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

Instructor: Moritz Hardt
Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz
Email: msimchow+ee227c@berkeley.edu

March 5, 2018

Abstract

This course explores some theory and algorithms for nonlinear optimization. We will focus on problems that arise in machine learning and modern data analysis, paying attention to concerns about complexity, robustness, and implementation in these domains. We will also see how tools from convex optimization can help tackle non-convex optimization problems common in practice.

Code examples are available at:

https://ee227c.github.io/.

Below are the course notes for EE227C (Spring 2018): Convex Optimization and Approximation, taught at UC Berkeley.

Contents

1 Lecture 12: Coordinate Descent

1.1 Why Coordinate Descent?

Many classes of functions where it is very cheap to compute directional derivatives along the directions $e_i, i \in [n]$. For example,

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

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

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

More generally, many objectives and reguarlizes have "group sparsity", that is,

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

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

1. Group sparsity penalties
2. Regularizes of the form $R(U^\top x)$, where R is coordinate-separable, and U has sparse columns (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".
4. ERM problems of the form

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

   where ϕ_i is a 1-d, and $w^{(i)}$ is non-zero except in a few coordinates.

(Draw Function-to-Coordinate Graph)

2 Coordinate Descent

For each round $t = 1, 2, \ldots$, choose an index i_t, and compute

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

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

$$\mathbb{E}[f(\frac{1}{T}\sum_{t=1}^{T} x_t)] - \min_{x \in \Omega} f(x) \leqslant \frac{2BR}{\sqrt{t}} \tag{6}$$

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

$$\mathbb{E}[\|g_t\|^2] = \sum_{i=1}^{n} \frac{1}{n} \cdot (n \cdot \partial_{x_i} f(x_t))^2 = n \cdot \|\nabla f(x_t)\|_2^2 \tag{7}$$

In particular if we assume that f is L Lipschitz, we have that $\mathbb{E}[\|g_t\|^2] \leqslant nL^2$. This implies the first result:

Proposition 2.1. *Coordinate descent with step size $\frac{1}{nR}$ has a convergence rate*

$$\mathbb{E}[f(\frac{1}{T}\sum_{t=1}^{T} x_t)] - \min_{x\in\Omega} f(x) \leqslant 2LR\sqrt{n/t} \tag{8}$$

2.1 Importance Sampling

In the above, we decided on using the uniform distribution. But suppose we have more fine-grained information. In particular, what if we knew that we could bound $\sup_{x\in\Omega} |(\nabla f(x))_i| \leqslant L_i$? 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

$$g_t = \frac{1}{p_{i_t}} \partial_{i_t} f(x_t) \text{ where } i_t \sim \mathrm{Cat}(p_1, \ldots, p_n) \tag{9}$$

Note then that $\mathrm{Exp}[g_t] = \nabla f_t$, but

$$\begin{aligned} \mathrm{Exp}[\|g_t\|^2] &= \sum_{i=1}^{n} (\partial_{i_t} f(x_t))^2 / p_i^2 && (10) \\ &\leqslant \sum_{i=1}^{n} L_i^2 / p_i^2 && (11) \end{aligned}$$

In this case, we can get rates

$$\mathrm{Exp}[f(\frac{1}{T}\sum_{t=1}^{T} x_t)] - \min_{x\in\Omega} f(x) \leqslant 2R\sqrt{1/T} \cdot \sqrt{\sum_{i=1}^{n} L_i^2 / p_i^2} \tag{12}$$

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

2.2 Importance Sampling For Smooth Coordinate Descent

In this section, we consider coordinate descent with an *biased* estimator of the gradient. Suppose that we have the inequality

$$|\partial_{x_i} f(x) - \partial_{x_i} f(x + \alpha e_i)| \leqslant \beta_i |\alpha| \tag{13}$$

where β_i are possibly heterogenous. Note if that f is twice-continuously differentiable, the above condition is equivalently to $\nabla_{ii}^2(f) \leqslant \beta_{ii}$, or that $\mathrm{Diag}(\nabla^2 f) \preceq \mathrm{diag}(\boldsymbol{\beta})$. Define the distribution p^γ via

$$p_i^\gamma = \frac{\beta_i^\gamma}{\sum_{j=1}^{n} \beta_j^\gamma} \tag{14}$$

We consider gradient descent with the rule called RCD(γ)

$$x_{t+1} = x_t - \frac{1}{\beta_{i_t}} \partial_{i_t}(x_t) e_{i_t}, \text{ where } i_t \sim p^\gamma \tag{15}$$

Note that this is *not generally* equivalent to SGD, because

$$\text{Exp}[\frac{1}{\beta_{i_t}} \partial_{i_t}(x_t) e_i] \quad = \quad \frac{1}{\sum_{j=1}^n \beta_j^\gamma} \cdot \sum_{i=1}^n \beta_i^{\gamma-1} \partial_i f(x_t) e_i = \frac{1}{\sum_{j=1}^n \beta_j^\gamma} \cdot \nabla f(x_t) \circ (\beta_i^{\gamma-1})_{i \in [n]} \tag{16}$$

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

Theorem 2.2. *Define the weighted norms*

$$\|x\|_{[\gamma]}^2 := \sum_{i=1}^n x_i^2 \beta_i^\gamma \text{ and } \|x\|_{[\gamma]}^{*2} := \sum_{i=1}^n x_i^2 \beta_i^{-\gamma} \tag{17}$$

and note that the norm are dual to one another. We then have that the rule RCD(γ) *produces iterates satisfying*

$$\text{Exp}[f(x_t) - \arg\min_{x \in \mathbb{R}^n} f(x)] \leqslant \frac{2R_{1-\gamma}^2 \cdot \sum_{i=1}^n \beta_i^\gamma}{t-1}, \tag{18}$$

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

Proof. Recall the inequality that for a general β_g-smooth convex function g, one has that

$$g(u - \frac{1}{\beta_g} \nabla g(u)) - g(u) \leqslant -\frac{1}{2\beta_g} \|\nabla g\|^2 \tag{19}$$

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 β_i smooth. Hence, we have

$$f(x - \frac{1}{\beta_i} \nabla f(x) e_i) - f(x) = g_i(0 - \frac{1}{\beta_g} g_i'(0; x)) - g(0; x) \leqslant -\frac{g_i'(u; x)^2}{2\beta_i} = -\frac{\partial_i f(x)^2}{2\beta_i} \tag{20}$$

Hence, if i p^γ, we have

$$\text{Exp}[f(x - \frac{1}{\beta_i} \partial_i f(x) e_i) - f(x)] \quad \leqslant \quad \sum_{i=1}^n p_i^\gamma \cdot -\frac{\partial_i f(x)^2}{2\beta_i} \tag{21}$$

$$= \quad -\frac{1}{2 \sum_{i=1}^n \beta_i^\gamma} \sum_{i=1}^n \beta^{\gamma-1} \partial_i f(x)^2 \tag{22}$$

$$= \quad -\frac{\|\nabla f(x)\|_{[1-\gamma]}^*}{2 \sum_{i=1}^n \beta_i^\gamma} \tag{23}$$

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

$$\delta_{t+1} - \delta_t \leqslant -\frac{\|\nabla f(x_t)\|_{[1-\gamma]}^{*2}}{2 \sum_{i=1}^n \beta_i^\gamma} \tag{24}$$

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

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

Putting things together implies that

$$\delta_{t+1} - \delta_t \leqslant -\frac{\delta_t^2}{2R^2_{1-\gamma} \sum_{i=1}^n \beta_i^\gamma} \tag{25}$$

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

Theorem 2.3. *If f is in addition α-strongly convex w.r.t to $\|\cdot\|_{[1-\gamma]}$, then we get*

$$\mathrm{Exp}[f(x_{t+1}) - \arg\min_{x \in \mathbb{R}^n} f(x)] \leqslant (1 - \frac{\alpha}{\sum_{i=1}^n \beta_i^\gamma})^t f(x_1) - f(x^*) \tag{26}$$

Proof. One can show that strong convexity impies that

Lemma 2.4. *Let f be an α-strongly convex function w.r.t to a norm $\|\cdot\|$. Then, $f(x) - f(x^*) \leqslant \frac{1}{2\alpha}\|\nabla f(x)\|_*^2$*

Proof.

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

■

Can someone finish the proof? ■

What's surprising is that RCD(γ) is a descent method, despite being random. This is not true of normal SGD.

2.3

When does RCD(γ) actually do better? If $\gamma = 1$, the savings are proportional to the ration of $\sum_{i=1} \beta_i / \beta \cdot (T_{coord}/T_{grad})$. When f is twice differentiable, this is the ratio of

$$\frac{\mathrm{tr}(\max_x \nabla^2 f(x))}{\|\max_x \nabla^2 f(x)\|_{\mathrm{op}}} (T_{coord}/T_{grad}) \tag{27}$$

2.4 Other Extensions

1. Non-Stochastic, Cyclic SGD
2. Sampling w/ Replacement
3. Strongly Convex + Smooth!?
4. Strongly Convex (generalize SGD)
5. Acceleration? See Tu et al. Breaking Locality Accelerates Block Gauss-Seidel

2.5 Duality and Coordinate Descent

2.6 The Fenchel Dual

Definition 2.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then its Fenchel dual is defined as

$$f^*(w) := \sup_{x \in \mathbb{R}^n} \langle w, x \rangle - f(x) \tag{28}$$

Observe that $f^*(w)$ is convex, being a supremum of affine functions. Moreover, since $x \mapsto \langle w, x \rangle - f(x)$ is concae, the inner supremum is convex, and is optimized if and only if

$$w \in \partial f(x) \tag{29}$$

Note moreover that, by Danksin's Theorem,

$$\partial f^*(w) := \{x \in \arg\sup \langle w, x \rangle - f(x)\} = \{x : w \in \partial f(x)\} \tag{30}$$

In particular, $\partial f^*(0)$ is the set of minimizers of f.

2.7 Duals of Coordinate Functions

Suppose that

$$f(x) = \sum_{i=1}^{N} \phi_i(c_i^T x) + R(x) \tag{31}$$

Consider the subsitution $z_i = c_i^T x$, so that $z = C^\top x$. Then, letting $\mathbb{I}(x)$ denote the convex indicator which is 0 if $x = 0$ and ∞ otherwise, can write

$$\begin{aligned}
\min_x f(x) &= \min_{x,z} \sum_{i=1}^n \phi_i(z_i) + \lambda R(x) : C^\top x = z \\
&= \min_{x,z} \sum_{i=1}^n \phi_i(z_i) + \lambda R(x) + \mathbb{I}(C^\top x - z) \\
&= \min_{x,z} \max_{w \in \mathbb{R}^n} \sum_{i=1}^n \phi_i(z_i) + R(x) - w^\top (C^\top x - z) \\
&\geqslant \max_{w \in \mathbb{R}^n} \arg\min_{x,z} \sum_{i=1}^n \phi_i(z_i) + R(x) - w^\top (C^\top x - z) \\
&= \max_w - \arg\min_{x,z} \sum_{i=1}^n \phi_i(z_i) - z_i w_i + R(x) - (Cw)^\top x \\
&= \max_w - \arg\max_z \sum_{i=1}^n z_i w_i - \phi_i(z_i) \arg\max_x +(Cw)^\top x - R(x) \\
&= \max_w - \sum_{i=1}^n \phi_i^*(w_i) + R^*(Cw)
\end{aligned}$$

We can call the above objective $D(w) := -\sum_{i=1}^n \phi_i^*(w_i) + \arg\max_x (Cw)^\top x - R(x)$. Observe that by weak duality, we have that for any pair of points $(w, x) \in \mathbb{R}^N \times \mathbb{R}^n$, we

$$f(x) \geqslant f(x^*) \geqslant D(w^*) \geqslant D(w) \tag{32}$$

Hence, if we can maintain a pair of points (w_t, x_t) such that $f(x_t) - D(w_t) \leqslant \epsilon$, then we get for free that $f(x_t) - f(x^*) \leqslant \epsilon$. Moreover, the objective $D(w)$ is *concave* so it can be efficiently optimized.

2.8

A natural pair of (w, x). In general, one should hope that the pair (w_t, x_t) are related by some easy to compute correspondence. Suppose that we have a rule $x(w)$, which maps any point w to a point x, so that $x_t = x(w_t)$. We should hope that $x(w_*) \in \arg\min_x f(x)$, for any $w^* \in \arg\max_w D(w)$. More generally, this can be solved by noting for a dual optimal pair (w^*, x^*), one must have that

$$x^* \in \arg\min_x R(x) - (Cw)^\top x \tag{33}$$

When $R(x)$ is differentiable, this implies that $\nabla R(x) = Cw$ and if $R(x)$ is strictly convex, $\nabla R(x)$ is invertible, and so we would generally take our pair ato be $(w, (\nabla R)^{-1}(Cw))$.

In general, $(\nabla R)^{-1}$ might be hard to compute. But for ridge-penalties, if $R(x) = \frac{\lambda}{2}\|x\|^2$, then $\nabla R(x) = \lambda x$, so we have that $\lambda x = Cw$, whence $x = \frac{1}{\lambda} Cw$, which isn't so bad.

Note that this implies that whenever you have a square loss penalty, any optimal x^* is always in the span of the data, which is powerful when the number of features n greatly exceeds the number of examples N.

2.9 SDCA

1. s

Lemma 2.6. *Suppose that ϕ is β-smooth. Then ϕ^* is $\alpha = 1/\beta$-strongly convex.*

Lemma 2.7. *Let g be a α-strongly convex (for $\alpha \geqslant 0$), and let $w_0, x_0 \in \mathbb{R}$. Then, for any $u \in \mathbb{R}^n$ and any*

$$\begin{aligned}\min_w \{g(w) + \frac{L}{2}(w - x_0)^2\} - \{g(w_0) - \frac{L}{2}(w_0 - x_0)^2\} \leqslant \\ s\{g(u) + u(w_0 - x_0) - g(w_0) - w_0(w_0 - x_0) - (\frac{\alpha(1-s) + Ls}{2})(w_0 - u)^2\}\end{aligned} \tag{34}$$

Proof. Observe that we have (this is another definition of strong convexity):

$$g(w_0 + s(u - w_0)) \leqslant (1-s)g(w_0) + sg(u) - \frac{\alpha}{2}s(1-s)(w_0 - u)^2 \tag{35}$$

Hence,

$$\begin{aligned}
& \min_w \{g(w) + \frac{L}{2}(w - x_0)^2\} - \{g(w_0) - \frac{\lambda}{2}(w_0 - x_0)^2\} && (36)\\
\leqslant\ & g(w_0 + s(u - w_0)) + \frac{L}{2}(w_0 + s(u - w_0) - x_0)^2 - -\{g(w_0) - \frac{L}{2}(w_0 - x_0)^2\} && (37)\\
=\ & g(w_0 + s(u - w_0)) - g(w_0) + \frac{L}{2}\{(s(u - w_0))^2 + s(u - w_0)(w_0 - x_0)\} && (38)\\
=\ & (1-s)g(w_0) + sg(u) - \frac{\alpha}{2}s(1-s)(w_0 - u)^2 - g(w_0) + \frac{L}{2}\{(s(u - w_0))^2 + s(u - w_0)(w_0 - x_0)\} && (39)\\
=\ & s\{g(u) + u(w_0 - x_0) - g(w_0) - w_0(w_0 - x_0) - (\frac{\alpha(1-s) + Ls}{2})(w_0 - u)^2\} && (40)
\end{aligned}$$

■

Theorem 2.8 (SCDA).

References