CME307/MS&E311 Suggested Course Project III: First-Order Algorithms for Conic Optimization

(You don't need to answer all the questions posted in the project)

February 21, 2017

1 Optimization over Convex Cones

We consider the following optimization problem in the non-nagative cone:

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) \\ \text{Subject To} & \mathbf{x} \geq 0. \end{array} \tag{1}$$

Here we assume that $f(\mathbf{x})$ is a convex or non-convex function in $\mathbf{x} \in R^n$ and the minimizer $\mathbf{x}^*$ is attainable. Furthermore, we make a standard Lipschitz assumption such that

$$f(\mathbf{x}+\mathbf{d}) - f(\mathbf{x}) \leq \nabla f(\mathbf{x})^T d + \frac{\beta}{2}\|\mathbf{d}\|^2,$$

where positive β is the Lipschitz parameter.

Note that any linear feasibility problem,

$$A\mathbf{x} = \mathbf{b};$$
$$\mathbf{x} \geq 0.$$

can be formulated as the model with $f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$ and β would be the largest eigenvalue of matrix A^TA.

The dual linear programming feasibility problem can be casted as

$$A^T\mathbf{y} + \mathbf{s} = \mathbf{c};$$
$$\mathbf{x} \geq 0.$$

Substitute $\mathbf{y} = (AA^T)^{-1}A(\mathbf{c} - \mathbf{s})$, then it becomes

$$(I - A^T(AA^T)^{-1}A)(\mathbf{s} - \mathbf{c}) = 0;$$
$$\mathbf{s} \geq 0.$$

can be casted as $f(\mathbf{s}) = \frac{1}{2}(I - A^T(AA^T)^{-1}A)(\mathbf{c} - \mathbf{s})$ and the Lipschitz parameter is 1.

2 Steepest-Descent Affine-Scaling Interior-Point Algorithm

Let an iterate solution $\mathbf{x}^k > 0$. Then, we can scale it to $\mathbf{e}$, the vector of all ones, by

$$\mathbf{x}' = (X^k)^{-1}\mathbf{x}$$

where X^k is the diagonal matrix of vector $\mathbf{x}^k$. This is called Affine Scaling, which preserves the non-negativity. Consider the function in the scaled space:

$$f'(\mathbf{x}') = f(X^k\mathbf{x}') \quad \text{and} \quad \nabla f'(\mathbf{x}') = X^k \nabla f(X^k\mathbf{x}').$$

The new SDM iterate in the scaled space would be

$$\mathbf{x}'(\alpha) = \mathbf{e} - \alpha_k \nabla f'(\mathbf{e}) = \mathbf{e} - \alpha X^k \nabla f(\mathbf{x}^k)$$

and the one in the original space is

$$\mathbf{x}(\alpha) = \mathbf{x}^k - \alpha (X^k)^2 \nabla f(\mathbf{x}^k),$$

for some step-size α.

If function f is β-Lipschitz, then so is f' with $\beta\|\mathbf{x}^k\|_\infty^2$:

$$\begin{aligned} f'(\mathbf{x}') - f'(\mathbf{y}') - \nabla f'(\mathbf{y}')(\mathbf{x}' - \mathbf{y}') &= f(X^k\mathbf{x}') - f(X^k\mathbf{y}') - \nabla f(X^k\mathbf{y}')X^k(\mathbf{x}' - \mathbf{y}') \\ &\le \tfrac{\beta}{2}\|X^k(\mathbf{x}' - \mathbf{y}')\|^2 \\ &\le \tfrac{\beta\|\mathbf{x}^k\|_\infty^2}{2}\|(\mathbf{x}' - \mathbf{y}')\|^2. \end{aligned}$$

In order to keep each iterate in the interior the non-negative cone, our selection would be

$$\alpha^k = \min\{\frac{1}{\beta\|\mathbf{x}^k\|_\infty^2}, \frac{1}{2\|X^k\nabla f(\mathbf{x}^k)\|}\}.$$

Question 1: Show that the step-size strategy would keep the next iterate positive.

(In practice, one can start $\alpha = \frac{1}{\beta\|\mathbf{x}^k\|_\infty^2}$. If the new iterate is not positive, then let $\alpha := \alpha/2$ till the new iterate to be positive.)

Question 2: Show that, assigning $\mathbf{x}^{k+1} = \mathbf{x}(\alpha^k) > 0$ one has

$$f(\mathbf{x}^{k+1} - f(\mathbf{x}^k) \le \frac{-1}{2\beta\|\mathbf{x}^k\|_\infty^2}\|X^k\nabla f(\mathbf{x}^k)\|^2$$

or

$$f(\mathbf{x}^{k+1} - f(\mathbf{x}^k) \le \frac{-1}{4}\|X^k\nabla f(\mathbf{x}^k)\|.$$

What is the convergence speed of the problem?

3 Steepest-Descent Potential Reduction Interior-Point Algorithm

We now consider the problem with the logarithmic barrier function:

$$\phi(\mathbf{x}) = f(\mathbf{x}) - \mu \sum_j \ln(x_j),$$

where μ is a fixed positive constant, and we assume that the potential value is bounded below by ϕ^*. Let us start from $\mathbf{x}^0 = \mathbf{e}$, the vector of all ones, and generate a sequence of points $\mathbf{x}^k > 0$, $k = 1, ...,$, whose potential value is strictly decreased. We now describe a first order steepest descent potential reduction algorithm.

Note that the gradient vector of the potential function of $\mathbf{x} > 0$ is

$$\nabla\phi(\mathbf{x}) = \nabla f(\mathbf{x}) - \mu X^{-1}\mathbf{e}.$$

Thus, the first-order optimality condition is

$$X\nabla f(\mathbf{x}) = \mu\mathbf{e}, \quad \nabla f(\mathbf{x}) > 0, \quad \mathbf{x} > 0.$$

The following lemma is well known in the literature of interior-point algorithms:

Lemma 1. *Let* $\mathbf{x}^k > 0$ *and* $\|(X^k)^{-1}\mathbf{d}\|_\infty \le \delta < 1$. *Then*

$$-\sum_j \ln(x_j^k + d_j) + \sum_j \ln(x_j^k) \le -\mathbf{e}^T(X^k)^{-1}\mathbf{d} + \frac{1}{2(1-\delta)}\|(X^k)^{-1}\mathbf{d}\|^2.$$

Again for any given $\mathbf{x}^k > 0$,

$$\begin{aligned} f(\mathbf{x}^k + d) - f(\mathbf{x}^k) &\le \nabla f(\mathbf{x}^k)^T\mathbf{d} + \tfrac{\beta}{2}\|\mathbf{d}\|^2 \\ &= \nabla f(\mathbf{x}^k)^T d + \tfrac{\beta}{2}\|(X^k)(X^k)^{-1}\mathbf{d}\|^2 \\ &\le \nabla f(\mathbf{x}^k)^T\mathbf{d} + \tfrac{\beta\|\mathbf{x}^k\|_\infty^2}{2}\|(X^k)^{-1}\mathbf{d}\|^2. \end{aligned}$$

Furthermore, if $\|(X^k)^{-1}\mathbf{d}\|_\infty \le \delta = 1/2$ so that $\mathbf{x}^+ = \mathbf{x}^k + \mathbf{d} = X^k(\mathbf{e} + (X^k)^{-1}\mathbf{d}) > 0$. Then, applying the above inequality and Lemma 1 we have

$$\begin{aligned} \phi(\mathbf{x}^+) - \phi(\mathbf{x}^k) &\le \nabla f(\mathbf{x}^k)^T\mathbf{d} + \tfrac{\beta\|\mathbf{x}^k\|_\infty^2}{2}\|(X^k)^{-1}\mathbf{d}\|^2 + \mu(-\mathbf{e}^T(X^k)^{-1}\mathbf{d} + \|(X^k)^{-1}\mathbf{d}\|^2) \\ &= \nabla\phi(\mathbf{x}^k)^T\mathbf{d} + \tfrac{\beta\|\mathbf{x}^k\|_\infty^2 + 2\mu}{2}\|(X^k)^{-1}\mathbf{d}\|^2. \end{aligned}$$

Now we let

$$\mathbf{d}^k = -\alpha^k(X^k)^2\nabla\phi(\mathbf{x}^k),$$

where

$$\alpha^k = \min\{\frac{\|X^k\nabla\phi(\mathbf{x}^k)\|}{\beta\|\mathbf{x}^k\|_\infty^2 + 2\mu},\ \frac{1}{2\|X^k\nabla\phi(\mathbf{x}^k)\|}\}.$$

Now we have

$$\nabla\phi(\mathbf{x}^k)^T\mathbf{d}^k = -\alpha^k\|X^k\nabla\phi(\mathbf{x}^k)\|^2,$$

so that if $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k$ we have

$$\phi(\mathbf{x}^{k+1}) - \phi(\mathbf{x}^k) \le -\min\{\frac{\|X^k \nabla\phi(\mathbf{x}^k)\|^2}{2(\beta\|\mathbf{x}^k\|_\infty^2 + 2\mu)}, \frac{\|X^k \nabla\phi(\mathbf{x}^k)\|}{4}\}.$$

Question 3: Show the following

Theorem 2. *Let $\mu = \epsilon$ and $\|\mathbf{x}^k\|_\infty$ be bounded above by R for the iterative sequence. Then, in no more than $O(\frac{\beta R^2 + 2\epsilon}{\epsilon^2}(\phi(\mathbf{x}^0) - \phi^*))$ iterations the steepest descent potential reduction algorithm generates a $\mathbf{x}^k > 0$ such that $\nabla f(\mathbf{x}^k)^T \mathbf{x}^k / n < 2\epsilon$ and $\nabla f(\mathbf{x}^k) \ge 0$.*

4 Affine-Scalling and Potential Reduction for SDP cone

Now consider the SDP cone where we solve for $X \in S^n$:

$$\begin{array}{ll} \text{Minimize} & f(X) \\ \text{Subject To} & X \succeq 0, \end{array} \tag{2}$$

We assume that $f(X)$ is β-Lipschitz, that is, for any $D \in S^n$,

$$f(X + D) - f(X) \le \nabla f(X) \bullet D + \frac{\beta}{2}\|D\|_f^2,$$

where $\|.\|_f$ is the Frobenius norm.

For example, the sensor network localization problem can be casted as such a problem with

$$f(X) = \frac{1}{2}\|\mathcal{A}X - \mathbf{b}\|^2$$

for given data $A_i \in S^n$ for $i = 1, ..., m$, and $\mathbf{b} \in R^m$. Recall that

$$\mathcal{A}X = \begin{pmatrix} A_1 \bullet X \\ ... \\ A_m \bullet X \end{pmatrix} \quad \text{and} \quad \mathcal{A}^T\mathbf{y} = \sum_{i=1} y_i A_i.$$

Note that $\nabla f(X) = \mathcal{A}^T(\mathcal{A}X - \mathbf{b})$ which is also a symmetric matrix.

Let an iterate $X^k \succ 0$. Then we can scale it to I (the identity matrix) by

$$X' = (X^k)^{-1/2} X (X^k)^{-1/2}$$

Then,

$$f'(X') = f((X^k)^{1/2} X' (X^k)^{1/2}) \quad \text{and} \quad \nabla f'(I) = (X^k)^{1/2} \nabla f(X^k)(X^k)^{1/2},$$

the new SDM iterate in the scaled space is

$$X'(\alpha) = I - \alpha(X^k)^{1/2} \nabla f(X^k)(X^k)^{1/2}$$

and in the original space

$$X(\alpha) = X^k - \alpha X^k \nabla f(X^k) X^k,$$

for some step-size α.

The optimization with the logarithmic barrier function for SDP cone would be

$$\phi(X) = f(X) - \mu \ln(\det(X)),$$

where μ is the fixed positive constant, and we assume that the potential value is bounded below by ϕ^*. Note that the gradient vector of the potential function of $X^k \succ 0$ is

$$\nabla\phi(X^k) = \nabla f(X^k) - \mu(X^k)^{-1}.$$

Question 4: Extend the two algorithms, the early described affine-scaling and potential reduction for the non-negative cone, to solving problem over the SDP cone, starting from $X^0 = I$ and generating interior-point matrices $X^k \succ 0$, $k = 1, ...,$. Produce similar results in **Question 1**, **Question 2** and **Question 3** for the SDP cone case.

Question 5: Implement the algorithm and perform numerical tests to solve

$$f(X) = \frac{1}{2}\|\mathcal{A}X - \mathbf{b}\|^2, \text{ s.t. } X \succeq 0.$$

Not that if a good step-size strategy is set, then no matrix inverse is never needed in computation which would be suitable for solving large-scale SDP optimization problems. Furthermore, if each data matrix A_i is rank-one, that is, $A_i = \mathbf{a}_i \mathbf{a}_i^T$ (as in sensor network localization), $X^k A_i X^k = X^k \mathbf{a}_i \mathbf{a}_i^T X^k$ so that you need only compute a matrix-vector multiplication, $X^k \mathbf{a}_i$, for each data matrix.

You may consider a one-time preconditioning of the problem to improve the Lipschitz constant. Let

$$M = \mathcal{A}\mathcal{A}^T = \begin{pmatrix} A_1 \bullet A_1 & ... & A_1 \bullet A_m \\ ... & ... & ... \\ A_m \bullet A_1 & ... & a_m \bullet A_m \end{pmatrix},$$

land $R^T R = M^{-1}$. Then let

$$\bar{A}_i = R A_i R^T, \ i = 1, ..., m,$$

and $\bar{\mathbf{b}} = M^{-1}\mathbf{b}$. Then you minimize

$$f(X) = \frac{1}{2}\|\bar{\mathcal{A}}X - \bar{\mathbf{b}})\|^2, \text{ s.t. } X \succeq 0,$$

References

[1] C. C. Gonzaga, Polynomial affine algorithms for linear programming, *Math. Programming* 49 (1990) 7–21.

[2] N. Karmarkar, A new polynomial-time algorithm for linear programming, *Combinatorica* 4 (1984) 373-395.

[3] S. Mehrotra. On the implementation of a primal–dual interior point method. *SIAM J. Optimization*, 2(4):575–601, 1992.

[4] Y. Ye, M. J. Todd, and S. Mizuno, An $O(\sqrt{n}L)$ - iteration homogeneous and self-dual linear programming algorithm, *Math. Oper. Res.* 19 (1994) 53–67.

[5] Y. Ye, On the complexity of approximating a KKT point of quadratic programming, *Math. Programming* 80 (1998) 195-211.