

Large-Scale Convex Optimization over Matrices for Multi-task Learning

Paul Tseng

Mathematics, University of Washington



Seattle

Optimization Seminar, Univ. Washington

January 27, 2009

Joint work with Ting Kei Pong and Jieping Ye (ASU)

Prologue

This story began with an innocent looking email..

A Question..

On Thu, 18 Sep 2008, Jieping Ye wrote:

Dr. Tseng,

I recently came across your interesting work on the block coordinate descent method for non-differentiable optimization. I wonder whether the convergence result will apply for the matrix case where each block is a positive definite matrix, $\min f(X, Y, Z)$, where X, Y, Z are positive definite matrices. I will appreciate it if you can provide some relevant references on this if any. Thanks!

Best,

Jieping

The Problem

$$\min_{Q \succ 0, W} f(Q, W) := \text{tr}(WQ^{-1}W^T) + \text{tr}(Q) + \|WA - B\|_F^2$$

$Q \in \Re^{n \times n}$, $W \in \Re^{m \times n}$ ($A \in \Re^{n \times p}$ and $B \in \Re^{m \times p}$ are given),
 $\|B\|_F = (\sum_{i,j} B_{ij}^2)^{1/2}$

The Problem

$$\min_{Q \succ 0, W} f(Q, W) := \text{tr}(WQ^{-1}W^T) + \text{tr}(Q) + \|WA - B\|_F^2$$

$Q \in \Re^{n \times n}$, $W \in \Re^{m \times n}$ ($A \in \Re^{n \times p}$ and $B \in \Re^{m \times p}$ are given),
 $\|B\|_F = (\sum_{i,j} B_{ij}^2)^{1/2}$

Note:

- $f(Q, W)$ is diff., convex in Q for each W , convex in W for each Q .
- The min is finite, but may not be attained (e.g., when $B = 0$).
- If min is attained, it's attained at a critical pt, i.e., $\nabla f(Q, W) = 0$.

On Wed, 24 Sep 2008, Jieping Ye wrote:

Dear Paul,

...

The multi-task learning problem comes from our biological application: Drosophila gene expression pattern analysis (funded by NSF and NIH).

...

Thanks,
Jieping

First Try

$$\nabla f(Q, W) = (-Q^{-1}W^T W Q^{-1} + I, 2WQ^{-1} + 2(WA - B)A^T)$$

So $\nabla f(Q, W) = 0$ implies

$$(WQ^{-1})^T W Q^{-1} = I, \quad WQ^{-1} + WM = BA^T$$

where $M := AA^T$.

First Try

$$\nabla f(Q, W) = (-Q^{-1}W^T W Q^{-1} + I, 2WQ^{-1} + 2(WA - B)A^T)$$

So $\nabla f(Q, W) = 0$ implies

$$(WQ^{-1})^T W Q^{-1} = I, \quad WQ^{-1} + WM = BA^T$$

where $M := AA^T$.

So $\text{rank}(W) = n$, $\text{rank}(BA^T) = n$, ... , and

$$(I + MQ)(I + QM) = AB^T BA^T$$

First Try

$$\nabla f(Q, W) = (-Q^{-1}W^T W Q^{-1} + I, 2WQ^{-1} + 2(WA - B)A^T)$$

So $\nabla f(Q, W) = 0$ implies

$$(WQ^{-1})^T W Q^{-1} = I, \quad WQ^{-1} + WM = BA^T$$

where $M := AA^T$.

So $\text{rank}(W) = n$, $\text{rank}(BA^T) = n$, ... , and

$$(I + MQ)(I + QM) = AB^T BA^T$$

Prop. 1: If f has a stationary pt, then $\text{rank}(BA^T) = n$, $M \succ 0$, and

$$Q = (M^{-1}AB^T BA^T M^{-1})^{\frac{1}{2}} - M^{-1}, \quad W = BA^T(QM + I)^{-1}Q$$

But..

Date: Sat, 25 Oct 2008 16:48:20 -0700

Dear Paul,

Thanks for the writeup. Very interesting.

...

Unfortunately, M is commonly not positive definite in our applications.

...

Thanks,
Jieping

Second Try

Suppose $M = AA^T$ is singular, so $r := \text{rank}(A) < n$.

Use SVD or QR decomp. of A :

$$A = R \begin{bmatrix} \tilde{A} \\ 0 \end{bmatrix} S^T$$

with $\tilde{A} \in \mathbb{R}^{r \times p}$, $R^T R = I$ and $S^T S = I$. Let $\tilde{B} := BS$.

Prop. 2:

$$\min_{Q \succ 0, W} f(Q, W) = \min_{\tilde{Q} \succ 0, \tilde{W}} \tilde{f}(\tilde{Q}, \tilde{W}),$$

where

$$\tilde{f}(\tilde{Q}, \tilde{W}) := \text{tr}(\tilde{W} \tilde{Q}^{-1} \tilde{W}^T) + \text{tr}(\tilde{Q}) + \left\| \tilde{W} \tilde{A} - \tilde{B} \right\|_F^2.$$

Then recover Q, W from \tilde{Q}, \tilde{W} .

Moreover, \tilde{f} has a stationary pt iff

$$(\widetilde{M}^{-1} \widetilde{A} \widetilde{B}^T \widetilde{B} \widetilde{A}^T \widetilde{M}^{-1})^{\frac{1}{2}} \succ \widetilde{M}^{-1}$$

where $\widetilde{M} := \widetilde{A} \widetilde{A}^T$.

Done?

But..

Date: Thu, 30 Oct 2008 10:44:56 -0700

Dear Tseng,

Thanks. I like the derivation.

It seems the condition in Eq. (4) is the key.

We need to somehow relax this condition.

Will perturbation solve this problem?

Best,

Jieping

Third Try

Assume w.l.o.g. $\text{rank} A = n$. Let

$$\begin{aligned} h(Q) &:= \inf_W f(Q, W) \\ &= \inf_W \text{tr}(WQ^{-1}W^T) + \text{tr}(Q) + \|WA - B\|_F^2 \\ &= \text{tr}(Q) + \text{tr}(E^T E(Q + C)^{-1}) + \text{const.} \end{aligned}$$

with $C := M^{-1} \succ 0$ and $E := BA^T C$. ($M = AA^T$)

Third Try

Assume w.l.o.g. $\text{rank} A = n$. Let

$$\begin{aligned} h(Q) &:= \inf_W f(Q, W) \\ &= \inf_W \text{tr}(WQ^{-1}W^T) + \text{tr}(Q) + \|WA - B\|_F^2 \\ &= \text{tr}(Q) + \text{tr}(E^T E(Q + C)^{-1}) + \text{const.} \end{aligned}$$

with $C := M^{-1} \succ 0$ and $E := BA^T C$. ($M = AA^T$)

Then $h(Q)$ is cont. over $Q \succeq 0$ (!) so

$$\min_{Q \succeq 0} h(Q) = \min_{Q \succ 0, W} f(Q, W).$$

Moreover, $(Q, W) \mapsto WQ^{-1}W^T$ is operator-convex, so f is convex, and hence h is convex.

Prop. 3: $\min_{Q \succeq 0} h(Q)$ is attained, and

$$\nabla h(Q) = I - (Q + C)^{-1} E^T E (Q + C)^{-1}$$

is Lipschitz cont. over $Q \succeq 0$.

Prop. 3: $\min_{Q \succeq 0} h(Q)$ is attained, and

$$\nabla h(Q) = I - (Q + C)^{-1} E^T E (Q + C)^{-1}$$

is Lipschitz cont. over $Q \succeq 0$.

Moreover, using Schur complement, $\min_{Q \succeq 0} h(Q)$ reduces to an SDP:

$$\begin{aligned} \min \quad & \text{tr}(Q) + \text{tr}(U) \\ \text{s.t.} \quad & Q \succeq 0, \quad \begin{bmatrix} Q & 0 \\ 0 & U \end{bmatrix} + \begin{bmatrix} C & E^T \\ E & 0 \end{bmatrix} \succeq 0 \end{aligned}$$

Recall $C \succ 0$ is $n \times n$ and E is $m \times n$. This SDP is solvable by existing IP solvers (SeDuMi, SDPT³, CSDP, Mosek, ..) for around $m + n \leq 500$.

But..

Date: Mon, 1 Dec 2008 09:33:13 -0700

...

In our application, n is around 1000-2000 and m is around 50-100.

...

It contains 1000-3000 rows depending on the feature extraction scheme. In general, X is dense. However, one of our recent feature extraction schemes produces sparse X . By the way, the columns of X correspond to biological images.

Best,
Jieping

For $m = 100$, $n = 2000$, (Q, W) comprises 2201000 variables. $A \in \mathbb{R}^{n \times p}$ may be dense.

!

Fourth Try

Lesson from my graduate student days:

“When stuck, look at the dual”

Consider the dual problem

$$\max_{\Lambda \succeq 0} \min_{Q \succeq 0, U} L(Q, U, \Lambda),$$

with Lagrangian ($\langle W, Z \rangle = \text{tr}(W^T Z)$)

$$\begin{aligned} L(Q, U, \Lambda) &:= \langle I, Q \rangle + \langle I, U \rangle - \left\langle \begin{bmatrix} \Lambda_1 & \Lambda_2^T \\ \Lambda_2 & \Lambda_3 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & U \end{bmatrix} + \begin{bmatrix} C & E^T \\ E & 0 \end{bmatrix} \right\rangle \\ &= \langle I - \Lambda_1, Q \rangle + \langle I - \Lambda_3, U \rangle - \langle \Lambda_1, C \rangle - 2\langle \Lambda_2, E \rangle \end{aligned}$$

Consider the dual problem

$$\max_{\Lambda \succeq 0} \min_{Q \succeq 0, U} L(Q, U, \Lambda),$$

with Lagrangian ($\langle W, Z \rangle = \text{tr}(W^T Z)$)

$$\begin{aligned} L(Q, U, \Lambda) &:= \langle I, Q \rangle + \langle I, U \rangle - \left\langle \begin{bmatrix} \Lambda_1 & \Lambda_2^T \\ \Lambda_2 & \Lambda_3 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & U \end{bmatrix} + \begin{bmatrix} C & E^T \\ E & 0 \end{bmatrix} \right\rangle \\ &= \langle I - \Lambda_1, Q \rangle + \langle I - \Lambda_3, U \rangle - \langle \Lambda_1, C \rangle - 2\langle \Lambda_2, E \rangle \end{aligned}$$

For dual feas., need $I \succeq \Lambda_1$, $I = \Lambda_3$, $\Lambda_1 \succeq \Lambda_2^T \Lambda_2$. Dual problem reduces to

$$\min_{I \succeq \Lambda_1 \succeq \Lambda_2^T \Lambda_2} \langle C, \Lambda_1 \rangle + 2\langle E, \Lambda_2 \rangle$$

Since $C \succ 0$, minimum w.r.t. Λ_1 is attained at $\Lambda_1 = \Lambda_2^T \Lambda_2$.

The dual problem reduces to (recall $\Lambda_2 \in \Re^{m \times n}$)

$$\min_{I \succeq \Lambda_2^T \Lambda_2} d_2(\Lambda_2) := \frac{1}{2} \langle C, \Lambda_2^T \Lambda_2 \rangle + \langle E, \Lambda_2 \rangle.$$

The dual problem reduces to (recall $\Lambda_2 \in \Re^{m \times n}$)

$$\min_{I \succeq \Lambda_2^T \Lambda_2} d_2(\Lambda_2) := \frac{1}{2} \langle C, \Lambda_2^T \Lambda_2 \rangle + \langle E, \Lambda_2 \rangle.$$

- No duality gap since the primal problem has interior soln.
- Recovers Q as Lagrange multiplier assoc. with $I \succeq \Lambda_2^T \Lambda_2$.
- $\nabla d_2(\Lambda_2) = \Lambda_2 C + E$ is Lipschitz cont. with constant $L = \lambda_{\max}(C)$.

The dual problem reduces to (recall $\Lambda_2 \in \Re^{m \times n}$)

$$\min_{I \succeq \Lambda_2^T \Lambda_2} d_2(\Lambda_2) := \frac{1}{2} \langle C, \Lambda_2^T \Lambda_2 \rangle + \langle E, \Lambda_2 \rangle.$$

- No duality gap since the primal problem has interior soln.
- Recovers Q as Lagrange multiplier assoc. with $I \succeq \Lambda_2^T \Lambda_2$.
- $\nabla d_2(\Lambda_2) = \Lambda_2 C + E$ is Lipschitz cont. with constant $L = \lambda_{\max}(C)$.

What about the constraint $I \succeq \Lambda_2^T \Lambda_2$?

Prop. 4: For any $\Lambda_2 \in \Re^{m \times n}$ ($m \leq n$) with SVD $\Lambda_2 = R \begin{bmatrix} D & 0 \end{bmatrix} S^T$,

$$\text{Proj}(\Lambda_2) := \arg \min_{I \succeq \Psi_2^T \Psi_2} \|\Lambda_2 - \Psi_2\|_F^2 = R \begin{bmatrix} \min \{D, I\} & 0 \end{bmatrix} S^T$$

Solving the reduced dual:

Coded 3 methods in Matlab: Frank-Wolfe, grad.-proj. with LS Goldstein, Levitin, Polyak, and accel. grad.-proj. Nesterov.

Accel. grad.-proj. seems most efficient.

0. Choose $I \succeq \Lambda_2^T \Lambda_2$. Set $\Lambda_2^{\text{prev}} = \Lambda_2$, $\theta^{\text{prev}} = \theta = 1$. Fix $L = \lambda_{\max}(C)$. Go to 1.

1. Set

$$\Lambda_2^{\text{ext}} = \Lambda_2 + \left(\frac{\theta}{\theta^{\text{prev}}} - \theta \right) (\Lambda_2 - \Lambda_2^{\text{prev}}).$$

Update $\Lambda_2^{\text{prev}} \leftarrow \Lambda_2$, $\theta^{\text{prev}} \leftarrow \theta$, and

$$\begin{aligned} \Lambda_2 &\leftarrow \text{Proj} \left(\Lambda_2^{\text{ext}} - \frac{1}{L} \nabla d_2(\Lambda_2^{\text{ext}}) \right) \\ \theta &\leftarrow \frac{\sqrt{\theta^4 + 4\theta^2} - \theta^2}{2}. \end{aligned}$$

If relative duality gap $\leq \text{tol}$, stop. Else to to 1.

Test Results (Preliminary)

Tested on random data: $A \sim U[0, 1]^{n \times p}$ and $B \sim U[0, 1]^{m \times p}$. $tol = .001$

```
n = 2000  m = 100  p= 1000  tol= 0.001
reduce A to have full row rank:
done reducing A, time: 38.9895
done computing C and E, time: 4.05682
termination due to negligible change in U = 3.4469e-11
iter= 10  dobj= -96.7469  dual feas= 8.88178e-15
          pobj= -96.7469  primal feas= 1.43293e-15
accel. grad-proj: iter= 10  total_time= 67.9021
fmin = 193.494  fval = 193.494
```

```
n = 2000  m = 100  p= 3000  tol= 0.001
done computing C and E, time: 31.5357
termination due to negligible change in U = 7.32917e-11
iter= 10  dobj= -137.14  dual feas= 6.21725e-15
          pobj= -137.14  primal feas= 3.06165e-15
accel. grad-proj: iter= 10  total_time= 190.652
fmin = 8632.32  fval = 8632.32
```

A Puzzle: When $n \approx p$, #iterations becomes very large.

Maybe finally..

Date: Sun, 11 Jan 2009 11:09:15 -0700

Dear Paul,

Sorry for the delay.

Some preliminary results prepared by my student are attached. Overall, it performs well, especially when the number of labels is large. We will conduct more extensive experimental studies and keep you updated.

Best,
Jieping

In preliminary result on *Drosophila* gene expression pattern annotation, a group of images are associated with variable number of terms using a controlled vocabulary.

k -means clustering and feature extractions are used to obtain a global histogram counting the number of features closest to the visual words obtained from the clustering algorithm (with 3000 clusters), etc.

- $n = 3000$ (dim. of data)
- $10 \leq m \leq 60$ (#terms/tasks)
- $2200 \leq p \leq 2800$ (#samples).

Measure	m	MTL	SVM	ML _{LS}
AUC (Area Under the Curve)	10	87.38	86.66	87.72
	20	84.57	83.25	84.61
	30	82.76	81.13	82.92
	40	81.61	79.56	77.25
	50	80.52	78.17	79.38
	60	80.17	77.18	78.11

Table 1: AUC Performance of MTL, SVM, ML_{LS} (multi-label).

Measure	m	MTL	SVM	ML _{LS}
macro F1	10	64.43	62.97	64.91
	20	49.98	50.45	49.90
	30	42.48	41.92	42.57
	40	34.72	35.26	32.26
	50	28.70	29.74	29.13
	60	24.78	25.49	25.18

Table 2: macro F1 Performance of MTL, SVM, ML_{LS} (multi-label).

Measure	m	MTL	SVM	ML _{LS}
micro F1	10	67.85	66.67	68.27
	20	58.25	55.66	57.62
	30	53.74	48.11	53.02
	40	50.68	44.26	46.92
	50	49.45	43.53	48.76
	60	48.79	42.84	48.02

Table 3: micro F1 Performance of MTL, SVM, ML_{LS} (multi-label).

Conclusions & Extensions

1. A seemingly nasty problem arising from application is tamed by a mix of convex/matrix analysis, and modern algorithms.

Conclusions & Extensions

1. A seemingly nasty problem arising from application is tamed by a mix of convex/matrix analysis, and modern algorithms.
2. Extension to related convex optimization problems in learning?

Conclusions & Extensions

1. A seemingly nasty problem arising from application is tamed by a mix of convex/matrix analysis, and modern algorithms.
2. Extension to related convex optimization problems in learning?
3. Better algorithms to handle the case of $p \approx n$?

Conclusions & Extensions

1. A seemingly nasty problem arising from application is tamed by a mix of convex/matrix analysis, and modern algorithms.
2. Extension to related convex optimization problems in learning?
3. Better algorithms to handle the case of $p \approx n$?

The End?