A Proof of Convergence for Stochastic Gradient Descent on Convex Cost Functions

Daniel Mourad

University of Maryland, College Park

Original paper by M. Fazylab, F. Gama, and S. Paternain

Fall 2017

Main Theorem

Theorem

In Stochastic Gradient Descent, given updates $X_{n+1} = X_n - \epsilon_n \widehat{\nabla} f_{n+1}$, *where*

- $X_n : \Omega \to \mathbb{R}^d$ *is a random variable representing the weights,*
- ϵ_n *is the step size,*
- $\widehat{\nabla} f_{n+1}$ *is the estimated gradient of* f, *a convex cost function with global mininum* $x^\star$,

$$X_n \to x^\star \ \ a.s.$$

Overview

Definitions

Setting: $(\Omega, \mathcal{F}, P)$ a probability space.

Definition (Sigma Algebra)

$\mathcal{F} \subset \mathcal{P}(\Omega)$ is a σ-algebra, meaning that it is closed under complementation, finite intersection, finite union, and that $\Omega \in \mathcal{F}$.

Definition (Random Variable)

$X : \Omega \to \mathbb{R}$ is a random variable in $\mathcal{F}$ $(X \in \mathcal{F})$ if

$$X^{-1}(B) \in \mathcal{F}$$

for all $B \in \mathcal{B}(R) := \{$the open sets in $\mathbb{R}$ closed under countable unions, intersections, and complementation$\}$.

Definitions (continued)

Setting: $(\Omega, \mathcal{F}, P)$ a probability space.

Definition (Expected Value)

For random variable $X : \Omega \to \mathbb{R}$,

$$E[X] = \int_{\Omega} X dP.$$

Definition (Conditional Expectation)

For $\mathcal{F}_0 \subset \mathcal{F}$, $E(X|\mathcal{F}_0)$ is the unique random variable $Y \in \mathcal{F}_0$ such that for all $A \in \mathcal{F}_0$,

$$\int_A X dP = \int_A Y dP.$$

Note: $E[E(X|\mathcal{F}_0)] = E[Y]$

Martingales

Definition

A sequence of random variables X_n along with an increasing sequence of σ-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ is a **martingale** if

(i) $E(|X_n|) < \infty$,

(ii) $X_i \in \mathcal{F}_i$ for each i,

(iii) $E(X_{i+1}|\mathcal{F}_i) = X_i$ for each i.

In a **supermartingale**, the equality in (iii) is replaced with $\leq$, and the reverse for **submartingales**.

Lemma (Lemma 1; [Durrett, 2010])

If X_n is a non-negative super martingale, there exists a random variable X such that $X_n \to X$ a.s. and $E(X) \leq E[X_0]$

Overview

Gradient Descent

We wish to minimize a loss function $f : \mathbb{R}^d \to \mathbb{R}$.
In deterministic gradient descent, we use the entire training set, so we can calculate the gradient exactly:

Definition (Deterministic Gradient Descent)

$$x_{n+1} = x_n - \epsilon_n \nabla f(x_n).$$

In Stochastic Gradient Descent, we don't use the entire training set. This introduces noise random variables, W_i.

Definition (Stochastic Gradient Descent)

$$X_{n+1} = X_n - \epsilon_n \widehat{\nabla} f_{n+1}$$

where

$$\widehat{\nabla} f_{n+1} = \nabla f(X_n) + W_{n+1}.$$

The Probability Space

We need a probability space for the random variables.

Setting

Our probability space is

$$\left(\prod_{i=1}^{\infty} \Omega_i, \bigcup_{i=1}^{\infty} \mathcal{F}_i, P \right)$$

where

- Ω_i *is the set of possible choices for test data at stage* i,
- $\mathcal{F}_i$ *represents the events determined by choices of data less at stages at and before* i,
- P *gives the uniform probability measure of at each* F_i.

Assumptions

Definition (Stochastic Gradient Descent)

$$X_{n+1} = X_n - \epsilon_n \widehat{\nabla} f_{n+1}$$

where

$$\widehat{\nabla} f_{n+1} = \nabla f(X_n) + W_{n+1}.$$

Note that each $X_n, \widehat{\nabla} f_n \in \mathcal{F}_n$ - that is, they are determined by the first n choices of training data.
We make some assumptions on $\widehat{\nabla} f_n$:

Assumptions

(i) $E(\widehat{\nabla} f_{n+1} | \mathcal{F}_n) = \nabla f(X_n)$.

(ii) *There is* $B \in \mathbb{R}$ *such that* $\left\| \widehat{\nabla} f_n \right\| \leq B$.

(iii) f *is convex.*

(iv) $\sum \epsilon_n = \infty, \quad \sum \epsilon_n^2 < \infty.$

Overview

Necessary Fact About Norms

Fact

For $a, b \in R^d$,

$$\|a-b\|^2 = \|a\|^2 - 2a^T b + \|b\|^2$$

Proof.

$$\begin{aligned}
\|a-b\|^2 &= \langle a-b, a-b\rangle \\
&= \langle a, a\rangle + \langle a, -b\rangle + \langle -b, a\rangle + \langle b, b\rangle \\
&= \|a\|^2 - a^T b - b^T a + \|b\|^2 \\
&= \|a\|^2 - 2a^T b + \|b\|^2
\end{aligned}$$

Checking $\|X_n - x^\star\|^2$

Is $\|X_n - x^\star\|^2$ a super martingale? We check condition (iii): that $E(Y_{i+1}|\mathcal{F}_i) \leq Y_i$ for each i.

Check

$$
\begin{aligned}
E(\|X_{n+1} - x^\star\|^2|\mathcal{F}_n) &= E(\left\|X_n - \epsilon_n \widehat{\nabla} f_{n+1} - x^\star\right\|^2|\mathcal{F}_n) \\
&= E(\left\|(X_n - x^\star) - \epsilon_n \widehat{\nabla} f_{n+1}\right\|^2|\mathcal{F}_n) \\
&= E(\|X_n - x^\star\|^2 - 2\epsilon_n(\widehat{\nabla} f_{n+1})^T(X_n - x^\star) + \epsilon_n^2\left\|\widehat{\nabla} f_n + 1\right\|^2|\mathcal{F}_n) \\
&\leq E(\|X_n - x^\star\|^2 - 2\epsilon_n(\widehat{\nabla} f_{n+1})^T(X_n - x^\star) + \epsilon_n^2 B^2|\mathcal{F}_n)
\end{aligned}
$$

Convergence of Y_n

Lemma (Lemma 2)

$$Yn := \|X_n - x^\star\|^2 + B^2 \sum_{i=n}^{\infty} \epsilon_i^2$$

is a nonnegative supermartingale. That is,

(i) $E(|Y_n|) < \infty$,

(ii) $Y_i \in \mathcal{F}_i$ *for each* i,

(iii) $E(Y_{i+1}|\mathcal{F}_i) \leq Y_i$ *for each* i.

By lemma 1, Y_n converges to some Y a.s.

Convergence of X_n

Lemma (Lemma 3)

$\|X_n - x^\star\|^2$ *converges almost surely to some integrable random variable.*

Proof.

$$Y = \lim_{n\to\infty} Y_n = \lim_{n\to\infty} \|X_n - x^\star\|^2 + B^2 \sum_{i=n}^{\infty} \epsilon_i^2$$
$$= \lim_{n\to\infty} \|X_n - x^\star\|^2.$$

□

Overview

Lemma (Lemma 4)

$$\liminf_{n\to\infty} E[(\nabla f(X_n))^T (X_n - x^\star)] = 0$$

Theorem (Dominated Convergence Theorem)

If

(i) $Y_n \to Y$ *and*

(ii) *there exists a* Z *such that* $|Y_n| \leq Z$ *for all* n *with* $E[Z]$ *finite,*

then

$$\lim_{n\to\infty} E[Y_n] = E[\lim_{n\to\infty} Y_n] \quad (= E[Y])$$

Lemma (Lemma 5)

If $g, h : \mathbb{R}^d \to \mathbb{R}$, and Z_n are a sequence of random variables s.t.

(i) *There is $\bar{x}$ such that $g(x) = 0, h(x) = 0 \leftrightarrow x = \bar{x}$*

(ii) $\lim_{\|x\| \to \infty} g(x), h(x) \neq 0$

(iii) *$g(Z_n)$ converges almost surely to some random variable*

(iv) $\liminf_{n \to \infty} E[h(Z_n)] = 0$

then

$$\lim_{n \to \infty} Z_n = \bar{x} \quad \text{almost surely.}$$

Facts Used

- *Convergence in L^p implies convergence in probability ([Durrett, 2010], Lemma 2.2.2, p. 54)*
- *Convergence in probability implies a subsequence with almost sure convergence ([Durrett, 2010] Theorem 2.3.2, p. 65)*

Final Result

Theorem

In Stochastic Gradient Descent with convex cost function f,

$$X_n \to x^\star \ \ a.s.$$

References

Durrett, R. (2010)

Probability Theory and Examples, fourth edition, Cambridge University Press, New York, NY.

Fazylab, M., Gama F., Paternian, S. (2012)

A Proof of Convergence for Stochastic Gradient Descent

URL:
https://pdfs.semanticscholar.org/4220/a820e237cb4eb66fb8aafb15d3c4d8bd2fd3.pdf