depends simply on the PARITY of the number of dual operations that act on the corresponding factor in the original expression. Here, U, V, W appear with $3, 2, 2$ layers of duals, respectively, whence by parity we obtain the $1, 0, 0$ layers of duals in the standardization.

In the common case where all the underlying spaces $U, \cdots, Z$ agree with a vector space V understood from context, we say that elements of the combined space are **tensors of type ^u_d**. We represent such tensors as grids with $u+d$ many axes, where each axis has a length equal to the dimension of V. Thus, elements of $((V^\star \otimes V)^\star \otimes V^\star)^\star$ are tensors of type $^2_1$. Numerically, they have $2 + 1 = 3$ axes: 2 that transform like vectors and 1 that transforms like a covector. The derivative

---

isomorphism, then defining $\iota'_X(b) = \kappa \circ \iota_X(b)$ and $\iota'_Z = \iota_Z$ for all $Z \ne X$ would turn a natural family ι of isomorphisms into a non-natural family ι' of isomorphisms.

[2]Here's an special case that contains all the ideas of a general proof. We'll define a natural isomorphism $\phi : U^\star \otimes V^\star \otimes W^\star \to (U \otimes V \otimes W)^\star$. By the defining quality of the domain, to specify ϕ is to specify a trilinear map from $U^\star, V^\star, W^\star$ into the codomain. By the defining quality of the codomain, to specify a map into the codomain is to specify a map into the space of trilinear real-valued maps on U, W, V. We accomplish all this by sending f, g, h to $((u, v, w) \mapsto f(u) \cdot g(v) \cdot h(w))$. We've thus defined a linear map ϕ. It is routine to check that ϕ sends no element except zero to zero, that ϕ's domain and codomain have equal and finite dimension, and thus that ϕ is an isomorphism.

of θ_T's covariance matrix with respect to the initialization θ_0 is an instance of such a type-$^2_1$ tensor.

As special cases, scalars are tensors of type $^0_0$, vectors are tensors of type $^1_0$, covectors are tensors of type $^0_1$, and linear maps from V to V are tensors of type $^1_1$.[1]

Whenever we have a tensor of type ^a_a, we may interpret it as a linear map from the a-fold tensor product $V \otimes V \otimes \cdots \otimes V$ to itself. In our work, this linear map will sometimes be invertible, in which case we use standard notation $(\cdot)^{-1}$ to denote the inverse. As a variation on this theme, a tensor (such as C) of type $^0_2$ we may view as a linear map from V to $V^\star$. This linear map is invertible whenever C is positive definite, in which case we use the notation C^{-1} to denote its inverse. Observe that this inverse, being a linear map from $V^\star$ to V, is a tensor of type $^2_0$, unlike C.

Contraction

Let's define a map $\epsilon_V : V \otimes V^\star \to \mathbb{R}$ by evaluation: $(v, f) \mapsto f(v)$. We may similarly define

$$\mathrm{id}_{(T \otimes U)} \otimes \epsilon_V : (T \otimes U) \otimes V \otimes V^\star \to (T \otimes U)$$

by $(x, v, f) \mapsto f(v) \cdot x$ for $x \in T \otimes U$. In general, we can map any tensor product of the form

$$(U \otimes V \otimes \cdots W) \otimes (U^\star \otimes Y^\star \otimes \cdots Z^\star)$$

to one of the form

$$(V \otimes \cdots W) \otimes (Y^\star \otimes \cdots Z^\star)$$

by applying the evaluation map ϵ_U.

Say we have a tensor of type $^{u+1}_{d+1}$. IF we specify a specific axis among the $u + 1$ vector-type axes as well as a specific axis among the $d + 1$ covector-type axes THEN we may apply the above linear map to get a tensor of type ^u_d. We call this operation **contraction**. Operationally, contraction just means summing across the two specified axes over the $\dim(V)$ many 'diagonal' index pairs (i, i).

For example, a linear map ϕ is the same as a tensor of type $^1_1$. We may contract it to get a tensor of type $^0_0$, that is, a real number. The number we get is the trace of ϕ.

As another example, consider the learning rate η (a tensor of type $^2_0$) and the hessian $H = \nabla G$ (a tensor of type $^0_2$). Then $\eta \otimes H$ is a tensor of type $^2_2$ and we may contract (in 2×2 many ways, here all equivalent due to η, H's symmetry) to obtain tensor of type $^1_1$: a linear map.

What is this linear map, concretely? It acts on a vector v by sending it to the vector w such that a gradient descent step $\theta \mapsto \theta - \eta \nabla f$ on the non-standard loss function

$$f(x) = G(x) \cdot v \in \mathbb{R}$$

displaces θ by $-w$. We see that the language of tensors concisely expresses otherwise cumbersome conceptual gymnastics. In our experience, this language also guides us through otherwise treacherously brittle algebraic manipulations.

---

[1]Here's an example that our paper doesn't use but that may aid geometric intuition. If V has dimension p, then to specify a linear-algebraic notion of volume in V is to specify a tensor of type ^0_p obeying certain antisymmetry properties. Intuitively, such a tensor produces for any given sequence of p input vectors a real number. We interpret this real number as the volume of a parallelepiped spanned by those p vectors.
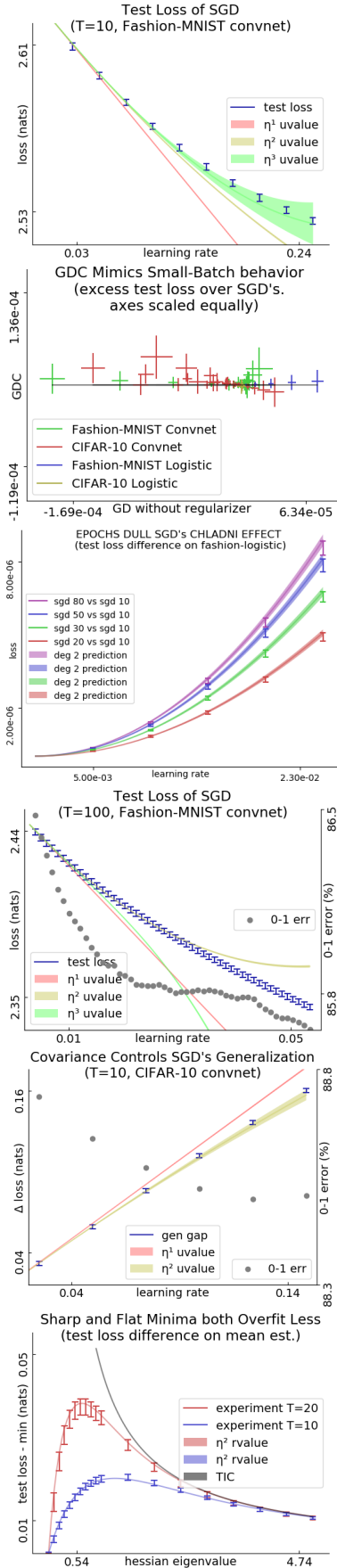

Figure 12: **Experiments on natural and artificial landscapes.** `rvalue` refers to Theorem 1's predictions, approximated as in Remark 1. `uvalues` are simpler but (see ▦) less accurate.

Left: Perturbation models SGD for small ηT. Fashion-MNIST convnet's testing loss vs learning rate. In this
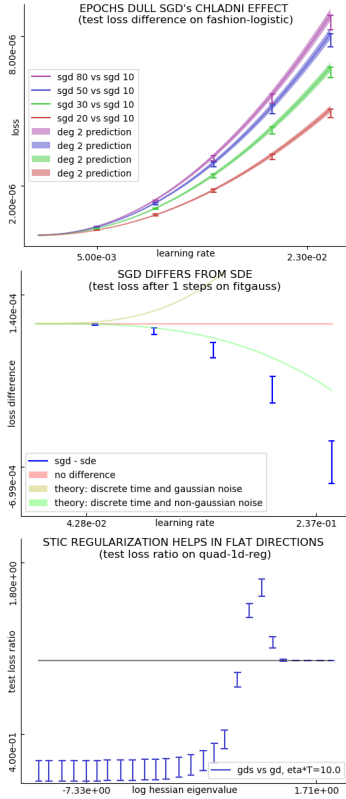
Figure 13: **Further experimental results**. **Left**: SGD with
$2, 3, 5, 8$ epochs incurs greater test loss than one-epoch SGD
(difference shown in I bars) by the predicted amounts (pre-
dictions shaded) for a range of learning rates. Here, all SGD
runs have $N = 10$; we scale the learning rate for E-epoch
SGD by $1/E$ to isolate the effect of inter-epoch correla-
tions away from the effect of larger ηT. **Center**: SGD's dif-
ference from SDE after $\eta T \approx 10^{-1}$ with maximal coarse-
ness on GAUSS. Two effects not modeled by SDE — time-
discretization and non-Gaussian noise oppose on this land-
scape but do not completely cancel. Our theory approxi-
mates the above curve with a correct sign and order of mag-
nitude; we expect that the fourth order corrections would
improve it further. **Right**: Blue intervals show regularization
using Corollary 6. When the blue intervals fall below the
black bar, this proposed method outperforms plain GD. For
MEAN ESTIMATION with fixed C and a range of Hs, initialized
a fixed distance *away* from the true minimum, descent on
an l_2 penalty coefficient λ improves on plain GD for most
Hessians. The new method does not always outperform GD,
because λ is not perfectly tuned according to STIC but in-
stead descended on for finite ηT.