EE 588: Optimization for the information and data sciences

University of Southern California

Lecturer: Mahdi Soltanolkotabi November 27, 2017

Lecture 21-22

1 Incremental and Stochastic Methods

Stochastic Gradient Method (SGM) is one of the most popular algorithms for contemporary data analysis and machine learning. It has a long history and has been invented many times as 'Least Mean Squares', 'Back Propogation', 'Randomized Kaczmarz method' and so on. The history of SGM goes back to Robbins and Monro (1950). The main insight of SGM is that gradient methods are very robust even if we encounter descent *on average*.

Assume that $\nabla f(x)$ is not known exactly but we know a random function $G(x, \xi)$ such that

$$\mathbb{E}_\xi [G(x)] = \nabla f(x)$$

Now pretend that G is the gradient and form the update

$$x_{t+1} = x_t - \mu_t G(x_t, \xi_t)$$

Note that on an *average* we have descent if we wait 'long enough'. An obvious complication of the above procedure is that even if $\nabla f(x^\star) = 0$, SGM might still move away from $x^\star$.

There are several variants of the function G that are available and we consider a few below.

- **Noisy Gradients**
  Assume that f is convex and we observe

  $$G(x, w) = \nabla f(x) + w$$

  where w is a suitable noise process, say Gaussian.

- **Incremental Gradient Method** (A special case of SGM)
  The incremental gradient method, a.k.a Perceptron or Back Propogation, is the most common usage of SGM.

  Assume that

  $$f(x) = \sum_{i=1}^{n} f_i(x)$$

  where n is large and so the full gradient computation is expensive. To tackle this problem, pick $\phi_t \in \{1, 2, \cdots, n\}$ at random and do

  $$x_{t+1} = x_t - \mu_t \nabla f_{\phi_t}(x_t) = x_t - \mu_t G(x_t, \phi_t)$$

By cyclicity, we hope to find a minimizer. Note that if ϕ_t is uniform then

$$\mathbb{E}_{\phi_t}\left[\nabla f_{\phi_t}(x)\right] = n^{-1}\sum_{i=1}^{n}\nabla f_i(x) = \nabla f(x)$$

We consider a few examples below.

1.1 Classification and the Perceptron

Consider n observations $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$. We want to find a $w \in \mathbb{R}^d$ such that

$$\langle w, x_i\rangle \begin{cases} > 0 \text{ if } y_i = 1 \\ < 0 \text{ if } y_i = -1 \end{cases} \quad \forall\ i = 1, 2, \cdots, n$$

Essentially, we want a half-space with the positives ($y_i = 1$) on one side and the negatives ($y_i = -1$) on the other. The Perceptron algorithm (1950's) solves this problem using one example at a time as follows:

- Initialize at some w_0
- At iteration t, choose data (x_{ϕ_t}, y_{ϕ_t})
- Update $w_{t+1} = (1-\gamma)w_t + \eta \begin{cases} y_{\phi_t}x_{\phi_t} \text{ ,if } sign\left(y_{\phi_t}\langle w_t, x_{\phi_t}\rangle\right) = -1 \\ 0 \text{ ,otherwise} \end{cases}$

The idea is to correct w_t in a direction so that $\langle w, x_i\rangle$ has the right sign($= 1$). This algorithm is equivalent to the following optimization problem

$$\min_{w\in\mathbb{R}^d} n^{-1}\sum_{i=1}^{n}\max\left(1 - y_i\langle w, x_i\rangle, 0\right) + \frac{\lambda}{2n}\|w\|_2^2$$

One can derive the Perceptron updates easily by noting

$$f_i(w) = \max\left(1 - y_i\langle w, x_i\rangle, 0\right) + \frac{\lambda}{2n}\|w\|_2^2$$
$$\nabla f_{\phi_t}(w) = -y_{\phi_t}x_{\phi_t} + \frac{\lambda}{n}w \text{ ,if } y_{\phi_t}\langle w, x_{\phi_t}\rangle < 0$$

with $\gamma = \eta\lambda n^{-1}$ and $\mu_t = \eta$.

2 Empirical Risk Minimization (ERM)

In machine learning, there are several instances of optimization problems called Empirical Risk Minimization. Many classification, regression, decision making tasks can be thought of as minimizing expected value of error over the data generating distributions. For example, given a loss function $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, we want to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that the risk

$$\mathcal{R}(f) = \mathbb{E}_{P(X,Y)}\left[\ell(f(X), Y)\right]$$

is minimized, where $P(X,Y)$ is the unknown joint probability distribution of (X,Y). Minimizing this risk is difficult without the knowledge of $P(X,Y)$ and even with $P(X,Y)$ known, the above minimization might still be hard. The idea of ERM is to use the data at hand $(x_i, y_i), i = 1, 2, \cdots, n$ and define the empirical risk

$$\mathcal{R}_{emp}(f) = n^{-1} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

and minimize this over an appropriate class of functions $\mathcal{F}$. Further, if we assume $(x_1, y_1), \cdots, (x_n, y_n)$ are i.i.d $P(X,Y)$ then

$$\mathbb{E}_{P(X,Y)} \left[R_{emp}(f) \right] = \mathcal{R}(f)$$

Thus by minimizing the empirical risk, we are approximately minimizing the true risk.

SGM and ERM are tied together. Instead of computing the full gradient, we sample a single data point and compute it's gradient. Let us see a few examples.

- **Computing mean**
  Consider the emprical risk function

  $$\mathcal{R}_{emp}(x) = \frac{1}{2n} \sum_{i=1}^{n} (x - w_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \ell(x, w_i)$$

  and minimize it with respect to x. Start with $x_1 = 0$ and use $\mu_t = 1/t$. Then

  $$\begin{aligned} x_2 &= w_1 \\ x_3 &= \frac{1}{2}(w_1 + w_2) \\ &\vdots \\ x_n &= \frac{1}{n-1} \sum_{i=1}^{n-1} w_i \end{aligned}$$

  The cost after n steps: empirical variance of w. In reality,

  $$\mathcal{R}(x) = \mathbb{E}_w \left[(x - w)^2 \right]$$

  and assuming $\mathbb{E}(w) = \mu, V(w) = \sigma^2$, we have

  $$\begin{aligned} \mathbb{E} \left[\mathcal{R}_{emp}(x_n) \right] &= \frac{\sigma^2}{2n} + \frac{\sigma^2}{2} \\ \text{and } \mathbb{E} \left[\mathcal{R}_{emp}(\mu) \right] - \mathbb{E} \left[\mathcal{R}_{emp}(x_n) \right] &= \frac{\sigma^2}{2n} \end{aligned}$$

  Thus a 'one at a time' procedure does not give us a better convergence than $O(n^{-1})$.

- **The Kaczmarz method**
  Consider the following problem

$$\min_x (2n)^{-1} \sum_{i=1}^{n} (\langle a_i, x \rangle - b_i)^2$$

and assume that there exists $x^\star$ such that $\langle a_i, x^\star \rangle - b_i = 0$ for all $i = 1, \cdots, n$. If we do the usual cyclic SGM with stepsize $\mu_t = 1$ then

$$x_n - x^\star = \prod_{i=1}^{n} \left(I - a_i a_i^{'} \right) (x_0 - x^\star)$$

In the above equation, note the expression for projection onto a one dimensional subspace. If two subsequent subspaces are close then we do not make any progress! Infact, we can design a sequence a_i such that we make no progress.

Next, let $w_n = \pi/n$ and set

$$a_t = \begin{pmatrix} cos(tw_n) \\ sin(tw_n) \end{pmatrix}, n \geq 3$$

If we use an SGM for m steps,

$$\begin{aligned} \mathbb{E}_\phi \left[\prod_{i=1}^{m} \left(I - a_{\phi_i} a_{\phi_i}^{'} \right) (x_0 - x^\star) \right] &= \left(I - m^{-1} \sum_{i=1}^{m} a_i a_i^{'} \right)^m (x_0 - x^\star) \\ &= 2^{-m} (x_0 - x^\star) \end{aligned}$$

So SGM converges linearly to the optimal solution. However, note that

$$\begin{aligned} \left\| \prod_{i=1}^{m} \left(I - a_i a_i^{'} \right) (x_0 - x^\star) \right\|_2 &= cos^{m-1}(w_n) |\langle a_1, x0 - x^\star \rangle| \\ &\leq \left(1 - \frac{1}{m} \right) |\langle a_1, x_0 - x^\star \rangle| \end{aligned}$$

Thus randomization helps. This is known as randomized Kaczmarz method. It is simply a SGM for solving least squares. This method was analyzed by signal processing researchers without knowing SGM.

3 Epochs

This is an important notion in SGM's. In an epoch, same number of iterations are run and then a decision is made whether to change the step size. So in the k^{th} epoch, $\mu_k = \mu \gamma^{k-1}$ where $\gamma \in (0.8, 0.9)$. The concept of epoch doubling is slightly different. In this case, the first T iterations may have a stepsize of μ, the next $2T$ with $\mu/2$, the next $4T$ with $\mu/4$ and so on. This scheme is nothing but a piece-wise approximation to $\mu_t = 1/t$.

4 Momentum

It is possible to run SGM with a momentum term $B \in (0.8, 0.95)$ so that

$$x_{t+1} = x_t - \mu_t G(x_t) + B(x_t - x_{t-1})$$

This scheme works well in practice although the theory suggests only minor improvements. It appears that the constants in the estimate of convergence rate can be improved upon by using a momentum based update scheme.

5 Analysis

Let $f : \mathbb{R}^d \to \mathbb{R}$ and $x^\star$ be the global optimal solution. We observe $G(x, \xi)$ such that

$$\mathbb{E}_\xi \left[G(x, \xi) \right] = \nabla f(x)$$

Suppose there exists L_G and B such that

$$\mathbb{E}_\xi \left[\|G(x, \xi)\|_2^2 \right] \leq L_G^2 \|x - x^\star\|_2^2 + B^2$$

We do not make any assumptions on the boundedness of G except in expectation. Consider the usual SGM update

$$x_{t+1} = x_t - \mu_t G(x_t, \xi_t)$$

- **Case:** $L_G = 0$ **and** $B > 0$
  Take

  $$f(x) = n^{-1} \sum_{i=1}^{n} \max\left(1 - y_i \langle x, z_i \rangle, 0\right)$$

  Then, as we saw earlier,

  $$G(x, \xi) = \begin{cases} -y_i z_i \text{ ,if } y_i \langle x, z_i \rangle < 0 \\ 0 \text{ ,otherwise} \end{cases}$$

  so that for this example,

  $$B = \sup_i \|z_i\|_2 \text{ and } L_G = 0$$

  Note that the case $L_G = 0$ does not cover m-convex functions because if f is m-convex then

  $$\|\nabla f(x)\|_2 \geq \frac{m}{2} \|x - x^\star\|_2$$
  $$\text{and } \|\nabla f(x)\|_2^2 \leq \mathbb{E}_\xi \left[\|G(x, \xi)\|_2^2 \right]$$

- **Case: $L_G > 0$ and $B = 0$**
  Take

$$f(x) = (2n)^{-1} \sum_{i=1}^{n} (\langle a_i, x \rangle - b_i)^2$$

  and assume that there exists $x^\star$ such that $\langle a_i, x^\star \rangle - b_i = 0$ for all $i = 1, \cdots, n$. We have

$$G(x, \xi) = a_\xi a_\xi^{'} (x - x^\star)$$

  and so

$$\begin{aligned} \mathbb{E}_\xi \left[\|G(x,\xi)\|_2^2 \right] &= \mathbb{E}_\xi \left[\|a_\xi\|_2^2 \left(\langle a_\xi, x - x^\star \rangle \right)^2 \right] \\ &\leq \mathbb{E}_\xi \left[\|a_\xi\|_2^4 \right] \|x - x^\star\|_2^2 \\ &= L_G^2 \|x - x^\star\|_2^2 \end{aligned}$$

- **Case: Additive Gaussian Noise**
  Say $G(x, w) = \nabla f(x) + w$ where $w \sim N(0, \sigma^2 I)$. Then

$$\mathbb{E}_w \left[\|G(x,w)\|_2^2 \right] = \|\nabla f(x)\|_2^2 + \sigma^2 d$$

  so that $L_G =$ smoothness constant of f and $B^2 = \sigma^2 d$.

5.1 How to analyze convergence?

Consider the usual SGM update. Now,

$$\begin{aligned} \|x_{t+1} - x^\star\|_2^2 &= \|x_t - \mu_t G(x_t, \xi_t) - x^\star\|_2^2 \\ &= \|x_t - x^\star\|_2^2 - 2\mu_t \langle G(x_t, \xi_t), x_t - x^\star \rangle + \mu_t^2 \|G(x_t, \xi_t)\|_2^2 \\ \text{so that } \mathbb{E}_{\xi_t} \left[\|x_{t+1} - x^\star\|_2^2 \right] &\leq \|x_t - x^\star\|_2^2 - 2\mu_t \langle \nabla f(x_t), x_t - x^\star \rangle + \mu_t^2 L_G^2 \|x - x^\star\|_2^2 + \mu_t^2 B^2 \end{aligned}$$

From here on, one uses properties of ∇f to obtain appropriate bounds on the convergence rate.