10-725: Optimization **Fall 2012**

Lecture 20: November 1st

Lecturer: Geoff Gordon *Scribes: Xiaolong Shen, Alex Beutel*

20.1 Administrivia

- HW3 back at end of class
- Last day for feedback survey
- All lectures now up on Youtube
- Reminder: midterm next Tuesday 11/6! in class test with 1 hr 20 mins, one sheet of notes.

20.2 Introduction and Defination

We are talking about Quadratic programs and cone programs. Starting with definition, this lecture will extend to the application of these optimization methods. Actually tons of research in Machine Learning and Statistics are trying to convert interesting problems to Quadratic Programs and Cone Programs. Some familiar problems is group lasso which is second order cone programs and SVM which is quadratic programs.

20.3 Quadratic Programs

- m constraints and n vars
  - $A : R^{m \times n}$ Constrain Matrix
  - $b : R^m$ Right hand side of constrains
  - $c : R^n$ Objective Vectors
  - $x : R^n$ Random Variables
  - $H : R^{n \times n}$ Quadratic Part of Objective
  - [min or max] $x^T Hx/2 + c^T x$
  - subject to $Ax \leq b$ or $Ax = b$
  - may have $x \geq 0$

- Typical Example:
  Maximize a convex function:

$$\begin{aligned} \max\ 2x + x^2 + y^2 s.t. \\ x + y \leq 4 \\ 2x + 5y \leq 12 \\ x + 2y \leq 5 \\ x, y \geq 0 \end{aligned}$$

- The Optimization Problem is Convex Problem if:
  - min $H \succeq 0$
  - max $H \preceq 0$

20.4 Cone Programs

- m constraints and n vars
  - $A : R^{m \times n}$ Constrain Matrix
  - $b : R^m$ Right hand side of constrains
  - $c : R^n$ Objective Vectors
  - $x : R^n$ Random Variables
  - Cones $K \subseteq R^m$ $L \subseteq R^n$
  - [min or max] $c^T x$
  - subject to $Ax + b \in K$, $x \in L$
  - convex problem if cones K and L are both convex
- Typical Example:
  - Typical Example of K: $K = \{0\}^p * R_+^q$ First p elements representing equality constrains and later q elements representing inequality constrains
  - Typical Example of L: $L = \{0\}^{p'} * R_+^{q'} * R^{q''}$

  Then with this constrains we get a ordinary LPs, if with more general constrains, we can have more general optimization problems.

20.4.1 SOCP

Mathematical Problem is min $c^T x$ s.t. $A_i x + b_i \in K_i, i = 1, 2, ...$ Where maximize is also possible.

- $A_i \in R^{m_i * n}$
- $x \in R^n$
- $b_i \in R^{m_i}$

- K_i should be some combination of equality constrains($\{0\}$), inequality constrains (R_+) and Secondary Order Cones (SOC)
  - $SOC : \{(y,t)|\|y\| \leq t\}$
  - $y : R^{m_i}$
  - $t : R$
  - $SOC : R^{m+1}$

As the object function is linear, we are only possible to get the optimal point at the boundary of the feasible region.

20.4.2 Relationship between QPs and SOCPs

20.4.2.1 QPs are reducible to SOCPs

The QP is

$$\begin{aligned} &\min\ x^T Hx/2 + c^T x \\ &\text{s.t.}\ \ \text{Some Constrains.} \end{aligned}$$

Then change the minimization problem to :

$$\begin{aligned} &\min\ t + c^T x \\ &\text{s.t.}\ t \geq \frac{x^T Hx}{2},\ \text{Some Constrains.} \end{aligned}$$

The goal is to implement the $t \geq \frac{x^T Hx}{2}$ using a Secondary Order Cone Constrains.
Start by get:

$$\begin{aligned} H &= R^T R \\ \text{then } x^T Hx &= (x^T R^T)(Rx) \\ &= \|Rx\|_2^2 \end{aligned}$$

So we need to constrain Rx,t,t+1

$$\begin{aligned} (Rx, t, t+1) &\in SOC \\ \text{which means } t+1 &\geq \sqrt{\|Rx\|_2^2 + t^2} \\ \text{square both side } t^2 + 2t + 1 &\geq \|Rx\|_2^2 + t^2 \\ t &\geq \frac{\|Rx\|_2^2}{2} - \frac{1}{2} \end{aligned}$$

as the $\frac{1}{2}$ is completely irrelevant and $\|Rx\|_2^2 = x^T Hx$

$$t \geq \frac{x^T Hx}{2}$$

20.4.2.2 Are all SOCPs are QPs?

The answer is actually no.
For example, the QCQP can be a reasonable instance. QCQP represents convex quadratic objective and constraints.
For a optimization problem:

$$\begin{aligned} \text{minimize } & a^2 + b^2 \\ \text{s.t. } & a \geq x^2, b \geq y^2 \\ & 2x + y = 4 \end{aligned}$$

This is not a QP as it has quadratic constraints. Actually from the nonlinear constraints we can reformulate the problem to:

$$\begin{aligned} \text{minimize } & x^4 + y^4 \\ \text{s.t. } & 2x + y = 4 \end{aligned}$$

It is a l-4 norm minimization problem which is different from a QP.
If we want to make it to SOCP, we are solving min $p + q$ where $p \geq a^2$ and $q \geq b^2$.

$$\begin{aligned} \text{So we need to constrain a,p,p+}\tfrac{1}{2} & \\ (a, p, p + \frac{1}{2}) &\in SOC \\ \text{which means } p + 0.5 &\geq \sqrt{a^2 + p^2} \\ \text{square both side } p^2 + p + 0.25 &\geq a^2 + p^2 \\ p &\geq a^2 - 0.25 \end{aligned}$$

as the $\frac{1}{4}$ is completely irrelevant, we can implement $p \geq a^2$. Similarly we can have $q \geq b^2$. Also we can use the intersecting of a cone and a plane to make the constrain of $a \geq x^2$ and $b \geq y^2$ in a similar way.

20.4.3 More Cone Programs: Semidefinite Program

First we make some definitions about semidefinite programs.

- Semidefinite Constraint:
  - variable $x \in R^n$
  - constant matrices $A_0, A_1, A_2, ... \in R^{m*m}$
  - constrain $A_0 + \sum_i x_i A_i \succeq 0$ and symmetric
  - constrain $A_0 + \sum_i x_i A_i \in S_+^m$
- Semidefinite program: min $c^T x$ with the constraints:
  - semidefinite constraints

- linear equalities
- linear inequalities

For a positive semidefinite cone S_+, it is self-dual:
Proof:
We surpress m in S_+^m to S_+:

$$S_+ : \{A | A = A^T, x^T Ax \geq 0 \text{ for all x } \}$$

To show that the dual is the same as the primal:

$$[x^T Ax \geq 0 \text{ for all x }] \Leftrightarrow [tr(B^T A) \geq 0 \text{ for all psd B}]$$

which demonstrate when $(A : B) \geq 0$ is true for all the elements in dual cone $K^* = psdB$, it is the same as being in the primal cone.
First, conduct $[x^T Ax \geq 0 \text{ for all x }] \Rightarrow [tr(B^T A) \geq 0 \text{ for all psd B}]$

$$\text{psd B} = \sum_i x_i x_i^T$$

Several matrix factorization methods like Singular Value Decomposition can make a positive semidefinite cone represented in this way.Then we will have:

$$\begin{aligned} tr(B^T A) &= \sum_i tr((x_i x_i^T)^T A) \\ &= \sum_i tr(x_i^T A x_i \\ &= \sum_i x_u^T A x_i \\ &\geq 0 \end{aligned}$$

Then, conduct $[x^T Ax \geq 0 \text{ for all x }] \Leftarrow [tr(B^T A) \geq 0 \text{ for all psd B}]$

$$B = xx^T$$

So we can have:

$$\begin{aligned} tr(B^T A) &= x^T Ax \\ &\geq 0 \end{aligned}$$

which shows that A is semidefinite. ■

20.5 Solving QPs and CPs

To solve a convex QP or CP is not much harder than LP as long as we have an efficient representation of the cone. So with a bit length of L and accuracy of ϵ we can get the QP or CP in time polynomial ploy(L, $\frac{1}{\epsilon}$).

But can these programs solved in time polynomial with bit length which is called strong polynomial ploy(L). Though people guess it's true, it still stays a open question which we can get our PhD degree by proving it!

For general QP or CP, they are NP-complete as to reduce the max cut to QP. So it shows us again that the convexity property is crucial in our problems.

20.6 Examples of QPs

- Euclidean projection
- Lasso–Mahalanobis projection
- Huber regression
- Support Vector Machine

20.6.1 Robuster Huber regression

Given points (x_i, y_i) where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ we have the following optimization problem (L_2 regression):

$$\min_w \sum_i (y_i - x_i^\top w)^2$$

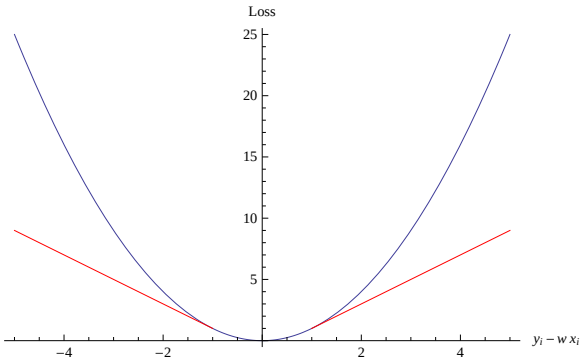
We would like to avoid overvfitting that could be caused by outliers, by using the Huber loss rather than the squared loss, so the optimization problem becomes

$$\min_w \sum_i \mathrm{Hu}(y_i - x_i^\top w)$$

where

$$\mathrm{Hu}(z) = \begin{cases} 2z - 1 & \text{if } z \geq 1 \\ -2z - 1 & \text{if } z \leq -1 \\ z^2 & \text{if } -1 < z < 1 \end{cases} \tag{20.1}$$

A plot comparing the two loss functions can be seen below where the red line is the Huber loss and the blue line is the squared loss.



We would like to now show that the Huber loss is in fact a quadratic program. We assume that we can compute it as the following QP and then show that it works out to the Huber loss function

$$\mathrm{Hu}(z_i) = \min_{a_i, b_i} (z_i + a_i - b_i)^2 + 2a_i + 2b_i \tag{20.2}$$

$$\text{s.t.} a_i, b_i \geq 0 \tag{20.3}$$

We find that taking the graident with respect to a and b we get

$$\nabla_a (z+a-b)^2 + 2a + 2b = 2(z+a-b) + 2 \tag{20.4}$$

$$\nabla_b (z+a-b)^2 + 2a + 2b = -2(z+a-b) + 2 \tag{20.5}$$

If we set these each to 0, we

When solving this we see there are four posibilities for the constraints:

1. Both constraints are tight $a = b = 0$ implies $\text{Hu}(z) = z^2$
2. If we assume both constraints are not tight then both of their derivatives must equal 0. We see adding the two derivatives that $4 = 0$ which is impossible and thus we can not have that both constraints are not tight.
3. $a > 0$ and $b = 0$ implies $a = z - 1$ and therefore $\text{Hu}(z) = -1 - 2z$. This is under the constraint that $a = z - 1 \geq 0$ and thus it is only true if $z \leq -1$.
4. Taking the symmetric case of $a = 0$ and $b > 0$ we get the other branch for $z \geq 1$.

20.6.2 Cone program examples

Sparse Group Lasso: We can show that (sparse) group lasso is a SOCP. Group lasso is defined as: given groups $g_i \subseteq \{1, \dots n\}$, we have the optimization problem

$$\min \|y - Xw\|_2^2 + \lambda \sum_i \|w_{g_i}\|_2 \tag{20.6}$$

where w_{g_i} is the subvector of w corresponding to the indices g_i.

Similarly sparse group lasso is defined as:

$$\min \|y - Xw\|_2^2 + \lambda \sum_i \|w_{g_i}\|_2 + \mu \|w\|_1. \tag{20.7}$$

We can formulate this as the following SOCP as:

$$\min\ t + \lambda \sum_i t_i \tag{20.8}$$

$$\text{s.t. } t \geq \|y - Xw\|_2^2 \tag{20.9}$$

$$t_i \geq \|w_{g_i}\|_2 \tag{20.10}$$

Inference in a discrete Markov random field can be relaxed to a SOCP [Kumar, Kolmogorov, Torr, JMLR 2008].

Minimum volume covering ellipsoid: We represent an ellipsoid as $\{x \mid \|Ax + b\| \leq 1\}$. We can write this as a second order cone constraint $\binom{Ax_i + b}{1} \in \text{SOC}$. The volume of the elliposed is $|\det A^{-1}|$, so we can set up our optimization problem as

$$\min\ \ln |\det A^{-1}| \tag{20.11}$$

This is a convex function, making this a second order cone program, but with a non-linear objective.

SDP: We can also frame a number of problems as SDPs:

- graphical LASSO (nonlinear objective):

$$\min \operatorname{tr}(S^\top X) - \log\det X + \lambda \sum_{i \neq j} |x_{ij}| \tag{20.12}$$

  where S is the empirical covariance, and requiring that $X \succeq 0$ and X is symmetric.

- Markowitz portofolio optimization
- Max-cut relaxation [Goemans,Williamson]
- Matrix completion
- Manifold learning through max variance unfolding

Matrix completition: We have a matrix A where we observe some elements A_{ij} for $(i,j) \in E$. We write

$$O_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if otherwise} \end{cases} \tag{20.13}$$

Our objective normally is:

$$\min_X \|(X - A) \circ O\|_F^2 + \lambda \|X\|_* \tag{20.14}$$

(As a recap, the nuclear norm term is in an effort to get a low rank X.) We would like to make this into a semidefinite program, by showing that $\lambda\|X\|_*$ is the same as $\lambda(\operatorname{tr}(P) + \operatorname{tr}(Q))/2$ subject to:

$$M = \begin{pmatrix} P & X \\ X^\top & Q \end{pmatrix} \succeq 0. \tag{20.15}$$

To show this we begin by decomposing $X = U\Sigma V^\top$. We know that $M \succeq 0$ is equivalent to $\operatorname{tr}(B^\top M) \geq 0$ for all $B \succeq 0$ (because PSDs are self dual). So if we take

$$B = \begin{pmatrix} UU^\top & -UV^\top \\ -VU^\top & VV^\top \end{pmatrix} \succeq 0. \tag{20.16}$$

then we find

$$0 \leq \operatorname{tr}(B^\top M) = \operatorname{tr}(UU^\top P) + \operatorname{tr}(VV^\top Q) - 2\operatorname{tr}(VU^\top X) \tag{20.17}$$

$$\operatorname{tr}(P) + \operatorname{tr}(Q) \geq 2\operatorname{tr}(U^\top XV) = 2\operatorname{tr}(\Sigma) = 2\|X\|_* \tag{20.18}$$

If we set $P = U\Sigma U^\top$ and $Q = V\Sigma V^\top$ then we find $\operatorname{tr}(P) = \operatorname{tr}(U^\top U\Sigma) = \operatorname{tr}(\Sigma)$ and similarly for Q. Then $\operatorname{tr}(P) + \operatorname{tr}(Q) = 2\|X\|_*$.

Finally we must show that this choice of P and Q make M PSD. We now have

$$M = \begin{pmatrix} U\Sigma U^\top & -U\Sigma V^\top \\ -V\Sigma U^\top & V\Sigma V^\top \end{pmatrix} \tag{20.19}$$

We can apply an orthogonal transformation and keep the trace norm:

$$\begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^\top \begin{pmatrix} U\Sigma U^\top & -U\Sigma V^\top \\ -V\Sigma U^\top & V\Sigma V^\top \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} \Sigma & -\Sigma \\ -\Sigma & \Sigma \end{pmatrix} \succeq 0 \tag{20.20}$$

From that we see that by setting M up as such, replacing $\lambda\|X\|_*$ by $\lambda(\operatorname{tr}(P)+\operatorname{tr}(Q))/2$, and requiring $M \succeq 0$ we have an equivalent SDP with a quadratic objective that solves the original matrix completion objective.