ECE 901: Large-scale Machine Learning and Optimization **Fall 2016**

Lecture 8 — 09/29

Lecturer: Dimitris Papailiopoulos *Scribe: Sijing Li*

8.1 Random Coordinate Descent

8.1.1 Coordinate Descent

Coordinate descent (CD) methods are among the oldest in the optimization literature. As their name suggests, they operate by taking steps along coordinate directions: one attempts to minimize the objective with respect to a single variable while all others are kept fixed, then other variables are updated similarly in an iterative manner.

The CD method for minimizing $f : \mathbb{R}^d \to \mathbb{R}$ is given by the iteration

$$x_{k+1} \leftarrow x_k - \gamma_k \nabla_{S_k} f(x_k) e_{S_k}$$

where $\nabla_{S_k} f(x_k) := \frac{\partial f}{\partial x^{S_k}}(x_k)$.

Also, x^{S_k} represents the S_k-th element of the parameter vector, and e_{S_k} represents the S_k-th coordinate vector for some $S_k \in \{1, \ldots, d\}$. In other words, the solution estimates x_{k+1} and x_k differ only in their S_k-th element as a result of a move in the S_k-th coordinate from x_k.

Concerning the choice of S_k, one could select it in each iteration in at least three different ways: by cycling through $\{1, \ldots, d\}$; by cycling through a random reordering of these indices (with the indices reordered after each set of d steps); or simply by choosing an index randomly with replacement in each iteration. Randomized Coordinate Descent (RCD) algorithms (represented by the latter two strategies) have superior theoretical properties than the cyclic method (represented by the first strategy) as they are less likely to choose an unfortunate series of coordinates.

In the following we denote $\nabla_i f(x) = \frac{\partial f}{\partial x_i}(x)$. Random Coordinate Descent (RCD) is defined as follows, with an arbitrary initial point $x_1 \in \mathbb{R}^n$,

$$x_{k+1} = x_k - \gamma [\nabla f(x_k)]_{S_k} e_{S_k},$$

where $S_k \sim unif\{1, \ldots, d\}$, and $e_{S_k} = [0 \ldots 010 \ldots 0]^T$.

8.1.2 Random Coordinate Descent (RCD) for Coordinate-Smooth Optimization

We call f is β_i-coordinate-wise smoothness if there exists $\beta_1, \ldots, \beta_n$ such that for any i, $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

$$|\nabla_i f(x + \alpha e_i) - \nabla_i f(x)| \leq \beta_i |\alpha|.$$

Then, based on this definition, we can have the following property,

$$f(x + \alpha e_i) \leq f(x) + \alpha [\nabla f(x)]_i + \frac{\beta_i \alpha^2}{2}.$$

8.1.3 Random Coordinate Descent (RCD) for Smooth and Strong Convex Optimization

If in addition to directional smoothness, one also assumes f is λ-strong convex, then RCD attains in fact a linear rate.

First, we introduce the following lemma,

Lemma 8.1. *Let f be λ-strong convex w.r.t. $||\cdot||$ on $\mathbb{R}^n$, then*

$$f(x) - f(y) \leq \frac{1}{\lambda} ||\nabla f(x)||^2.$$

This lemma can be derived using the property of strong convexity. Then, we have the following theorem and can prove it by using this lemma,

Theorem 8.2. *Let $\gamma > 0$. Suppose that the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is continuous differentiable, strong convex with constant $\lambda > 0$. For all $k \in \mathbb{N}$, the iteration $x_{k+1} = x_k - \gamma [\nabla f(x_k)]_{S_k} e_{S_k}$ yields*

$$\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - \frac{\lambda}{2d\beta_{max}})^k (f_0 - f^*)$$

where $\beta_{max} = \max_i \beta_i$.

Proof: *Let $g_k = [\nabla f(x_k)]_{S_k} e_{S_k}$, then we can write $x_{k+1} = x_k - \gamma g_k$. Then, by using the lemma, we have*

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma ||g_k||^2 + \frac{\gamma^2 \beta_{S_k}}{2} ||g_k||^2 \\ &= f(x_k) - \gamma (1 - \frac{\gamma \beta_{S_k}}{2}) ||g_k||^2 \end{aligned}$$

where $\beta_{max} = \max_i \beta_i$ *and* $||g_k||^2 = |[\nabla f(x_k)]_{S_k}|^2$.

The expectation of g_k *is*

$$\mathbb{E}(g_k) = \frac{1}{d}\nabla f(x_k),$$

and the expectation of $\mathbb{E}||g_k||^2$ *is*

$$\begin{aligned}\mathbb{E}||g_k||^2 &= \sum_{i=1}^{d} \frac{1}{d}|[\nabla f(x_k)]_i|^2 \\ &= \frac{1}{d}||\nabla f(x)||^2.\end{aligned}$$

Then, we have

$$\begin{aligned}\mathbb{E}_{S_1,\ldots,S_k}(f(x_{k+1})) &\leq \mathbb{E}_{S_1,\ldots,S_k}(f(x_k)) - \frac{\lambda\gamma}{d}(1 - \frac{\gamma\beta_{max}}{2})(\mathbb{E}_{S_1,\ldots,S_k}(f_k) - f^*) \\ \mathbb{E}_{S_1,\ldots,S_k}(f(x_{k+1})) - f^* &\leq \mathbb{E}_{S_1,\ldots,S_k}(f(x_k)) - f^* - \frac{\lambda\gamma}{d}(1 - \frac{\gamma\beta_{max}}{2})(\mathbb{E}_{S_1,\ldots,S_k}(f(x_k)) - f^*) \\ &\leq [1 - \frac{\lambda\gamma}{d}(1 - \frac{\gamma\beta_{max}}{2})](\mathbb{E}_{S_1,\ldots,S_k}(f(x_k)) - f^*) \\ &\leq [1 - p(\gamma)](\mathbb{E}_{S_1,\ldots,S_k}(f(x_k)) - f^*)\end{aligned}$$

We want to find the appropriate γ *to let the step size* $1 - p(\gamma)$ *to be bounded by 1, which gives us* $\gamma \leq \frac{1}{\beta_{max}}$. *If we plug in* $\gamma = \frac{1}{\beta_{max}}$ *into the equation, we can get*

$$\begin{aligned}\mathbb{E}_{S_1,\ldots,S_k}(f(x_{k+1})) - f^* &\leq [1 - \frac{\lambda}{2d\beta_{max}}](\mathbb{E}_{S_1,\ldots,S_k}(f(x_k)) - f^*) \\ &\leq [1 - \frac{\lambda}{2d\beta_{max}}]^k(f_0 - f^*)\end{aligned}$$

□

8.1.4 Uniform Sampling vs. Importance Sampling

When choosing the coordinate, we can have different methods.

For uniform sampling from $\{1, \ldots, d\}$, we have the complexity as

$$T_\epsilon^{unif} = \mathcal{O}(\frac{d\beta_{max}}{\lambda}\log(\frac{f_0 - f^*}{\epsilon})),$$

where $d\beta_{max} = d||\beta||_\infty$.

Instead of uniform sampling, if we assign the weights for i-th coordinate based on smoothness as $P(S_k = i) = \frac{\beta_i}{\sum_{j=1}^{n} \beta_j}$, which we call it importance sampling. Then we have the complexity as

$$T_\epsilon^{importance} = \mathcal{O}(\frac{||\beta||_1}{\lambda} \log(\frac{f_0 - f^*}{\epsilon}))$$

Here we have $||\beta||_1 \leq d||\beta||_\infty$.

8.1.5 Random Coordinate Descent (RCD) for Least Gradient

Least Squares problem has the form

$$f(x) = ||Ax - b||^2$$

We want to $\min_X f(x) = \min_X ||Ax - b||^2$, which we can write in the matrix notation as,

$$\nabla f(x) = 2A^T Ax - 2A^T(Ax - b)$$

By what we learned from above, we hope to have the form of $|\nabla_i f(x + \alpha e_i) - \nabla_i f(x)| \leq \beta_i |\alpha|$. Then, we write

$$|2[A]_{\cdot,i}^T A(x - x^{(i)})| \leq 2||A_{\cdot,i}|| \cdot ||A(x - x^{(i)})||$$

where $x^{(i)} = x + \alpha e_i$. We have no way to bound $||A_{\cdot,i}||$, but we can bound the right-hand side as

$$2||A_{\cdot,i}|| \cdot ||A(\alpha \times e_i)|| \leq 2||A_{\cdot,i}||^2 |\alpha|,$$

We know that β_i should be more tighter and bounded as $\beta_i \leq 2||A_{\cdot,i}||^2$, and in some cases, we can let $\beta_i = 2||A_{\cdot,i}||^2$.

Consider the two different sampling methods,

$$d||\beta||_\infty = 2 \max_i ||A_{\cdot,i}||^2$$
$$||\beta||_1 = 2 \sum_i ||A_{\cdot,i}||^2 = 2||A||_F^2$$

which $||\cdot||_F^2$ represent the Fubini's norm.

If we let $A = [1_{n-1}|I_{n-1}]_{n \times n}^{n=d}$, then in this case we will have

$$||A||_F^2 = \mathcal{O}(d),$$
$$d||A_{\cdot,i}||^2 = \mathcal{O}(d^2).$$

So we can see that the importance sampling is good for this case when the support is not uniform.