SGD
What drives convergence and divergence?

Vivak Patel

University of Chicago
OMS 2017 Havana, Cuba

December 12, 2017
Outline

Background. What is SGD? What is the important terminology?

Motivation. Convergence results exist, so why does this matter?

Mechanism. What is causing the motivating phenomenon?
Background: Problem

In data analysis (e.g., statistics and machine learning), we are interested in identifying parameters by solving

$$\min_x F(x), \text{ where } F(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x).$$  \hspace{1cm} (1)

When $N$ is large (depends on context), speed benefits are obtained by using gradient information from randomly subsampled $f_i$ (Bertsekas, 2011).
Stochastic gradient descent (SGD): Given $x_0$, SGD generates a sequence $\{x_k : k \in \mathbb{N}\}$ defined iteratively by

$$x_{k+1} = x_k - \alpha_k \sum_{j=kJ+1}^{k(J+1)} \nabla f_{Z_j}(x_k),$$

where $\{\alpha_k\}$ is the step size; $\{Z_j\}$ are i.i.d. with $\mathbb{P}[Z_1 = n] = p_n > 0$ for $n = 1, \ldots, N$. 
Background: Terminology

**Batch Size.** $J$ is the batch-size.

**Expected Objective.** $F_{\mathbb{E}}(x) = \mathbb{E}[f_{Z_i}(x)]$

**Homogeneous Minimizer.** A minimizer is common to all $f_i$.

**Flat Minimizer.** At $x^*$, $\nabla^2 F_{\mathbb{E}}$ has small eigenvalues.
Motivation

On experiments with deep neural networks, Keskar et al. (2016) observed that small batch (SB) stochastic methods consistently converged to flat minimizers of the training function while large batch (LB) stochastic methods converged to sharper minimizers of the training function.
Motivation

Adapted from Figure 1 from Keskar et al. (2016).
The optimization method is somehow selecting the minimizer. This is unprecedented for classical optimization techniques.
Mechanism

What is driving SB stochastic gradient methods to prefer flat minimizers?
The de facto hypothesis in machine learning is that because SB methods are more stochastic than LB methods, their random search directions will eventually force them out of basins of attractions for sharp minimizers.
Mechanism: Stochastic

Expected Objective

Flat Minimum  Sharp Minimum
Mechanism: Counter Evidence

Expected Objective

Flat Minimum

Sharp Minimum
Mechanism: Quadratic Sums

Consider the Quadratic Sums problem which is defined as follows. Let $\beta^* \in \mathbb{R}^d$, $Q_1, \ldots, Q_N \in \mathbb{R}^{d \times d}$ be symmetric, positive semi-definite matrices, and $r_1, \ldots, r_N \in \mathbb{R}^d$ such that

$$\beta^* \in \text{argmin}_{x \in \mathbb{R}^d} \frac{1}{2} x'Q_n x + x'r_n,$$

for $n = 1, \ldots, N$. The (homogeneous) quadratic sums objective is

$$F(x) = \sum_{n=1}^{N} \frac{1}{2} x'Q_n x + x'r_n.$$
Mechanism: Gradient Descent

A classical result on the convergence and divergence of Gradient Descent (SGD with $J = \infty$) for Quadratic Sums:

**Theorem**

Let $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$ be non-zero eigenvalues of $\nabla^2 F_{\mathbb{E}}(\beta^*)$. When $\alpha_k > 2/(\lambda_r)$ or $\alpha_k < 0$ then $F_{\mathbb{E}}(x_{k+1}) > F_{\mathbb{E}}(x_k)$. When $0 < \alpha_k < 2/(\lambda_1)$ then $F_{\mathbb{E}}(x_k) > F_{\mathbb{E}}(x_{k+1})$. 
Mechanism: SGD Divergence

Theorem (Patel (2017) Theorem 16)

Let \( s(F_E) > 0 \). Define

\[
M = \{ m : s > J\lambda_{m-1} \lambda_m \} \quad \text{and} \quad M = \begin{cases} 
\min M & M \neq \emptyset \\
r + 1 & M = \emptyset 
\end{cases}
\]

If either \( \alpha_k < 0 \) or

\[
\alpha_k > \frac{2\lambda_{M-1}}{\lambda_{M-1}^2 + s/J},
\]

then \( F_E(x_{k+1}) > F_E(x_k) \).
Mechanism: SGD Convergence

Theorem (Patel (2017) Theorem 17)

Let $g(F_E) > 0$. If

$$t \leq J \lambda_1 \lambda_r \text{ and } 0 < \alpha_k < 2 \lambda_1 \left[ \lambda_1^2 + \frac{t}{J} \right]^{-1}, \text{ or}$$

$$t > J \lambda_1 \lambda_r \text{ and } 0 < \alpha_k < 2 \lambda_r \left[ \lambda_r^2 + \frac{t}{J} \right]^{-1},$$

then $F_E(x_{k+1}) < F_E(x_k)$. 
Mechanism: SGD

In Patel (2017), we show that these results extend to the inhomogeneous case. They also readily extend to the infinite $N$ linear regression problem.

In addition, we numerically calculate the convergence and divergence bounds for two non-convex stochastic optimization problems and demonstrate that SGD-$J$ will converge or diverge to minima depending on the learning rate with respect to these bounds.
SGD-J converges or diverges from minimizers based on the expected local geometry and batch size, which are extensions of the classical GD criteria. Moreover, our results explain how flat regions and sharp regions mediate the convergence and divergence of SGD.
Summary

Observation. Keskar et al. (2016) observed that SB methods converge to flat minimizers in comparison to LB methods.

De Facto Mechanism. Stochasticity of SB methods allows them to escape sharp minimizers.

Our Mechanism. Deterministic analogues to classical properties actually mediate this behavior.
Bertsekas, D. P.  


Patel, V.  
Acknowledgements

Mihai Anitescu for his general guidance.

NSF RTG 1547396 for its financial support.

Rob Webber for his helpful discussions.
Thank You
www.vivakpatel.org