

A Perturbative Analysis of Stochastic Descent

RQE Slides

Sam Tenka

July 29, 2020

People

Sam Tenka — favorite animal is COW. 2nd year grad student. Now working in program induction, but this project is about gradient descent theory.

I would like to thank SHO YAIDA, DAN A. ROBERTS, and JOSH TENENBAUM for their patient guidance. It was DAN A. ROBERTS who recognized that interesting questions lie in the analysis of epoch number and batch size; it was he who introduced me to much prior work. Only with SHO YAIDA's advice to re-sum did the theory attain its most precise and conceptual form. I appreciate the time and energy that ANDY BANBURSKI, BEN R. BRAY, JEFF LAGARIAS, and WENLI ZHAO spent to critique my drafts. In particular, WENLI ZHAO nudged me to consider and discuss connections to physics, and BEN R. BRAY taught me that gradient noise is rarely isotropic, homogeneous, or Gaussian.

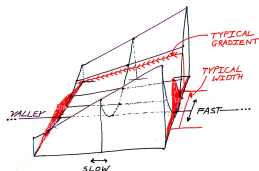
Overview

Disciplines such as particle physics and crystallography make aggressive use of specialized diagrams to solve problems. Diagrams depict the essential structure of a problem, streamlining computation and inspiring intuition. We will use diagrams to answer questions about stochastic gradient descent (SGD) such as:

What effect does non-gaussian noise have on eventual test loss?

Does SGD overfit more in the presence of flat or sharp minima?

How does SGD select from among a valley of minima?



Our main result expresses the expected test loss after T steps of SGD as a sum over diagrams, each interpretable as an interaction of weights and data. For example, gradient noise may push θ_t up the valley's walls ("fast"); then θ_{t+1} will slide toward the valley's flatter regions ("slow").


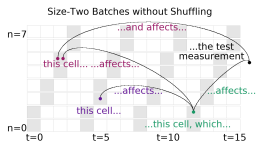
We quantify this using the diagram .

Diagram-based computation

Theorem (Informal)

SGD's expected test loss is a sum over weight-data interactions drawable as diagrams. Summing the smallest diagrams suffices for small ηT .



A single diagram summarizes the effects of many related weight-data interactions (see Left).

Each diagram embeds in many ways into the grid of (n, t) pairs s.t. the t th update involves the n th training point. Above: an $N = 8$, $T = 16$ grid and the diagram

Intuitively, fuzzy outlines depict correlations (noise; how ∇l depends on x); black edges depict differentiations (curvature; how ∇l depends on θ). Thus, a diagram shows how gradient information and noise flows forward in time toward the test measurement. (Due to time ordering, we insist that diagrams are rooted trees, drawn with root rightmost).

Diagram-based computation


We evaluate diagrams as follows: For each **node**, write an l_x . For each **black edge** between two nodes, differentiate the two nodes and contract them using η . Group the resulting factors within expectation brackets according to **fuzzy outlines**.

$$\text{Diagram 1} \rightarrow \mathbb{E}[\nabla l_x] \eta \mathbb{E}[\nabla \nabla l_x] \eta \mathbb{E}[\nabla l_x]$$

$$\text{Diagram 2} \rightarrow \mathbb{E}[\nabla l_x \eta \nabla \nabla l_x] \eta \mathbb{E}[\nabla l_x]$$



Example (*Does skewed noise affect test loss?*)

To leading order, the test loss due to skewed noise is , which for large T and isotropic H evaluates to $-\frac{\eta^3}{3!} \frac{S_{\mu\nu\lambda} J_{\mu\nu\lambda}}{3\|\eta H\|_2}$. Here, we used the skewness and jerk

$$S = \mathbb{E}(\nabla l_x(\theta_0) - G)^3 = \text{Diagram 3} \quad J = \mathbb{E}(\nabla \nabla \nabla l_x(\theta_0)) = \text{Diagram 4}$$

and G, H are the expected gradient and hessian.

Problem setup

Fix a data distribution \mathcal{D} , a manifold \mathcal{H} of weights, and a loss landscape $l : |\mathcal{D}| \rightarrow \mathcal{H} \rightarrow \mathbb{R}$, considered as a random function. For an initialization $\theta_0 \in \mathcal{H}$ and a sequence $x_t \sim \mathcal{D} : 0 \leq t < T$, we consider the iteration

$$\theta_{t+1} = \theta_t - \eta \nabla l_{x_t}(\theta_t)$$

We use such **stochastic gradient descent** in learning, as an approximate optimizer. Compare to $T \rightarrow \infty$ limits: fixing ηT , we recover **ODE**; fixing $\eta\sqrt{T}$, we recover **SDE**. More generally, if $\mathcal{S} = (x_n : 0 \leq n < N) \sim \mathcal{D}^N$ and we update with $l_{\mathcal{B}_t} = \sum_{n \in \mathcal{B}_t} l_{x_n} / |\mathcal{B}_t|$, then **GD** is SGD with $\mathcal{B} = \mathcal{S}$.

Question

How does SGD's dynamics on a curved and noisy landscape affect optimization and generalization? How does SGD differ from GD, SDE?

We wish to express $\mathbb{E}_{\mathcal{S}} l_x$ and $\mathbb{E}_{\mathcal{D}} l_x - \mathbb{E}_{\mathcal{S}} l_x$ (at θ_T) in terms of l 's statistics.

Theory

The Theorem below expresses SGD's test loss as a sum over diagrams. A diagram with d edges scales as $O(\eta^d)$, so the following is a series in η . We later truncate the series to small d , focusing on few-edged diagrams.

Theorem (Re-summation)

For $|\mathcal{B}| = 1$ and any T : for η small enough, SGD has expected test loss

$$\sum_{\substack{D \text{ an irreducible diagram}}} \sum_{\substack{f \text{ an embedding of } D}} \frac{(-1)^{|\text{edges}(D)|}}{|Aut_f(D)|} rvalue_f(D)$$

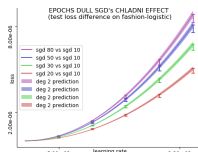
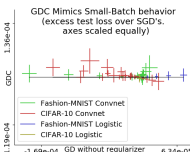
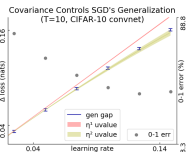
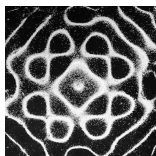
We may approximate sums by integrals and $(I - \eta H)^t$ by $\exp(-\eta H t)$, reducing to a routine integration of exponentials with error factor $1 + o(\eta)$.

Theorem (Convergence)

If θ_* is a non-degenerate local minimum of l (i.e. $G(\theta_*) = 0$ and $H(\theta_*) > 0$), then for SGD initialized sufficiently close to θ_* , the d th-order truncation of Theorem 2 converges as $T \rightarrow \infty$.

The above $T \rightarrow \infty$ limit might not measure any well-defined limit of SGD, since the limit might not commute with the infinite sum. We see no such pathologies in practice, so we freely speak of “SGD in the large- T limit.”

High- C regions repel SGD



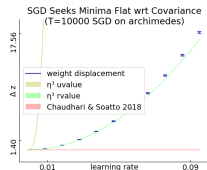
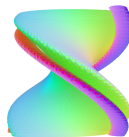
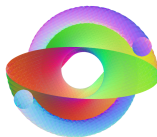
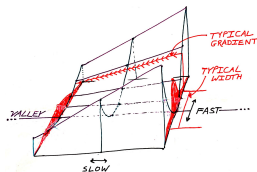
If two neighboring regions of weight space have high and low levels of gradient noise, respectively, then we expect the rate at which θ jumps from the former to the latter to exceed the opposite rate. Hence, net movement toward regions of small C ! More precisely, the drift is in the direction of $-\nabla C$. The effect is strongest when gradient noise is not averaged out by large batch sizes, but we may counter this effect via an artificial loss term:¹

Corollary ()

SGD avoids high- C regions more than GD:

$I_C \triangleq \frac{N-1}{4N} \nabla^\mu C_\nu^\nu = \mathbb{E}(\theta_{GD} - \theta_{SGD})^\mu - o(\eta^2)$. If \hat{l}_c is a smooth unbiased estimator of I_C , then GD on $I + \hat{l}_c$ has an expected test loss that agrees with SGD's to order η^2 .

SGD prefers minima flat with respect to C



Gradient noise may push θ_t up the valley's walls ("fast"); then θ_{t+1} will slide toward the valley's flatter regions ("slow").

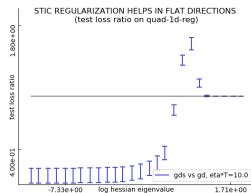
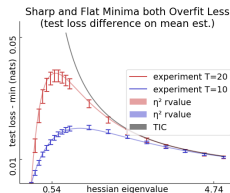
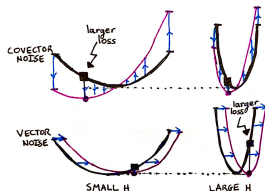
Corollary (Computed from )

Run SGD for $T \gg 1/\eta H$ from a non-degenerate test minimum. Written in an eigenbasis of ηH , θ has an expected displacement of

$$-\frac{\eta^3}{2} \sum_{\mu\nu} C_{\mu\nu} \frac{1}{\eta(H_{\mu\mu} + H_{\nu\nu})} J_{\mu\nu\lambda} \frac{1}{H_{\lambda\lambda}} + o(\eta^2)$$

$CH/2 + o(C)$ is the increase upon convolving I with a C-shaped Gaussian: SGD descends on a C -smoothed landscape. This landscape changes as C does, so SGD's velocity field generically has *curl*. Indeed, while $\nabla(CH)$ is a total derivative, $C\nabla H$ is not. Contrast to Wei and Schwab [2019]'s assumptions constant C analysis.

Both flat and sharp minima overfit less






It is obvious that SGD overfits less near **sharp** minima. Prior work empirically supports this claim; after all, l_2 regularization acts by increasing the Hessian.

It is also obvious that SGD overfits less near **flat** minima. Prior work empirically supports this claim; after all, flat minima are stable as the weight changes.

More balanced view: Sharp minima are robust to slight changes in the average *gradient* and flat minima are robust to slight *displacements* in weight space. As we take $T \rightarrow \infty$: overfitting $\rightarrow \text{TIC} = C/2NH$. Prior works such as Dixon and Ward [2018] use TIC to estimate overfitting but require arbitrary cutoffs for singular H . Counter to intuition! Our theory shows that the implicit regularization of finite- T descent tames these singularities: the overfitting formula tends to 0 as H shrinks.

Contributions

We presented a diagram-based method for studying stochastic optimization on short timescales or near minima. The corollaries offer insight into SGD's success in training deep networks: SGD avoids curvature and noise, and curvature and noise control generalization.

Analyzing , we proved that **flat and sharp minima both overfit less** than medium minima. Intuitively, flat minima are robust to vector noise, sharp minima are robust to covector noise, and medium minima robust to neither. We thus proposed a regularizer enabling gradient-based hyperparameter tuning. Inspecting , we extended Wei and Schwab [2019] to nonconstant, nonisotropic covariance to reveal that **SGD descends on a landscape smoothed by the current covariance C** . As C evolves, the smoothed landscape evolves, resulting in non-conservative dynamics. Examining , we showed that **GD may emulate SGD**, as conjectured by Roberts [2018]. This is significant because, while small batch sizes can lead to better generalization, Bottou [1991] modern infrastructure increasingly rewards large batch sizes. Goyal et al. [2018]

Related work; limitations

SGD-trained networks generalize despite their ability to shatter large sets [Zhang et al., 2017], so generalization must arise from the aptness-to-data of not only architecture but also **optimization** [Neyshabur et al., 2017].

Approaches via **stochastic differential equations** assume uncorrelated, Gaussian noise in continuous time [Chaudhari and Soatto, 2018, Li et al., 2017]; per Yaida [2019], they cannot treat SGD noise correctly. Prior **perturbative approaches** were limited to specific neural architectures [Dyer and Gur-Ari, 2019] or to computing Gaussian statistics over $T = 2$ [Roberts, 2018]. We do not assume **information-geometric** relations between C and H , so we may model VAEs.

Our predictions depend only on loss data near θ_0 , so they only apply for long times (large ηT) near an isolated minimum or for short times (small ηT) in general. Meteorologists understand how warm and cold fronts interact **despite long-term intractability** [§C.1]; we quantify curvature's and noise's counter-intuitive effects in each short-term interval of SGD.

Future directions: Lagrangians and curved backgrounds

Our diagrams invite exploration of Lagrangian formalisms and curved backgrounds:

Question

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

Some *higher-order* methods — such as the H -based update $\theta \leftarrow \theta - (\eta^{-1} + \lambda \nabla \nabla l_t(\theta))^{-1} \nabla l_t(\theta)$ parameterized by small η, λ — admit diagrammatic analysis when we represent the λ term as a second type of node.

Our work assumes a flat metric $\eta^{\mu\nu}$, but it might generalize to curved spaces.² Prior work focuses on *optimization* on curved weight spaces, in machine learning we also wish to analyze *generalization*.

Conjecture (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 generalization 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 $c > 0$.

Bibliography

- L. Bottou. Stochastic gradient learning in neural networks. *Neuro-Nîmes*, 1991.
- P. Chaudhari and S. Soatto. Sgd performs variational inference, converges to limit cycles for deep networks. *ICLR*, 2018.
- M.F. Dixon and T. Ward. Takeuchi information as a form of regularization. *Arxiv Preprint*, 2018.
- E. Dyer and G. Gur-Ari. Asymptotics of wide networks from feynman diagrams. *ICML Workshop*, 2019.
- P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd. *Data @ Scale*, 2018.
- Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms i. *PMLR*, 2017.
- B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *Chapter 4 from Intel CRI-CI: Why and When Deep Learning Works Compendium*, 2017.
- D.A. Roberts. Sgd implicitly regularizes generalization error. *NeurIPS: Integration of Deep Learning Theories Workshop*, 2018.
- Mingwei Wei and D.J. Schwab. How noise affects the hessian spectrum in overparameterized neural networks. *Arxiv Preprint*, 2019.