Stochastic and Incremental Gradient Methods

Benjamin Recht

Department of Computer Sciences, University of Wisconsin-Madison
1210 W Dayton St, Madison, WI 53706
email: brecht@cs.wisc.edu

November 22, 2010

Abstract

abstract here.

Keywords. none.

1 Problem Set-up

The problem set up is more the less the same as in [1] as are the derivations. We want to minimize a function

$$\text{minimize}_x \, \mathbb{E}_\xi[F(x,\xi)] + P(x) \tag{1.1}$$

but we only get access to subgradients $g(x,\xi)$ of $F(x,\xi)$ with ξ sampled at random. Examples of this set-up include

1. **Noisy gradients.** We want to minimize a smooth function $f(x)$. At every iteration, we compute or gain access to a *noisy* gradient $g_k = \nabla f(x_k) + \omega_k$ where ω_k is some zero-mean noise process which is independent of x_k.

2. **Incremental gradients.** We want to minimize a function of the form

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

   At every iteration, we choose a random index i_k uniformly at random from $\{1,\dots,m\}$, and we take a step along the gradient of f_{i_k} rather than of the full function f. This is obviously faster to compute when m is large. When does this approach find a minimum of f?

Throughout we assume

1. $f(x) := \mathbb{E}_\xi[F(x,\xi)]$ is differentiable and strongly convex. So there exists a constant $\ell > 0$ such that

$$f(z) \geq f(x) + \nabla f(x)^*(z-x) + \frac{\ell}{2}\|z-x\|^2 \,. \tag{1.2}$$

2. ∇f is Lipschitz so that $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$.

3. $P(x)$ is a convex extended real valued function.

Note tlhat the results apply to the case where there is only one value of ξ. That is, the non-stochastic setting. In this case we would have a differentiable convex function plus an arbitrary convex function. Also note that we can enforce the constraint $x \in X$ for some convex set X by letting $P(x) = 0$ for $x \in X$ and $P(x) = \infty$ for $x \notin X$.

Let us define a stochastic projected gradient scheme to solve this problem. Let

$$\mathrm{prox}_{\nu P}(z) = \arg\min_x \|x - z\|^2 + \nu P(x) \tag{1.3}$$

Let $\gamma_0, \ldots, \gamma_T, \ldots,$ be a sequence of positive numbers. Choose $x_0 \in X$, and iterate

$$x_{k+1} = \mathrm{prox}_{\gamma_k P}(x_k - \gamma_k G(x_k, \xi_k))\,. \tag{1.4}$$

2 Analysis of Unconstrained Stochastic Gradient

First, let's examine the case with $P = 0$ and let's make no assumptions about strong convexity. Assume $\|G(x, \xi)\| \le M$ for all x and ξ. Let x_* denote any optimal solution of (1.1). Then we have

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x_*\|^2] &= \mathbb{E}[\|x_k - \gamma_k G(x_k, \xi_k)) - x_*\|^2] && \text{(2.1a)}\\
&= \mathbb{E}[\|x_k - x_*\|^2] - 2\gamma_k\, \mathbb{E}[\langle G(x_k, \xi_k), x_k - x_*\rangle] + \gamma_k^2\, \mathbb{E}[\|G(x_k, \xi_k)\|^2] && \text{(2.1b)}\\
&\le \mathbb{E}[\|x_k - x_*\|^2] - 2\gamma_k\, \mathbb{E}[\langle G(x_k, \xi_k), x_k - x_*\rangle] + \gamma_k^2 M^2 && \text{(2.1c)}\\
&= \mathbb{E}[\|x_k - x_*\|^2] - 2\gamma_k\, \mathbb{E}[\langle \nabla f(x_k), x_k - x_*\rangle] + \gamma_k^2 M^2 && \text{(2.1d)}\\
&\le \mathbb{E}[\|x_k - x_*\|^2] - 2\gamma_k\, \mathbb{E}[f(x_k) - f(x_*)] + \gamma_k^2 M^2 && \text{(2.1e)}
\end{aligned}$$

(2.1d) follows because

$$\begin{aligned}
\mathbb{E}[\langle G(x_k, \xi_k), x_k - x_*\rangle] &= \mathbb{E}_{\xi_0,\ldots,\xi_{k-1}}[\mathbb{E}_{\xi_k}[\langle G(x_k, \xi_k), x_k - x_*\rangle \mid \xi_0, \ldots, \xi_{k-1}] && (2.2)\\
&= \mathbb{E}_{\xi_0,\ldots,\xi_{k-1}}[\langle \nabla f(x_k), x_k - x_*\rangle \mid \xi_0, \ldots, \xi_{k-1}] && (2.3)\\
&= \mathbb{E}[\langle \nabla f(x_k), x_k - x_*\rangle] && (2.4)
\end{aligned}$$

by the law of iterated expectation. (2.1e) is a consequence of the inequality

$$\langle \nabla f(x_k), x_k - x_*\rangle \ge (x_k) - f(x_*) \tag{2.5}$$

which holds because f is convex.

Arranging the bound, we have for any n

$$\frac{1}{\sum_{k=0}^n \gamma_k} \sum_{k=0}^n \gamma_k\, \mathbb{E}[f(x_k)] - f(x_*) \le \frac{D^2 + M^2 \sum_{k=0}^n \gamma_k^2}{2\sum_{k=0}^n \gamma_k} \tag{2.6}$$

where $D = \|x_0 - x_*\|^2$. This bound can be computed by summing the inequalities for $k = 0, \ldots, n$ and then dividing by the sum of the γ_k. Let

$$\bar{x} := \frac{1}{\sum_{k=0}^n \gamma_k} \sum_{k=0}^n \gamma_k x_k \tag{2.7}$$

Then, by convexity, we have

$$\mathbb{E}[f(\bar{x})] - f(x_*) \leq \frac{D^2 + M^2 \sum_{k=0}^n \gamma_k^2}{2 \sum_{k=0}^n \gamma_k} \tag{2.8}$$

This is precisely the bound rate of convergence we have seen for deterministic subgradient descent.

3 Analysis of Projected Stochastic Gradient

Let x_* denote the optimal solution of (1.1). x_* is unique because of strong convexity. Observe that

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x_*\|^2] &= \mathbb{E}[\|\Pi_{\gamma_k}(x_k - \gamma_k G(x_k, \xi_k)) - \Pi_{\gamma_k}(x_* - \gamma_k \nabla f(x_*))\|^2] && \text{(3.1a)}\\
&\leq \mathbb{E}[\|x_k - \gamma_k G(x_k, \xi_k) - x_* + \gamma_k \nabla f(x_*)\|^2] && \text{(3.1b)}\\
&= \mathbb{E}[\|x_k - \gamma_k \nabla f(x_k) + \gamma_k(\nabla f(x_k) - G(x_k, \xi_k)) - x_* + \gamma_k \nabla f(x_*)\|^2] && \text{(3.1c)}\\
&= \mathbb{E}[\|x_k - \gamma_k \nabla f(x_k) - x_* + \gamma_k \nabla f(x_*)\|^2] && \text{(3.1d)}\\
&\qquad + 2\gamma_k \,\mathbb{E}[\langle \nabla f(x_k) - G(x_k, \xi_k), x_k - \gamma_k \nabla f(x_k) - x_* + \gamma_k \nabla f(x_*) \rangle]\\
&\qquad\quad + \gamma_k^2 \,\mathbb{E}[\|\nabla f(x_k) - G(x_k, \xi_k)\|^2]\\
&= \mathbb{E}[\|x_k - \gamma_k \nabla f(x_k) - x_* + \gamma_k \nabla f(x_*)\|^2] + \gamma_k^2 \,\mathbb{E}[\|\nabla f(x_k) - G(x_k, \xi_k)\|^2] && \text{(3.1e)}
\end{aligned}$$

Here, the first equality follows by the definition of x_{k+1} and because x_* is optimal. (3.1b) follows because the proximity operator is non-expansive. (3.1c) follows because $\mathbb{E}[G(z, \xi_k)] = \nabla f(z)$ for all z and $G(z, \xi_k)$ is independent of ξ_k. Thus we have

$$\begin{aligned}
&\mathbb{E}[\langle \nabla f(x_k) - G(x_k, \xi_k), x_k - \gamma_k \nabla f(x_k) - x_* + \gamma_k \nabla f(x_*)\rangle] && \text{(3.2a)}\\
=&\mathbb{E}_{\xi_0,\ldots,\xi_{k-1}}[\mathbb{E}_{\xi_k}[\langle \nabla f(x_k) - G(x_k, \xi_k), x_k - \gamma_k \nabla f(x_k) - x_* + \gamma_k \nabla f(x_*)\rangle | \xi_0, \ldots, \xi_{k-1}]] && \text{(3.2b)}\\
=&0 && \text{(3.2c)}
\end{aligned}$$

Note that the first term in (3.1e) is completely independent of ξ_k while the second term is a variance term concerning the second moments of the subgradients at the current iterate and at the optimum. We can bound each of these terms separately. First, since f is strongly convex and has a Lipschitz continuous gradient, it follows that

$$\mathbb{E}[\|x_k - \gamma_k \nabla f(x_k) - x_* + \gamma_k \nabla f(x_*)\|^2] \leq \max\{|1 - \gamma_k L|, |1 - \gamma_k \ell|\}^2 \,\mathbb{E}[\|x_k - x_*\|^2]\,. \tag{3.3}$$

For the second term, we must make some assumption about the statistics of the random function

$$\varphi(z; \xi_k) := G(z, \xi_k) - \nabla f(z) \tag{3.4}$$

Let's explore some possibilities.

3.1 $\varphi \equiv 0$

In the case when there is no randomness at all and we are just following the gradient, we only need upper bound (3.3). In this case, setting $\gamma_k = \frac{2}{L+\ell}$ for all k, we find that

$$\|x_{k+1} - x_*\| \leq \left(\frac{L - \ell}{L + \ell}\right) \|x_k - x_*\| \tag{3.5}$$

or, letting $\kappa = \frac{L}{\ell}$ and $D_0 = \|x_0 - x_*\|$,

$$\|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k D_0 \tag{3.6}$$

That is, a constant step-size policy converges at a linear rate.

3.2 φ bounded

The simplest non-trivial assumption is that the deviations are bounded:

$$\|\varphi(z; \xi_k)\| \leq M \tag{3.7}$$

for some universal constant M. In this case, we have the upper bound

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \max\{|1 - \gamma_k L|, |1 - \gamma_k \ell|\}^2\, \mathbb{E}[\|x_k - x_*\|^2] + \gamma_k^2 M^2 \tag{3.8a}$$
$$\leq (1 - 2\gamma_k \ell + \gamma_k^2 L^2)\, \mathbb{E}[\|x_k - x_*\|^2] + \gamma_k^2 M^2\,. \tag{3.8b}$$

With such a bound, we can achieve the so-called "optimal" $O(1/k)$ rate by choosing

$$\gamma_k = \frac{1}{k\ell}\,. \tag{3.9}$$

Indeed, in this case, it follows by induction that

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \frac{M^2 + D_0^2 L^2}{k\ell^2} \tag{3.10}$$

where D_0 again equals $\|x_0 - x_*\|$. To verify this inequality, note that for $k = 0$, the right hand is greater than D_0^2. Assuming that the inequality holds for $k \leq K$, observe

$$E[\|x_{K+1} - x_*\|^2] \leq (1 - \frac{2}{K} + \frac{L^2}{K^2\ell^2})\, \mathbb{E}[\|x_K - x_*\|^2] + \frac{M^2}{K^2\ell^2} \tag{3.11a}$$
$$\leq (1 - \frac{2}{K})\, \mathbb{E}[\|x_{K-1} - x_*\|^2] + \frac{M^2 + L^2\, \mathbb{E}[\|x_{K-1} - x_*\|^2]}{K^2\ell^2} \tag{3.11b}$$
$$\leq (1 - \frac{2}{K})\frac{M^2 + D_0^2 L^2}{K\ell^2} + \frac{M^2 + D_0^2 L^2}{K^2\ell^2} \tag{3.11c}$$
$$= \frac{K^2 - 1}{K^2} \cdot \frac{M^2 + D_0^2 L^2}{(K+1)\ell^2} \leq \frac{M^2 + D_0^2 L^2}{(K+1)\ell^2}\,. \tag{3.11d}$$

3.3 φ Lipschitz

If we add additional assumptions about the behavior of φ, we can derive considerably faster convergence. In particular, consider the unconstrained case and suppose φ is Lipschitz in expectation:

$$E[\|\varphi(x; \xi)\|^2] \leq \beta^2 \|x - x_*\|^2 \quad \forall x \tag{3.12}$$

In this case, we have a bound of the form

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \left(\max\{|1 - \gamma_k L|, |1 - \gamma_k \ell|\}^2 + \gamma_k^2 \beta^2\right) \mathbb{E}[\|x_k - x_*\|^2] \tag{3.13}$$

Now we can always select a constant γ that provides a linear convergence rate. Indeed,if $\beta < \sqrt{\ell L}$, then setting $\gamma_k = \frac{2}{\ell+L}$ gives

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{4(\kappa - \beta^2/\ell^2)}{(\kappa+1)^2}\right)^k D_0^2 \tag{3.14}$$

Otherwise, setting $\gamma_k = \frac{\ell}{\ell^2+\beta^2}$, we achieve

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \left(1 + \ell^2/\beta^2\right)^{-k} D_0^2\,. \tag{3.15}$$

References

[1] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.