University of California, Berkeley Lecturer: Benjamin Recht

CS294/STAT260 The Mathematics of Information and Data

September 13, 2013 Scribe: Jonathan Terhorst

Lecture 1

1 Introduction

Today's lecture is about solving the following optimization problem:

$$\min_{x\in\mathscr{X}} E_\xi\left[f(x,\xi)\right] \equiv \min_{x\in\mathscr{X}} \mathcal{F}(x),$$

where $\mathscr{X}$ is a (relatively) simple set and ξ is some well-behaved random variable. Some well-known instances of this problem are:

- Training an SVM;
- Lasso;
- Matrix-completion;
- M-estimation (of which above three are special cases);
- Graph problems:
  - Matching
  - Cuts
  - Flows

Remark (Ben). Is "big data" just M-estimation?

Remark (Tamara). Is variational Bayes an M-estimation problem? Variational Bayes objective:

$$\max_{\theta\in\Theta} E_{q(\theta)}\left[\log\frac{p(x,\theta)}{q(\theta)}\right].$$

Here the setup is a bit different in that the density $q(\theta)$ depends on the variable being optimized.

Goal of today: identify a general algorithm to solve (1).

2 Sample average approximation

A natural algorithm is:

1. Sample i.i.d. $\xi_1,\ldots,\xi_n \sim p$ for some density p. (Assume that $p(\xi)$ is *not* a function of $x\in\mathscr{X}$.)
2. Solve

$$\hat{x} \in \arg\min_{x\in\mathscr{X}} \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i). \tag{1}$$

This is known as *sample average approximation.*

Example 1. $\xi_i \sim \mathcal{N}(0,1)$, $f(x,\xi) = (x-\xi)^2$. The optimum is $x^* = 0$, and it is easy to see that $\hat{x} = \frac{1}{n}\sum_{i=1}^n \xi_i$, the sample average. So the SAA converges to the truth at rate $\mathcal{O}(n^{-1})$.

What are good error metrics to use?

1. Objective distance: $\|\mathcal{F}(\hat{x}) - \mathcal{F}(x^*)\|$
2. Parameter distance: $\|\hat{x} - x^*\|$.
3. Some other arbitrary function $\mathrm{err}(\hat{x}, x^*)$.

All three of these are functions of the sampled data, so you have to, for example, look at them in expectation or bound large deviations.

Note that the optimization (1) introduces another source of error. We actually obtain $\tilde{x}$, where $\mathcal{F}(\tilde{x}) \approx \max \mathcal{F}(x)$. So our actual error "decomposes" as

$$\text{error} = \underbrace{\|\tilde{x} - \hat{x}\|}_{\text{optimization error}} + \underbrace{\|\hat{x} - x^*\|}_{\text{statistical error}}$$

In instructor's experience, error matters in practice. This is why you randomly re-order the samples and re-run the algorithm. Example: matrix completion can require 10-20 passes over the data. The actual data does not meet the assumptions of the analysis; in particular the data are not i.i.d.

3 Stochastic Gradient

The basic idea:

$$x_{k+1} \equiv x_k - \alpha_k \nabla f(x_k, \xi_k).$$

Idea is the same as gradient descent—follow ∇f downhill—only now you have a noisy idea of what downhill is.

Exercise 2. Let $\alpha_k = \frac{1}{k+1}$, $x_1 = 0$, $f(x,\xi) = \frac{1}{2}(x-\xi)^2$. Show that SGD chooses $x_k = \frac{1}{k}\sum_{i=1}^k \xi_i$.

3.1 Adaptive filtering

In *adaptive filtering*, $\alpha_k = \text{const}$. The Kurtz et al. paper says that in this case, $x_{k+1} = x_k + \alpha g(x_k, \xi)$ looks like a finite difference approximation to an ODE. So they integrate the ODE and solve; this is known as the ODE method. Step sizes satisfy $\sum \alpha_k \to \infty$ but $\alpha_k \to 0$. The analysis doesn't require any convexity assumptions on f.

3.2 Robust stochastic approximation

See Nemirovski, Juditsky, Lan, and Shapiro (2009). Assumptions:

- $\mathcal{F}$ is strongly convex:

$$\mathcal{F}(z) \geq \mathcal{F}(x) + \nabla\mathcal{F}(x)^T(z-x) + \frac{\ell}{2}\|x-z\|_2^2$$

  for some $\ell > 0$.

- Gradients are Lipschitz:

$$\|\nabla\mathcal{F}(z) - \nabla\mathcal{F}(x)\| \leq L\|z-x\|.$$

  (Equivalent statement for subgradients is that they are bounded.)

- $E\left[\|\nabla f(x,\xi)\|^2\right] \leq M^2$ for some M.

- $\nabla\mathcal{F}(x^*) = 0$, i.e. $x^* \in \mathscr{X}^o$.

Proof sketch: let

$$x_{k+1} = \Pi_{\mathscr{X}}\left[x_k - \alpha_k \nabla f(x_k, \xi_k)\right]$$

where $\Pi_{\mathscr{X}}(y)$ projects the point y into $\mathscr{X}$. This gives:

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|\Pi_{\mathscr{X}}(x_k - \alpha_k\nabla f(x_k,\xi_k) - \Pi_{\mathscr{X}}(x^*)\|^2 \\
&\leq \|x_k - \alpha_k\nabla f(x_k,\xi_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - 2\alpha_k\langle x_k - x^*, \nabla f(x_k,\xi_k)\rangle + |\alpha_k|^2\|\nabla f(x_k,\xi_k)\|^2
\end{aligned}$$

where the inequality follows since $\Pi_{\mathscr{X}}$ is contractive. Taking expectations, we get

$$\begin{aligned}
a_{k+1} &\leq a_k - \alpha_k E\langle x_k - x^*, \nabla f(x_k,\xi_k)\rangle + |\alpha_k|^2\frac{M^2}{2} \\
&= a_k - \alpha_k E\langle x_k - x^*, \nabla\mathcal{F}(x_k)\rangle + |\alpha_k|^2\frac{M^2}{2}.
\end{aligned}$$

where we have defined $a_k \equiv E\left[\|x_{k+1} - x^*\|^2\right]$. The equality of the middle terms follows because $x_k \in \sigma(\xi_1, \ldots, \xi_{k-1})$:

$$\begin{aligned}
E\langle x_k - x^*, \nabla f(x_k,\xi_k)\rangle &= E\left[E_{\xi_k}\left(\langle x_k - x^*, \nabla f(x_k,\xi_k)\rangle \mid \xi_1,\ldots,\xi_{k-1}\right)\right] \\
&= E\langle x_k - x^*, \nabla\mathcal{F}(x_k)\rangle.
\end{aligned}$$

To finish the proof note that, by applying the first and last assumptions,

$$\begin{aligned}
E\left[\langle\nabla\mathcal{F}(x_k), x_k - x^*\rangle\right] &\geq E\left[\mathcal{F}(x_k) - \mathcal{F}(x^*) + \frac{\ell}{2}\|x_k - x^*\|^2\right] \\
&\geq \ell\|x_k - x^*\|^2.
\end{aligned}$$

Putting it all together,

$$a_{k+1} \leq (1 - 2\ell\alpha_k)a_k + \frac{1}{2}\alpha_k^2 M^2.$$

With $\alpha_k = 1/k$ we get $a_k \to 0$. With $\alpha_k = C$ we get $a_k \to \frac{\alpha M^2}{4\ell}$: the expected mean squared error does not go to zero. With slowly changing but constant step sizes you can get faster convergence while still letting $a_k \to 0$.

Now let $\bar{x}_k = \sum_{i=1}^k x_i$. We have

$$\begin{aligned} E\mathcal{F}(\bar{x}_k) - \mathcal{F}(x^*) &\leq E\left[\frac{1}{k}\sum_{i=1}^k \mathcal{F}(x_i)\right] - \mathcal{F}(x^*) \\ &\leq E\left[\frac{1}{k}\sum_{i=1}^k \langle \mathcal{F}(x_i), x^* - x_i\rangle\right] \\ &= \frac{a_0 - a_k}{K\alpha} + \frac{1}{2}\alpha M^2 \\ &\leq \frac{\|x_0 - x^*\|^2}{K\alpha} + \frac{1}{2}\alpha M^2 \end{aligned}$$

(see dual averaging paper by Nestorov).

4 Kaczmarz Algorithm

We now assume $f(x,\xi) = a_i^T x - b_i$. This implies that

$$\|\nabla f(x,\xi) - \nabla f(x^*,\xi)\| \leq M\|x - x^*\|.$$

In the analysis of the previous section we now obtain the upper bound $a_{k+1} \leq (1 - 2\ell\alpha + \alpha^2 M^2)a_k$, so you get the exponential (aka linear) rate of convergince.

5 Research questions

Some potential research questions that came up during lecture:

1. Alternative methods of estimating error. For example, can you use bootstrap to estimate large deviations of the error metric? See Rakhlin, Karthik, Hazan in COLT.

2. Replace bounds in the proof of robust stochastic approximation with conditions on the correlation matrix of the ξ_i.