Spectral bundle method: part 1 of > 1

March 20, 2018

Main citation: [Helmberg and Rendl, 2000]

1 Problem formulation

- Main problem and Lagrange dual of standard linear semidefinite optimization problem (SDP)

$$\begin{array}{lll} \underset{X}{\text{maximize}} & \text{tr}(CX) & \underset{y,Z}{\text{minimize}} \quad b^T y \\ \text{subject to} & \mathcal{A}(X) = b & \text{subject to} \quad Z = \mathcal{A}^T(y) - C \\ & X \succeq 0 & Z \succeq 0 \end{array} \tag{1}$$

  Here, X is $n \times n$ symmetric. The notation $X \succeq 0$ means X is positive semidefinite, i.e. $u^T X u \geq 0$ for all u. The linear operators

$$\mathcal{A}(X)_k = \text{tr}(A_k X),\ k = 1, \ldots, m. \quad \mathcal{A}^*(y) = \sum_{i=1}^{m} y_k A_k$$

  You can verify that $\mathcal{A}$ and $\mathcal{A}^*$ are *adjoint operators*, e.g.

$$\langle \mathcal{A}(X), y \rangle = \langle X, \mathcal{A}^*(y) \rangle$$

  and here $\langle X, Y \rangle = \mathbf{vec}(X)^T \mathbf{vec}(Y) = \text{tr}(X^T Y)$ is a (standard) way of defining an inner product over matrices.

- For Lagrange duals, X^* and Z^* admit a simultaenous eigendecomposition. That is, if the eigendecomposition of X^* is

$$X^* = U \text{diag}(\lambda_X) U^T$$

  then

$$Z^* = U \text{diag}(\lambda_Z) U^T$$

  and by complementary conditions, $(\lambda_X)_i (\lambda_Z)_i = 0$ for $i = 1, \ldots, n$. Since for strong duality we have strict complementarity, then at least one is always nonzero.

- **Background** Without thinking too hard, there are two main ways of solving either problem in (1):
  - Interior point method. Then at each iteration we compose a KKT system, which solves a system of equations at each point equal to # variables (n^2) + constraints (m). Recall that a direct solve of a system $Ax = b$ is cubic in the number of rows /cols of A, so the per iteration complexity is $(n^2 + m)^3$, which, well, if n is like a million, is prohibitive. [1]

[1] In reality, anyone proposing to use an interior point method to solve an SDP thought long and hard about sparsity and structure, so there exist pretty good solvers out there. But, they're still not like super scalable.

– A first order method like projected gradient, or ADMM, or Douglas-Rachford (very similar to ADMM). Here, you're bottlenecked either by projecting on the linear constraint, or by projecting on the postive semidefinite cone $\{X : X \succeq 0\}$. Taking the eigendecomposition $X = U\Lambda U^T$,

$$\text{proj}_{\{X:X\succeq 0\}} = \sum_{i=1}^{n} \max\{\lambda_i, 0\} u_i u_i^T$$

where

$$U = [u_1, \ldots, u_n], \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

For SDPs, unless m is super huge, usually this is the thing that prohibits you. Based on my past experience, if you don't like the smell of burning metal, $n <\approx 10000$ for dense problems.

– Now, if you assume low rank, you have much better tools available to you. Specifically, maybe you only care about the top r eigenvalue / eigenvector pairs, rather than a full eigendecomposition. So, maybe we do something like the power method: for a matrix X, pick a random vector z^0, and do

$$y^k := Xz^k, \quad z^{k+1} = \frac{1}{\|y^k\|_2} y^k$$

until it converges $z^k \to z$. Then $z^T X z$ is the top eigenvalue of X and z is its associated eigenvector. This happens pretty fast, and if X is sparse, is easy to do. Moral of the story: eig (full decomposition) = bad, eigs (very few eigenvalues) = good.

- **Reformulations of** (1) Replace $Z \succeq 0$ with $\lambda_{\max}(-Z) \leq 0$. By strict complementarity, since $Z \succ 0 \iff X = 0$. So, by ensuring $X^* \neq 0$ we can assume $\lambda_{\max}(-Z) = 0$.

  (1) is therefore equivalent to

$$\begin{array}{ll} \underset{y,Z}{\text{minimize}} & b^T y \\ \text{subject to} & \lambda_{\max}(C - \mathcal{A}^T(y)) = 0 \end{array} \tag{2}$$

  Now suppose that $\mathcal{A}(X)$ includes a trace constraint. Say, $\text{tr}(A_0 X) = \text{tr}(X) = b_0$. Then $\mathcal{A}^*(y) = y_0 I + \sum_{k=1}^m y_k A_k$ and the Lagrangian of (2) is

$$a\lambda_{\max}(C - \sum_{k=1}^m y_k A_k - y_0 I) + b_0 y_0 + \sum_{k=1}^m b_k y_k = a\lambda_{\max}(C - \sum_{k=1}^m y_k A_k) + \sum_{k=1}^m b_k y_k + (b_0 y_0 - a y_0)$$

  which, when minimized over y, will be $-\infty$ unless $a = b_0$. Therefore, taking $a = b_0$, (1) is equivalent to

$$\min_y \; a\lambda_{\max}(C - \mathcal{A}^*(y)) + b^T y. \tag{3}$$

  So the remainder of the paper is about solving (3).

2 Subdifferentials and ϵ-subdifferentials

- **Definition:** A *subgradient* of f at x is defined as the set of w where

$$f(y) - f(y) \geq w^T (y - x), \; \forall y \in \text{dom} f. \tag{4}$$

  We denote the set of subgradients of f at x as $\partial f(x)$.

  – When $w = \nabla f(x)$ then property (4) is the equivalent first order condition of convexity of f. So, nothing surprising.

  – As long as f is convex, this set always exists and is nonempty.

- If $f(x)$ is smooth, then $\partial f(x) = \{\nabla f(x)\}$ (a singleton).
- If $f(x)$ is nonsmooth, then $\partial f(x)$ is a convex set. A popular example is $f(x) = |x|$ (absolute value). Then

$$\partial f(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ [-1, 1] & x = 0 \end{cases}$$

- **Claim:** The subdifferential of $\lambda_{\max}(X)$ is

$$\partial \lambda_{\max}(X) = G := \{W : \langle W, X \rangle \geq \lambda_{\max}(X),\ \mathrm{tr}(W) = 1,\ W \succeq 0\}$$

Proof: If X's top eigenvalue is isolated, then it is clear from the definition of a maximum eigenvalue that $\partial \lambda_{\max}(X) = \{uu^T\}$ where $u = \mathrm{evec}_{\max}(X)$ (eigenvector associated with top eigenvalue).

If X's top eigenvalue has multiplicity $r > 1$, then

$$\partial \lambda_{\max}(X) = \mathrm{conv}(u_i u_i^T),\ i = 1, \ldots, r = G.$$

[2]

- **Definition:** An ϵ subgradient is defined as a w such that

$$f(y) - f(x) \geq w^T(y - x) - \epsilon, \forall y \in \mathrm{dom} f$$

Note that this guy is with a prespecified $\epsilon > 0$.

- **Claim:** The ϵ subdifferential of $\lambda_{\max}(X)$ is

$$\partial \lambda_{\max}(X) = G_\epsilon := \{W : \langle W, X \rangle \geq \lambda_{\max}(X) - \epsilon,\ \mathrm{tr}(W) = 1,\ W \succeq 0\}$$

I personally don't believe that there is an easy way of deriving this subdifferential, but some of the efforts is included in the end. (Not a full proof, though.)

However, the authors are nice enough to point to [Ye, 1993] for verification.

- Note also that the ϵ subgradient of a linear function must be just the subgradient, since you can always choose y and $\tilde{y}$ such that $y - x = -(\tilde{y} - x)$. Therefore

$$b^T y - b^T x \geq g^T(y - x) - \epsilon \iff g = b$$

- Now taking $f(y) := a\lambda_{\max}(C - \mathcal{A}^*(y)) + b^T y$,

$$\partial_\epsilon f(y) = \{W : \mathrm{tr}(W(C - \mathcal{A}^*(y))) \geq \lambda(C - \mathcal{A}^*(y)) - \epsilon/a,\ \mathrm{tr}(W) = 1,\ W \succeq 0\}.$$

- Big idea, quoted from paper: "We will see that the ϵ-subdifferential of eigenvalue problems has the form of the feasible set of a semidefinite program. This suggests to use, instead of the traditional polyhedral cutting plane method, a semidefinite cutting plane model that is formed by restricting the feasible set of ϵ-subgradients to an appropriate face[3] of the semidefinite cone. This specialization of the cutting plane model is the main contribution of the paper."

3 Next time...

Traditional bundle method of Kiwiel [Kiwiel, 1990], with serious vs null step, aggregation, etc.

Would looking at the Pataki result be of interest as well, to motivate low rank solutions?

[2] For your viewing pleasure, I have included a way more involved proof at the end.

[3] A face of a PSD cone $\{X \succeq 0\}$ refers to a lower dimensional "corner", e.g. $\{PVP^T : V \succeq 0\}$ and P is tall and prespecified.

4 A way more insane proof for deriving subdifferentials

- **Lemma 1**:

$$\lambda_{\max}(Y) \geq \text{tr}(WY),\ \forall Y \in S_n \quad \iff \quad \text{tr}(W) = 1,\ W \succeq 0$$

Proof: ($\Rightarrow$) Pick an adversarial Y. Define $Y = cI$. Then

$$\text{tr}(WY) = c\text{tr}(W) = \lambda_{\max}(Y)\text{tr}(W)$$

If $c > 0$ then this implies $\text{tr}(W) \leq 1$. If $c < 0$, then $\text{tr}(W) \geq 1$. Therefore it must be that $\text{tr}(W) = 1$.

Now pick a different adversarial $Y = U\text{diag}(\sigma)U^T$ where $W = U\text{diag}(\lambda)U^T$ is the eigendecomposition of W. Define

$$\sigma_i = \begin{cases} 0 & \lambda_i \geq 0 \\ -1 & \lambda_i < 0. \end{cases}$$

Then

$$\text{tr}(WY) = -\sum_{i:\lambda_i<0} \lambda_i \geq 0$$

Suppose W has a single negative eigenvalue. Then $\lambda_{\max}(Y) = -1 < 0$ which is a contradiction. Then the only way this is satisfied is if W has no negative eigenvalues, and thus defined in this way, $Y = 0$

($\Leftarrow$) If $W \succeq 0$ and $\text{tr}(W) = 1$ then $\|W\|_* = \text{tr}(W) = 1$. Using Cauchy Schwartz we have

$$\text{tr}(WY) \leq \|W\|_*\|Y\|_2 = \|Y\|_2 \text{ and } \text{tr}(WY) \leq \|-W\|_*\|-Y\|_2 = \|-Y\|_2$$

Since $\lambda_{\max} \geq \min\{\|Y\|_2, \|-Y\|_2\}$, then the claim is proven.

- **Lemma 2**: For some $X \succeq 0$, $\lambda_{\max}(X) > 0$, $\epsilon > 0$

$$\lambda_{\max}(Y) - \lambda_{\max}(X) \geq \text{tr}(W(Y-X)) - \epsilon,\ \forall Y \Rightarrow W \text{ is not negative definite.}$$

Proof: If $W \prec 0$ then $\text{tr}(XW) \leq 0$. Pick

$$Y = cI - X, \quad c = \frac{-(1+\epsilon)}{|\text{tr}(W)|} < 0.$$

Then $\text{tr}(W) < 0$ and

$$\lambda_{\max}(Y) - \text{tr}(WY) = 0 + c|\text{tr}(W)| + \text{tr}(WX) = -(1+\epsilon) + \text{tr}(WX)$$

and

$$\lambda_{\max}(Y) - \text{tr}(WY) \geq \lambda_{\max}(X) - \text{tr}(WX) - \epsilon \iff -1 \geq \lambda_{\max}(X) - 2\text{tr}(WX) \geq \lambda_{\max}(X)$$

which can't be true as long as $\lambda_{\max}(X) > 0$.

- **Lemma 3**: For some $X \succeq 0$, $\lambda_{\max}(X) > 0$, $\epsilon \geq 0$

If W is not negative definite,

$$\lambda_{\max}(Y) - \lambda_{\max}(X) \geq \text{tr}(W(Y-X)) - \epsilon,\ \forall Y \Rightarrow \|W\|_* = 1$$

where $\|W\|_* = \sum_i |\lambda_i|$ where λ_i are the eigenvalues of W.

Proof: Using the same eigendecomposition $W = U\text{diag}(\lambda)U^T$, pick again 1 $Y = cU\text{diag}(\mathbf{sign}(\lambda))U^T$. to get

$$c(1 - \|W\|_*) \geq \lambda_{\max}(X)(1 - \|W\|_*) - \epsilon$$

by Cauchy Schwartz.

Consider $1 \neq \|W\|_*$, and pick c as

$$c = \frac{\lambda_{\max}(X)(1 - \|W\|_*) - \epsilon - 1}{1 - \|W\|_*}$$

Then

$$c(1 - \|W\|_*) \geq \lambda_{\max}(X)(1 - \|W\|_*) - \epsilon \iff -1 \geq 0.$$

Therefore $\|W\|_* = 1$.

- **Subdiff of** $\lambda_{\max}$: If $X \succeq 0$ and $\lambda_{\max}(X) > 0$ then the subdifferential of $\lambda_{\max}$ is

$$\partial\lambda_{\max}(X) = G := \{W : \langle W, X\rangle \geq \lambda_{\max}(X),\ \mathrm{tr}(W) = 1,\ W \succeq 0\}$$

Proof: Lemma 1 gives $G \subseteq \partial\lambda_{\max}(X)$.

Lemma 2,3 shows that if $W \in \partial\lambda_{\max}(X)$, then $\|W\|_* = 1$.

Then

$$\mathrm{tr}(WX) \leq \|W\|_*\lambda_{\max}(X) = \lambda_{\max}(X)$$

which implies

$$\lambda_{\max}(Y) \geq \mathrm{tr}(WY),\ \forall Y \iff W \in G.$$

Invoking lemma 1 again gives the final result.

- ϵ **subdiff of** $\lambda_{\max}$: The ϵ-subdifferential for $\lambda_{\max}(X)$ is

$$\partial_\epsilon\lambda_{\max}(X) = G_\epsilon := \{W : \langle W, X\rangle \geq \lambda_{\max}(X) - \epsilon,\ \mathrm{tr}(W) = 1,\ W \succeq 0\}$$

Proof: [4]

Clearly if $W \in G_\epsilon$, then by lemma 1, $W \in \partial_\epsilon(\lambda_{\max}(X))$.

Again, we can use Lemma 2, 3 to assist in the other direction. But we can't use Lemma 1 anymore...

5 Further reading

References

[Helmberg and Rendl, 2000] Helmberg, C. and Rendl, F. (2000). A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696.

[Kiwiel, 1990] Kiwiel, K. C. (1990). Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical programming*, 46(1-3):105–122.

[Ye, 1993] Ye, D. (1993). Sensitivity analysis of the greatest eigenvalue of a symmetric matrix via the - subdifferential of the associated convex quadratic form. *Journal of optimization theory and applications*, 76(2):287–304.

[4] Ye. Sensitivity analysis of the greatest eigenvalue of a symmetric matrix via the ϵ-subdifferential of the associated convex quadratic form.