Optimization Methods in Sparse Approximation

With Applications to Basis Pursuit and Gaussian Graphical Models

Syed Rahman

Department of Statistics
University of Florida

Basis Pursuit Denoising

- The basis pursuit problem is as follows:

$$\min_{\beta} \|\beta\|_1 \text{ s.t. } y = X\beta$$

- In the presence of noise, we can reformulate this problem as

$$\min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- The focus of this talk will be to optimization methods for this problem

Gaussian Graphical Models: Sparsity in Ω

- Let Y be a p-dimensional random vector with a $N_p(0, \Sigma = \Omega^{-1})$ distibution
- $\Omega = ((\omega_{ij}))_{1 \leqslant i,j \leqslant p}$
- $\omega_{ij} = Cov(Y_i, Y_j \mid Y_{-(i,j)})$
- $\omega_{ij} = 0$ if and only if the i^{th} and j^{th} variables are conditionally independent given the other variables
- Zeros in Ω encode conditional independence under Gaussianity

Concentration Graphical Models: Connections with graphs

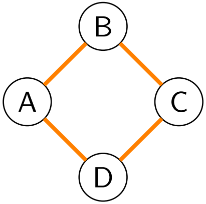
- Obtain a sparse estimate for Ω by minimizing the constrained objective function:

$$\hat{\Omega} = \underset{\Omega \succ 0}{\operatorname{argmin}} \left(\underbrace{\operatorname{trace}(\Omega S) - \log|\Omega|}_{\text{log-likelihood}} + \underbrace{\lambda \|\Omega\|_1}_{\substack{\text{penalty term to} \\ \text{induce sparsity/zeros}}} \right) \tag{1}$$

- The sparsity pattern in Ω can be represented by a graph, $G = (V, E)$.

- $V = \{1, \ldots, p\}$ and set E of edges is such that $\omega_{ij} \neq 0 \Leftrightarrow (i, j) \in E$.

$$\Omega = \begin{array}{cc} & \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ & \left(\begin{array}{cccc} 4.29 & 0.65 & 0 & 0.8 \\ 0.65 & 4.25 & 0.76 & 0 \\ 0 & 0.76 & 4.16 & 0.8 \\ 0.80 & 0 & 0.80 & 4 \end{array} \right) \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \end{array}$$



Subgradient Descent

- Hence the goal is to solve problems of the form:

$$\underset{\beta}{\operatorname{argmin}}\, F(\beta) = \underset{\beta}{\operatorname{argmin}}\, g(\beta) + h(\beta)$$

  where $g(\beta)$ is convex and differentiable and $h(\beta)$ is convex, but non-differentiable.

- The basic update in a subgradient algorithm in such a case is

$$\beta^{k+1} = \beta^{k} - t_k \partial F(\beta^{k})$$

  where t_k is the step size and $\partial F(\beta)$ is the subgradient of $F(\beta)$

What is a subgradient?

- Recall that a gradient of **differentiable** $F : \mathbb{R}^n \to \mathbb{R}$ at β satisfies for all $\eta \in \mathbb{R}^n$

$$F(\eta) \geqslant F(\beta) + \nabla F(\beta)^t(\eta - \beta)$$

- A subgradient of convex $F : \mathbb{R}^n \to \mathbb{R}$ at β is any $g \in \mathbb{R}^\ltimes$ such that for all $\eta \in \mathbb{R}^n$ we have

$$F(\eta) \geqslant F(\beta) + g^t(\eta - \beta)$$

- If F is differentiable, $g = \nabla F$
- When $h(\beta) = ||\beta||_1$, the subgradient is equal to s, where

$$s_i = \begin{cases} \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}$$

Subgradient methods for BP:

- For the BP problem, the subgradient is $-X^t(y - X\beta) + \lambda s$ with

$$s_i = \begin{cases} \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ 0 & \text{if } \beta_i = 0 \end{cases}$$

- Hence the basic update is:

$$\beta^k = \beta^{k-1} + t_k(X^t(y - X\beta^{k-1}) - \lambda s^{k-1})$$

- For back-tracking line search, fix $\eta \in (0, 1)$. At each iteration, while

$$F(\beta - t\partial F(\beta)) > F(\beta) - \frac{t}{2}\|\partial F(\beta)\|^2$$

let $t = \eta t$.

Subgradient methods for Gaussian Graphical Models:

- For the *glasso* problem, the subgradient is $S - \Omega^{-1} + \lambda\Gamma$ with

$$\Gamma_{ij} = \begin{cases} \text{sign}(\Omega_{ij}) & \text{if } \Omega_{ij, i \neq j} \neq 0 \\ 0 & \text{if } \Omega_{ij} = 0, \Omega_{ij, i=j} \end{cases}$$

- Hence the basic update is:

$$\Omega^k = \Omega^{k-1} + t_k(S - (\Omega^{k-1})^{-1} + \lambda\Gamma^{k-1})$$

Convergence of subgradient methods:

Theorem

For fixed step sizes, the sudgradient method satisfies

$$\lim_{k\to\infty} F(\beta^k) \leqslant F(\beta^*) + \frac{L^2 t}{2}$$

with convergenve rate of $O(\frac{1}{\sqrt{k}})$.

where $\left|F(\beta^1) - F(\beta^2)\right| \leqslant L \left|\left|\beta^1 - \beta^2\right|\right|$

Theorem

For diminshing step sizes, the sudgradient method satisfies

$$\lim_{k\to\infty} F(\beta^k) = F(\beta^*)$$

with convergenve rate of $O(\frac{1}{\sqrt{k}})$.

Proximal Gradient Methods:

- The prioximal operator for $h(\beta)$ is:

$$\operatorname{prox}_t(\beta) = \underset{\eta}{\operatorname{argmin}} \frac{1}{2t} ||\beta - \eta||_2^2 + h(\eta)$$

- The proximal gradient method to minimize $F(\beta) = g(\beta) + h(\beta)$ is:

$$\beta^{(k)} = \operatorname{prox}_{t_k h}(\beta^{(k-1)} - t_k \nabla g(\beta^{(k-1)}))$$

- To see why this works, note that

$$\begin{aligned}
\beta^{+} &= \underset{\eta}{\operatorname{argmin}}(h(\eta) + \frac{1}{2t} ||\eta - \beta + t\nabla g(\beta)||_2^2) \\
&= \ldots \\
&= \underset{\eta}{\operatorname{argmin}}(h(\eta) + g(\beta) + \nabla g(\beta)^t(\eta - \beta) + \frac{1}{2t} ||\eta - \beta||_2^2)
\end{aligned}$$

How Proximal Gradient methods work:

- Recall, the 2^{nd} order Taylor series approximation to $g(\eta)$ near β is

$$\begin{aligned} g(\eta) =& g(\beta) + \nabla g(\beta)^t(\eta - \beta) + (\eta - \beta)^t \nabla^2 g(\beta)(\eta - \beta) \\ &\leqslant g(\beta) + \nabla g(\beta)^t(\eta - \beta) + L(\eta - \beta)^t(\eta - \beta) \end{aligned}$$

  where the function $\nabla g(\beta)$ has Lipschitz constant L.

- Hence, we are essentially minimizing $h(\eta)$ plus a simple local model of $g(\eta)$ around β.

ISTA for BP:

- The proximal operator for the ℓ_1 penalty, $h(\beta) = \lambda \|\beta\|_1$ is

$$\begin{aligned} \text{prox}_t(\beta) &= \underset{\eta}{\text{argmin}} \frac{1}{2t} \|\beta - \eta\|_2^2 + \lambda \|\eta\|_1 \\ &= S_{\lambda t}(\beta) \end{aligned}$$

where $[S_{\lambda t}(\beta)]_i = \text{sign}(x_i) * \max\{|\beta_i| - \lambda t, 0\}$

- In addition, $\nabla g(\beta) = -X^t(y - X\beta)$
- Hence the ISTA update is:

$$\beta^k = S_{\lambda t_k}(\beta^{k-1} + t_k X^t(y - X\beta^{k-1}))$$

Choice of step-size:

- Note that for BP, we have that

$$\begin{aligned}\|\nabla g(\beta_1) - \nabla g(\beta_2)\|_2 &\leqslant L\,\|\beta_1 - \beta_2\|_2 \\ &= \lambda_{max}(X^t X)\,\|\beta_1 - \beta_2\|_2\end{aligned}$$

- Hence set $t_k = 1/L$
- If L is difficult to attain, use back-tracking line-search

FISTA for BP:

- In 1983, Nesterov proposed the following Accelarated gradient descent algorithm for convex, differentiable functions $g(\beta)$:

$$\beta^{k+1} = \eta^k - t_k \nabla g(\eta^k)$$
$$\eta^{k+1} = (1-\gamma_k)\beta^{k+1} + \gamma_k \beta^k$$

  with convergence rate $O(\frac{1}{k^2})$.

- FISTA is essentially this method combined with ISTA. The basic updates are as follows:

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$$
$$\gamma = \beta^{k-1} + \frac{t_k - 1}{t_{k+1}}(\beta^{k-1} - \beta^{k-2})$$
$$\beta^k = S_{\lambda t_k}(\gamma + t_k X^t(y - X\gamma))$$

ISTA/FISTA for Gaussian Graphical Models:

- Recall that we want to minimize

$$\hat{\Omega} = \underset{\Omega \succ 0}{\operatorname{argmin}} \left(\operatorname{trace}(\Omega S) - \log|\Omega| + \lambda \|\Omega\|_1 \right)$$

- Now $\nabla(\operatorname{trace}(\Omega S) - \log|\Omega|) = S - \Omega^{-1}$

- Hence the basic graphical-ISTA update is:

$$\Omega^{k+1} = S_{\lambda t_k}(\Omega^k + t_k(S - (\Omega^k)^{-1}))$$

- And the basic graphical-FISTA update is:

$$\begin{aligned} \Omega^{k+1} &= S_{\lambda t_k}(\zeta^k + t_k(S - (\zeta^k)^{-1})) \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ \zeta^k &= \Omega^{k+1} + \frac{t_k - 1}{t_{k+1}}(\Omega^{k+1} - \Omega^k) \end{aligned}$$

Convergence for ISTA/FISTA:

Theorem

Let β^k be a sequence generated by either of the ISTA algorithms as described above. Then for any $k \geqslant 1$

$$F(\beta_k) - F(\beta^*) \leqslant \frac{\alpha L(g) \|\beta_0 - \beta^*\|_2}{2k}$$

where $\alpha = 1$ for constant step size and $\alpha = \eta$ for back-tracking line search.

Theorem

Let β^k be a sequence generated by either of the FISTA algorithms as described above. Then for any $k \geqslant 1$

$$F(\beta_k) - F(\beta^*) \leqslant \frac{\alpha L(g) \|\beta_0 - \beta^*\|_2}{(k+1)^2}$$

where $\alpha = 1$ for constant step size and $\alpha = \eta$ for back-tracking line search.

ADMM for BP:

- ADMM mixes the decomposability of the *dual ascent method* with the superior convergence properties of the *method of multipliers*.

- Recall the BP problem:

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- This is equivalent to:

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\gamma\|_1 \ \text{ s.t. } \ \beta = \gamma$$

- The augmented Lagrangian in this case is:

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\gamma\|_1 + \frac{\rho}{2}\|\beta - \gamma\|_2^2 \ \text{ s.t. } \ \beta = \gamma$$

ADMM for BP continued:

- We can rewrite this as:

$$L(\beta, \gamma, \eta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\gamma\|_1 + \frac{\rho}{2}\|\beta - \gamma\|_2^2 + \eta^t(\beta - \gamma)$$

- To minimize this we need to following updates:

$$\begin{aligned}
\beta^k &= \underset{\beta}{\operatorname{argmin}}\, L(\beta^{k-1}, \gamma, \eta) \\
\gamma^k &= \underset{\gamma}{\operatorname{argmin}}\, L(\beta, \gamma^{k-1}, \eta) \\
\eta^k &= \eta^{k-1} + \rho(\beta - \gamma)
\end{aligned}$$

where the last step is the *dual ascent step*

Convergence for ADMM:

- Define $r^k = \beta^k - \eta^k$. Then $r^k \to 0$ as $k \to \infty$
- $\frac{1}{2}\left|\left|y - X\beta^k\right|\right|_2^2 + \lambda\left|\left|\gamma^k\right|\right|_1 \to p^*$ as $k \to \infty$ where p^* is the optimal value
- $\eta^k \to \eta^*$ as $k \to \infty$ where η^* is a dual optimal point

ADMM for Gaussian Graphical Models:

- Recall that we want to solve:

$$\min_{\Omega \succ 0} \operatorname{trace}(S\Omega) - \log|\Omega| + \lambda \|\Omega\|_1$$

- This is equivalent to:

$$\min_{\Omega, Z} \operatorname{trace}(S\Omega) - \log|\Omega| + \lambda \|Z\|_1 \ \text{ s.t } \Omega = Z$$

- The augmented Lagrangian in this case is:

$$\min_{\Omega, Z, Y} \operatorname{trace}(S\Omega) - \log|\Omega| + \lambda \|Z\|_1 + Y^t(\Omega - Z) + \frac{\rho}{2}\|\Omega - Z\|_F^2$$

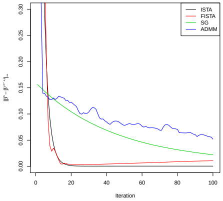
- Solving this involves doing the following at each iteration:

$$
\begin{aligned}
\min_{\Omega} \operatorname{trace}(S\Omega) - \log|\Omega| + Y^t(\Omega - Z) + \frac{\rho}{2}\|\Omega - Z\|_F^2 \\
\min_{Z} \lambda\|Z\|_1 + Y^t(\Omega - Z) + \frac{\rho}{2}\|\Omega - Z\|_F^2 \\
\max_{Y} Y^t(\Omega - Z)
\end{aligned}
$$

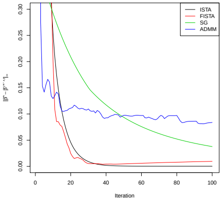
Data for BP Experiments:

- We set $p = 200$ and $n = \{20, 50, 100, 500\}$.
- Number of non-zero elements of β^* was set equal to 20.
- $X_{ij} \overset{iid}{\sim} \mathcal{N}(0,1); i = 1, ..., n; j = 1, ..., p,\ E_i \overset{iid}{\sim} \mathcal{N}(0,1); i = 1, ..., n$ and $y = X\beta^* + E$.
- λ was picked through 5-fold cross-validation
- We compared $\left|\left|\beta^k - \beta^{k-1}\right|\right|_\infty$ at each step, timing and $\frac{||\hat{\beta}-\beta^*||_2^2}{||\beta^*||_2^2}$ for all the methods
- In the above case, X had a condition number of 3.1677. We repeated the experiments with with X having a condition number of 101.9279. The performance of the subgradient method was very poor showing how this algorithm lacks stability. ISTA/FISTA's performance was pretty good, but inconsistent. The most reilable was the ADMM algorithm, whose performance hardly changed.

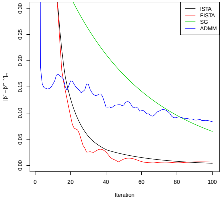
Convergence Plots for Basis Pursuit:



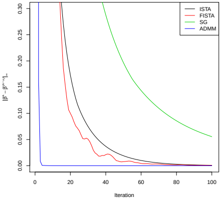
(a) $n = 20$



(b) $n = 50$

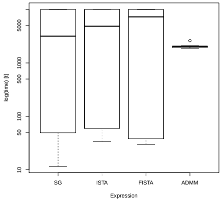


(c) $n = 100$

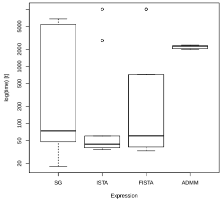


(d) $n = 500$

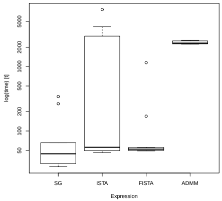
Timing plots for Basis Pursuit:



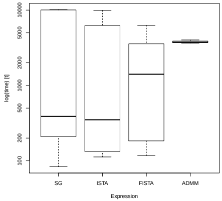
(a) $n = 20$



(b) $n = 50$



(c) $n = 100$

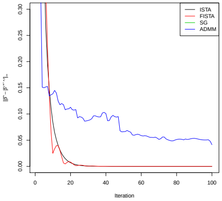


(d) $n = 500$

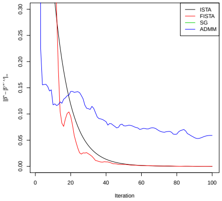
Relative Norm Error ($\frac{||\hat{\beta}-\beta*||_2^2}{||\beta*||_2^2}$):

	Method	$n = 20$	$n = 50$	$n = 100$	$n = 500$
1	SG	0.0184	0.0164	0.0095	0.0007
2	ISTA	0.0188	0.0146	0.0036	0.00008
3	FISTA	0.0197	0.0149	0.0038	0.00008
4	ADMM	0.0223	0.0174	0.0053	0.00008

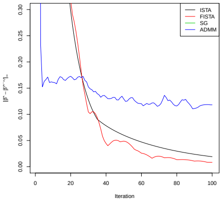
Convergence Plots for Basis Pursuit for ill-conditioned X:



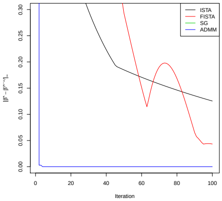
(a) $n = 20$



(b) $n = 50$

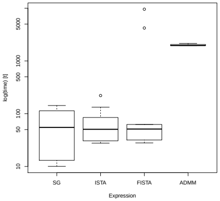


(c) $n = 100$

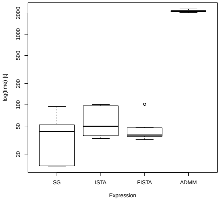


(d) $n = 500$

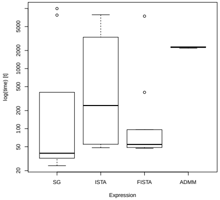
Timing plots for Basis Pursuit for ill-conditioned X:



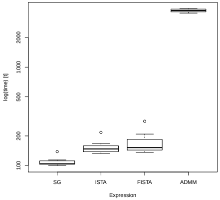
(a) $n = 20$



(b) $n = 50$



(c) $n = 100$



(d) $n = 500$

Relative Norm Error ($\frac{||\hat{\beta}-\beta*||_2^2}{||\beta*||_2^2}$):

	Method	$n = 20$	$n = 50$	$n = 100$	$n = 500$
1	SG	1.473	∞	∞	∞
2	ISTA	0.015	0.013	0.0009	0.0011
3	FISTA	0.015	0.013	0.0008	0.0001
4	ADMM	0.023	0.016	0.0025	0.00001

Data for Covariance Esimation Experiments:

- We set $p = 500$ and $n = 1000$.
- Approximately 95% of the entries in Ω^* were set to 0.
- Generate $X_i \overset{iid}{\sim} \mathcal{N}_p(0, \Omega^{-1})$ for $i = 1, ..., n$. Let X_i be the i^{th} row of X.
- We compared $\frac{||\hat{\Omega} - \Omega^*||_F}{||\Omega^*||_F}$ for all the methods: SG(0.463), ISTA(0.463), FISTA(0.463), ADMM(0.553). These were all inferior to glasso(0.346) discussed last week in class.

References

[1] Beck, A. and Teboulle, M. (2009), "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *Siam J. Imaging Sciences* Vol. 2, No. 1, pp. 183-202

[2] Boyd, S. and Parikh, N. and Chu, E. and Peleato, B. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, Vol. 3, No. 1, 1-122

[3] Boyd, S. and Vandenberghe, L. (2009), "Convex Optimization,"