# A Perturbative Analysis of Stochastic Descent

## RQE Slides

Sam Tenka

July 26, 2020

# Problem setup

Fix a data distribution $\mathcal{D}$, a manifold $\mathcal{H}$ of weights, and a loss landscape $l : |\mathcal{D}| \to \mathcal{H} \to \mathbb{R}$, considered as a random function. For an initialization $\theta_0 \in \mathcal{H}$ and a sequence $\mathcal{S} = x_t \sim \mathcal{D} : 0 \leq t < T$, 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 $T \to \infty$ limits: with fixed $\eta T$, recover **ODE**; with fixed $\eta\sqrt{T}$, recover **SDE**.

### 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 $\theta_T$) in terms of $l$'s statistics.

# *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 $\eta T$.*

**Example** (*How does skewed noise affect SGD's test loss?*)

The relevant diagram is , which for large $T$ and isotropic hessian evaluates to $-\frac{\eta^3}{3!}\frac{S_{\mu\nu\lambda}J_{\mu\nu\lambda}}{3\|\eta H\|_2}$. This is the leading order test loss due to skewed noise! Here, we used the jerk $J = \mathbb{E}(\nabla\nabla\nabla l_x(\theta_0)) =$  and the skewness $S = \mathbb{E}(\nabla l_x(\theta_0) - G)^3 =$  at initialization. $G, H$ are the expected gradient and hessian.

# Related work; limitations

Approaches via **stochastic differential equations** assume uncorrelated, Gaussian noise in continuous time. Prior **perturbative approaches** were limited to specific neural architectures or to computing Gaussian statistics over $T = 2$. We do not assume **information-geometric** relationships between $C$ and $H$, so we may model VAEs.

Our predictions depend only on loss data near $\theta_0$, so they only apply for long times (large $\eta T$) near an isolated minimum or for short times (small $\eta T$) in general. Meteorologists understand how warm and cold fronts interact despite long-term intractability; we quantify curvature's and noise's counter-intuitive effects in each short-term interval of SGD.

# *Main result*

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# SGD prefers minima flat with respect to C

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# *Both flat and sharp minima overfit less*

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# *High-C regions repel SGD*

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# Non-gaussian noise affects SGD but not SDE

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# Contributions

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# Future direction: Lagrangians

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# Future direction: Curved backgrounds

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

## *Bird's eye view*

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# References