Noise in Stochastic Gradient Descent

Kenny Chen

December 15, 2023

1 Introduction

Stochastic gradient descent (SGD) and its variants (Adam, Momentum, Adagrad) are widely used in machine learning, especially in non-convex optimization tasks such as training neural networks. Stochastic gradient descent itself is a modification of the gradient descent algorithm. Recall that the gradient descent (GD) algorithm starts at point $x_0 \in \mathbb{R}^d$ and we iteratively update as:

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

where η_t is the step-size (also know as the learning rate).

From class, we've seen various extensions and alterations of this algorithm such as the accelerated gradient descent. Stochastic gradient descent is one alternative approach to this algorithm that is computed using a mini-batch of the data set. The mini-batch is a fixed number of training examples that is less than the actual data set. Another view [6] of it is as Gradient Descent with an unbiased noise inserted at every iteration, called the gradient noise. Kleinberg et al. [4] in their paper defines it as:

$$x_{t+1} = x_t - \eta_t v_t$$

where v_t is the stochastic gradient that satisfies $E(v_t) = \nabla f(x_t)$.

It has become quite popular due to having a nice trade off between accuracy and efficiency; it requires more iterations to converge but fewer gradient evaluations per iteration compared to the Gradient Descent. Motivated by the work done in Kleinberg et al. [4], the goal of this paper is to examine how noise in Stochastic Gradient Descent can lead to better convergence results, especially in a non-convex optimization context. It is my goal to briefly dive into the exploration of the role of noise and its impact in various settings. Noise has been shown to help escape saddle points, provide better generalizations, and guarantee polynomial hitting time of good local minima under some assumptions. Due to time and page constraints, this literature review will briefly examine [4] and connect those ideas with a few contemporary and emerging ideas in the analysis of noise in Stochastic Gradient Descent to better understand how noise can help in non-convex optimization settings like neural networks.

2 Escaping Local Minima

Noise seems to play a crucial role in non-convex optimization, but back in 2018, it was unclear why Stochastic Gradient Descent could converge to better local minima than Gradient Descent in non-convex optimization problems. To tackle this issue, Kleinberg et al. [4] demonstrated in their paper that Stochastic Gradient Descent was able to escape local minima under certain properties and empirically showed those properties are common in modern neural networks. As a brief summary, they take the alternative view of Stochastic Gradient Descent that it is working on a convolved version of the loss function. And when the convolved function is one point convex [1] with respect to the final solution x^*, Stochastic Gradient Descent could escape all other local minima and stay around x^* with constant probability.

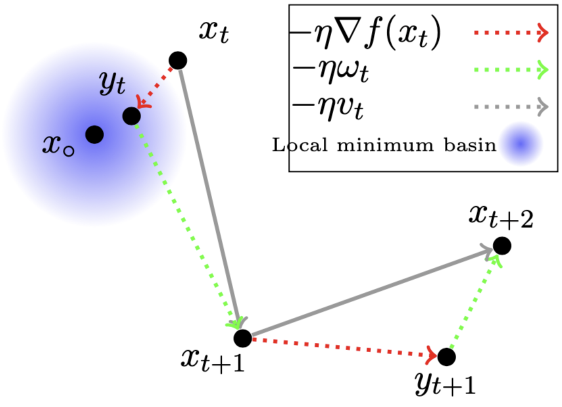

Figure 1: SGD path $x_t \to x_{t+1}$ can be decomposed into $x_t \to y_t \to x_{t+1}$. If the local minimum basin has small diameter, the gradient at x_{t+1} will point away from the basin.

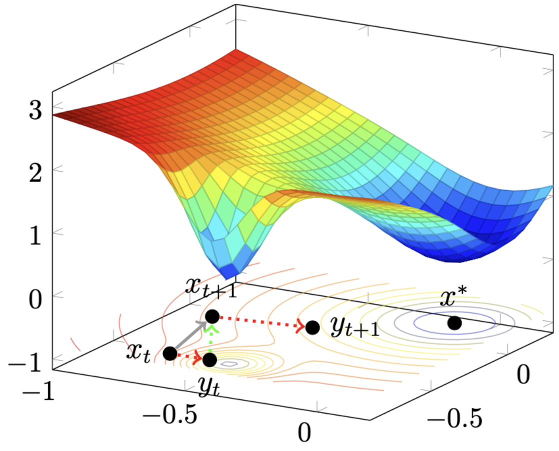

Figure 2: 3D version of Figure 1: SGD could escape a local minimum within one step.

Consider the following example from their paper based on Figure 1 for some x_t, that instead of pointing to the solution x^* (not shown), its negative gradient points to a bad local

[1] Kleinberg et al. define one-point convexity as "If f is δ-one point strongly convex around x^* in a convex domain D, then x^* is the only local minimum point in D."

minimum x_o giving us $y_t = x_t - \eta \nabla f(x_t)$. But since we are performing Stochastic Gradient Descent, the actual direction we take is $-\eta v_t = -\eta(\nabla f(x_t) + \omega_t)$ where ω_t is the noise with expectation 0 and follows some distribution $W(x_t)$. If we take a large enough step size, η, we might get out of the basin region with the help of noise (from y_t to x_{t+1} where getting out of the basin region means x_{t+1} does not point to x_o. Kleinberg et al. [4] continue by considering the sequence $y_t \rightarrow y_{t+1}$ where y_t is defined as above. Note that the Stochastic Gradient Descent algorithm never computes these vectors y_t but they are being used as an analysis tool. And from the equation $x_{t+1} = y_t - \eta\omega_t$, we get the following update rule:

$$y_{t+1} = y_t - \eta\omega_t - \eta\nabla f(y_t - \eta\omega_t) \tag{1}$$

Since the random vector $\eta\omega_t$ has expectation 0, if we take the expectation of both sides, we get that $E_{\omega_t}[y_{t+1}] = y_t - \eta\nabla E_{\omega_t}[f(y_t - \eta\omega_t)]$. They then define $g_t(y) = E_{\omega_t}[f(y_t - \eta\omega_t)]$ which is the original function f convolved with the η-scaled gradient noise, then one can view the sequence y_t as approximately doing gradient descent on the sequence of functions, g_t.

Using this alternative view can help us better understand why Stochastic Gradient Descent converges to a good local minimum even when f has many other sharp local minima [2]. They argue that intuitively, the sharp local minima are eliminated by the convolution operator that transforms f to g_t, since convolutions have the effect of smoothing out short-range fluctuations. Instead of imposing convexity or one-point convexity requirements on f itself, they only require those properties to hold for the smoothed functions obtained from the convolved f. They formalize this argument under the following assumption:

Assumption 1. *For a fixed point x^*, noise distribution $W(x)$, step size η, the function f is c-one point strongly convex with respect to x^* after convolved with noise. That is, for any x, y in domain D such that $y = x - \eta\nabla f(x)$,*

$$\langle -\nabla E_{\omega \in W(x)} f(y - \eta\omega), x^* - y\rangle \geq c\|x^* - y\|_2^2$$

They continue by arguing that for a point y, since the direction $x^* - y$ points to x^*, by having a positive inner product with $x^* - y$, we known the direction $-\eta\nabla f(y_t - \eta\omega_t)$ in (1) approximately points to x^* in expectation. And so, with decent probability, y_t will converge to x^*:

[2] As someone who's extremely rudimentary in my machine learning knowledge and terminology. I like this definition for sharp and flat local minima that works for our purposes from Dinh et al. [1], "While the concept of flat minima is not well defined, having slightly different meanings in different works, the intuition is relatively simple. If one imagines the error as a one dimensional curve, a minimum is flat if there is a wide region around it with roughly the same error, otherwise the minimum is sharp."

Theorem 1. *Assume f is smooth, for every $x \in D$, $W(x)$ s.t.* $\max_{\omega \sim W(x)} \|\omega\|_2 \leq r$. *Also assume η is bounded by a constant, and assume Assumption 1 with x^*, η, c. For $T_1 \geq \tilde{O}(\frac{1}{\eta^c})$*[3]*, and any $T_2 > 0$, with probability at least* $1/2$ *we have* $\|x^* - y\|_2^2 \leq O(\log(T_2)\frac{\eta r^2}{c}$ *for any t s.t. $T_1 + T_2 \geq t \geq T_1$.*

This theorem says that the Stochastic Gradient Descent will get close to x^*, but also stays with constant probability that the Stochastic Gradient descent will stay close to x^* for future T_2 steps. Since sharp local minima have smaller loss values and smaller diameters than flat minima, after convolved with the noise kernel, they disappear which means Assumption 1 holds. On the other hand, flat local minima have larger diameters, so they still exist after the convolution. In that instance, the theorem that states the Stochastic Gradient Descent algorithm will more likely converge to the flat local minima and this is important because it is hypothesized that flat local minima may lead to better generalizations [3]. However, this has been a contentious topic as Dinh et al. [1] argues based on the idea that the definition behind flatness is more nuanced that we had initially described. As well as Zhou et al. [8] who claims "local optima do not necessarily guarantee generalization".

They also provide another simple example that better illustrates this alternative view. Consider the following Figure 3 taken from their paper. The function f in the first row and first column is approximately convex but spiky. So Gradient Descent gets stuck at various local minima (row 2, col 1). By taking the alternative view that Stochastic Gradient Descent works on the convolved f, (plots in row 1, column 2,3, and 4), we see that these functions are smoother and have less local minima.

[3] In their paper, they use $\tilde{O}$ to hide the log terms here so I'm not quite sure what the complexity is.

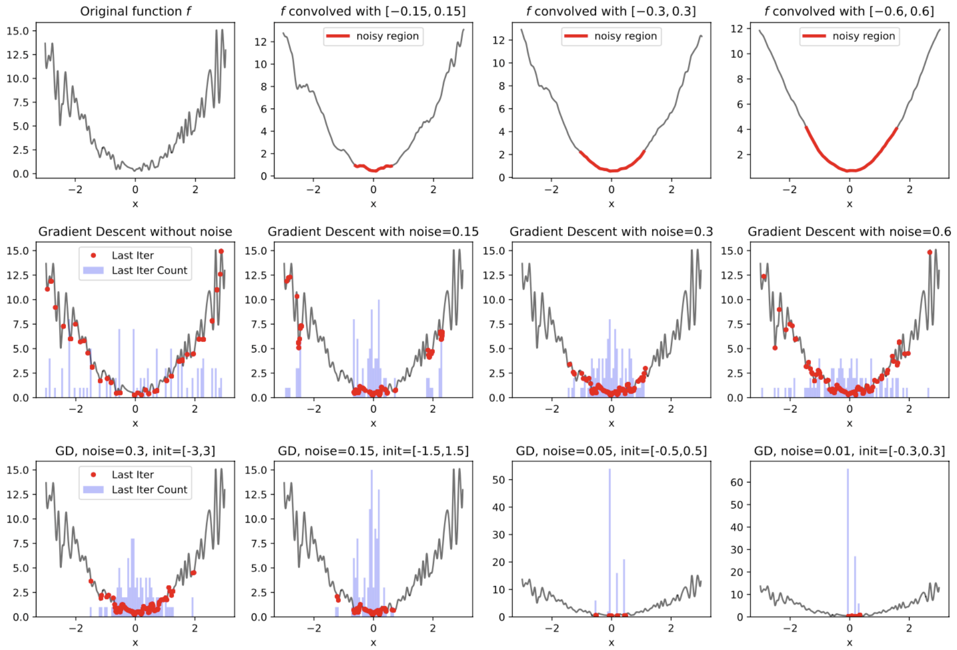

Figure 3: Running SGD on a spiky function f. **Row 1:** f gets smoother after convolving with uniform random noise. **Row 2:** Run SGD with different noise levels. Every figure is obtained with 100 trials with different random initializations. Red dots represent the last iterates of these trials, while blue bars represent the cumulative counts. GD without noise easily gets stuck at various local minima, while SGD with appropriate noise level converges to a local region. **Row 3:** In order to get closer to x^*, one may run SGD in multiple stages with shrinking learning rates.

They lastly provide some empirical observations and suggest that modern neural networks have many nice one point convex properties needed for their theorem to explain why Stochastic Gradient Descent works well in practice.

3 Other works and remarks

While Kleinberg et al. [4] showed Stochastic Gradient Descent works well under certain assumptions, Zhou et al. [8] argue that this is an ”unconventional assumption” and they provide no theoretical evidence showing that this ”complicated assumption” holds when using Stochastic Gradient Descent in a non-convex optimization problem. Prompted by this critique, we were interested to see under what assumption of alternative views have been proposed to demonstrate the utility of Stochastic Gradient descent in non-convex optimization problems.

More recently, Ibayashi et al. [2] in 2021 demonstrated that Stochastic Gradient Descent escapes from sharp minima exponentially fast even before Stochastic Gradient Descent reaches a stationary distribution. In their paper, they proposed a theory to tackle the question as to why Stochastic Gradient Descent finds generalizable solutions in complicated models such as neural networks. They frame their question in terms of ”escape efficiency from sharp minima”. This ”escape efficiency” is a measure for how fast the Stochastic Gradient Descent moves out of the neighborhood of the minima. In other words, Stochastic Gradient Descent can find generalizable minima because it has high escape efficiency. The concept of high escape efficiency is realized by the concept of ”anisotropic noise” in Stochastic Gradient Descent [9] (the authors define it as the noise with the various magnitudes among directions). Additionally, Xie et al. [7] proposed a density diffusion theory that Stochastic Gradient Descent favors flat minima exponentially more than sharp minima. I did not dive too in-depth into these papers as I wanted to simply understand the research contour and what people are doing in this field to either prove or empirically demonstrate the usefulness of Stochastic Gradient Descent.

While noise has been studied and shown to have nice properties, step size also players a role in the popularity of Stochastic Gradient Descent and it would be remiss of me to not at least make note of it. In optimization, small step sizes can help refine the network and converge to a local minima while large step sizes can help in escaping local minima and point towards better ones. Kleinberg et al. [4] in addition to exploring the effect of gradient noise, also show the importance of picking a good step size or having a training schedule that can shrink the step size to escape bad local minima. And Mohtashami et al. [5] in response to Kleinberg et al. [4] argue that the gradient noise is not completely sufficient enough to explain its good convergence properties and that even after the noise, there still might be regions that can only be avoided using a high learning rate. Xie et al. [7] also discussed in their work the impact that a small step size or large batch training can impact training with an observation that large batches have difficulty in searching flat minima efficiently under realistic computational restraints.

Overall, there has been a lot of work to describe Stochastic Gradient Descent with mathematical rigor. Often time, novel theories and strong assumptions have to be made to achieve these properties and empirical observations are often used to back up the theories. In this literature review, we discussed how viewing Stochastic Gradient Descent through a convolution framework can help us better understand why Stochastic Gradient Descent may lead to better generalizations through converging to flatter local minima. While this framework requires strong assumptions that might not be met, its a step in the right direction in terms of shedding light onto what these algorithms are doing in non-convex optimization problems. We also discovered that noise was not the only impactful thing in Stochastic Gradient Descent but a careful consideration of step size or learning rate and batch size are also required.

As we've learned in class, many optimizers consider an adaptive step size that gets smaller as we get closer to the local minima to prevent it from bouncing around too much as well as different batch sizes leading to different convergence results. A lot of work remains in enumerating the properties of Stochastic Gradient Descent and understanding how it works, but in light of that fact, this literature review hopes to provide one perspective as to why Stochastic Gradient Descent is so popular.

References

[1] Laurent Dinh et al. *Sharp Minima Can Generalize For Deep Nets*. 2017. DOI: 10.48550/ARXIV.1703.04933. URL: https://arxiv.org/abs/1703.04933.

[2] Hikaru Ibayashi and Masaaki Imaizumi. *Exponential escape efficiency of SGD from sharp minima in non-stationary regime*. 2021. DOI: 10.48550/ARXIV.2111.04004. URL: https://arxiv.org/abs/2111.04004.

[3] Nitish Shirish Keskar et al. *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*. 2016. DOI: 10.48550/ARXIV.1609.04836. URL: https://arxiv.org/abs/1609.04836.

[4] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. *An Alternative View: When Does SGD Escape Local Minima?* 2018. DOI: 10.48550/ARXIV.1802.06175. URL: https://arxiv.org/abs/1802.06175.

[5] Amirkeivan Mohtashami, Martin Jaggi, and Sebastian Stich. *Special Properties of Gradient Descent with Large Learning Rates*. 2022. DOI: 10.48550/ARXIV.2205.15142. URL: https://arxiv.org/abs/2205.15142.

[6] Jingfeng Wu et al. *On the Noisy Gradient Descent that Generalizes as SGD*. 2019. DOI: 10.48550/ARXIV.1906.07405. URL: https://arxiv.org/abs/1906.07405.

[7] Zeke Xie, Issei Sato, and Masashi Sugiyama. *A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima*. 2020. DOI: 10.48550/ARXIV.2002.03495. URL: https://arxiv.org/abs/2002.03495.

[8] Mo Zhou et al. "Toward Understanding the Importance of Noise in Training Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 7594–7602. URL: https://proceedings.mlr.press/v97/zhou19d.html.

[9] Zhanxing Zhu et al. *The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects*. 2018. DOI: 10.48550/ARXIV.1803.00195. URL: https://arxiv.org/abs/1803.00195.