Lecture Notes for Week 5

Proof of Proposition 2 for the case of SDP:

Since $(\mu, \epsilon) \in \mathcal{N}(\gamma)$, there exists X satisfying $\mathcal{A}X = b$ and

$$\|Y^{1/2}XY^{1/2} - \epsilon I\|_F \leq \gamma\epsilon, \tag{38}$$

where $Y = \mathcal{A}^*\mu - C \succ 0$. To show that $(\mu_+, \epsilon_+) \in \mathcal{N}(\gamma)$, it suffices to find an X_+ satisfying analogous conditions. Letting $w = \mu_+ - \mu$ and using (28), we can rewrite the Newton equation (36) as

$$b + \epsilon\mathcal{A}\nabla\psi(Y) + \epsilon\mathcal{A}\nabla^2\psi(Y)\mathcal{A}^*w = 0, \tag{39}$$

where $Y = \mathcal{A}^*\mu - C \succ 0$. Letting $V = \mathcal{A}^*w$ and using the formula for $\nabla\psi_{\text{SDP}}$ and $\nabla^2\psi_{\text{SDP}}$ from week 2, this expands to

$$b - \epsilon\mathcal{A}Y^{-1} + \epsilon\mathcal{A}Y^{-1}VY^{-1} = 0$$

or $b = \mathcal{A}(\epsilon Y^{-1} - \epsilon Y^{-1}VY^{-1})$. This suggests taking

$$X_+ = \epsilon Y^{-1} - \epsilon Y^{-1}VY^{-1}, \tag{40}$$

which then satisfies $\mathcal{A}X_+ = b$. Letting $Y_+ = Y + V$ (so $Y = \mathcal{A}^*\mu_+ - C$) and $\bar{V} = Y^{-1/2}VY^{-1/2}$, we have

$$\begin{aligned}
\|Y_+^{1/2}X_+Y_+^{1/2} - \epsilon I\|_F^2 &= \operatorname{tr}\left[(Y_+^{1/2}X_+Y_+^{1/2} - \epsilon I)^2\right] \\
&= \operatorname{tr}\left[Y_+^{1/2}X_+Y_+X_+Y_+^{1/2} - 2\epsilon Y_+^{1/2}X_+Y_+^{1/2} + \epsilon^2 I\right] \\
&= \operatorname{tr}\left[X_+Y_+X_+Y_+ - 2\epsilon X_+Y_+ + \epsilon^2 I\right] \\
&= \operatorname{tr}\left[(X_+Y_+ - \epsilon I)^2\right] \\
&= \operatorname{tr}\left[\left((\epsilon Y^{-1} - \epsilon Y^{-1}VY^{-1})(Y+V) - \epsilon I\right)^2\right] \\
&= \operatorname{tr}\left[\left(\epsilon Y^{-1}VY^{-1}V\right)^2\right] \\
&= \epsilon^2 \operatorname{tr}\left[\left(Y^{-1/2}\bar{V}^2Y^{1/2}\right)^2\right] \\
&= \epsilon^2 \operatorname{tr}\left[\bar{V}^4\right] \\
&= \epsilon^2 \sum_{i=1}^n \lambda_i(\bar{V})^4 \\
&\leq \epsilon^2 \left(\sum_{i=1}^n \lambda_i(\bar{V})^2\right)^2 \\
&= \epsilon^2 \left(\operatorname{tr}\left[\bar{V}^2\right]\right)^2 \\
&= \epsilon^2 \|\bar{V}\|_F^4,
\end{aligned} \tag{41}$$

where the third, eighth equalities use $\operatorname{tr}[AB] = \operatorname{tr}[BA]$ $(A, B \in \mathbb{R}^{n\times n})$; the fifth equality uses (42); the seventh equality uses $Y^{-1} = Y^{-1/2}Y^{-1/2}$ and $Y^{-1/2}Y^{1/2} = I$; the nineth and tenth equalities use $\operatorname{tr}[A^k] = \sum_i \lambda_i(A^k) = \sum_i \lambda_i(A)^k$ for any even integer k; and the inequality uses $\sum_i a_i^2 \leq (\sum_i |a_i|)^2$.

Now we bound $\|\bar{V}\|_F$. Letting $U = X_+ - X$, we have $\mathcal{A}U = 0$. (Equivalently, (40) and $\mathcal{A}U = 0$, $V = \mathcal{A}^*w$ are the linearization of (33).) Multiplying left and right by $Y^{1/2}$, (40) can be written as

$$Y^{1/2}UY^{1/2} + \epsilon Y^{-1/2}VY^{-1/2} = \epsilon I - Y^{1/2}XY^{1/2}. \tag{42}$$

Letting $\bar{U} = Y^{1/2}UY^{1/2}$ and $R = \epsilon I - Y^{1/2}XY^{1/2}$, this simplifies to

$$\bar{U} + \epsilon\bar{V} = R.$$

Using $\text{tr}[AB] = \text{tr}[BA]$ $(A, B \in \mathbb{R}^{n\times n})$ and $\mathcal{A}U = 0$ yields

$$\langle \bar{U}, \bar{V}\rangle = \text{tr}[Y^{1/2}UY^{1/2}Y^{-1/2}VY^{-1/2}] = \text{tr}[UV] = \langle U, V\rangle = \langle U, \mathcal{A}^*w\rangle = \langle \mathcal{A}U, w\rangle = 0.$$

Thus

$$\|R\|_F^2 \;=\; \|\bar{U} + \epsilon\bar{V}\|_F^2 \;=\; \|\bar{U}\|_F^2 + 2\epsilon\langle\bar{V}, \bar{U}\rangle + \epsilon\|\bar{V}\|_F^2 \;=\; \|\bar{U}\|_F^2 + \epsilon^2\|\bar{V}\|_F^2.$$

Since, by (38), $\|R\|_F \le \gamma\epsilon$, this implies

$$\|\bar{V}\|_F \le \frac{\|R\|_F}{\epsilon} \le \gamma. \tag{43}$$

Using (37) and the triangle inequality yields

$$\begin{aligned}
\|Y_+^{1/2}X_+Y_+^{1/2} - \epsilon_+ I\|_F &= \|Y_+^{1/2}X_+Y_+^{1/2} - \epsilon(1-\theta)I\|_F \\
&\le \|Y_+^{1/2}X_+Y_+^{1/2} - \epsilon I\|_F + \epsilon\theta\|I\|_F \\
&\le \epsilon\gamma^2 + \epsilon\theta\sqrt{n} \\
&= \gamma\epsilon(1-\theta) \\
&= \gamma\epsilon_+,
\end{aligned}$$

where the second inequality uses (41) and (43). Finally, (43) implies $\sum_{i=1}^n \lambda_i(\bar{V})^2 = \|\bar{V}\|_F^2 \le \gamma^2$, so $|\lambda_i(\bar{V})| \le \gamma < 1$ for all i, implying $\lambda_i(I + \bar{V}) = 1 + \lambda_i(\bar{V}) > 0$ for all i, so $I + \bar{V} \succ 0$ and hence $Y_+ = Y^{1/2}(I + \bar{V})Y^{1/2} \succ 0$. ■

If C and $A_1, \ldots, A_m$ are diagonal, then so is Y, and X can be taken to be diagonal. The SDP then reduces to an LP. There are still the questions of how to find an initial (μ, ϵ) and how to efficiently solve the Newton equation (36) or, equivalently, (39). This is problem specific and often requires exploiting the problem structure.

Consider the SDP relaxation of the MaxCut problem we saw in week 1:

$$\begin{array}{ll}
\max_X & \displaystyle\sum_{i,j\in\mathcal{N}} c_{ij}(1 - x_{ij}) \\
\text{s.t.} & X \succeq 0, \quad x_{ii} = 1 \;\forall i \in \mathcal{N}.
\end{array}$$

Letting $C = [c_{ij}]_{i,j\in\mathcal{N}}$ (C is symmetric) and A_i be the matrix with 1 in its (i,i)th entry and zero elsewhere, we can rewrite this problem in the primal conic form (18):

$$\begin{array}{ll}
\max_X & \langle -C, X\rangle \\
\text{s.t.} & \langle A_i, X\rangle = 1, \quad i = 1, \ldots, n, \\
 & X \succeq 0.
\end{array}$$

Notice that $X = I$ is feasible and $I \succ 0$. The dual problem is

$$\begin{array}{ll} \min_{\mu} & \sum_{i=1}^{n} \mu_i \\ \text{s.t.} & \sum_{i=1}^{n} A_i \mu_i + C \succeq 0. \end{array}$$

Take $\mu_i = \epsilon$ for all i. Then

$$Y = (\sum_i A_i)\epsilon + C = I\epsilon + C \succeq 0 \quad \Longleftrightarrow \quad \epsilon I \succeq -C \quad \Longleftrightarrow \quad \epsilon \geq \lambda_{\max}(-C).$$

Moreover,

$$\left\| \frac{Y^{1/2} X Y^{1/2}}{\epsilon} - I \right\|_F = \left\| \frac{\epsilon I + C}{\epsilon} - I \right\|_F = \frac{\|C\|_F}{\epsilon} \leq \gamma \quad \Longleftrightarrow \quad \frac{\|C\|_F}{\gamma} \leq \epsilon.$$

Thus, for $\epsilon \geq \max\left\{\lambda_{\max}(-C), \frac{\|C\|_F}{\gamma}\right\}$, we have $(\mu, \epsilon) \in \mathcal{N}(\gamma)$.

For SDP, (39) reduces to

$$b - \epsilon \mathcal{A} Y^{-1} + \epsilon \mathcal{A} Y^{-1} (\mathcal{A}^* w) Y^{-1} = 0.$$

or, using the forms of $\mathcal{A}$ and $\mathcal{A}^*$,

$$b_i - \epsilon \langle A_i, Y^{-1} \rangle + \epsilon \langle A_i, Y^{-1} \sum_{j=1}^{m} A_j w_j Y^{-1} \rangle = 0, \quad i = 1, \ldots, m.$$

Thus, letting $m_{ij} = \langle A_i, Y^{-1} A_j Y^{-1} \rangle$ and $r_i = \epsilon \langle A_i, Y^{-1} \rangle - b_i$, this reduces to the $m \times m$ linear equation

$$\epsilon M w = r. \tag{44}$$

Moreover, it's not hard to verify that M is symmetric, positive definite (check?), so one can solve it using direct method like Cholesky factorization or iterative method like preconditioned conjugate gradient method (typically with a diagonal preconditioner). On the other hand, as $\epsilon \to 0$, this equation becomes increasingly ill-conditioned (due in part to $\lambda_{\min}(Y) \to 0$) which means the equation cannot be solved to high accuracy. This in turn limits how small ϵ can be in practice. The primal-dual method that we will see next turns out to be better suited for achieving high solution accuracy.

The dual method has one key advantage when solving SDP because $Y = \sum_{i=1}^{m} A_i \mu_i + C$ inherits the sparsity pattern in $A_1, \ldots, A_m, C$. For the SDP relaxation of the MaxCut problem, A_{ii} has only one nonzero on the ith diagonal and C is sparse (i.e., few nonzeros) whenever the graph is sparse (i.e., few edges). This sparsity pattern does not change with the iteration, which allows the coefficients m_{ij} to be computed efficiently by computing a Cholesky factorization of Y, i.e.,

$$PYP^T = LL^T,$$

where L is lower triangular and P is a permutation matrix chosen at the beginning to reduce the nonzero fill-in, such as minimum-degree ordering (see, e.g., the 1981 book Computer Solution of Large Sparse Definite Systems by George and Liu or google "sparse linear" for more recent work.)

Often L is sparse, even though Y^{-1} is typically dense, and L can be computed as fast or faster than Y^{-1} (to good accuracy). Then we compute m_{ij} using L. For the SDP relaxation of the MaxCut problem, $A_{ii} = e^i(e^i)^T$, where e^i is the ith unit coordinate vector in $\mathbb{R}^n$. Thus

$$\begin{aligned} m_{ij} &= \operatorname{tr}\left[e^i(e^i)^T Y^{-1} e^j (e^j)^T Y^{-1}\right] \\ &= \operatorname{tr}\left[(e^i)^T Y^{-1} e^j (e^j)^T Y^{-1} e^i\right] \\ &= ((e^i)^T Y^{-1} e^j)^2 \\ &= ((e^i)^T (P^T L L^T P)^{-1} e^j)^2 \\ &= ((L^{-1} P e^i)^T (L^{-1} P e^j))^2, \end{aligned}$$

where we use $P^T = P^{-1}$ in the last equality. Thus, at each iteration, we compute L (in $O(n^3)$ operations or less), solve $Lu^i = Pe^i$ for u^i (in $O(n^2)$ operations or less) for $i = 1, \ldots, n$, and compute $m_{ij} = ((u^i)^T u^j)^2$ (in $O(n)$ operations) for $i, j = 1, \ldots, n$. Thus the total work to compute M is $O(n^3)$. r is computed similarly. Solving (44) takes $O(n^3)$ operations using, say, Cholesky factorization. Thus the total computational cost per iteration is $O(n^3)$ operations, which is pretty good. Notice that M has the same sparsity pattern as Y^{-1}. In fact, M is the Hadamard (entrywise) product of Y^{-1} with itself.

4.2 Primal path-following method

The primal problem (18) can be written as

$$\begin{array}{ll} \min_\mu & \langle c, x\rangle \\ \text{s.t.} & \mathcal{A}x = b, \quad x \in K. \end{array}$$

For $K \in \{K_{\text{LP}}, K_{\text{SOCP}}, K_{\text{SDP}}\}$, the log-barrier problem is

$$\begin{array}{ll} \min_x & \langle c, x\rangle + \epsilon\psi(x) \\ \text{s.t.} & \mathcal{A}x = b \end{array} \qquad (\epsilon > 0),$$

where $\psi \in \{\psi_{\text{LP}}, \psi_{\text{SOCP}}, \psi_{\text{SDP}}\}$. A primal path-following method can be developed analogously as the dual path-following method, whereby at each iteration it solves

$$\begin{array}{ll} \min_u & \langle c, u\rangle + \epsilon\langle \nabla\psi(x), u\rangle + \frac{\epsilon}{2}\langle \nabla^2\psi(x)u, u\rangle \\ \text{s.t.} & \mathcal{A}u = 0 \end{array}$$

and updates $x_+ = x + u$ and $\epsilon_+ = \epsilon(1-\theta)$. However, this primal method does not appear to offer any advantage in practice. Unlike the dual method, it cannot exploit the sparsity of the data on SDP (because X is typically dense, even if C and $A_1, \ldots, A_m$ are sparse).

4.3 Primal-dual path-following method

Primal-dual methods, first suggested by Adler and Monteiro for LP in the late 80s, have the advantages of fewer iterations (in practice) and can be accelerated to achieve local superlinear convergence. These methods, instead of maintaining dual variables μ and linearizing the optimality equation the "dual form" (30) and (33), they maintain both primal variables x and dual variables μ and linearize the same equation written in the "primal-dual form" (31) and (34). This seemingly small change turns out to make a big difference in practice! A further practical improvement is to use a wider neighborhood, which allows the methods to take larger steps and converge faster (albeit at the expense of a somewhat worse iteration complexity).

Specifically, (31) and (34) can be written in the form

$$\mathcal{A}x = b, \quad y = \mathcal{A}^*\mu - c, \quad x \circ y = \epsilon e, \tag{45}$$

where $\circ$ is a suitable operator from $\mathbb{H} \times \mathbb{H}$ to $\mathbb{H}$ and e is the identity element with respect to $\circ$. For LP, we see from (31) that the "natural" choice of $\circ$ in (45) is

$$x \circ y = [x_i y_i]_{i=1}^n, \tag{46}$$

and e is the vector of 1's (so $x \circ e = e \circ x = x$ for all $x \in \mathbb{R}^n$). For SDP, a common choice of $\circ$ in (45) is

$$X \circ Y = X^{1/2} Y X^{1/2} \tag{47}$$

(or $X \circ Y = Y^{1/2} X Y^{1/2}$) and $e = I$. (Another choice is the Jordan product associated with semidefinite matrices, namely, $X \circ Y = (XY + YX)/2$. But this turns out to be more difficult to work with. For example, $X \succ 0$ and $Y \succ 0$ do not imply $(XY + YX)/2 \succ 0$.)

Primal-Dual Path-Following method

- Choose a neighborhood $\mathcal{N}$ and an initial $(x, y, \mu, \epsilon) \in \mathcal{N}$ and $\epsilon^{\text{final}} > 0$.
- While $\epsilon > \epsilon^{\text{final}}$:

  Solve the Newton equation

  $$\mathcal{A}u = 0, \quad v = \mathcal{A}^* w, \quad L_x u + L_y v = \sigma\epsilon e - x \circ y \tag{48}$$

  for u, v, w, where $0 \le \sigma \le 1$ and the linear mappings L_x, L_y (to be specified) are in some sense the partial derivatives of $x \circ y$ with respect to x and y. Set

  $$x[\alpha] = x + \alpha u, \quad y[\alpha] = y + \alpha, \quad \mu[\alpha] = \mu + \alpha w, \quad \epsilon[\alpha] = \epsilon(1 - \alpha\theta), \tag{49}$$

  where $0 \le \theta \le 1$ depends on σ and $\mathcal{N}$. Choose $\alpha > 0$ such that $(x[\alpha], y[\alpha], \mu[\alpha], \epsilon[\alpha]) \in \mathcal{N}$. Update

  $$(x_+, y_+, \mu_+, \epsilon_+) \quad = \quad (x[\alpha], y[\alpha], \mu[\alpha], \epsilon[\alpha]) \tag{50}$$

In popular implementations of the primal-dual method, the initial ϵ is chosen to satisfy $\epsilon = \frac{\langle x, y \rangle}{n}$, and this is maintained at subsequent iterations by setting

$$\epsilon[\alpha] = \frac{\langle x[\alpha], y[\alpha] \rangle}{n}. \tag{51}$$

We will see that (51) corresponds to (49) with $\theta = 1 - \sigma$. In this case, $\sigma < 1$ is needed for convergence. This choice has good practical performance especially when $\mathcal{N}$ is the so-called wide neighborhood.