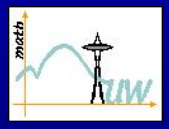
Large-Scale Convex Optimization over Matrices for Multi-task Learning

Paul Tseng
Mathematics, University of Washington
Seattle



Optima, Univ. Illinois, Urbana-Champaign
March 26, 2009
Joint work with Ting Kei Pong (UW) 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 optimizatic
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

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

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

The Problem

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

$Q \in \Re^{n \times n}$, $W \in \Re^{n \times m}$ ($A \in \Re^{p \times n}$ and $B \in \Re^{p \times m}$ 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., $\nabla f(Q, W) = 0$.
- m = # tasks. Q^{-1} = covar. matrix.

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

$$\nabla f(Q, W) = \left(-Q^{-1}WW^TQ^{-1} + I, 2Q^{-1}W + 2A^T(AW - B)\right)$$

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

$$(W^TQ^{-1})^TW^TQ^{-1} = I, \quad Q^{-1}W + MW = A^TB$$

where $M := A^TA$.

First Try

$$\nabla f(Q, W) = \left(-Q^{-1}WW^TQ^{-1} + I, 2Q^{-1}W + 2A^T(AW - B)\right)$$

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

$$(W^TQ^{-1})^TW^TQ^{-1} = I, \quad Q^{-1}W + MW = A^TB$$

where $M := A^TA$.

So $\mathrm{rank}(W) = n$, $\mathrm{rank}(A^TB) = n$, ... , and

$$(I + MQ)(I + QM) = A^TBB^TA$$

First Try

$$\nabla f(Q, W) = \left(-Q^{-1}WW^TQ^{-1} + I, 2Q^{-1}W + 2A^T(AW - B)\right)$$

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

$$(W^TQ^{-1})^TW^TQ^{-1} = I, \quad Q^{-1}W + MW = A^TB$$

where $M := A^TA$.

So $\mathrm{rank}(W) = n$, $\mathrm{rank}(A^TB) = n$, ... , and

$$(I + MQ)(I + QM) = A^TBB^TA$$

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

$$Q = (M^{-1}A^TBB^TAM^{-1})^{\frac{1}{2}} - M^{-1}, \quad W = (M + Q^{-1})^{-1}A^TB$$

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

Use SVD or QR decomp. of A:

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

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

Prop. 2:

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

where

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

Then recover Q, W from $\widetilde{Q}, \widetilde{W}$.

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

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

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

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. $\operatorname{rank} A = n$. Let

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

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

Third Try

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

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

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

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

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

Moreover, $(Q, W) \mapsto W^T Q^{-1} W$ 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:

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

Recall $C \succ 0$ is $n \times n$ and E is $m \times n$. This SDP is solvable by existing IP solvers (SeDuMi, SDPT^3, CSDP, ..) 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 correpspond to biological images.

Best,
Jieping

For $m = 100$, $n = 2000$, (Q, W) comprises 2201000 variables. $A \in \Re^{p \times n}$ 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 ($\langle W, Z \rangle = \mathrm{tr}(W^T Z)$)

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

Consider the dual problem

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

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

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

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 $\Lambda_2 = R \begin{bmatrix} D & 0 \end{bmatrix} S^T$,

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

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

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

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 $\leq tol$, stop. Else to to 1.

Test Results (Preliminary)

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

```
n = 2000  m = 100  p= 1000  tol= 0.001
reduce A to have full column 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
```

However: When $n = p$, L is large ($\approx 10^6$) and #iterations is 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 in the codebook obtained from the clustering algorithm (with 3000 clusters), etc. Hard (soft) assignment: a feature assigned to one (multiple) word.

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

m	MTL_{hard}	MTL_{soft}	SVM_{hard}	SVM_{soft}	PMK_{star}	$\text{PMK}_{\text{clique}}$
10	77.22±0.63	78.86±0.58	74.89±0.68	78.51±0.60	71.80±0.81	71.98±0.87
20	78.95±0.82	80.90±1.02	76.38±0.84	78.94±0.86	72.01±1.01	71.70±1.20
10	52.57±1.19	54.89±1.24	52.25±0.98	55.64±0.69	46.20±1.18	47.06±1.16
20	33.15±1.37	37.25±1.25	35.62±0.99	39.18±1.18	28.21±1.00	28.11±1.09
10	59.92±1.04	60.84±0.99	55.74±1.02	59.27±0.80	53.25±1.15	53.36±1.20
20	55.33±0.88	56.79±0.72	51.70±1.17	54.25±0.93	49.59±1.24	48.14±1.34

Table 1: Performance (top: AUC, middle: macro F1, bottom: micro F1) of MTL, SVM, PMK on data sets in stage range 9-10 ($m = 10, 20$ and $p = 919, 1015$).

m	MTL_{hard}	MTL_{soft}	SVM_{hard}	SVM_{soft}	PMK_{star}	$\text{PMK}_{\text{clique}}$
10	84.06±0.53	86.18±0.45	83.05±0.54	84.84±0.57	78.68±0.58	78.52±0.55
30	81.83±0.46	83.85±0.36	79.18±0.51	81.31±0.48	71.85±0.61	71.13±0.53
50	80.56±0.53	82.87±0.53	76.19±0.72	78.75±0.68	69.66±0.81	68.80±0.68
10	60.30±0.92	64.00±0.85	60.37±0.88	62.61±0.82	54.61±0.68	55.19±0.62
30	35.20±0.85	39.15±0.83	35.32±0.75	37.38±0.95	22.30±0.70	24.85±0.63
50	23.07±0.86	26.67±1.05	23.46±0.60	26.26±0.65	14.07±0.48	15.04±0.46
10	66.89±0.79	68.92±0.68	65.67±0.60	66.73±0.74	62.06±0.54	61.84±0.51
30	55.66±0.64	56.70±0.68	48.87±0.85	51.52±0.96	47.08±0.81	44.81±0.66
50	52.92±0.78	54.54±0.70	47.18±0.84	47.97±0.90	44.25±0.65	42.49±0.70

Table 2: Performance (top: AUC, middle: macro F1, bottom: micro F1) of MTL, SVM, PMK on data sets in stage range 11-12 ($10 \leq m \leq 50$, $1622 \leq p \leq 2070$).

m	MTL_{hard}	MTL_{soft}	SVM_{hard}	SVM_{soft}	PMK_{star}	$\text{PMK}_{\text{clique}}$
10	87.38±0.36	89.43±0.31	86.66±0.35	88.42±0.35	82.07±0.41	82.53±0.62
30	82.76±0.36	85.86±0.34	81.13±0.46	83.45±0.38	73.34±0.46	73.73±0.52
60	80.17±0.40	83.32±0.45	77.18±0.46	79.75±0.47	67.15±0.57	67.11±0.64
10	64.43±0.77	67.42±0.78	62.97±0.68	66.38±0.71	57.37±0.91	58.42±0.94
30	42.48±0.87	47.39±0.91	41.92±0.76	45.07±0.68	29.62±0.67	31.04±0.82
60	24.78±0.67	29.84±0.62	25.49±0.55	28.72±0.57	15.65±0.46	16.13±0.48
10	67.85±0.60	70.50±0.58	66.67±0.45	68.79±0.60	60.98±0.74	61.87±0.77
30	53.74±0.45	57.04±0.69	48.11±0.90	51.19±0.83	43.50±0.70	44.14±0.78
60	48.79±0.60	51.35±0.58	42.84±0.76	44.48±0.84	37.28±0.81	38.29±0.78

Table 3: Performance (top: AUC, middle: macro F1, bottom: micro F1) of MTL, SVM, PMK on data sets in stage range 13-16 ($10 \leq m \leq 60$, $2228 \leq p \leq 2754$).

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?

Nooo..

On Sun, 22 Mar 2009, Jieping Ye wrote:

Dear Paul,

Thanks for the updated version.
...
There are two tex files: introduction.tex and experiment.tex
...

Thanks,
Jieping

The introduction shows the original problem is a reformulation of

$$\min_{W} \ \|W\|_* + \|AW - B\|_F^2$$

with $\|W\|_* = \sum_i \sigma_i(W)$ ("trace/nuclear-norm").

This can be solved by accel. gradient method (one SVD per iter.) too.

Which is faster? Will see..