

# Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: [hardt+ee227c@berkeley.edu](mailto:hardt+ee227c@berkeley.edu)

Graduate Instructor: Max Simchowitz

Email: [msimchow+ee227c@berkeley.edu](mailto:msimchow+ee227c@berkeley.edu)

March 6, 2018

## 12 Lecture 12: Coordinate Descent

### 12.1 Why Coordinate Descent?

There are many classes of functions for which it is very cheap to compute directional derivatives along the standard basis vectors  $e_i, i \in [n]$ . For example,

$$f(x) = \|x\|^2 \text{ or } f(x) = \|x\|_1 \quad (1)$$

This is especially true of common regularizers, which often take the form

$$R(x) = \sum_{i=1}^n R_i(x_i) . \quad (2)$$

More generally, many objectives and regularizers exhibit “group sparsity”; that is,

$$R(x) = \sum_{j=1}^m R_j(x_{S_j}) \quad (3)$$

where each  $S_j, j \in [m]$  is a subset of  $[n]$ , and similarly for  $f(x)$ . Examples of functions with block decompositions and group sparsity include:

1. Group sparsity penalties;

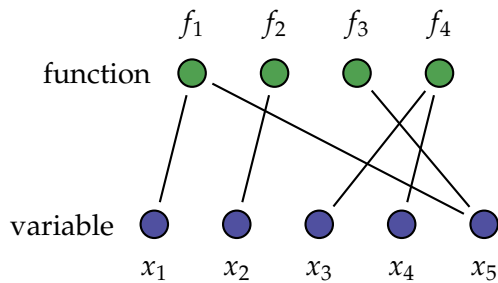


Figure 1: Example of the bipartite graph between component functions  $f_i, i \in [m]$  and variables  $x_j, j \in [n]$  induced by the group sparsity structure of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ . An edge between  $f_i$  and  $x_j$  conveys that the  $i$ th component function depends on the  $j$ th coordinate of the input.

2. Regularizes of the form  $R(U^\top x)$ , where  $R$  is coordinate-separable, and  $U$  has sparse columns and so  $(U^\top x) = u_i^\top x$  depends only on the nonzero entries of  $U_i$ ;
3. Neural networks, where the gradients with respect to some weights can be computed “locally”; and
4. ERM problems of the form

$$f(x) := \sum_{i=1}^n \phi_i(\langle w^{(i)}, x \rangle) \quad (4)$$

where  $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ , and  $w^{(i)}$  is zero except in a few coordinates.

## 12.2 Coordinate Descent

Denote  $\partial_i f = \frac{\partial f}{\partial x_i}$ . For each round  $t = 1, 2, \dots$ , the coordinate descent algorithm chooses an index  $i_t \in [n]$ , and computes

$$x_{t+1} = x_t - \eta_t \partial_{i_t} f(x_t) \cdot e_{i_t}. \quad (5)$$

This algorithm is a special case of stochastic gradient descent. For

$$\mathbb{E}[x_{t+1} | x_t] = x_t - \eta_t \mathbb{E}[\partial_{i_t} f(x_t) \cdot e_{i_t}] \quad (6)$$

$$= x_t - \frac{\eta_t}{n} \sum_{i=1}^n \partial_i f(x_t) \cdot e_i \quad (7)$$

$$= x_t - \eta_t \nabla f(x_t). \quad (8)$$

Recall the bound for SGD: If  $\mathbb{E}[g_t] = \nabla f(x_t)$ , then SGD with step size  $\eta = \frac{1}{BR}$  satisfies

$$\mathbb{E}[f(\frac{1}{T} \sum_{t=1}^T x_t)] - \min_{x \in \Omega} f(x) \leq \frac{2BR}{\sqrt{T}} \quad (9)$$

where  $R^2$  is given by  $\max_{x \in \Omega} \|x - x_1\|_2^2$  and  $B = \max_t \mathbb{E}[\|g_t\|_2^2]$ . In particular, if we set  $g_t = n \partial_{x_{i_t}} f(x_t) \cdot e_{i_t}$ , we compute that

$$\mathbb{E}[\|g_t\|_2^2] = \frac{1}{n} \sum_{i=1}^n \|n \cdot \partial_{x_i} f(x_t) \cdot e_i\|_2^2 = n \|\nabla f(x_t)\|_2^2. \quad (10)$$

If we assume that  $f$  is  $L$ -Lipschitz, we additionally have that  $\mathbb{E}[\|g_t\|^2] \leq nL^2$ . This implies the first result:

**Proposition 12.1.** *Let  $f$  be convex and  $L$ -Lipschitz on  $\mathbb{R}^n$ . Then coordinate descent with step size  $\eta = \frac{1}{nR}$  has convergence rate*

$$\mathbb{E}[f(\frac{1}{T} \sum_{t=1}^T x_t)] - \min_{x \in \Omega} f(x) \leq 2LR\sqrt{n/T} \quad (11)$$

### 12.3 Importance Sampling

In the above, we decided on using the uniform distribution to sample a coordinate. But suppose we have more fine-grained information. In particular, what if we knew that we could bound  $\sup_{x \in \Omega} \|\nabla f(x)_i\|_2 \leq L_i$ ? An alternative might be to sample in a way to take  $L_i$  into account. This motivates the ‘‘importance sampled’’ estimator of  $\nabla f(x)$ , given by

$$g_t = \frac{1}{p_{i_t}} \cdot \partial_{i_t} f(x_t) \text{ where } i_t \sim \text{Cat}(p_1, \dots, p_n). \quad (12)$$

Note then that  $\mathbb{E}[g_t] = \nabla f(x_t)$ , but

$$\mathbb{E}[\|g_t\|_2^2] = \sum_{i=1}^n (\partial_{i_t} f(x_t))^2 / p_i^2 \quad (13)$$

$$\leq \sum_{i=1}^n L_i^2 / p_i^2 \quad (14)$$

In this case, we can get rates

$$\mathbb{E}[f(\frac{1}{T} \sum_{t=1}^T x_t)] - \min_{x \in \Omega} f(x) \leq 2R\sqrt{1/T} \cdot \sqrt{\sum_{i=1}^n L_i^2 / p_i^2} \quad (15)$$

In many cases, if the values of  $L_i$  are heterogenous, we can optimize the values of  $p_i$ .

## 12.4 Importance Sampling For Smooth Coordinate Descent

In this section, we consider coordinate descent with a *biased* estimator of the gradient. Suppose that we have, for  $x \in \mathbb{R}^n$  and  $\alpha \in \mathbb{R}$ , the inequality

$$|\partial_{x_i} f(x) - \partial_{x_i} f(x + \alpha e_i)| \leq \beta_i |\alpha| \quad (16)$$

where  $\beta_i$  are possibly heterogenous. Note that if that  $f$  is twice-continuously differentiable, the above condition is equivalent to  $\nabla_{ii}^2 f(x) \leq \beta_i$ , or  $\text{Diag}(\nabla^2 f(x)) \preceq \text{diag}(\beta)$ . Define the distribution  $p^\gamma$  via

$$p_i^\gamma = \frac{\beta_i^\gamma}{\sum_{j=1}^n \beta_j^\gamma} \quad (17)$$

We consider gradient descent with the rule called RCD( $\gamma$ )

$$x_{t+1} = x_t - \frac{1}{\beta_{i_t}} \cdot \partial_{i_t}(x_t) \cdot e_{i_t}, \text{ where } i_t \sim p^\gamma \quad (18)$$

Note that as  $\gamma \rightarrow \infty$ , coordinates with larger values of  $\beta_i$  will be selected more often. Also note that this is *not generally* equivalent to SGD, because

$$\mathbb{E} \left[ \frac{1}{\beta_{i_t}} \partial_{i_t}(x_t) e_i \right] = \frac{1}{\sum_{j=1}^n \beta_j^\gamma} \cdot \sum_{i=1}^n \beta_i^{\gamma-1} \partial_i f(x_t) e_i = \frac{1}{\sum_{j=1}^n \beta_j^\gamma} \cdot \nabla f(x_t) \circ (\beta_i^{\gamma-1})_{i \in [n]} \quad (19)$$

which is only a scaled version of  $\nabla f(x_t)$  when  $\gamma = 1$ . Still, we can prove the following theorem:

**Theorem 12.2.** *Define the weighted norms*

$$\|x\|_{[\gamma]}^2 := \sum_{i=1}^n x_i^2 \beta_i^\gamma \text{ and } \|x\|_{[\gamma]}^{*2} := \sum_{i=1}^n x_i^2 \beta_i^{-\gamma} \quad (20)$$

*and note that the norms are dual to one another. We then have that the rule RCD( $\gamma$ ) produces iterates satisfying*

$$\mathbb{E}[f(x_t) - \arg \min_{x \in \mathbb{R}^n} f(x)] \leq \frac{2R_{1-\gamma}^2 \cdot \sum_{i=1}^n \beta_i^\gamma}{t-1}, \quad (21)$$

where  $R_{1-\gamma}^2 = \sup_{x \in \mathbb{R}^n: f(x) \leq f(x_1)} \|x - x^*\|_{[1-\gamma]}$ .

*Proof.* Recall the inequality that for a general  $\beta_g$ -smooth convex function  $g$ , one has that

$$g \left( u - \frac{1}{\beta_g} \nabla g(u) \right) - g(u) \leq -\frac{1}{2\beta_g} \|\nabla g\|^2 \quad (22)$$

Hence, considering the functions  $g_i(u; x) = f(x + ue_i)$ , we see that  $\partial_i f(x) = g'_i(u; x)$ , and thus  $g_i$  is  $\beta_i$  smooth. Hence, we have

$$f\left(x - \frac{1}{\beta_i} \nabla f(x) e_i\right) - f(x) = g_i\left(0 - \frac{1}{\beta_g} g'_i(0; x)\right) - g(0; x) \leq -\frac{g'_i(u; x)^2}{2\beta_i} = -\frac{\partial_i f(x)^2}{2\beta_i}. \quad (23)$$

Hence, if  $i \sim p^\gamma$ , we have

$$\mathbb{E}[f(x - \frac{1}{\beta_i} \partial_i f(x) e_i) - f(x)] \leq \sum_{i=1}^n p_i^\gamma \cdot -\frac{\partial_i f(x)^2}{2\beta_i} \quad (24)$$

$$= -\frac{1}{2 \sum_{i=1}^n \beta_i^\gamma} \sum_{i=1}^n \beta_i^{\gamma-1} \partial_i f(x)^2 \quad (25)$$

$$= -\frac{\|\nabla f(x)\|_{[1-\gamma]}^{*2}}{2 \sum_{i=1}^n \beta_i^\gamma} \quad (26)$$

Hence, if we define  $\delta_t = \mathbb{E}[f(x_t) - f(x^*)]$ , we have that

$$\delta_{t+1} - \delta_t \leq -\frac{\|\nabla f(x_t)\|_{[1-\gamma]}^{*2}}{2 \sum_{i=1}^n \beta_i^\gamma} \quad (27)$$

Moreover, with probability 1, one also has that  $f(x_{t+1}) \leq f(x_t)$ , by the above. We now continue with the regular proof of smooth gradient descent. Note that

$$\begin{aligned} \delta_t &\leq \nabla f(x_t)^\top (x_t - x_*) \\ &\leq \|\nabla f(x_t)\|_{[1-\gamma]}^* \|x_t - x_*\|_{[1-\gamma]} \\ &\leq R_{1-\gamma} \|\nabla f(x_t)\|_{[1-\gamma]}^*. \end{aligned}$$

Putting these things together implies that

$$\delta_{t+1} - \delta_t \leq -\frac{\delta_t^2}{2R_{1-\gamma}^2 \sum_{i=1}^n \beta_i^\gamma} \quad (28)$$

Recall that this was the recursion we used to prove convergence in the non-stochastic case. ■

**Theorem 12.3.** *If  $f$  is in addition  $\alpha$ -strongly convex w.r.t to  $\|\cdot\|_{[1-\gamma]}$ , then we get*

$$\mathbb{E}[f(x_{t+1}) - \arg \min_{x \in \mathbb{R}^n} f(x)] \leq \left(1 - \frac{\alpha}{\sum_{i=1}^n \beta_i^\gamma}\right)^t (f(x_1) - f(x^*)). \quad (29)$$

*Proof.* We need the following lemma:

**Lemma 12.4.** *Let  $f$  be an  $\alpha$ -strongly convex function w.r.t to a norm  $\|\cdot\|$ . Then,  $f(x) - f(x^*) \leq \frac{1}{2\alpha} \|\nabla f(x)\|_*^2$ .*

*Proof.*

$$\begin{aligned}
f(x) - f(y) &\leq \nabla f(x)^\top (x - y) - \frac{\alpha}{2} \|x - y\|_2^2 \\
&\leq \|\nabla f(x)\|_* \|x - y\|^2 - \frac{\alpha}{2} \|x - y\|_2^2 \\
&\leq \max_t \|\nabla f(x)\|_* t - \frac{\alpha}{2} t^2 \\
&= \frac{1}{2\alpha} \|\nabla f(x)\|_*^2.
\end{aligned}$$

■

Lemma 12.4 shows that

$$\|\nabla f(x_s)\|_{[1-\gamma]}^{*2} \geq 2\alpha\delta_s.$$

On the other hand, Theorem 12.2 showed that

$$\delta_{t+1} - \delta_t \leq -\frac{\|\nabla f(x_t)\|_{[1-\gamma]}^{*2}}{2\sum_{i=1}^n \beta_i^\gamma} \quad (30)$$

Combining these two, we get

$$\delta_{t+1} - \delta_t \leq -\frac{\alpha\delta_t}{\sum_{i=1}^n \beta_i^\gamma} \quad (31)$$

$$\delta_{t+1} \leq \delta_t \left(1 - \frac{\alpha}{\sum_{i=1}^n \beta_i^\gamma}\right). \quad (32)$$

Applying the above inequality recursively and recalling that  $\delta_t = \mathbb{E}[f(x_t) - f(x^*)]$  gives the result.

■

## 12.5 Random Coordinate vs. Stochastic Gradient Descent

What's surprising is that  $\text{RCD}(\gamma)$  is a descent method, despite being random. This is not true of normal SGD. But when does  $\text{RCD}(\gamma)$  actually do better? If  $\gamma = 1$ , the savings are proportional to the ratio of  $\sum_{i=1} \beta_i / \beta \cdot (T_{\text{coord}} / T_{\text{grad}})$ . When  $f$  is twice differentiable, this is the ratio of

$$\frac{\text{tr}(\max_x \nabla^2 f(x))}{\|\max_x \nabla^2 f(x)\|_{\text{op}}} (T_{\text{coord}} / T_{\text{grad}}) \quad (33)$$

## 12.6 Other Extensions to Coordinate Descent

1. Non-Stochastic, Cyclic SGD
2. Sampling with Replacement
3. Strongly Convex + Smooth!?
4. Strongly Convex (generalize SGD)
5. Acceleration? See [TVW<sup>+</sup>17]

## References

[TVW<sup>+</sup>17] Stephen Tu, Shivaram Venkataraman, Ashia C Wilson, Alex Gittens, Michael I Jordan, and Benjamin Recht. Breaking locality accelerates block gauss-seidel. In *Proc. 34th ICML, 2017*.