4 Interior-Point Methods

For LP, the feasible set is polyhedral and it can be solved by the simplex method, moving successively from a vertex to a neighboring vertex of lower cost. This does not work for SOCP or SDP since their feasible sets are not polyhedral. There has been some study to try to extend the simplex method to solve SOCP or SDP, but none successfully. Interior-point methods do not move along the boundary of the feasible set but in the "interior" and can solve a large class of convex optimization problems. In fact, even for LP, interior-point method can be more efficient than the simplex method on certain large scale problems.

Interior-point (abbreviated as "IP") methods were popularized by Narendra Karmakar who in his 1984 paper proposed a projective scaling method for LP, proved its polynomial-time complexity, and claimed superior practical performance over simplex methods. (The latter turned out to be not quite accurate.) Soon it was recognized that Karmark's method is closely related to the log-barrier method from the late 1960s, based on which many improvements were made, and this is how the method is often presented now. In the early 1990s, these methods were extended to solve SDP and later also SOCP. IP methods are not the only methods for convex optimization, but they are special in that they enjoy the best iteration complexity and in practice the number of iterations are typically below 50. The main computational cost per iteration lies in solving a certain linear equation that becomes increasingly ill-conditioned as the method converges. Whether an IP method is efficient depends on how efficiently this linear equation can be solved per iteration.

The idea underlying IP methods is to add a log-barrier function to the objective to make it unconstrained and solve it approximately by Newton's method. Why log? Roughly speaking, the 1st- and 2nd-derivatives of logarithm have nice scaling properties (2nd-derivative is square of the 1st-derivative), which is not true for other functions. In what follows, we assume the cone is

$$K_{\text{LP}} = [0,\infty)^n \quad \text{or} \quad K_{\text{SOCP}} = \text{SOC}^{n_1} \times \cdots \times \text{SOC}^{n_N} \quad \text{or} \quad K_{\text{SDP}} = \{X \in \mathbb{S}^n \mid X \succeq 0\}$$

with corresponding log-barrier functions

$$\begin{aligned}
\psi_{\text{LP}}(x) &= \begin{cases} -\sum_{i=1}^n \log x_i & \text{if } x \in \text{int}K_{\text{LP}} \\ \infty & \text{else} \end{cases} \\
\psi_{\text{SOCP}}(x) &= \begin{cases} -\sum_{i=1}^N \log\left(x_{n_i,i}^2 - (x_{1,i}^2 + \cdots + x_{n_i-1,i}^2)\right) & \text{if } x \in \text{int}K_{\text{SOCP}} \\ \infty & \text{else} \end{cases} \\
\psi_{\text{SDP}}(X) &= \begin{cases} -\log\det X & \text{if } X \in \text{int}K_{\text{SDP}} \\ \infty & \text{else} \end{cases}
\end{aligned}$$

In applications, these three cones arise most often. They are also symmetric cones, with associated Jordan algebra and spectral decomposition, which can be useful in computation. The type of IP method we study here are called path-following, for reasons we will see below. (These are not the

only IP methods, but they are the most popular these days.) There are primal, dual, and primal-dual path-following methods. The first two methods can be readily generalized to other cones using suitable barrier functions (so-called self-concordant barrier functions), while the primal-dual method is trickier to generalize even for $K_{\rm SDP}$. Primal-dual path-following method tends to converge faster and yields more accurate solution, but in the case of SDP, its linear equation can be computationally more expensive to solve than dual path-following method.

Consider the primal and dual problems (18) and (19). We will simplify the notation by introducing the linear mapping $\mathcal{A} : \mathbb{H} \to \mathbb{R}^m$ defined as

$$\mathcal{A}x = [\langle a_i, x\rangle]_{i=1}^m.$$

Its adjoint $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{H}$ is

$$\mathcal{A}^*\mu = \sum_{i=1}^m a_i\mu.$$

(Here * denotes the adjoint, not an optimal solution.) We assume without loss of generality that $a_1, \ldots, a_m$ are linearly independent, so that

$$\mathrm{Null}(\mathcal{A}^*) = \{0\}. \tag{23}$$

4.1 Dual path-following method

For $K \in \{K_{\rm LP}, K_{\rm SOCP}, K_{\rm SDP}\}$, we have $K^\circ = -K$, so the dual problem (19) can be rewritten as

$$\begin{array}{ll} \min_\mu & b^T\mu \\ \text{s.t.} & \mathcal{A}^*\mu - c \in K. \end{array} \tag{24}$$

The log-barrier problem is

$$\min_\mu \quad d_\epsilon(\mu) := b^T\mu + \epsilon\psi(\mathcal{A}^*\mu - c) \qquad (\epsilon > 0), \tag{25}$$

where $\psi \in \{\psi_{\rm LP}, \psi_{\rm SOCP}, \psi_{\rm SDP}\}$. Since $\mathrm{dom}\psi = \mathrm{int}K$, (25) is feasible if and only if $\mathcal{F}(D) \cap (\mathrm{int}K^\circ \times \mathbb{R}^m) \neq \emptyset$. However, an optimal solution may not exist even when (24) has an optimal solution, e.g. $b = c = 0$ and $\mathcal{A} = 1$ so any $\mu \geq 0$ solves (24) but $\min_\mu -\epsilon\log(\mu) = -\infty$. A sufficient condition for (25) to have an optimal solution (which is unique since ψ is strictly convex and $\mathrm{Null}(\mathcal{A}^*) = \{0\}$) is $\mathcal{F}(P) \cap \mathrm{int}K \neq \emptyset$.

Proposition 1. *Suppose that $\mathcal{F}(P) \cap \mathrm{int}K \neq \emptyset$ and $\mathcal{F}(D) \cap (\mathrm{int}K^\circ \times \mathbb{R}^m) \neq \emptyset$. Then (25) has an optimal solution (which is unique).*

Proof. We prove this for $\mathbb{H} = \mathbb{S}^n$, $K = K_{\rm SDP}$, and $\psi = \psi_{\rm SDP}$. The case of $\psi = \psi_{\rm LP}, \psi_{\rm SOCP}$ can be treated similarly. It suffices to show that the level set

$$\{\mu \mid b^T\mu - \epsilon\log\det(\mathcal{A}^*\mu - C) \leq v\}$$

is bounded for any $v \in \mathbb{R}$. Then an optimal solution exists by the Bolzano-Weierstrass theorem. We argue this by contradiction. Suppose this level set is unbounded for some $v \in \mathbb{R}$. Then there

exists a sequence μ_k such that $\|\mu_k\|_2 \to \infty$ and $b^T\mu_k - \epsilon \log\det(\mathcal{A}^*\mu - C) \le v$ for all k. Dividing both sides by $\|\mu_k\|$ yields

$$b^T \frac{\mu_k}{\|\mu_k\|_2} + \epsilon \sum_{i=1}^{n} \frac{-\log \lambda_i(\mathcal{A}^*\mu_k - C)}{\|\mu_k\|_2} \le \frac{v}{\|\mu_k\|_2}. \tag{26}$$

We can rewrite the ith term in the sum of (26) as

$$\frac{-\log \lambda_i(\mathcal{A}^*\mu_k - C)}{\lambda_i(\mathcal{A}^*\mu_k - C)} \cdot \frac{\lambda_i(\mathcal{A}^*\mu_k - C)}{\|\mu_k\|_2}.$$

The first term tends to 0 if $\lambda_i(\mathcal{A}^*\mu_k - C) \to \infty$ (since $\lim_{t\to\infty} \frac{\log t}{t} = 0$). It is bounded from below if $\lambda_i(\mathcal{A}^*\mu_k - C)$ is bounded as $k \to \infty$. The second term is bounded since $\lambda_i(\cdot)$ is Lipschitz continuous on $\mathbb{S}^n$.[9] Moreover, it tends to 0 if $\lambda_i(\mathcal{A}^*\mu_k - C)$ is bounded as $k \to \infty$. Thus the limit-infimum of the this ith term is nonnegative. It follows from (26) that every cluster point u of $\mu_k/\|\mu_k\|_2$ as $k \to \infty$ satisfies $b^T u \le 0$ and $u \ne 0$. On the other hand, since $\mathcal{A}^*\mu_k - C \succ 0$ so that $(\mathcal{A}^*\mu_k - C)/\|\mu_k\|_2 \succ 0$ for all k, we also have $\mathcal{A}^*u \succeq 0$. Also, $\mathcal{A}^*u \ne 0$ (since $u \ne 0$ and $\mathrm{Null}(\mathcal{A}^*) = \{0\}$). Since $\mathcal{F}(P) \cap \mathrm{int}K \ne \emptyset$, there exists $\bar{X} \succ 0$ satisfying $\mathcal{A}\bar{X} = b$, implying $b^T u = \langle \bar{X}, \mathcal{A}^*u\rangle > 0$. (In general, $\langle x, y\rangle < 0$ for any $x \in \mathrm{int}K, 0 \ne y \in K^\circ$.) ■

The exact solution of (25) traces the so-called central path as ϵ ranges over $(0, \infty)$. Intuitively, as $\epsilon \to 0$, any cluster point of this exact solution is an optimal solution of the dual (24). The dual path-following method solves (25) inexactly using Newton method and decreases ϵ after each iteration. The key to the complexity analysis is the notion of inexact solution. Since d_ϵ is convex differentiable on its effective domain, the minimum is attained at

$$\nabla d_\epsilon(\mu) = 0. \tag{27}$$

By the chain rule for differentiation, we have

$$\nabla d_\epsilon(\mu) = b + \epsilon\mathcal{A}\nabla\psi(y), \qquad \nabla^2 d_\epsilon(\mu) = \epsilon\mathcal{A}\nabla^2\psi(y)\mathcal{A}^* \quad \text{with} \quad y = \mathcal{A}^*\mu - c. \tag{28}$$

Thus, the minimum is attained when $b + \epsilon\mathcal{A}\nabla\psi(y) = 0$. This looks like primal feasibility, especially if we let

$$x = -\epsilon\nabla\psi(y). \tag{29}$$

- For $\mathbb{H} = \mathbb{R}^n$ and $\psi = \psi_{\mathrm{LP}}$, this yields

  $$\mathcal{A}x = b, \quad y = \mathcal{A}^*\mu - c > 0, \quad x = \epsilon[1/y_i]_{i=1}^n. \tag{30}$$

  This can rewritten more symmetrically in x and y:

  $$\mathcal{A}x = b, \quad y = \mathcal{A}^*\mu - c, \quad [x_i y_i]_{i=1}^n = \epsilon e, \tag{31}$$

  where e denotes the vector of 1s. Our notion of inexact solution is based on a relaxation of the third equation, which is nonlinear.

  $$\mathcal{N}(\gamma) = \{(\mu, \epsilon) \mid \exists\ x \text{ with } \mathcal{A}x = b,\ y = \mathcal{A}^*\mu - c > 0,\ \|[x_i y_i]_{i=1}^n - \epsilon e\|_2 \le \gamma\epsilon\} \tag{32}$$

  with $0 < \gamma < 1$. This is often called the "narrow neighborhood" of the central path, to distinguish it from "wide neighborhood" that we will see later.

[9] See Matrix Analysis by Horn & Johnson, Cor. 6.3.4 or Matrix Analysis by Bhatia, p. 63.

- For $\mathbb{H} = \mathbb{S}^n$ and $\psi = \psi_{\text{SDP}}$, this yields

$$\mathcal{A}X = b, \quad Y = \mathcal{A}^*\mu - C \succ 0, \quad X = \epsilon Y^{-1}. \tag{33}$$

By expressing $Y = Y^{1/2}Y^{1/2}$, this can rewritten more symmetrically in X and Y:

$$\mathcal{A}X = b, \quad Y = \mathcal{A}^*\mu - C \succ 0, \quad Y^{1/2}XY^{1/2} = \epsilon I. \tag{34}$$

The corresponding narrow neighborhood of the central path is

$$\mathcal{N}(\gamma) = \left\{(\mu, \epsilon) \mid \exists\ X \text{ with } \mathcal{A}X = b, \quad Y = \mathcal{A}^*\mu - C \succ 0,\ \|Y^{1/2}XY^{1/2} - \epsilon I\|_F \le \gamma\epsilon\right\} \tag{35}$$

Here we use $Y^{1/2}XY^{1/2}$ (which is symmetric and positive definite whenever $X, Y \succ 0$) instead of XY because XY may not be symmetric when X, Y is off the central path. An alternative is to use $(XY+YX)/2$, which is symmetric (but not necessarily positive definite). However, the complexity analysis becomes much harder.

For any $(\mu, \epsilon) \in \mathcal{N}(\gamma)$, there exists X with $\mathcal{A}X = b$ and $\|Y^{1/2}XY^{1/2} - \epsilon I\|_F \le \gamma\epsilon$, where $Y = \mathcal{A}^*\mu - C \succ 0$ (so $(-Y, \mu) \in \mathcal{F}(D)$). We show below that X is primal feasible and the duality gap is $O(\epsilon)$. (For illustration, we consider SDP, but the same argument works for LP and SOCP.) We have from $\|A\|_F = \left(\text{tr}(A^2)\right)^{1/2} = \left(\sum_{i=1}^n \lambda_i(A)^2\right)^{1/2} \ge \max_i |\lambda_i(A)|$ for any $A \in \mathbb{S}^n$ that

$$\max_i |\lambda_i(Y^{1/2}XY^{1/2} - \epsilon I)| \le \|Y^{1/2}XY^{1/2} - \epsilon I\|_F \le \gamma\epsilon.$$

Since $\lambda_i(Y^{1/2}XY^{1/2} - \epsilon I) = \lambda_i(Y^{1/2}XY^{1/2}) - \epsilon$, this yields

$$-\gamma\epsilon \le \lambda_i(Y^{1/2}XY^{1/2}) - \epsilon \le \gamma\epsilon \quad \forall i.$$

Hence

$$(1-\gamma)\epsilon \le \lambda_i(Y^{1/2}XY^{1/2}) \le (1+\gamma)\epsilon \quad \forall i.$$

Since $(1-\gamma)\epsilon > 0$, this shows $Y^{1/2}XY^{1/2} \succ 0$ and hence $X \succ 0$. Thus $X \in \mathcal{F}(P)$. Moreover,

$$0 \le \langle X, Y\rangle = \text{tr}[XY] = \text{tr}[Y^{1/2}XY^{1/2}] = \sum_{i=1}^n \lambda_i(Y^{1/2}XY^{1/2}) \le n(1+\gamma)\epsilon \le 2n\epsilon$$

while

$$\langle X, Y\rangle = \langle X, \mathcal{A}^*\mu - C\rangle = \langle \mathcal{A}X, \mu\rangle - \langle X, C\rangle = \langle b, \mu\rangle - \langle X, C\rangle.$$

Thus we obtain bounds on the duality gap: $0 \le b^T\mu - \langle C, X\rangle \le 2n\epsilon$. As $\epsilon \to 0$, a continuity argument shows that any cluster point of X, μ (which can be shown to exist by a similar argument as in the proof of Proposition 1) solves the primal and the dual and there is no duality gap. In the case of LP, it has been shown that there is a unique limit which is the "analytic center" of the optimal face in the primal and dual. For SOCP and SDP, this may or may not be true. The 2002 book de Klerk (Chap. 3) discusses known results for SDP.

The dual path-following method starts with any $(\mu, \epsilon) \in \mathcal{N}(\gamma)$ and solves the optimality equation (27) using Newton's method, with ϵ decreased after each iteration. This method is often called "short-step" because it uses the narrow neighborhood. It is not as practical as the

"long-step" method we will see later, but enjoys the best iteration complexity (in the worst-case sense).

Dual Path-Following method

- Choose any initial $(\mu, \epsilon) \in \mathcal{N}(\gamma)$ and $\epsilon^{\text{final}} > 0$.
- While $\epsilon > \epsilon^{\text{final}}$:

  Solve the Newton equation

$$\nabla d_\epsilon(\mu) + \nabla^2 d_\epsilon(\mu)(\mu_+ - \mu) = 0 \tag{36}$$

  for μ_+, and update

$$\epsilon_+ = \epsilon(1-\theta) \qquad \text{with} \qquad \theta = \frac{\gamma(1-\gamma)}{\gamma + \sqrt{n}} \tag{37}$$

Thus, if ϵ^0 denotes the initial ϵ, we have $\epsilon = (1-\theta)^k \epsilon^0$ after k iterations, so that the method terminates when $(1-\theta)^k \epsilon^0 \leq \epsilon^{\text{final}}$ or, equivalently,

$$\begin{aligned}
\frac{\epsilon^0}{\epsilon^{\text{final}}} \leq \frac{1}{(1-\theta)^k} &\iff \log\left(\frac{\epsilon^0}{\epsilon^{\text{final}}}\right) \leq -k\log(1-\theta) \\
&\iff \frac{\log\left(\frac{\epsilon^0}{\epsilon^{\text{final}}}\right)}{-\log(1-\theta)} \leq k \\
&\iff k = \left\lceil \frac{\log\left(\frac{\epsilon^0}{\epsilon^{\text{final}}}\right)}{-\log(1-\theta)} \right\rceil
\end{aligned}$$

Since $\log(1-\theta) \leq -\theta$, this yields

$$k \leq \left\lceil \theta^{-1} \log\left(\frac{\epsilon^0}{\epsilon^{\text{final}}}\right) \right\rceil .$$

Since $\theta^{-1} = O(\sqrt{n})$, this yields $O\left(\sqrt{n}\log\left(\frac{\epsilon^0}{\epsilon^{\text{final}}}\right)\right)$ iterations. Note that θ is maximized at around $\gamma = \frac{1}{2}$. We can accelerate convergence of the method by, say, setting $\mu[\alpha] = \mu + \alpha(\mu_+ - \mu)$ and $\epsilon[\alpha] = \min\{\epsilon \mid (\mu[\alpha], \epsilon) \in \mathcal{N}(\gamma)\}$ and do a 1-dimensional search in α to minimize $\epsilon[\alpha]$ (which has closed form in terms of α).

Proposition 2. *If $(\mu, \epsilon) \in \mathcal{N}(\gamma)$, then (μ_+, ϵ_+) given by (36) and (37) is also in $\mathcal{N}(\gamma)$.*

Proof. We prove this for SDP only, i.e., $\psi = \psi_{\text{SDP}}$ and $\mathcal{N}(\gamma)$ given by (35). LP and SOCP can be recovered as special cases.