Large Scale Optimization: Lecture 16

Fall 2014

The University of Texas at Austin

Recap

- Dual of *Semi-Definite Programming (SDP)* problems with linear objective function are derived as

$$\begin{aligned} \min \quad & -\langle G, Z\rangle \\ \text{s.t.} \quad & \langle F_i, Z\rangle = c_i,\ \forall i \\ & z \succeq 0 \end{aligned} \tag{1}$$

- Application of linear SDP optimization:
  - Find a matrix with largest eigenvalue,
  - Find sum of r-largest eigenvalues of a given matrix and
  - Find sum of singular values of a symmetric but not PSD matrix.

Recap

- Study of SDP is then extended to non-linear objective, called *Log Determinant Optimization*. The general form is as follows:

$$
\begin{aligned}
\min_{x} \quad & c^T x - \log \det\, G(x) \\
\text{s.t.} \quad & G(x) \succeq 0 \\
& F(x) \geq 0
\end{aligned}
\tag{2}
$$

  where $-\log \det\, G(x)$ is proven to be convex function.

Recap

- Following problems are formulated as Log Determinant Optimization problem:
  - Find the minimal-volume ellipsoid that contains all given points,
  - Find the maximum-volume ellipsoid enclosed within a given polyhedron,
  - Find the most likely parameters that generates a given set of samples
  - Find the variance matrix of gaussian channel with maximum capacity.

Motivations for Proximal Gradient Descent

- Motivation 1: Iterative Shrinkage-Thresholding Algorithm (ISTA)
- Motivation 2

Iterative Shrinkage-Thresholding Algorithm (ISTA)

Consider the unconstrained optimization problem with l_1 regularization

$$\min_x \quad \frac{1}{2}\underbrace{||y-Ax||_2^2}_{\text{smooth/"nice"}} + \underbrace{\lambda||x||_1}_{\text{not smooth/"nice", but "special"}} \tag{3}$$

If $A = I$, then

$$\begin{aligned} &\min_x \quad \frac{1}{2}||y-x||_2^2 + \lambda||x||_1 \\ = &\min_x \quad \sum_i \{\frac{1}{2}(y_i - x_i)^2 + \lambda|x_i|_1\} \end{aligned} \tag{4}$$

Iterative Shrinkage-Thresholding Algorithm (ISTA)

Now we derive closed-form solution for each "separated" problem:

$$
\begin{aligned}
& x_i - y_i + \lambda && \text{if } x_i > 0 && (5)\\
& x_i - y_i - \lambda && \text{if } x_i < 0 && (6)\\
& -y_i + r\lambda, r \in (-1, 1) && \text{if } x_i = 0 && (7)
\end{aligned}
$$

Suppose $y_i > 0$, then it obvious that $x_i \geq 0$.

- if $x_i > 0$, then $\exists x_i = y_i - \lambda$.
- if $x_i = 0$, then

Similarly, if $y_i < 0$, then exists $x_i = y_i + \lambda$ if $x_i < 0$.

Motivation 2

Consider another optimization problem

$$\begin{aligned} \min_{x} \quad & \underbrace{g(x)}_{\text{"nice"}} \\ \text{s.t.} \quad & x \in Q \end{aligned} \tag{8}$$

where Q is a convex set and is easy to project onto, e.g. $Q : \{x | ||x||_\infty \leq 1\}$.
The above optimization problem equates to

$$\min_{x} \quad g(x) + I_Q(x) \tag{9}$$

where

$$I_Q(x) = \begin{cases} 0 & \text{if } x \in Q \\ \infty & \text{if } x \notin Q \end{cases} \tag{10}$$

Hence, we have

$$[P_Q(y)]_i = \begin{cases} y_i & \text{if } |y_i| \leq 1 \\ sign(y_i) & \text{otherwise} \end{cases} \tag{11}$$

Proximal Gradient Algorithm

- Unconstrained problem with sum of two functions:

$$\text{minimize } f(x) = g(x) + h(x) \tag{12}$$

$g(x)$: *'nice'*, (ex) convex, L-lipschitz gradient

$h(x)$: *'special'*, (ex) convex, not smooth, **prox** is inexpensive

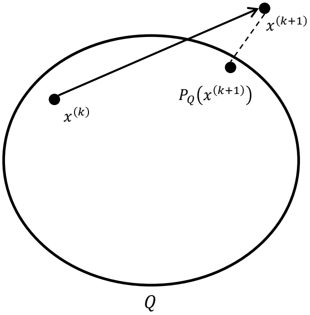
Definition

- **Prox function/ Proximal operator**

The prox function (or proximal operator) of a function $h(\cdot)$ is defined as:

$$\mathbf{prox}_{th}(v) = \arg\min_x \left(h(x) + \frac{1}{2t}\|x - v\|_2^2 \right) \tag{13}$$

Prox function/ Proximal operator



- Ex) Projection on set Q

$$x^{(k+1)} = P_Q\left(x^{(k)} - t\nabla g(x^{(k)})\right)$$

$$\begin{aligned}\mathbf{prox}_h(x) &= \arg\min_u \left(I_Q(u) + \frac{1}{2}\|u-x\|_2^2 \right) \\ &= \arg\min_{u\in Q} \left(\frac{1}{2}\|u-x\|_2^2 \right) \\ &= P_Q(x)\end{aligned}$$

$$I_Q(x) = \begin{cases} 0 & \text{if } x \in Q \\ \infty & \text{if } x \notin Q \end{cases}$$

Proximal Gradient Algorithm

- **Proximal Gradient Algorithm**

$$x_+ \leftarrow \mathbf{prox}_{th}\left(x - t\nabla g(x)\right)$$

- Uses two black boxes for the task:

$$\text{minimize } f(x) = g(x) + h(x)$$

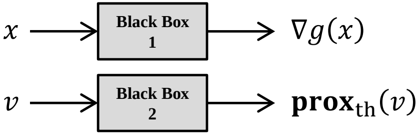


Figure : Using two black boxes for proximal gradient

Proximal Gradient Algorithm

- Form of quadratic approximation of $g(u)$ around x

$$
\begin{aligned}
x_+ \leftarrow & \mathbf{prox}_{th}(x - t\nabla g(x)) \\
&= \arg\min_u \left(h(u) + \frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2 \right) \\
&= \arg\min_u \left(h(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 + \frac{t}{2}\|\nabla g(x)\|_2^2 \right) \\
&= \arg\min_u \left(h(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 + g(x) \right)
\end{aligned}
\tag{14}
$$

Convergence Analysis of Proximal Gradient

Theorem

If g is convex and g has L-lipschitz gradient, i.e. $\|\nabla g(x) - \nabla g(y)\|_2 \le L\|x-y\|_2$, using fixed size $t < 1/L$ on proximal gradient algorithm gives $O(1/\varepsilon)$ of convergence to the optimal point x^*, where $x^* = \arg\min_x (g(x) + h(x))$.

- Introducing $G_t(x)$

$$
\begin{aligned}
x_+ \leftarrow x - tG_t(x) &= \mathbf{prox}_{th}(x - t\nabla g(x)) \\
\Leftrightarrow G_t(x) &\triangleq \frac{1}{t}\left(x - \mathbf{prox}_{th}(x - t\nabla g(x))\right)
\end{aligned}
$$

Convergence Analysis of Proximal Gradient - Preliminaries for proof

Claim

$G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$

Lemma

For any point z, $f(x_+) \leq f(z) + \langle G_t(x), x - z \rangle - \frac{t}{2} \| G_t(x) \|_2^2$

Proof of Claim and Lemma: Later

Proof of Theorem

Putting $z = x$ in the Lemma gives,

$$f(x_+) \leq f(x) - \frac{t}{2}\|G_t(x)\|_2^2$$

This gives the decreasing property $f(x_+) \leq f(x)$
And putting $z = x^*$ gives,

$$\begin{aligned} f(x_+) &\leq f^* - \frac{t}{2}\|G_t(x)\|_2^2 + \langle G_t(x), x - x^* \rangle \\ \Leftrightarrow f(x_+) - f^* &\leq \frac{1}{2t}\left[\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2\right] \\ &= \frac{1}{2t}\left[\|x - x^*\|_2^2 - \|x_+ - x^*\|_2^2\right] \end{aligned}$$

Proof of Theorem

Adding up over T iterations gives

$$\sum_{k=1}^{T}\left(f(x^{(k)})-f^*\right)\leq\frac{1}{2t}\left[\|x^{(0)}-x^*\|_2^2-\|x^{(T)}-x^*\|_2^2\right]$$
$$\leq\frac{1}{2t}\|x^{(0)}-x^*\|_2^2$$

From the fact that $f(x_+)\leq f(x)$,

$$f(x^{(k)})-f^*\geq f(x^{(T)})-f^* \text{ for } k=1,\cdots T$$

This gives,

$$f(x^{(T)})-f^*\leq\frac{1}{2tT}\|x^{(0)}-x^*\|_2^2$$

$\Rightarrow$ $O(1/\varepsilon)$ convergence
(requires $k=O(1/\varepsilon)$ of iteration number for $f(x^{(k)})-f^*\leq\varepsilon$)

Proof of Lemma

From the L-lipschitz continuous of gradient on g,

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$$

So, for $y = x_+ = x - tG_t(x)$

$$g(x_+) \leq g(x) - \langle \nabla g(x), tG_t(x) \rangle + \frac{L}{2}t^2\|G_t(x)\|_2^2 \tag{15}$$

Recall the definition of subgradient,

$$a \in \partial h(x_+) \iff h(z) \geq h(x_+) + \langle a, z - x_+ \rangle$$

By Claim, $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$

$$\begin{aligned} & h(z) \geq h(x_+) + \langle G_t(x) - \nabla g(x), z - x_+ \rangle \\ \Leftrightarrow \;\; & h(x_+) \leq h(z) - \langle G_t(x) - \nabla g(x), z - x_+ \rangle \qquad (16) \end{aligned}$$

Proof of Lemma

Add (15), (16) and use $f(x_+) = g(x_+) + h(x_+)$

$$
\begin{aligned}
f(x_+) &\leq g(x) - \langle \nabla g(x), tG_t(x)\rangle + \frac{L}{2}t^2 \|G_t(x)\|_2^2 \\
&\quad + h(z) - \langle G_t(x) - \nabla g(x), z - x_+\rangle \\
&\leq g(z) - \langle \nabla g(x), z - x\rangle - \langle \nabla g(x), tG_t(x)\rangle + \frac{L}{2}t^2 \|G_t(x)\|_2^2 \\
&\quad + h(z) - \langle G_t(x) - \nabla g(x), z - x_+\rangle \\
&\qquad\qquad (\because g(x) \leq g(z) - \langle \nabla g(x), z - x\rangle \text{ from convexity of } g) \\
&= f(z) + \frac{L}{2}t^2 \|G_t(x)\|_2^2 - \langle G_t(x), z - x + tG_t(x)\rangle \\
&\leq f(z) + \langle G_t(x), x - z\rangle + \frac{t}{2}\|G_t(x)\|_2^2 - t\|G_t(x)\|_2^2 \quad (\because t < 1/L) \\
&= f(z) + \langle G_t(x), x - z\rangle - \frac{t}{2}\|G_t(x)\|_2^2
\end{aligned}
$$

Proof of Claim

Claim. $G_t(x) - \nabla g(x) \in \partial h(x_+)$

- Property of proximal operator

$$u = \mathbf{prox}_{th}(x) \Leftrightarrow \frac{1}{t}(x - u) \in \partial h(u) \tag{17}$$

(Proof: Use that $\mathbf{prox}_{th}(x)$ is a minimizer)

Now put $u = x_+ = \mathbf{prox}_{th}(x - \nabla g(x))$ in equation (17),

$$\begin{aligned}
&\Rightarrow \frac{1}{t}(x - \nabla g(x) - x_+) \in \partial h(x_+) \\
&\Rightarrow \frac{1}{t}(x - \nabla g(x) - x + tG_t(x)) \in \partial h(x_+) \\
&\Rightarrow G_t(x) - \nabla g(x) \in \partial h(x_+)
\end{aligned}$$

Next lecture

- Review of oracle complexity of algorithm
- Nesterov's accelerated gradient descent

Thank you!