**Columbia University in the City of New York**
**Optimization Methods for Machine Learning**
Instructors: Satyen Kale
Authors: Gurpreet Singh
Email: gurpreet.s@columbia.edu

SCRIBE
10

# Stochastic Gradient Descent

## 1. Stochastic Optimization

Often an optimization problem requires us to find the minimum of an expectation over a distribution. This is particularly common in Machine Learning algorithms where we are trying to minimize the expected loss over a distribution (though it is usually estimated using a set of samples). This optimization objective looks like as given below.

$$\min_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x}) \;=\; \min_{\mathbf{x}\in\mathcal{K}} \mathop{\mathbb{E}}_{\xi\sim\mathcal{D}} \left[\, g(\mathbf{x},\xi) \,\right]$$

In such cases, it is often quite cumbersome to compute the expectation at every descent step. Each descent step, therefore, is expensive and, henceforth, the convergence process is slow. The concept behind Stochastic Optimization is to estimate the gradients, *i.e.* compute stochastic (or noisy) gradients and take descent steps using these stochastic gradients. The idea of stochastic optimization originated from Robbins-Monro algorithm and is now the most important optimization method in machine learning.

Before looking at the stochastic gradient descent algorithm, we discuss the formal definition of stochastic gradients.

**Definition 10.1** Stochastic Gradients are noisy gradients that are "unbiased" estimates of the actual gradients. Formally, the stochastic gradient for the expectation $f(\mathbf{x}) = \mathop{\mathbb{E}}_{\xi\sim\mathcal{D}} \left[\, g(\mathbf{x},\xi) \,\right]$ is given by $\nabla_{\mathbf{x}} g(\mathbf{x},\xi')$ where $\xi'$ is a sample from the distribution $\mathcal{D}$.

By unbiased, we mean that the expectation of the gradient is equal to the actual gradient at that point. This is trivially true as the gradient is with respect to $\mathbf{x}$ and can be propogated into the expectation term.

## 2. Stochastic Gradient Descent

The assumption for the optimization problem is that $g$ (which implies $f$ is convex as well) is a convex function for every $\xi$ and, as asual, $\mathcal{K}$ is a convex set. The SGD algorithm for the above problem is given in algorithm 1.

One thing to note is that because $\xi_t$ are sampled from $\mathcal{D}$ and therefore both $\xi_t$ and $\mathbf{x}_t$ are random

**Algorithm 1**: Stochastic Gradient Descent

```
1. Start with an arbitrary point $\mathbf{x}_0 \in \mathcal{K}$
2. For $t = 1, 2 \dots T$
    (a) Draw $\xi_t \sim \mathcal{D}$
    (b) Set $\mathbf{y}_t = \mathbf{x}_{t-1} - \eta \nabla g(\mathbf{x}_{t-1}, \xi_t)$
    (c) Update $\mathbf{x}_t = \Pi_{\mathcal{K}}(\mathbf{y})$
3. Output the final estimate as some combination of $\{\mathbf{x}_0, \mathbf{x}_1 \dots \mathbf{x}_T\}$
```

variables (for $t \geq 1$). However, we can conditionally compute the expectation of the gradient of $g$. We can write the following terms based on conditional expectations which will be useful for the convergence analysis of SGD discussed in later sections.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \nabla g(\mathbf{x}, \xi) \right] &= \nabla f(\mathbf{x}) \\
\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \nabla g(\mathbf{x}_t, \xi) \,\middle|\, \xi_0, \xi_1 \dots \xi_{t-1} \right] &= \nabla f(\mathbf{x}_t) \\
\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \nabla g(\mathbf{x}_t, \xi) \,\middle|\, \mathbf{x}_t \right] &= \nabla f(\mathbf{x}_t) \\
\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \mathbf{y}_{t+1} \,\middle|\, \mathbf{x}_t \right] &= \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)
\end{aligned}$$

## 3. Convergence Analysis of SGD

Suppose we have a convex function $g$ which also satisfies the following constraint

$$\forall \mathbf{x} \in \mathcal{K}, \quad \mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \| \nabla g(\mathbf{x}, \xi) - \nabla f(\mathbf{x}) \|_2^2 \right] \leq \sigma^2$$

This is equivalent to saying that the variance of our unbiased estimator of the gradient *i.e.* $\nabla g(\mathbf{x}, \xi)$ for any $\mathbf{x} \in \mathcal{K}$ is lesser than $\sigma^2$. Note that because $\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \nabla g(\mathbf{x}, \xi) \right] = \nabla f(\mathbf{x})$, we have $\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \| \nabla g(\mathbf{x}, \xi) - \nabla f(\mathbf{x}) \|_2^2 \right] = \mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \| \nabla g(\mathbf{x}, \xi) \|^2 \right] - \| \nabla f(\mathbf{x}) \|_2^2$, and so the above assumption implies that $\mathop{\mathbb{E}}_{\xi \sim \mathcal{D}} \left[ \| \nabla g(\mathbf{x}, \xi) \|^2 \right] \leq \| \nabla f(\mathbf{x}) \|_2^2 + \sigma^2$.

We further assume that the function $f$ is $L$-Lipschitz. Therefore, we have $\| \nabla f(\mathbf{x}) \| \leq L$ for all $\mathbf{x} \in \mathcal{K}$.

From the analysis of projected gradient descent, we already know the following result

$$\begin{aligned}
\| \mathbf{x}_{t+1} - \mathbf{x}^* \|^2 &\leq \| \mathbf{x}_t - \mathbf{x}^* \|^2 + \eta^2 \| \nabla g(\mathbf{x}_t, \xi_t) \|^2 - 2\eta \nabla g(\mathbf{x}_t, \xi_t)^{\mathrm{T}} (\mathbf{x}_t - \mathbf{x}^*) \\
\implies \nabla g(\mathbf{x}_t, \xi_t)^{\mathrm{T}} (\mathbf{x}_t - \mathbf{x}^*) &\leq \frac{1}{2\eta} \left( \| \mathbf{x}_t - \mathbf{x}^* \|^2 - \| \mathbf{x}_{t+1} - \mathbf{x}^* \|^2 \right) + \frac{\eta}{2} \| \nabla g(\mathbf{x}_t, \xi_t) \|^2
\end{aligned}$$

Up on taking an expectation conditioned on $\mathbf{x}_t$ (i.e. $\underset{\xi_t \sim \mathcal{D}}{\mathbb{E}} \left[ \ldots \,|\, \mathbf{x}_t \right]$) on both sides, we get

$$\mathbb{E}\left[ \nabla g(\mathbf{x}_t, \xi_t)^{\mathrm{T}} (\mathbf{x}_t - \mathbf{x}^*) \,|\, \mathbf{x}_t \right] \;\leq\; \mathbb{E}\left[ \frac{1}{2\eta} \left( \| \mathbf{x}_t - \mathbf{x}^* \|^2 - \| \mathbf{x}_{t+1} - \mathbf{x}^* \|^2 \right) + \frac{\eta}{2} \| \nabla g(\mathbf{x}_t, \xi_t) \|^2 \,|\, \mathbf{x}_t \right]$$

The LHS is simply $\nabla f(\mathbf{x}_t)^{\mathrm{T}}(\mathbf{x}_t - \mathbf{x}^*)$. Also, $\| \nabla g(\mathbf{x}_t, \xi_t) \|^2$ is smaller than $\| \nabla f(\mathbf{x}_t) \|^2 + \sigma^2$ which is smaller than $L^2 + \sigma^2$. Hence, we have

$$\nabla f(\mathbf{x}_t)^{\mathrm{T}}(\mathbf{x}_t - \mathbf{x}^*) \;\leq\; \frac{1}{2\eta} \left( \| \mathbf{x}_t - \mathbf{x}^* \|^2 - \frac{1}{2\eta} \mathbb{E}\left[ \| \mathbf{x}_{t+1} - \mathbf{x}^* \|^2 \,|\, \mathbf{x}_t \right] \right) + \frac{\eta}{2} L^2 + \frac{\eta}{2} \sigma^2$$

From the convexity of $f$, we have

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \;\leq\; \nabla f(\mathbf{x}_t)^{\mathrm{T}}(\mathbf{x}_t - \mathbf{x}^*)$$

Therefore, we can write

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \;\leq\; \frac{1}{2\eta} \left( \| \mathbf{x}_t - \mathbf{x}^* \|^2 - \frac{1}{2\eta} \mathbb{E}\left[ \| \mathbf{x}_{t+1} - \mathbf{x}^* \|^2 \,|\, \mathbf{x}_t \right] \right) + \frac{\eta}{2} \| \nabla f(\mathbf{x}_t) \|^2 + \frac{\eta}{2} \sigma^2$$

Taking expectation and averaging over $t = 0$ to $T - 1$ on both sides, we get

$$\mathbb{E}\left[ \frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \right] - f(\mathbf{x}^*) \;\leq\; \frac{1}{2\eta T} \| \mathbf{x}_0 - \mathbf{x}^* \|^2 + \frac{\eta}{2} \left( L^2 + \sigma^2 \right)$$

In order to make the RHS as small as possible, we choose $\eta$ such that $\frac{1}{2\eta T} \| \mathbf{x}_0 - \mathbf{x}^* \|^2 = \frac{\eta}{2} \left( L^2 + \sigma^2 \right)$. Therefore, we get

$$\eta \;=\; \frac{\| \mathbf{x}_0 - \mathbf{x}^* \|}{\sqrt{T \left( L^2 + \sigma^2 \right)}}$$

Also, from the convexity of $f$, we have $\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) \geq f\left( \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t \right) = f(\bar{\mathbf{x}}_T)$. Therefore, we can write

$$\mathbb{E}\left[ f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \right] \;=\; \frac{\| \mathbf{x}_0 - \mathbf{x}^* \| \sqrt{L^2 + \sigma^2}}{\sqrt{T}}$$

Notice that we can only comment on the convergence in expectation. Therefore, we are “expecting” our function value to converge to the optimal value. We can also get a high probability convergence result by bounding the variance. That, however, is outside of the scope of this class.