Fast first-order methods for convex optimization with line search

Katya Scheinberg
Lehigh University
(joint work with X. Bai, D. Goldfarb and S. Ma)

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

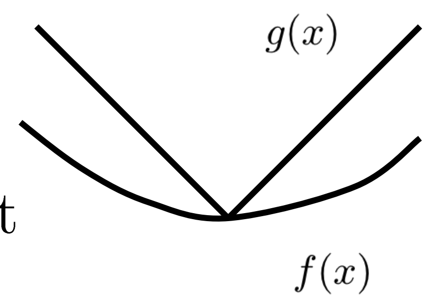
- At the core of many statistical machine learning problems lies an optimization problem, often convex, from a well-studied class (LP, QP, SDP).
- These problems are very large and dense in terms of data.
- IPMs are often too expensive to use. ML community initially assumed that traditional optimization 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.

Problem under consideration

- Problem:

$$\min_x F(x) = f(x) + g(x),$$

- Assumptions:



- $f(x)$ is convex with Lipschitz continuous gradient

$$|\nabla f(x) - \nabla f(y)| \leq L||x - y||, \text{ for some } L, \ \forall x, y$$

- $g(x)$ is convex, possibly nonsmooth and "easy" (in some sense).

Many applications involve optimization of this form

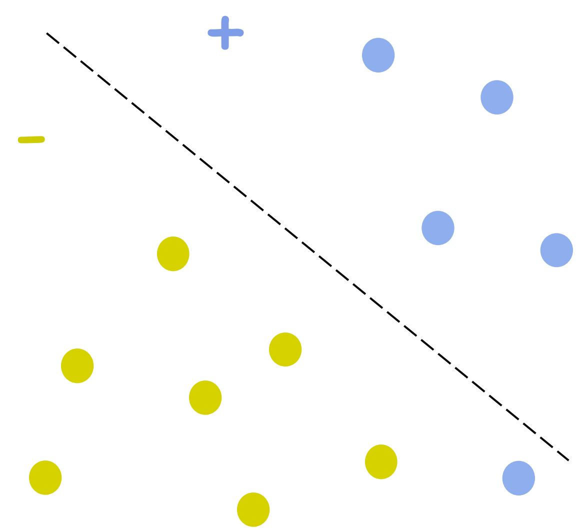
Support vector machine classification

- Breast cancer diagnostics
  - Test results of a group of patients, some have been diagnosed with cancer, other do not have it. Find how the test can predict high risk patients.
- Spam filter
  - From a list of spam and nonspam labeled emails learn to detect spam automatically.
- Genetic disease
  - find away to identify high risk individuals based on gene expression data.
- Target customer groups
  - By demographic data and past purchases find customers most likely to buy certain products.

Support Vector Machines

- Problem:

$$\min_{w,\beta} \rho ||w||^2 + \sum_i \min\{0, (1 - b_i(w^\top a_i + \beta))\}$$

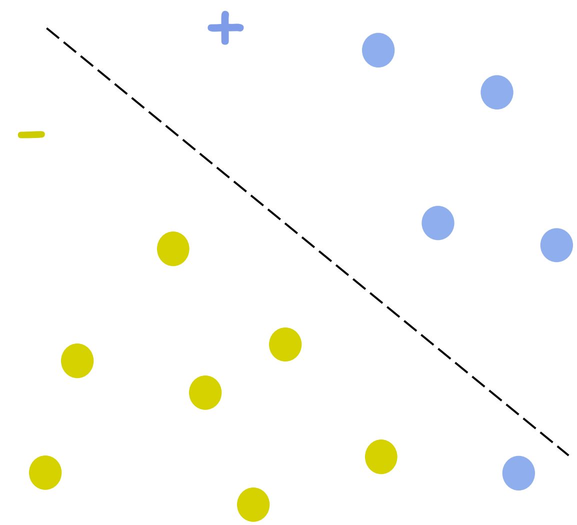


- a_i - data points
- $b_i = \{+1, -1\}$ - data label
- $w^\top x + \beta$ - the separating hyperplane

Linear Support Vector Machines

- Problem:

$$\min_{w,\beta} \rho||w||_1 + \sum_i \min\{0, (1 - b_i(w^\top a_i + \beta))\}$$



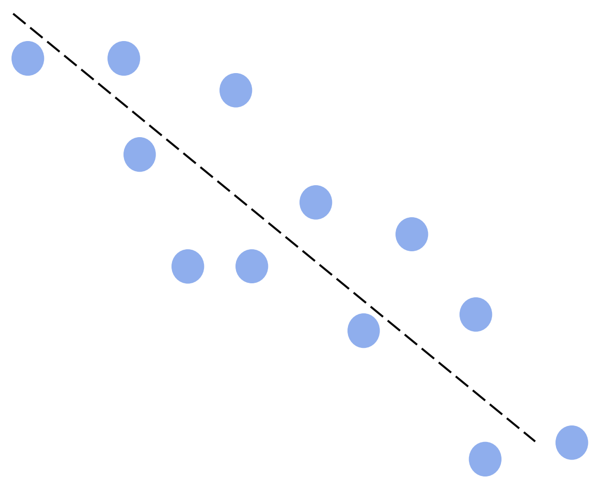
- a_i - data points
- $b_i = \{+1, -1\}$ - data label
- $w^\top x + \beta$ - the separating hyperplane

Sparse models

- Compresses sensing, MRI
  - Recover sparse signal *x*, which satisfies $Ax=b$.
- Sparse least square regression (Lasso)
  - Find linear regression models while selecting important features.
- Regression models using polynomials with variable selection
  - birthweight dataset from Hosmer and Lemeshow (1989), weight of 189 babies and 8 variables per mother. Predictive models for birthweight.

Lasso regression

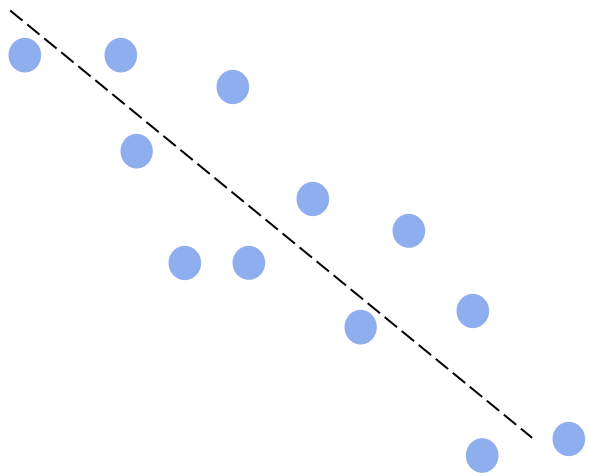
- Problem: $$\min_x \quad \frac{1}{2}||Ax - b||^2 + \rho||x||_1$$



- Rows of A, a_i - data points
- $b_i \in R$ - labels
- $x^\top a = \beta$ - linear model
- x is sparse

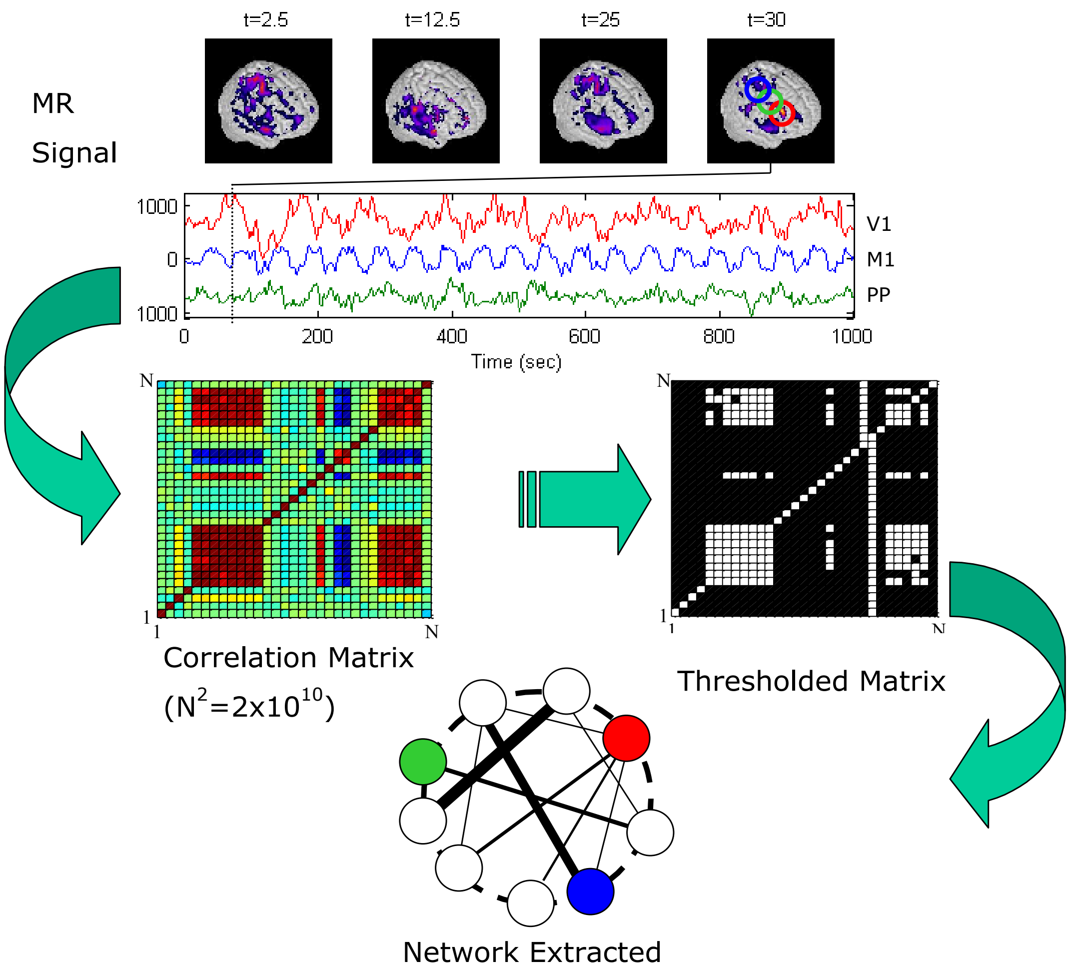
Group Lasso regression

- Problem: $$\min_x \quad \frac{1}{2}||Ax-b||^2 + \rho \sum_i ||x_i||_2$$



- Assume that columns of A form groups of correlated features.
- Find sparse vector x where nonzeros are selected according to groups
- x_i is a subvector of x corresponding to the i-th group of features.

FMRI Analysis and schizophrenia prediction



Measuring blood oxigination in voxels of the brain.

Construct predictive models based on FMRI data, use to predict/diagnose schizophrenia or classify "states of mind".

Sparse Inverse Covariance Selection

- Problem:

Given n random varibles $\boldsymbol{p} = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_n\}$

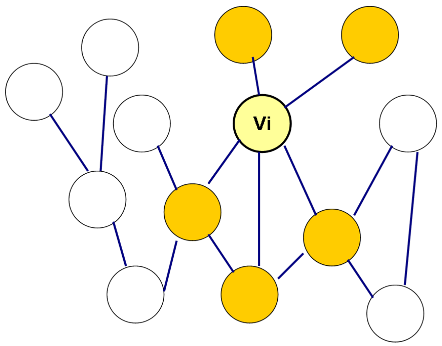
Find multivariate Gaussian probability density function:

$$P(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{p}-\boldsymbol{\mu})\right)$$

- Formulation:

$$\max_X \frac{m}{2}(\log\det X - Tr(AX)) - \rho||X||_1$$

$$(\|X\|_1 = \sum_{ij} |X_{ij}|, \quad A = \frac{1}{m} BB^\top)$$



- $\mu = 0$
- $(\Sigma^{-1})_{ij}$ is zero if p_i and p_j are conditionally independent.

Summary and add'l examples

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

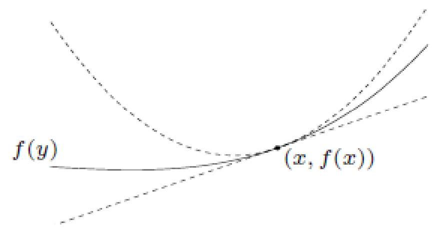
First-order methods applied to problems of the form $f(x)+g(x)$

Prox method with nonsmooth term

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

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

- Quadratic upper approximation



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

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

ISTA/prox gradient projection

$$\min_x F(x) = f(x) + g(x)$$

- Minimize quadratic upper approximation on each iteration

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

- O(L/ϵ) complexity: If $\beta/L \leq \mu \leq 1/L$ then in k iterations finds solution

$$x^k : \; F(x^k) - F(x^*) \leq \frac{2L\|x^k - x^*\|^2}{k}$$

Fast first-order method

Nesterov, Beck & Teboulle

$$\min_x F(x) = f(x) + g(x)$$

- Minimize upper approximation at a "shifted" point.

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

$$t_{k+1} := (1 + \sqrt{1 + 4t_k^2})/2$$

$$y^{k+1} := x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

- $O(\sqrt{L/\epsilon})$ complexity: If $\beta/L \leq \mu \leq 1/L$ then in k iterations finds solution

$$x^k : \; F(x^k) - F(x^*) \leq \frac{2L\|x^k - x^*\|^2}{k^2}$$

Specifically for CS setting and Lasso

$$\min_x \|Ax - b\|^2 + \rho\|x\|_1$$

$$\underbrace{}_{f(x)} \quad \underbrace{}_{g(x)}$$

$$\nabla f(x) = A^\top(Ax - b)$$

$$x^{k+1} = \min_y \frac{1}{2\mu}\|(y^k - \mu A^\top(Ay^k - b)) - y\|^2 + \rho\|y\|_1$$

2 matrix/vector multiplications + shrinkage operator per iteration

$\sqrt{\frac{2\|x^0 - x^*\|^2}{\mu\epsilon}}$ iteration bound

Choosing prox parameter via backtracking

Iterative Shrinkage Threshholding Algorithm (ISTA)

- Minimize quadratic upper relaxation on each iteration

$$x^{k+1} = \mathrm{argmin}_y Q_{f,\mu_k}(x^k, y) = f(x^k) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2\mu_k}||x^k - y||^2 + g(y)$$

- Using backtracking find μ_k such that

$$F(x^{k+1}) \leq Q_{f,\mu_k}(x^k, x^{k+1})$$

- In k iterations finds solution

$$F(x^k) - F(x^*) \leq \frac{2\|x^k - x^*\|^2}{\sum_k \mu_k} = \frac{2\|x^k - x^*\|^2}{\bar{\mu}(k)k} \quad \bar{\mu}(k) = \frac{\sum \mu_k}{k}$$

Beck&Teboulle, Tseng, Auslender&Teboulle, 2008

Fast Iterative Shrinkage Threshholding Algorithm (FISTA)

Minimize quadratic upper relaxation on each iteration

$$x^{k+1} = \text{argmin}_y Q_{f,\mu_k}(y^k, y) = f(y^k) + \nabla f(y^k)^\top (y - y^k) + \frac{1}{2\mu_k}||y^k - y||^2 + g(y)$$

Using backtracking find $\mu_k \leq \mu_{k-1}$ such that

$$F(x^{k+1}) \leq Q_{f,\mu_k}(y^k, x^{k+1})$$

$$t_{k+1} := (1 + \sqrt{1 + 4t_k^2})/2$$

$$y^{k+1} := x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

In *k* iterations finds solution

$$x^k : \; F(x^k) - F(x^*) \leq \frac{2L\|x^k - x^*\|^2}{k^2}$$

Beck&Teboulle, Tseng, 2008

Fast Iterative Shrinkage Threshholding Algorithm (FISTA)

Minimize quadratic upper relaxation on each iteration

$$x^{k+1} = \text{argmin}_y Q_{f,\mu_k}(y^k, y) = f(y^k) + \nabla f(y^k)^\top (y - y^k) + \frac{1}{2\mu_k} ||y^k - y||^2 + g(y)$$

Using backtracking find $\mu_k \leq \mu_{k-1}$ such that

$$t_k := \quad (1 + \sqrt{1 + 4t_{k-1}^2})/2$$

$$y^k := \quad x^k + \frac{t_{k-1} - 1}{t_k}[x^k - x^{k-1}]$$

$$F(x^{k+1}) \leq Q_{f,\mu_k}(y^k, x^{k+1})$$

In *k* iterations finds solution

$$x^k : \; F(x^k) - F(x^*) \leq \frac{2L\|x^k - x^*\|^2}{k^2}$$

Find $\mu_k \leq \mu_{k-1}$ such that

$$t_k := \quad (1 + \sqrt{1 + 4t_{k-1}^2})/2$$

$$y^k := \quad x^k + \frac{t_{k-1} - 1}{t_k}[x^k - x^{k-1}]$$

Cycle to find μ_k

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

$$F(x^{k+1}) \leq Q_{f,\mu_k}(y^k, x^{k+1})$$

Need to compute *Ax-b*

Convergence rate:

$$F(x^k) - F(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\mu_k t_k^2}$$

To allow for larger μ_k we need to reduce t_k

and vice versa

Find μ_k such that

$$\mu_{k-1} t_{k-1}^2 \geq \mu_k t_k (t_k - 1)$$

$$y^k := x^k + \frac{t_{k-1} - 1}{t_k}[x^k - x^{k-1}]$$

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

$$F(x^{k+1}) \leq Q_{f,\mu_k}(y^k, x^{k+1})$$

$$\Downarrow$$

$$F(x^k) - F(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\mu_k t_k^2}.$$

Goldfarb & S. 2010

FISTA with full backtracking

Find μ_k such that

Cycle to find μ and t

$$t_{k+1} := (1 + \sqrt{1 + 4\frac{\mu_k}{\mu_{k-1}} t_k^2})/2$$

$$y^k := \quad x^k + \frac{t_{k-1} - 1}{t_k}[x^k - x^{k-1}]$$

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

$$F(x^{k+1}) \leq Q_{f,\mu_k}(y^k, x^{k+1})$$

$$\Downarrow$$

$$F(x^k) - F(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\mu_k t_k^2}$$

Goldfarb & S. 2010

FISTA with full backtracking

Find μ_k such that

Cycle to find
μ and t

Bla-bla-bla….

$$\mu_k t_k^2 \geq (\sum_{i=1}^{k} \sqrt{\mu_i}/2)^2 = \bar{\mu}(k)k^2$$

$$\bar{\mu}(k) = ((\sum_{i=1}^{k} \sqrt{\mu_i})/k)^2$$

$$F(x^k) - F(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\bar{\mu}(k)k^2}$$

Goldfarb & S. 2010

Computational Results

Computational results

- Spear1 (1024×512), with $\rho = 1$.
- Dynamic range is 3.02e+4.
- Sparsity is 18 (i.e. 18 nonzero elements in the true solution).
- Optimal tolerance is set to be 1e-12. FISTA(100) = 5.3839e+5. FISTA(500) = 1.2799e+5. FISTA(1000) = 1.0035e+5.

solver	iter	mult	iter	mult	iter	mult	final iter	mult	final Obj.
FISTA	100	206	500	1006	1000	2006	1361	2728	9.996e+4
FISTA_bt	69	170	283	619	627	1343	711	1527	9.996e+4
SpaRSA	98	196	1487	2974	1689	3378	1704	3408	9.996e+4
YALL1	18	55	30	91	89	268	197	592	9.997e+4

Computational results

- Spear1 (1024×512), with $\rho = 0.01$.
- Dynamic range is 3.02e+4.
- Sparsity is 18 (i.e. 18 nonzero elements in the true solution).
- Optimal tolerance is set to be 1e-12. FISTA(100) = 6.0980e+3. FISTA(500) = 5.8943e+3. FISTA(1000) = 5.4176e+3.

solver	iter	mult	iter	mult	iter	mult	final iter	mult	final Obj.
FISTA	100	206	500	1006	1000	2006	19872	39750	999.4
FISTA_bt	79	190	372	804	746	1607	13655	30005	999.4
SpaRSA	30	60	687	1374	5024	10048	-	-	-
YALL1	65	196	65	196	66	199	257	772	1015.3

Computational results

- Spear3 (1024×512), with $\rho = 0.1$.
- Dynamic range is 2.7535e+4.
- Sparsity is 6 (i.e. 6 nonzero elements in the true solution).
- Optimal tolerance is set to be 1e-12. FISTA(100) = 1.1825e+4. FISTA(500) = 1.1793e+4. FISTA(1000) = 1.1784e+4.

solver	iter	mult	iter	mult	iter	mult	final iter	mult	final Obj.
FISTA	100	211	500	1011	1000	2011	28539	57089	7.33e+3
FISTA_bt	220	490	265	580	316	687	6077	14069	7.33e+3
SpaRSA	5	10	264	528	1215	2430	-	-	Failed
YALL1	541	1624	541	1624	541	1624	1661	4984	7.33e+3

Computational results

- Bdata1 (1036×1036), with $\rho = 0.1$
- Dynamic range is 5.9915.
- Sparsity is 16 (i.e. 16 nonzero elements in the true solution).
- Optimal tolerance is set to be 1e-12. FISTA(10) = 3.490836e+002. FISTA(50) = 3.490804e+002. FISTA(100) = 3.490804e+002.

solver	iter	mult	iter	mult	iter	mult	final iter	mult	final Obj.
FISTA	10	23	50	103	100	203	105	213	349.08
FISTA_bt	8	21	44	111	76	181	90	212	349.08
SpaRSA	4	8	34	68	70	140	80	160	349.08
YALL1	11	34	108	325	233	700	2263	6790	349.08

Computational results

- Sparco 2(2048×1024), with $\rho = 0.1$
- Dynamic range is 2.
- Sparsity is 2 (i.e. 16 nonzero elements in the true solution).
- Optimal tolerance is set to be 1e-12. FISTA(10) = 13.07180. FISTA(50) = 8.187212. FISTA(100) = 2.710062.

solver	iter	mult	iter	mult	iter	mult	final iter	mult	final Obj.
FISTA	10	24	50	104	100	204	207	418	2.22278
FISTA_bt	6	18	38	97	78	187	173	387	2.22278
SpaRSA	7	14	11	22	72	144	99	198	2.22278
YALL1	10	31	10	31	19	58	262	787	2.22278

Complexity bounds on alternating linearization methods

Alternating directions method

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

$$\Updownarrow$$

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

- Relax constraints via Augmented Lagrangian technique

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

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

Sparse Inverse Covariance Selection

$$\max_{X \succ 0} (\ln\det(X) - Tr(AX)) - \rho \|X\|_1$$

f(x) *g(x)*

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

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

$$Y^{k+1} := \text{argmin}_Y \{ g(Y) + \frac{1}{2\mu_{k+1}} \|Y - (X^{k+1} - \mu_{k+1}(A - (X^{k+1})^{-1}))\|_F^2 \}$$

Shrinkage $O(n^2)$ ops

Lasso or group Lasso

$$\min_x \|Ax - b\|^2 + \rho\|x\|_1$$

f(x) *g(x)*

$$x^{k+1} := \text{argmin}_x\{f(x) + \frac{1}{2\mu_{k+1}}\|x - (y^k + \mu_{k+1}\lambda^k)\|^2\}$$

Matrix inverse, can take $O(n^3)$ ops. But can also be *O(n ln n)* for special A.

$$y^{k+1} := \text{argmin}_y\{g(y) + \frac{1}{2\mu_{k+1}}\|y - (x^{k+1} - \mu_{k+1}A^\top(Ax - b))\|^2\}$$

Shrinkage $O(n^2)$ ops

Alternating direction method (ADM)

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

Widely used method without complexity bounds

A slight modification of ADM

Goldfarb, Ma, S, '09-'10

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

This turns out to be equivalent to……

Alternating linearization method (ALM)

Goldfarb, Ma, S, '09-'10

- $x^{k+1} = \min_x Q_{g,\mu_g}(x, y^k)$
- $y^{k+1} = \min_y Q_{f,\mu_f}(x^{k+1}, y)$

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

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

Fast ALM (FALM)

Goldfarb, Ma, S, '09-'10

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

Complexity results

FISTA

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

FALM

$$F(x^k) - F(x^*) \le \frac{2L(f)L(g)\|x^0 - x^*\|^2}{(L(f) + L(g))k^2}.$$

Experiments on SICS

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

Experiments in CS

Comparison of algorithms on image recovery problem.
Here matrix inverse take O(n ln n) ops, as do mat-vec multiplications.

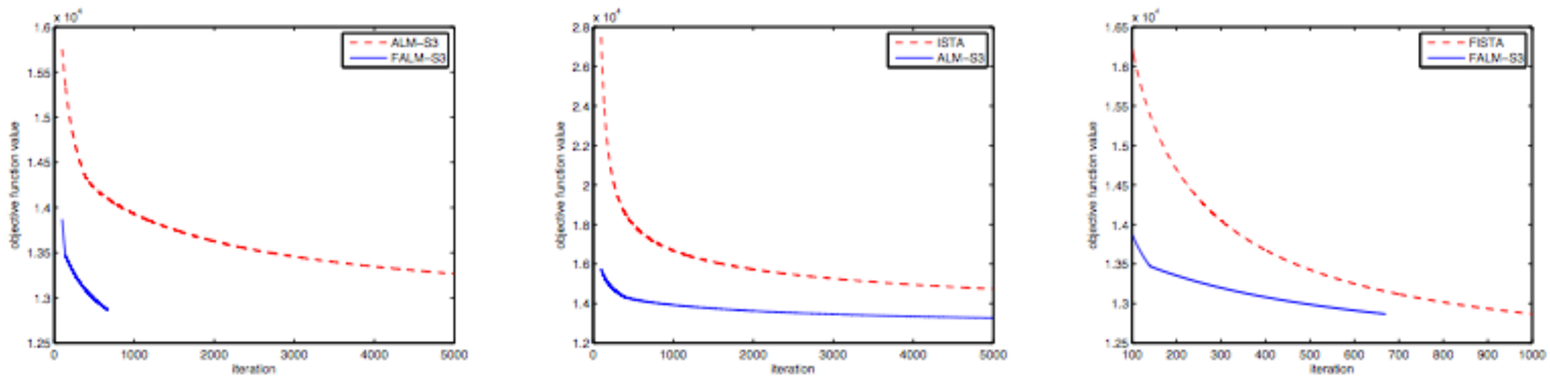


FIG. 4.1. *The comparison of objective function values versus number of iterations for Algorithms ISTA, FISTA, ALM-S3, FALM-S3 for* $\rho = 0.001$

FALM with backtracking

Goldfarb, S, ‘10

- $x^{k+1} := \text{argmin}_x Q_{g,\mu_k^g}(x, z^k),\ F(x^{k+1}) \leq Q_{f,\mu_k}(z^k, x^{k+1})$

- $y^{k+1} := \text{argmin}_y Q_{f,\mu_k^f}(x^{k+1}, y)\ F(y^{k+1}) \leq Q_{f,\mu_k}(x^{k+1}, y^{k+1})$

- $t_{k+1} := (1 + \sqrt{1 + 4\frac{\bar{\mu}_{k+1}}{\bar{\mu}_k} t_k^2})/2$

- $z^{k+1} := y^{k+1} + \frac{t_k - 1}{t_{k+1}}[y^{k+1} - y^k]$

$$F(x^k) - F(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\bar{\mu}(k)k^2}.$$

$$\sqrt{\bar{\mu}(k)} = (\textstyle\sum_{i=1}^k \sqrt{\bar{\mu}_i})/k,\ \ \bar{\mu}_i = \frac{\mu_i^f + \mu_i^g}{2}$$

Conclusion and Future work

- Performing backtracking carefully is possible and desirable in accelerated first order methods.
- The trade-offs are different and need to be explored for particular applications beyond CS.
- Accelerated alternating direction methods can utilize the same ideas.
- Careful implementation is being considered.
- Combining backtracking with inexact evaluations may be beneficial.
- Seeking problems where average behavior differs greatly from the worst case.

Thank you!