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

April 23, 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 Primal Dual IPM

Let $x \in \mathbb{R}^n$ be the decision variable, and $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$

$$\min c^\top x \quad \text{s.t.} \quad Ax = b, x \geqslant 0 \tag{1}$$

$\geqslant$ is pointwise.

1.1 Deriving the dual problem

Observe that we can always write the constraint as

$$
\begin{aligned}
& \min_{x \geqslant 0} c^\top x + \max_z z^\top (b - Ax) && (2)\\
= \;& \min_{x \geqslant 0} \max_z c^\top x + \max_z z^\top (b - Ax) && (3)\\
\geqslant \;& \max_z z^\top b - \min_{x \geqslant 0} (c - A^\top z)^\top x && (4)\\
= \;& \max_z z^\top b - \infty \cdot \mathbb{I}(A^\top z > c) && (5)\\
& && (6)
\end{aligned}
$$

Hence, the dual problem is as follows

$$\max b^\top z \quad \text{s.t.} \quad A^\top z \leqslant c$$

This is equivalent to

$$\max b^\top z \quad \text{s.t.} \quad A^\top z + s = c, \quad s \geqslant 0 \tag{7}$$

where we introduce the slack variable $s \in \mathbb{R}^n$. If (x, z, s) are only *feasible*, then

$$Ax = b \quad A^\top z + s = c \quad x, s \geqslant 0 \tag{8}$$

Moreover, we can compute that for feasible (x, z, s),

$$0 \leqslant \langle x, s \rangle = x^\top (c - A^\top z) = \langle x, c \rangle - \langle Ax, z \rangle = \langle x, c \rangle - \langle b, z \rangle \,. \tag{9}$$

This is a proof of weak duality, namely that for any feasible x and z,

$$\langle x, c \rangle \geqslant \langle b, z \rangle \tag{10}$$

and therefore

$$\langle x^*, c \rangle \geqslant \langle b, z^* \rangle \tag{11}$$

Moreover, if there exists an feasible (x^*, z^*, s^*), with $\langle x^*, s^* \rangle = 0$ then we have

$$\langle x^*, c \rangle = \langle b^*, z \rangle \,, \tag{12}$$

which is strong duality.

Duality is also useful to bound the suboptimality gap, since in fact if (x, z, s) is feasible, then

$$\langle x, s \rangle = \langle x, c \rangle - \langle b, z \rangle \geqslant \langle x, c \rangle - \langle x^*, c \rangle = \langle x - x^*, c \rangle \tag{13}$$

1.2

This suggests the following approach. Consider iterates (x_k, z_k, s_k), and define

$$\mu_k := \frac{1}{n} \cdot \langle x_k, s_k \rangle = \frac{\langle x_k, c\rangle - \langle b, z_k \rangle}{n} \geqslant \frac{\langle x_k - x^*, c\rangle}{n} \tag{14}$$

Define the strictly feasible set

$$\mathcal{F}^o := \{Ax = b \quad A^\top z + s = c \quad x, s > 0\} \tag{15}$$

This is minimizing a *bilinear* objective over a linear constraint set. The goal is to general iterates $(x^{k+1}, z^{k+1}, s^{k+1})$ just that

$$\mu_{k+1} \leqslant (1 - Cn^{-\rho})\mu_k \tag{16}$$

Implies that

$$\langle x_k - x^*, c\rangle \leqslant \epsilon \text{ in } k = \mathcal{O}\left(n^\rho \log(n/\epsilon)\right) \text{steps.} \tag{17}$$

1.3 Central Path

The goal is to find a pair of s, z, x such that $\mu \approx 0$. We consider the following approach. Define

$$F_\tau(x, z, s) := \begin{bmatrix} Ax - b \\ A^\top z + s - c \\ x \circ s - \tau \mathbf{1} \end{bmatrix} \tag{18}$$

Then the goal is to approx solve $F_0(x_0, z_0, s_0) = \mathbf{0}$ over $\mathcal{F}^0$. We see that this can be obtained by computing the solutions (x_τ, z_τ, z_τ) to $F_\tau(x, z, s) = \mathbf{0}$. We call the curve $\tau \mapsto (x_\tau, z_\tau, z_\tau)$ the "central path". Note that, on the central path, $x_i s_i = \tau$ for some $\tau > 0$. To ensure we stay close to the central path, we consider

$$\mathcal{N}_{-\infty}(\gamma) \quad := \quad \{(x, z, s) \in \mathcal{F}_0 : \min_i x_i s_i \geqslant \gamma\mu(x, s)\}$$

What we would like to do is take iterates (x_k, z_k, s_k) such that μ_k decreases, and $(x_k, z_k, s_k) \in \mathcal{N}_\infty(\gamma)$ for approriate constants γ. $\mathcal{N}_\infty(\gamma)$ ensures the nonnegativity contraints.

Input: Parameters $\gamma \in (0, 1)$, $0 < \sigma_{\min} < \sigma_{\max} < 1$, and initialization $(x^0, z^0, t^0) \in \mathcal{N}_{-\infty}(\gamma)$.
for $t = 0, 1, 2, \ldots$ **do**

Choose $\sigma_k \in [\sigma_{\min}, \sigma_{\max}]$;

Run Newton step on $F_{\sigma_k \mu_k}$ (to be defined). Let $(\Delta x^k, \Delta z^k, \Delta s^k)$ denote the Newton step

$$(\Delta x^k, \Delta z^k, \Delta s^k) = -\nabla^2 F_{\tau_k}(w^k)^{-1} \cdot \nabla F_{\tau_k}(w^k), \tag{19}$$

$$\text{where } \tau_k = \sigma_k \mu_k \text{ and } w^k = (x^k, z^k, s^k)\,. \tag{20}$$

Let $\alpha_k \in (0, 1]$ be the largest step such that

$$\alpha_k = \max\{\alpha \in (0, 1] : (x^k, z^k, s^k) + \alpha(\Delta x^k, \Delta z^k, \Delta s^k) \in \mathcal{N}_\infty(\gamma)\} \tag{21}$$

Set $(x^{k+1}, z^{k+1}, s^{k+1}) \leftarrow (x^k, z^k, s^k) + \alpha_k(\Delta x^k, \Delta z^k, \Delta s^k)$.

end

Algorithm 1: PGD

1.4 What is the Newton Step?

The Newton Step for solving fixed point equations $F(w) = 0$. Indeed

$$F(w + d) = F(w) + J(w) \cdot d + o(\|d\|) \tag{22}$$

The Newton's method then chooses $w \leftarrow w + d$,

$$J(w)d = -F(w) \tag{23}$$

Which implies that $F(w + d) = o(\|d\|)$ for w sufficiently closed to the fixed point. This gives you the quick converge. Note that, if F is a linear map, then in fact *one Newton step suffices*. This can be seen from the Taylor expansion.

1.5

Our function F_{τ_k} is nearly linear, but not quite. Let's compute the Netwon Step. We observe that the Jacobian is the linear operator

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^\top & I \\ \mathrm{Diag}(S) & 0 & \mathrm{Diag}(X) \end{bmatrix} \tag{24}$$

Moreover, since $(x^k, z^k, s^k) \in \mathcal{F}^0$, we have that

$$F_{\tau_k}(x^k, z^k, s^k) = \begin{bmatrix} Ax^k - b \\ A^\top z^k + s^k - c \\ x^k \circ s^k - \tau_k \mathbf{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ x^k \circ s^k - \tau_k \mathbf{1} \end{bmatrix} \tag{25}$$

Let's drop subscripts. Then, on one can verify that the Newton satisfies

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^\top & I \\ \mathrm{Diag}(S) & 0 & \mathrm{Diag}(X) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \\ \Delta s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -x \circ s + \tau \mathbf{1} \end{bmatrix} \tag{26}$$

Some remarks

1. $A\Delta x = 0$, and that $A^\top \Delta z + \Delta_s = 0$.
2. This implies that $(x^+, z^+, s^+) := (x + \Delta x, z + \Delta z, s + \Delta s)$ satisfies

$$Ax^+ - b = 0 \quad \text{and} \quad A^\top z^+ + s - c = 0 \tag{27}$$

$$\tag{28}$$

3. We also have

$$s \circ \Delta x + x \circ \Delta s = -x \circ s + \tau \mathbf{1} \tag{29}$$

and thus

$$\begin{aligned} x^+ \circ s^+ &= x \circ s + (\circ \Delta x + x \circ \Delta s) + \Delta x \circ \Delta s && (30) \\ &= x \circ s - x \circ s + \tau \mathbf{1} + \Delta x \circ \Delta s && (31) \end{aligned}$$

4. Thus,

$$F_\tau(x^+ \circ s^+) = \begin{bmatrix} 0 \\ 0 \\ \Delta x \circ \Delta s \end{bmatrix} \tag{32}$$

In other words, if we can argue that $\Delta x \circ \Delta s$ is "neglibible", then the Newton step produces an almost exact solution.

A more concrete analysis would be to study the term

$$\begin{aligned} n\mu(x + \alpha\Delta x, s + \alpha\Delta x) &= \langle x + \Delta x, s + \Delta s\rangle \\ &= \langle x, s\rangle + \alpha\left(\langle x, \Delta s\rangle + \langle s, \Delta x\rangle\right) + \alpha^2\langle \Delta s, \Delta x\rangle \end{aligned}$$

The last term in the above display vanishes, as shown by the above

$$0 = \Delta x^\top(A^\top \Delta z + \Delta s) = (A\Delta x)^\top z + \langle \Delta x, \Delta s\rangle = \langle \Delta x, \Delta\rangle \tag{33}$$

Moreover, since $s \circ \Delta x + x \circ \Delta s = -x \circ s + \tau\mathbf{1}$, we have by summing that

$$\langle x, \Delta s\rangle + \langle s, \Delta x\rangle = -\langle x, x\rangle + \tau n = -(1-\sigma)\langle x, s\rangle \tag{34}$$

where the last line uses $n\tau = n\sigma\mu = \sigma n\tau = \sigma\langle x, s\rangle$. Hence,

$$n\mu(x + \alpha\Delta x, s + \alpha\Delta x) = n\mu(x, s)(1 - \alpha(1-\sigma)) \tag{35}$$

Hence, if one can show that $(1-\alpha)\alpha \geqslant C(n) > 0$ for some constant depending on the dimension, then we see that

$$n\mu(x^{k+1}) \leqslant (1 - C(n))^k n\mu(x^0) \tag{36}$$

giving us the rate of decrease. One can then show with more technical work that $\alpha = \Omega(1/n)$ while maintaining the $\mathcal{N}_{-\infty}(\gamma)$ invariant.