Frist order optimization methods for sparse inverse covariance selection

Katya Scheinberg

Lehigh University

ISE Department

(joint work with D. Goldfarb, Sh. Ma, I. Rish)

Introduction

- The field of convex optimization has been extensively developed since Khachian showed in 1979 that ellipsoid method has polynomial complexity when applied to LP.
- General theory of interior point algorithms for convex optimization was developed by Nesterov and Nemirovskii.
- Any convex optimization problem can be solved in polynomial time by an IPM. For some known classes (LP, QP, SDP) the IPMs are readily available.
- For decades optimization methods relied of the fact that the problem data, when large, is typically sparse.
- Second-order methods (IPM) have good convergence rate, but high per iteration complexity. They exploit sparsity structure to facilitate linear algebra.
- First-order methods (gradient based) have slow convergence and were considered inefficient.

Introduction

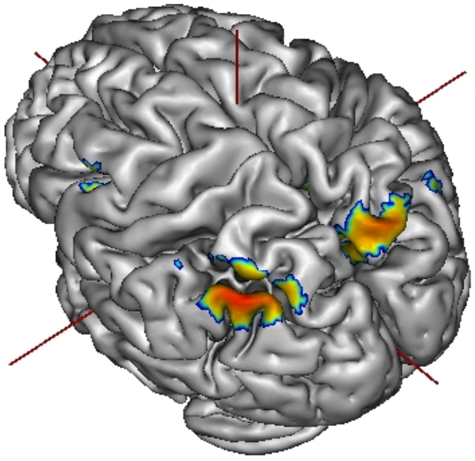
- At the core of many statistical machine learning problems lies an optimization problem, often convex, from a well-studied class (LP, QP, SDP).
- These problems are very large and dense in terms of data.
- IPMs are often too expensive to use. ML community initially assumed that traditional optimization methods have to be abandoned.
- However, often structure (sparsity) is present in the solution.
- This structure can be well exploited by first-order approaches to convex optimization.
- Recent advances in complexity results give rise to very significant interest in first-order methods.

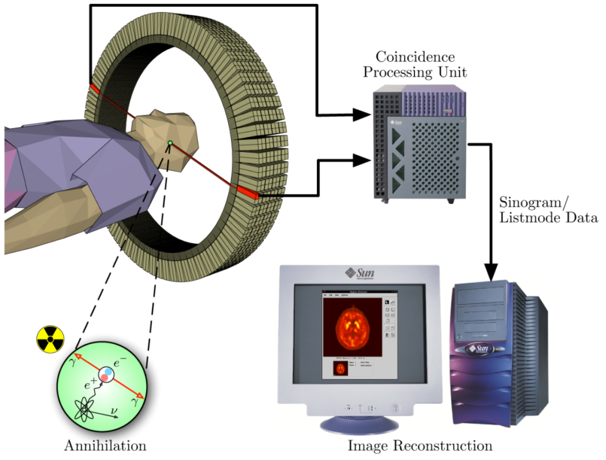
Sparse inverse covariance selection

Neuroimaging: MRI, fMRI, PET

Detecting Alzheimer, Schizophrenia, etc by constructing and exploring connectivity network of the brain.

- **AD**: Grady et al. 2001, Heun et al. 2006, Celone et al 2006, Rombouts et al. 2005, Lustig et al. 2006.
- **Schizophrenia:** Cecchi, G. et al (2009), Carroll, M. K., et al(2009) Neuroimage, M. Plaze et al. (2006), Schizophrenia Research, . V.M. Eguiluz et al(2005), Phys. Rev. Letters , Y. Liu et al. (2008). Brain, Feb. 2008.

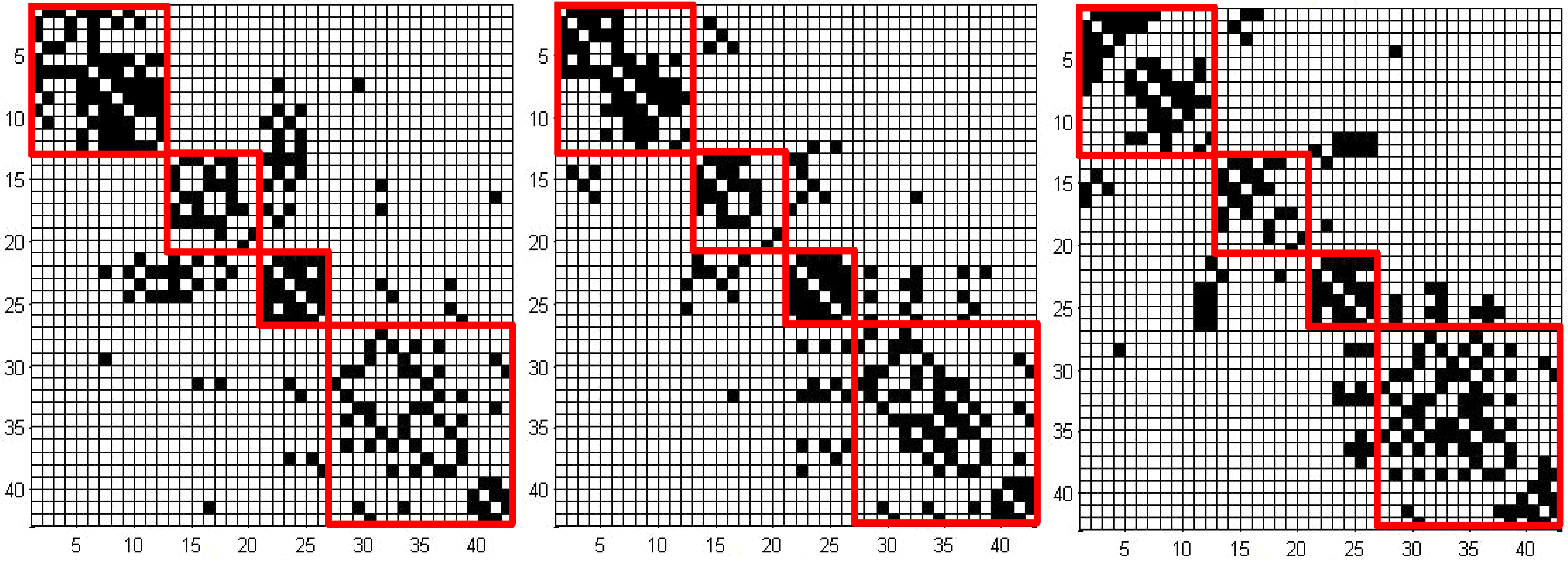




From Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, and Eric Reiman. Learning Brain Connectivity of Alzheimer's Disease by Sparse Inverse Covariance Estimation. **NeuroImage**, 50, 935-949, 2010.

- There is significant, quantifiable difference in brain connectivity between AD and normal brains.

frontal, parietal, occipital, and temporal lobes in order



AD

MCI

NC

Sparse Inverse Covariance Estimation

- Given the observations x_i~N(μ, Σ), the empirical covariance matrix $S \in$ S^p, is

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

- We can estimate $\Theta = \Sigma^{-1}$ by solving the following maximum likelihood problem

$$\max_{\Theta \succ 0} f = \log\det\Theta - \mathrm{tr}(S\Theta)$$

- By penalizing the L1-norm, we can obtain the sparse inverse covariance matrix

$$\max_{\Theta \succ 0} f = \log\det\Theta - \mathrm{tr}(S\Theta) - \lambda\left\|\mathrm{vec}(\Theta)\right\|_1$$

From Jieping Ye, KDD'09 presentation

Why Sparse Inverse Covariance?

- Employ sparse inverse covariance estimation for brain region connectivity identification.

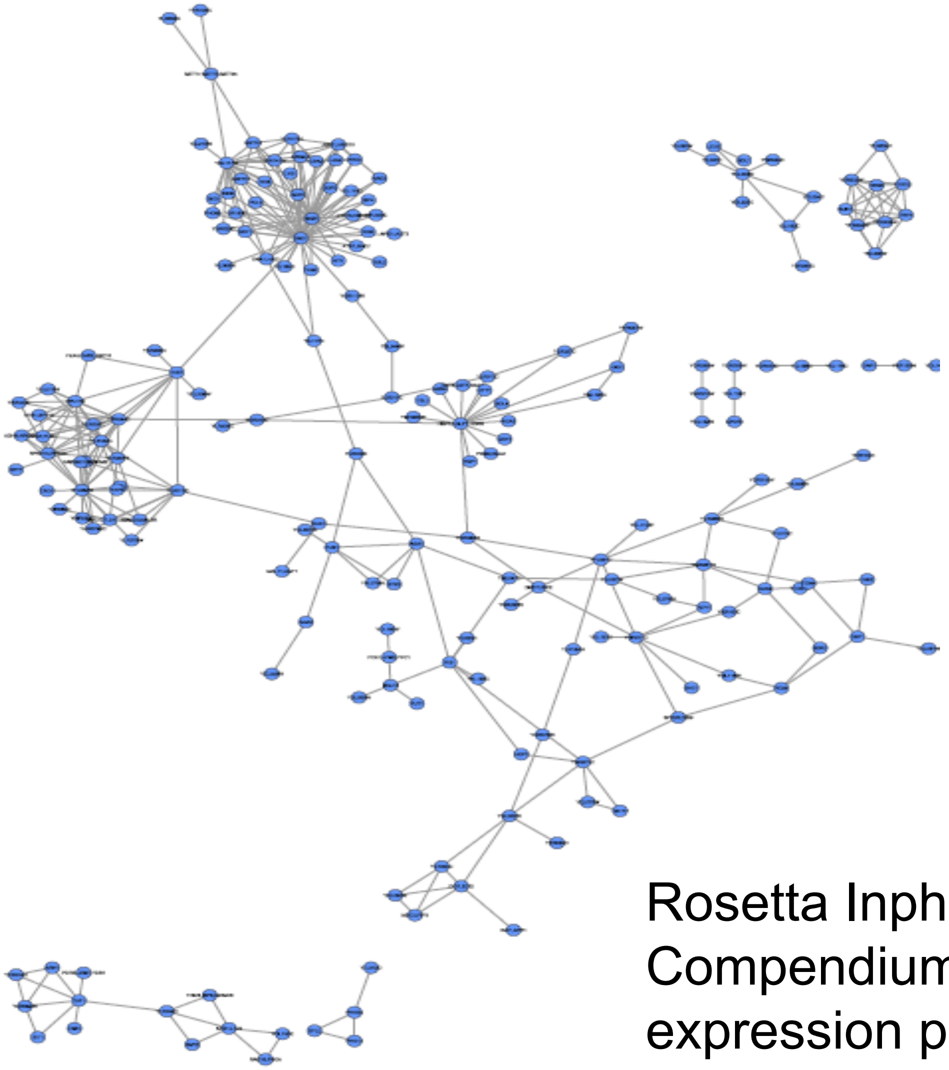
- The covariance matrix can be estimated robustly when many entries of the inverse covariance matrix are zero.

- The sparse inverse covariance matrix can be interpreted from the perspective of undirected graphical model.
  - If the ijth component of Θ is zero, then variables i and j are conditionally independent, given the other variables in the multivariate Gaussian distribution.

- Many real-world networks are sparse.
  - Gene interaction network

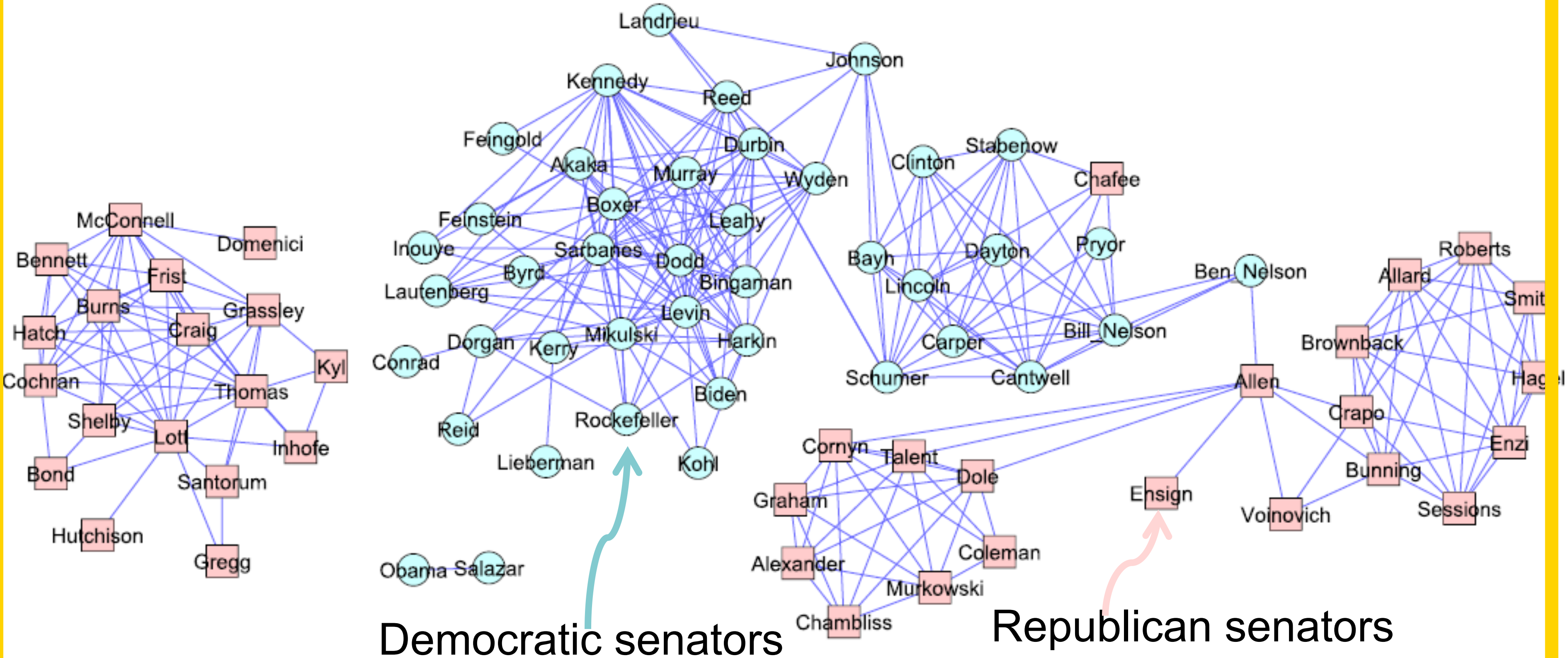
From Jieping Ye, KDD'09 presentation

Example: Gene Network



Rosetta Inpharmatics Compendium of gene expression profiles described by Hughes et al. (2000)

Example: Senate Voting Records Data (2004-06)



Democratic senators

Republican senators

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research, 9:485–516,* 2008.

How to ensure that Θ is sparse?

- Maximum likelihood formulation

$$\Theta = \arg\max_C \left(\frac{n}{2} (\log \det C - Tr(SC)) \right)$$

- Strictly convex problem, hence unique solution

$$\Theta = S^{-1}$$
$$S = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

- But never sparse!

Similar to sparse signal recovery

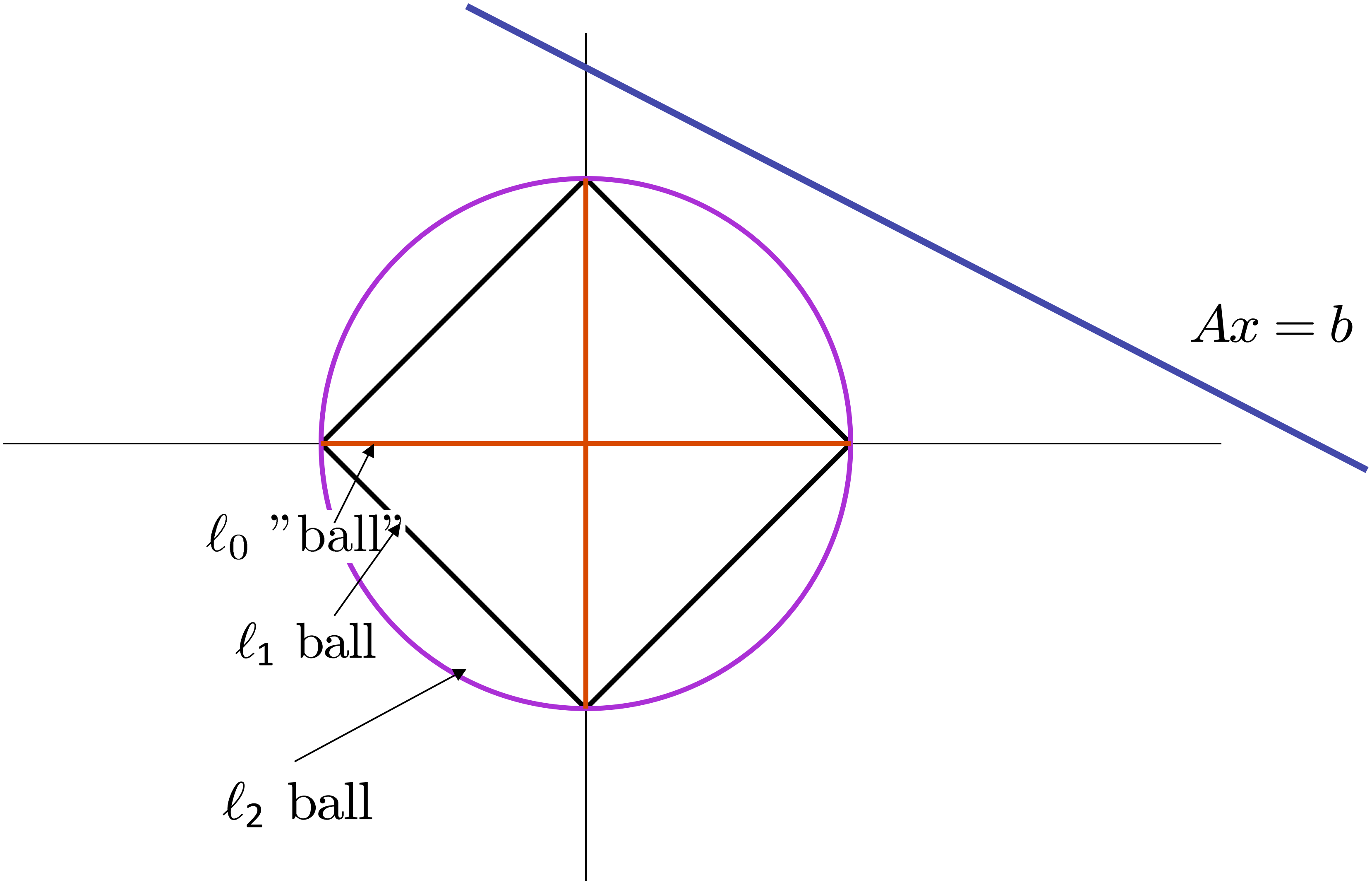
- We seek sparse signal x, which satisfies underdetermined system Ax=b.
- Recently it was shown by Candes & Tao and Donoho that under certain conditions on matrix A the sparse signal

$$\begin{aligned} \min \quad & card(x) \\ s.t. \quad & Ax = b. \end{aligned}$$

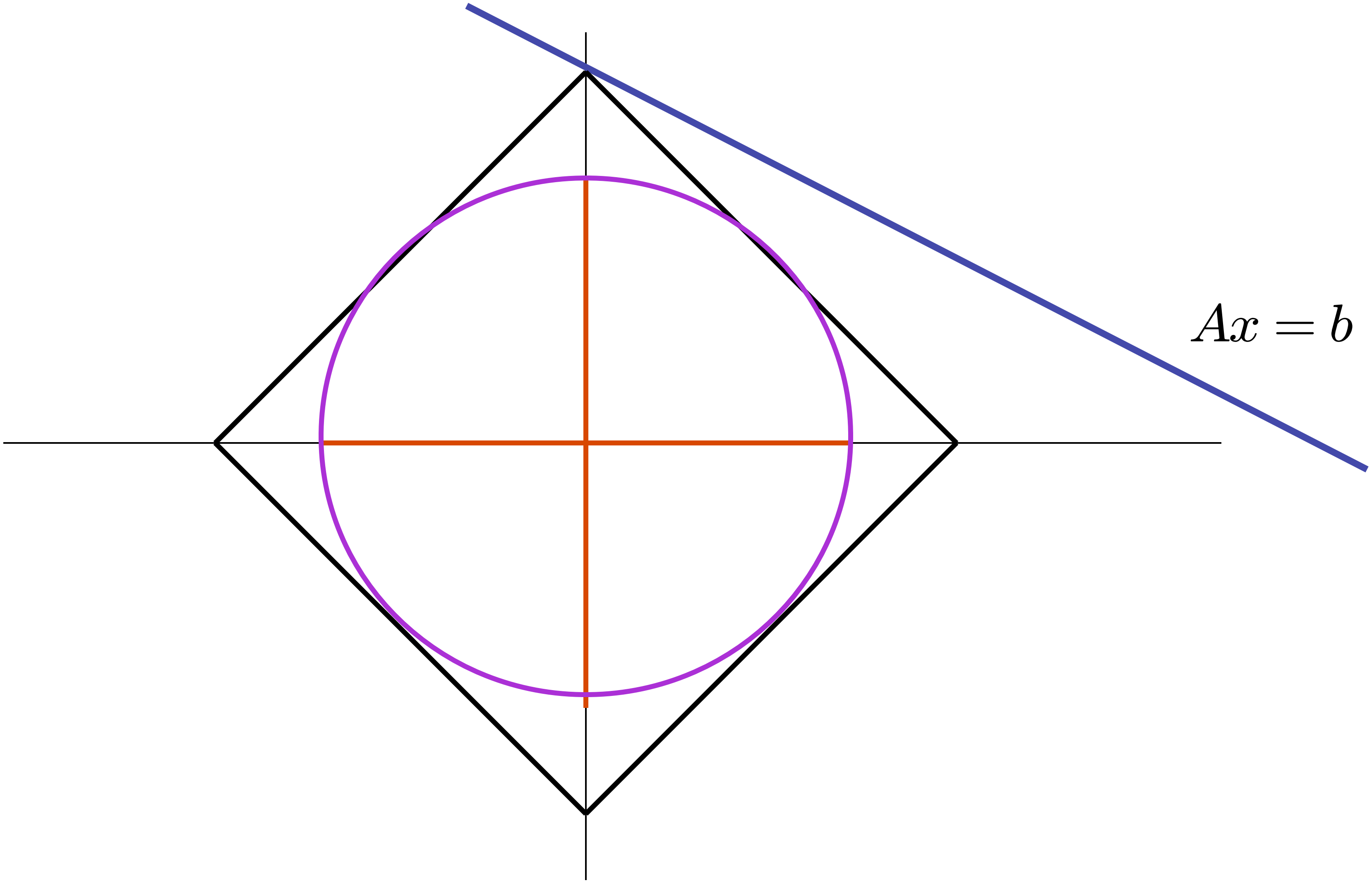
is recovered exactly by solving the convex relaxation

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & Ax = b. \end{aligned}$$

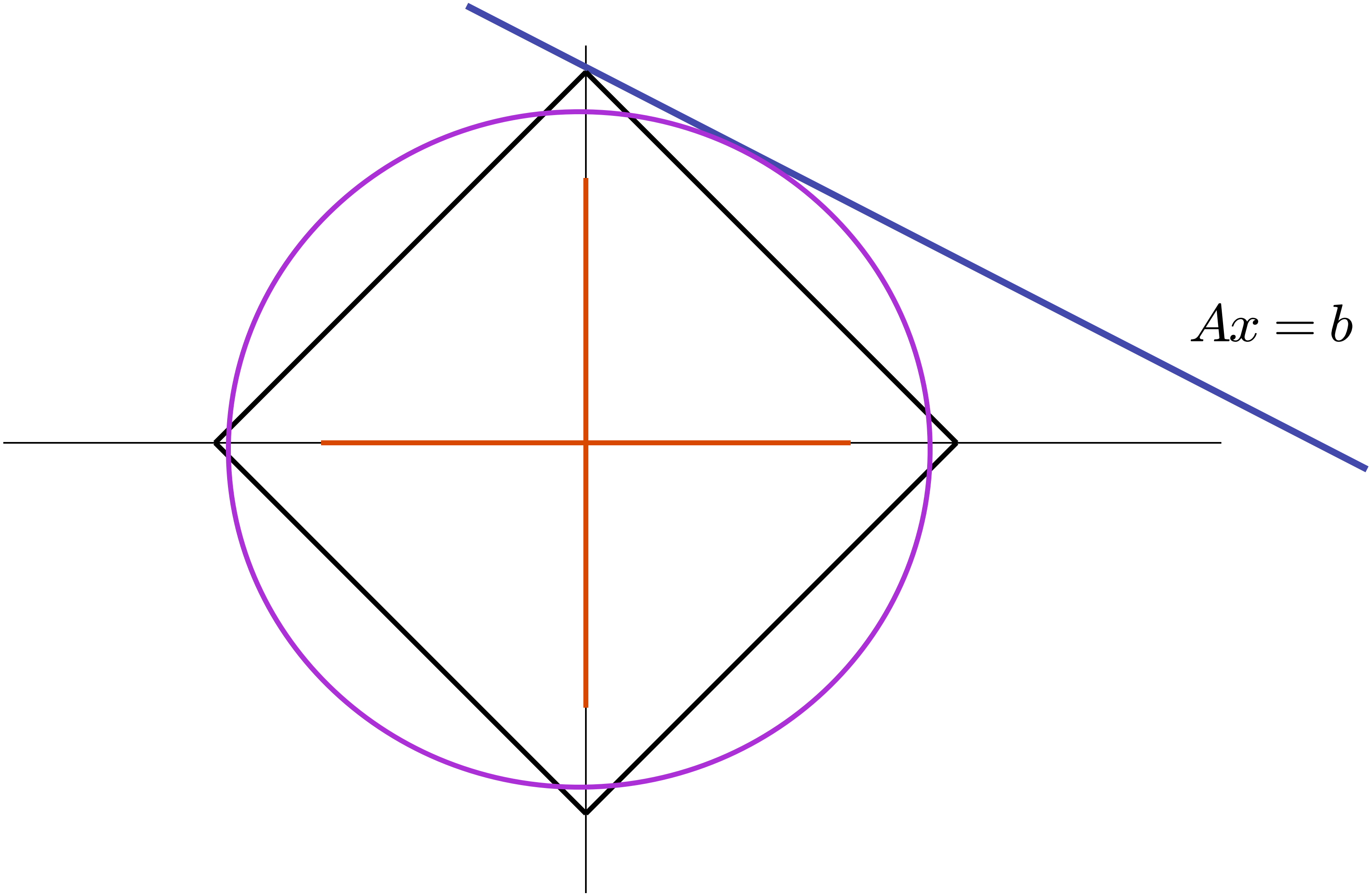
Why $||\cdot||_1$ norm?



Why $||\cdot||_1$ norm?



Why $||\cdot||_1$ norm?



Sparsity inducing formulation

- NP-hard formulation

$$\Theta = \arg\max_C \left(\frac{n}{2}(\log\det C - Tr(SC)) - \lambda Card(C)\right)$$

- Convex relaxation

$$\Theta = \arg\max_C \frac{n}{2}(\log\det C - Tr(SC)) - \lambda||C||_1$$

$$(\|C\|_1 = \sum_{ij} |C_{ij}|)$$

- Convex optimization problem with unique solution for each λ
- Number of variables is $p^2/2$.

Convex constrained formulation

Primal problem

$$\max_{C\succ 0} \tfrac{n}{2}(\mathrm{ln det}(C) - Tr(SC)) - \lambda\|C\|_1$$

Reformulate as a smooth convex constrained problem

$$\begin{aligned} &\max_{C',C''} && \frac{n}{2}[\ln\det(C'-C'') - Tr(S(C'-C''))] - \lambda Tr(E(C'+C'')), \\ &\mathrm{s.\ t.} && C' \geq 0,\ C'' \geq 0,\ C'-C'' \succ 0 \end{aligned}$$

Primal-dual pair of problems

Primal problem

$$\max_{C \succ 0} \tfrac{n}{2}(\ln\det(C) - Tr(SC)) - \lambda \|C\|_1$$

Dual problem

$$\max_{W \succ 0}\{\tfrac{n}{2}\ln(\det(W)) - np/2 : \text{ s.t. } \tfrac{n}{2}||(W - S)||_\infty \leq \lambda\}$$

Interior point method – O(p^6) operations/iter

Similar to Lasso and sparse signal recovery

Primal-Dual pair of problems

$$\min \quad \frac{1}{2}||Ax - b||^2 + \lambda||x||_1$$

$$\begin{aligned} \min \quad & \frac{1}{2}x^\top A^\top Ax \\ s.t. \quad & \|A^\top(Ax - b)\|_\infty \leq \lambda \end{aligned}$$

First Order Methods

First-order proximal gradient methods

- Consider: $\boxed{\min_x f(x)}$

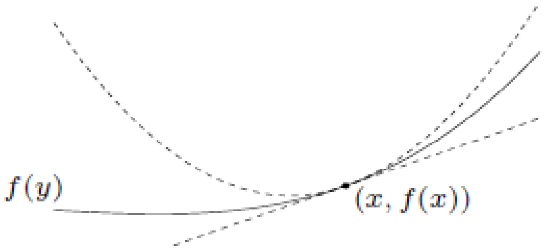
$$|\nabla f(x) - \nabla f(y)| \leq L||x - y||$$

- Linear lower approximation

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

- Quadratic upper approximation

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 = Q_{f,\mu}(x, y)$$



$$f(y) \leq f(x) - \mu^2\|\nabla f(x)\|^2 + \frac{1}{2\mu}||x - \mu\nabla f(x)^\top - y||^2 = Q_{f,\mu}(x, y)$$

First-order proximal gradient method

$$\boxed{\min_x f(x)}$$

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \text{argmin}_y Q_{f,\mu}(x^k, y)$$

$$\Updownarrow$$

$$x^{k+1} = x^k - \mu \nabla f(x^k)$$

- If $\mu \leq$ *1/L* then

$$f(x^{k+1}) \leq f(x^k) + \frac{1}{2\mu} ||x^k - \mu \nabla f(x^k)^\top - x^{k+1}||^2 = Q_{f,\mu}(x^k, x^{k+1})$$

Complexity of proximal gradient method

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \mathrm{argmin}_y Q_{f,\mu}(x^k, y)$$

$$\Updownarrow$$

$$x^{k+1} = x^k - \mu \nabla f(x^k)$$

- If $\mu \leq$ *1/L* then in $O(L||x^0\text{-}x^*||/\epsilon)$ iterations finds solution

$$x^k : \; f(x^k) \leq f(x^*) + \epsilon$$

Compare to ***O(log(L/ϵ))*** of interior point methods.

Can we do better?

Accelerated first-order method

Nesterov, ’83, ‘00s,

Beck&Teboulle ‘09

$$\boxed{\min_x f(x)}$$

- Minimize upper approximation at an intermediate point.

$$x^{k+1} = y^k - \mu \nabla f(y^k)$$

$$y^{k+1} := \quad x^k + \tfrac{k-1}{k+2}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then

$$f(x^k) - f(x^*) \leq \frac{L\|x^0 - x^*\|}{2k^2}$$

First order methods for composite functions

Examples

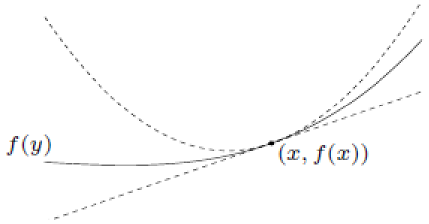
- Lasso or CS: $\min_x \quad \frac{1}{2}||Ax - b||^2 + \lambda||x||_1$
- Group Lasso or MMV $\min_x \quad \frac{1}{2}||Ax - b||^2 + \lambda \sum_{j \in J} ||x_j||$
- Matrix Completion $\min_{X \in \mathrm{R}^{n \times m}} \lambda \sum_{(i,j) \in I} (X_{ij} - M_{ij})^2 + ||X||_*$
- Robust PCA $\min_{X \in \mathrm{R}^{n \times m}} \lambda \| X_{ij} - M_{ij} \|_1 + ||X||_*$
- SICS $\max_X \frac{n}{2}(\log \det X - Tr(SX)) - \lambda||X||_1$

Prox method with nonsmooth term

- Consider: $\boxed{\min_x F(x) = f(x) + g(x)}$

$$|\nabla f(x) - \nabla f(y)| \leq L||x-y||$$

- Quadratic upper approximation



$$f(y)+g(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2\mu}||y-x||^2 + g(y) = Q_{f,\mu}(x,y)$$

$$F(y) \leq f(x) + \frac{1}{2\mu}||x - \mu \nabla f(x)^\top - y||^2 + g(y) = Q_{f,\mu}(x,y)$$

Assume that *g(y)* is such that the above function is easy to optimize over *y*

Example 1 (Lasso and SICS)

$$\boxed{\min_x f(x) + ||x||_1}$$

- Minimize upper approximation function $Q_{f,\mu}(x,y)$ on each iteration

$$\min_y Q_{f,\mu}(x,y) = \min_y f(x) + \frac{1}{2\mu}||x - \mu\nabla f(x)^\top - y||^2 + ||y||_1$$

$$\Updownarrow$$

$$\sum_i \min_{y_i}\left[\frac{1}{2\mu}(y_i - r_i)^2 + |y_i|\right]$$

$$\Updownarrow$$

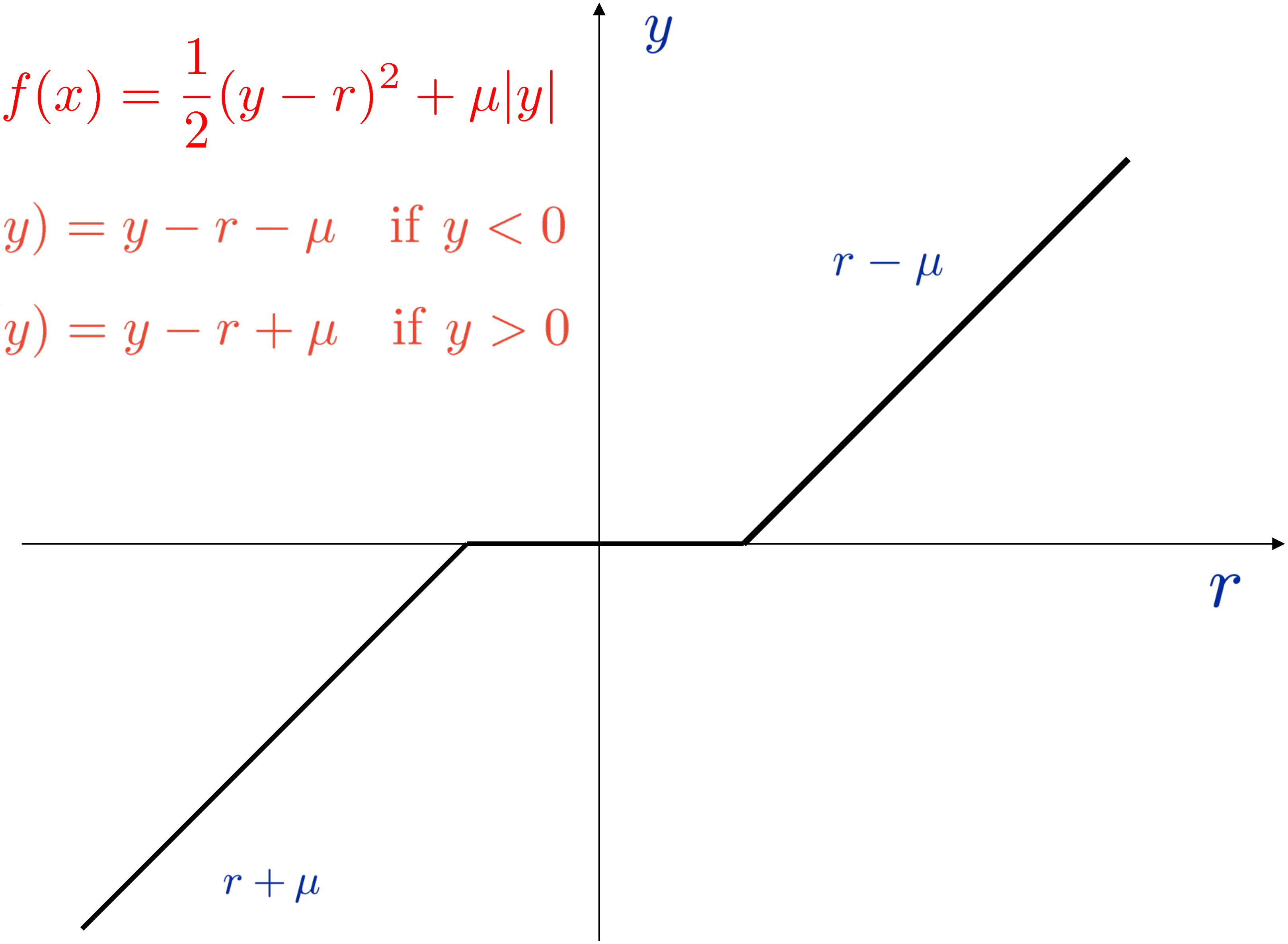
Closed form solution! *O(n)* effort

$$\min_{y_i} \frac{1}{2}(y_i - r_i)^2 + \mu|y_i| \rightarrow y_i^* = \begin{cases} r_i - \mu & \text{if } r_i > \mu \\ 0 & \text{if } -\lambda \le r_i \le \mu \\ r_i + \mu & \text{if } r_i < -\mu \end{cases}$$

$$f(x) = \frac{1}{2}(y - r)^2 + \mu|y|$$

$$f'(y) = y - r - \mu \quad \text{if } y < 0$$

$$f'(y) = y - r + \mu \quad \text{if } y > 0$$



ISTA/Gradient prox method

$$\boxed{\min_x F(x) = f(x) + g(x)}$$

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \text{argmin}_y Q_f(x^k, y)$$

$$Q_{f,\mu}(x, y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

- Problem: does not work for SICS: *log det(X)* is not defined when *X* is not positive definite, hence there is an additional constraint on x^{k+1}

Splitting, alternating linearization and alternating direction methods

Augmented Lagrangian

$$\begin{aligned} \min \quad & f_0(x), \\ \text{s.t.} \quad & f_i(x) = 0,\ i = 1, \dots, m \end{aligned}$$

Augmented Lagrangian function

$$L(x, y) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{m} \frac{1}{2\mu_i} \|f_i(x)\|^2$$

Augmented Lagrangian method

$$\boxed{\begin{aligned} & \text{For } k = 1, 2, \dots \\ & x^k = \text{argmin}_x L(x, \lambda^k) \\ & \lambda_i^{k+1} = \lambda_i^k - \frac{1}{\mu_i} f_i(x^k),\ i = 1, \dots, m \end{aligned}}$$

Alternating directions (splitting) method

- Consider: $\boxed{\min_x F(x) = f(x) + g(x)}$

$$\Updownarrow$$

$$\begin{array}{ll} \min_{x,y} & f(x) + g(y) \\ \text{s.t.} & y = x \end{array}$$

- Relax constraints via Augmented Lagrangian technique

$$\min_{x,y} f(x) + g(y) + \lambda^\top (y - x) + \frac{1}{2\mu} ||y - x||^2 = Q_\lambda(x, y)$$

Assume that *f(x)* and *g(y)* are both such that the above functions are easy to optimize in x or y

Alternating direction method (ADM)

- $x^{k+1} = \min_x Q_\lambda(x, y^k)$
- $y^{k+1} = \min_y Q_\lambda(x^{k+1}, y)$
- $\lambda^{k+1} = \lambda^k + \frac{1}{\mu}(y^{k+1} - x^{k+1})$

Widely used method without complexity bounds

Combettes and Wajs, ‘05

Eckstein and Bertsekas, ‘92,

Eckstein and Svaiter, ’08

Glowinski and Le Tallec, ‘89

Kiwiel, Rosa, and Ruszczynski, ’99

Lions and Mercier ‘79

A slight modification of ADM

- $x^{k+1} = \min_x Q_\lambda(x, y^k)$
- $\lambda^{k+\frac{1}{2}} = \lambda^k + \frac{1}{\mu}(y^k - x^{k+1})$
- $y^{k+1} = \min_y Q_\lambda(x^{k+1}, y)$
- $\lambda^{k+1} = \lambda^{k+\frac{1}{2}} + \frac{1}{\mu}(y^{k+1} - x^{k+1})$

This turns out to be equivalent to……

Goldfarb, Ma and S, ' 10

Alternating linearization method (ALM)

- $x^{k+1} = \min_x Q_g(x, y^k)$
- $y^{k+1} = \min_y Q_f(x^{k+1}, y)$

$$Q_g(x, y) = f(x) + \nabla g(y)^\top (x - y) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

$$\lambda^k = \nabla g(y^k)$$

$$Q_f(x, y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

$$\lambda^{k+1} = \nabla f(x^{k+1})$$

Goldfarb, Ma, S, ‘10

Convergence rate for ALM

- $x^{k+1} = \min_x Q_\lambda(x, y^k)$
- $\lambda^{k+\frac{1}{2}} = \lambda^k + \frac{1}{\mu}(y^k - x^{k+1})$
- $y^{k+1} = \min_y Q_\lambda(x^{k+1}, y)$
- $\lambda^{k+1} = \lambda^{k+\frac{1}{2}} + \frac{1}{\mu}(y^{k+1} - x^{k+1})$

Th: If $\mu \leq 1/L$ then in $O(L/\epsilon)$ iterations finds ϵ -optimal solution

No need to compute the gradient for *f* and *g*

Goldfarb, Ma, S, ’10

Can accelerate ALM

- $x^k := \min_x Q_g(x, z^k)$
- $y^k := \min_y Q_f(x^k, y)$
- $t_{k+1} := (1 + \sqrt{1 + 4t_k^2})/2$
- $z^{k+1} := y^k + \frac{t_k - 1}{t_{k+1}}[y^k - y^{k-1}]$

Th: If $\mu \leq 1/L$ then in $O(\sqrt{L/\epsilon})$ iterations finds ϵ -optimal solution

But now need to compute the gradient of *g* at *z^k*

Goldfarb, Ma, S, ' 10

Applications of alternating linearization method to SICS

Sparse Inverse Covariance Selection

$$\boxed{\max_{X \succ 0}(\underbrace{\ln\det(X) - Tr(SX)}_{f(x)}) - \underbrace{\rho\|X\|_1}_{g(x)}}$$

$$X^{k+1} := \mathrm{argmin}_X \{f(X) + \tfrac{1}{2\mu_{k+1}}\|X - (Y^k + \mu_{k+1}\Lambda^k)\|_F^2\}$$

Eigenvalue decomposition O(p^3) ops. Same as one gradient of *f(X)*

$$Y^{k+1} := \mathrm{argmin}_Y \{g(Y) + \tfrac{1}{2\mu_{k+1}}\|Y - (X^{k+1} - \mu_{k+1}(S - (X^{k+1})^{-1}))\|_F^2\}$$

Shrinkage O(p^2) ops

Sparse Inverse Covariance Selection

$$\boxed{\max_{X\succ 0}(\underbrace{\ln\det(X) - Tr(SX))}_{f(x)} - \underbrace{\lambda\|X\|_1}_{g(x)}}$$

$$X^{k+1} := \operatorname{argmin}_X \{f(X) + \frac{1}{2\mu_{k+1}}\|X - (Y^k + \mu_{k+1}\Lambda^k)\|_F^2\}$$

$V\mathrm{Diag}(d)V^\top$ - the spectral decomposition of $Y^k + \mu_{k+1}(\Lambda^k - S)$

$$\gamma_i = \left(d_i + \sqrt{d_i^2 + 4\mu_{k+1}}\right)/2, \quad i = 1, \ldots, p$$

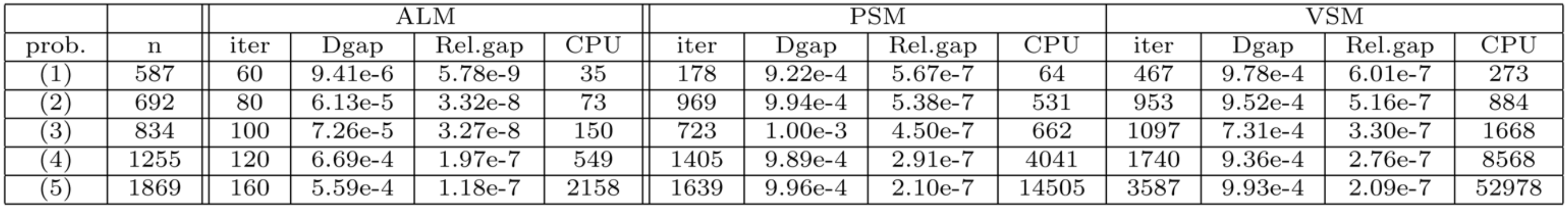
$$X^{k+1} := V\mathrm{Diag}(\gamma)V^\top$$

Eigenvalue decomposition O(p^3) ops. Same as one gradient of *f(X)*

Numerical comparisons

Gene expression networks using the five data sets from Li and Toh(2010)
(1) Lymph node status
(2) Estrogen receptor;
(3) Arabidopsis thaliana;
(4) Leukemia;
(5) Hereditary breast cancer.

PSM by Duchi et al (2008) and VSM by Lu (2009)



		ALM				PSM				VSM			
prob.	n	iter	Dgap	Rel.gap	CPU	iter	Dgap	Rel.gap	CPU	iter	Dgap	Rel.gap	CPU
(1)	587	60	9.41e-6	5.78e-9	35	178	9.22e-4	5.67e-7	64	467	9.78e-4	6.01e-7	273
(2)	692	80	6.13e-5	3.32e-8	73	969	9.94e-4	5.38e-7	531	953	9.52e-4	5.16e-7	884
(3)	834	100	7.26e-5	3.27e-8	150	723	1.00e-3	4.50e-7	662	1097	7.31e-4	3.30e-7	1668
(4)	1255	120	6.69e-4	1.97e-7	549	1405	9.89e-4	2.91e-7	4041	1740	9.36e-4	2.76e-7	8568
(5)	1869	160	5.59e-4	1.18e-7	2158	1639	9.96e-4	2.10e-7	14505	3587	9.93e-4	2.09e-7	52978

Coordinate descent approach

Updating one-two elements of X at a time

Consider changing *only* X_{ij}=X_{ji} at each step for some *i* and *j*.

$$X(\theta) = X + \theta(e_i e_j^\top + e_j e_i^\top)$$

Then the objective function becomes a function of one variable, θ

$$f(\theta) = \log\det(X + \theta e_i e_j^\top + \theta e_j e_i^\top) -$$
$$\langle S, X + \theta e_i e_j^\top + \theta e_j e_i^\top \rangle - \lambda |X_{ij} + \theta e_i e_j^\top + \theta e_j e_i^\top|.$$

Using SMW formula we can write

$$\det(X + \theta e_i e_j^\top + \theta e_j e_i^\top) = \det(X)(1 + 2\theta W_{ij} + \theta^2(W_{ij}^2 - W_{ii}W_{jj}))$$

$$W = X^{-1}$$

Main idea for primal coordinate descent

Given, current *X* and *W*, or each *i* and *j* maximize *f(θ)* as a function of θ

Each step takes *O(1)* operations, hence in *O*(p^2) ops we can try all pairs of *i,j*.

After we try all pairs we can choose the "best" (biggest improvement???)

Perform the update of *X* and *W* in *O(p^2)* operations

$$\bar{W} = W - \theta(\kappa_1 W_i W_j^\top + \kappa_2 W_i W_i^\top + \kappa_3 W_j W_j^\top + \kappa_1 W_j W_i^\top)$$

Each step introduces one nonzero at most, can be easily parallelized.

Conclusions and future directions

- First order optimization methods can be highly effective for structured large scale convex optimization problems.
- They can exploit specific structure of the problem and the solution.
- Second order methods can improve the convergence substantially, but have to be applied with care probably as a second stage. This avenue can be further exploited.
- Further improvement should be achieved via parallelization of these methods.
- New optimization models for graphical models can be explored and the existing methods extended.

Thank you!

Block coordinate ascent

Update one row and one column of the dual matrix W at each step

$$W = \begin{bmatrix} W_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

$$\max_{W \succ 0}\{\tfrac{m}{2}\ln(\det(W)) - mp/2 : \text{ s.t. } \tfrac{m}{2}||W - A||_\infty \le \lambda\}$$

$$\ln\det W = \ln(\det(W_{11})(w_{22} - {w_{12}}^T W_{11}^{-1} w_{12}))$$

Block coordinate ascent subproblem

Update one row and one column of the dual matrix W at each step

$$W = \begin{bmatrix} W_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

$$\begin{array}{ll} \max_{w_{12},w_{22}} & \ln(w_{22} - {w_{12}}^T W_{11}^{-1} w_{12})) \\ \text{s.t.} & \|w_{12} - a_{12}\|_\infty \leq \frac{2}{m}\lambda, |w_{22} - a_{22}| \leq \frac{2}{m}\lambda \end{array}$$

$$\min_{w_{12}} \{ w_{12}^\top W_{11}^{-1} w_{12} : \quad \text{s.t.} \quad \|w_{12} - a_{12}\|_\infty \leq \frac{2}{m}\lambda,$$

Subproblem reformulation

$$\min_{w_{12}}\{w_{12}^\top W_{11}^{-1} w_{12} : \quad \text{s.t.} \quad \|w_{12} - a_{12}\|_\infty \le \frac{2}{m}\lambda,$$

$$w_{12} = W_{11}\beta$$

$$\min_{\beta}\{\beta^\top W_{11}\beta : \quad \text{s.t.} \quad \|W_{11}\beta - a_{12}\|_\infty \le \frac{2}{m}\lambda\}$$

Dual subproblem

$$\min_{w_{12}}\{w_{12}^{\top} W_{11}^{-1} w_{12} : \quad \text{s.t.} \quad \|w_{12} - a_{12}\|_{\infty} \le \frac{2}{m}\lambda,$$

$$w_{12} = W_{11}\beta$$

$$\min_{\beta}\{\beta^{\top} W_{11}\beta : \quad \text{s.t.} \quad \|W_{11}\beta - a_{12}\|_{\infty} \le \frac{2}{m}\lambda\}$$

$$\min_{\beta}\{\|W_{11}^{1/2}\beta - W_{11}^{-1/2}a_{12}\|^2 + \frac{4}{m}\lambda\|\beta\|_1$$

The dual subproblem is the Lasso problem

Remember coordinate descent for Lasso

$$\min_{x_i} \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$$

Choose one variable x_i and column A_i.
Let $\bar{x}$ and $\bar{A}$ correspond to the fixed part

$$\min_{x_i} \quad \frac{1}{2}(A_i x_i + \bar{A}\bar{x} - b)^2 + \lambda|x_i|$$

Soft-thresholding operator

$$\min_{x_i} \frac{1}{2}(x_i - r)^2 + \lambda|x| \rightarrow x_i = \begin{cases} r - \lambda & \text{if } r > \lambda \\ 0 & \text{if } -\lambda \leq r \leq \lambda \\ r + \lambda & \text{if } r < -\lambda \end{cases}$$

$$r = -A_i^\top(\bar{A}\bar{x} - b)/||A_i||^2, \ \lambda \rightarrow \lambda/||A_i||^2$$

Remember coordinate descent for Lasso

$$\min_{x_i} \quad \frac{1}{2}\|W_{11}^{1/2}\beta - W_{11}^{-1/2}a_{12}\|^2 + \lambda\|\beta\|_1$$

$$\min_{\beta_i} \frac{1}{2}(\beta_i - r)^2 + \lambda|x| \rightarrow \beta_i = \begin{cases} r - \lambda & \text{if } r > \lambda \\ 0 & \text{if } -\lambda \leq r \leq \lambda \\ r + \lambda & \text{if } r < -\lambda \end{cases}$$

$$r = -((W_{11})_i{}^\top \bar{\beta} - (a_{12})_i)/(W_{11})_{ii}, \ \lambda \rightarrow \lambda/(W_{11})_{ii}$$

No need to compute $W^{1/2}$