Supplement: proofs of the theorems

The IMA Summer Course 2016: First-Order Algorithms

Shuzhong Zhang

August 1, 2016

1 Proof of Theorem 1

By the mean-value theorem we have

$$f(y+\alpha d) = f(y) + \alpha \nabla f(y)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(\bar{y}) d$$

where $\bar{y}$ lies between y and $y+\alpha d$.

First, by letting $y = x^*$, $\alpha = 1$ and $d = x - x^*$ we have

$$\frac{m}{2}\|x - x^*\|^2 \le f(x) - f(x^*) \le \frac{M}{2}\|x - x^*\|^2.$$

Then, by letting $y = x$, $\alpha = 1$ and $d = x^* - x$

$$\begin{aligned} f(x^*) &= f(x) + \nabla f(x)^\top (x^* - x) + \frac{1}{2}(x^* - x)^\top \nabla^2 f(\bar{x})(x^* - x) \\ &\ge f(x) - \|\nabla f(x)\| \cdot \|x^* - x\| + \frac{m}{2}\|x^* - x\|^2. \end{aligned}$$

This way, we get

$$m\|x - x^*\| \le \|\nabla f(x)\|.$$

Similarly we get both lower and upper bounds:

$$m\|x - x^*\| \le \|\nabla f(x)\| \le M\|x - x^*\|.$$

If we let $\alpha = \frac{1}{M}$ and $d = -\nabla f(x)$, then

$$f(x + \alpha d) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|^2.$$

On the other hand,

$$\begin{aligned} f(x) - f(x^*) &\le \|\nabla f(x)\| \cdot \|x - x^*\| - \frac{m}{2}\|x - x^*\|^2 \\ &\le \frac{1}{2m}\|\nabla f(x)\|^2, \end{aligned}$$

where in the last inequality we used the fact that $at - \frac{m}{2}t^2 \le \frac{a^2}{2m}$.

Substituting this into an earlier inequality yields

$$\begin{aligned} f(x + \alpha d) &\le f(x) - \frac{1}{2M}\|\nabla f(x)\|^2 \\ &\le f(x) - \frac{m}{M}(f(x) - f(x^*)). \end{aligned}$$

In other words, we have the Q-linear convergent rate

$$f(x+\alpha d) - f(x^*) \leq \left(1 - \frac{m}{M}\right)(f(x) - f(x^*)).$$

It is possible to improve the rate a little bit. In fact, one can improve the rate to

$$f(x+\alpha d) - f(x^*) \leq \left(\frac{1 - \frac{m}{M}}{1 + \frac{m}{M}}\right)^2 (f(x) - f(x^*)).$$

The above estimation is indeed tightest possible.

2 Proof of Theorem 2

Observe that if

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

for all $x, y \in \mathcal{X}$, then

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2.$$

Therefore, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &\leq f(x^k) + t_k \nabla f(x^k)^\top (y^{k+1} - x^k) + \frac{L}{2} t_k^2 D^2. \end{aligned}$$

Now, the optimality condition gives us

$$\nabla f(x^k)^\top (y^{k+1} - x^k) \leq \nabla f(x^k)^\top (x^* - x^k) \leq f(x^*) - f(x^k).$$

We have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + t_k \nabla f(x^k)^\top (y^{k+1} - x^k) + \frac{L}{2} t_k^2 D^2 \\ &\leq f(x^k) + \Big(f(x^*) - f(x^k)\Big) t_k + \frac{LD^2 t_k^2}{2}. \end{aligned}$$

The claimed bound can be shown by induction on k.

We have $f(x^2) - f(x^*) \leq LD^2$. Suppose the bound holds for k. Then,

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq (f(x^k) - f(x^*))(1 - t_k) + \frac{LD^2 t_k^2}{2} \\ &\leq \frac{2LD^2}{k} \times \left(1 - \frac{2}{k}\right) + \frac{2LD^2}{k^2} \\ &< \frac{2LD^2}{k+1}. \end{aligned}$$

3 Proof of Theorem 3

Consider iteration ℓ. The optimality condition yields

$$\left(d^{\ell+1} - \frac{x^\ell - x^{\ell+1}}{t_\ell}\right)^\top (x - x^{\ell+1}) \geq 0,\ \forall x \in \mathcal{X},$$

where $d^{\ell+1} \in \partial f_\ell(x^{\ell+1})$. By the subgradient inequality we have

$$f_\ell(x) - f_\ell(x^{\ell+1}) \geq \left(\frac{x^\ell - x^{\ell+1}}{t_\ell}\right)^\top (x - x^{\ell+1})$$

for any given $x \in \mathcal{X}$. We have

$$\begin{aligned} \|x^\ell - x\|^2 &= \|x^\ell - x^{\ell+1} + x^{\ell+1} - x\|^2 \\ &= \|x^{\ell+1} - x\|^2 + 2(x^\ell - x^{\ell+1})^\top (x^{\ell+1} - x) + \|x^\ell - x^{\ell+1}\|^2 \\ &\geq \|x^{\ell+1} - x\|^2 + 2t_\ell(f_\ell(x^{\ell+1}) - f_\ell(x)) + \|x^{\ell+1} - x^\ell\|^2. \end{aligned}$$

Summarizing, we have

$$\|x^{\ell+1} - x\|^2 \leq \|x^\ell - x\|^2 - 2t_\ell(f_\ell(x^{\ell+1}) - f_\ell(x)) - \|x^{\ell+1} - x^\ell\|^2,$$

for all $\ell = 0, 1, 2, ...$

Let us construct $f_\ell(x) := f(x^\ell) + \nabla f(x^\ell)^\top (x - x^\ell)$ and $t_\ell = 1/L$. We have $f(x) \leq f_\ell(x) + \frac{L}{2}\|x - x^\ell\|^2$. So

$$\begin{aligned} f(x^{\ell+1}) &\leq f_\ell(x^{\ell+1}) + \frac{L}{2}\|x^{\ell+1} - x^\ell\|^2 \\ f(x^*) &\geq f_\ell(x^*). \end{aligned}$$

This gives us

$$\begin{aligned} \|x^{\ell+1} - x^*\|^2 &\leq \|x^\ell - x^*\|^2 - 2t_\ell(f_\ell(x^{\ell+1}) - f_\ell(x^*)) - \|x^{\ell+1} - x^\ell\|^2 \\ &\leq \|x^\ell - x^*\|^2 - \frac{2}{L}\left(f(x^{\ell+1}) - f(x^*)\right). \end{aligned}$$

On the other hand, we know that proximal point algorithm yields monotonically improving iterates:

$$f(x^{\ell+1}) \leq f_\ell(x^{\ell+1}) + \frac{L}{2}\|x^{\ell+1} - x^\ell\|^2 \leq f_\ell(x^\ell) = f(x^\ell).$$

Therefore, adding up the inequalities

$$\|x^{\ell+1} - x^*\|^2 \leq \|x^\ell - x^*\|^2 - \frac{2}{L}\left(f(x^{\ell+1}) - f(x^*)\right)$$

from $\ell = 0$ to $k - 1$, we obtain

$$f(x^k) - f(x^*) \leq \frac{\sum_{\ell=0}^{k-1}\left(f(x^{\ell+1}) - f(x^*)\right)}{k} \leq \frac{L\|x^0 - x^*\|^2}{2k}.$$

4 Proof of Theorem 5

Consider an iteration ℓ of the ISTA. Take any $h'(x^{\ell+1}) \in \partial h(x^{\ell+1})$.

By the optimality condition, we have

$$\left[\nabla f(x^\ell) + L(x^{\ell+1} - x^\ell) + h'(x^{\ell+1})\right]^\top (x - x^{\ell+1}) \geq 0, \ \forall x \in \mathcal{X}.$$

Therefore, for any $x \in \mathcal{X}$, we have

$$\begin{aligned} h(x^{\ell+1}) &\leq h(x) - h'(x^{\ell+1})^\top (x - x^{\ell+1}) \\ &\leq h(x) + \nabla f(x^\ell)^\top (x - x^{\ell+1}) + L(x^{\ell+1} - x^\ell)^\top (x - x^{\ell+1}). \end{aligned}$$

Moreover,

$$\begin{aligned} f(x^{\ell+1}) &\leq f(x^\ell) + \nabla f(x^\ell)^\top (x^{\ell+1} - x^\ell) + \frac{L}{2}\|x^{\ell+1} - x^\ell\|^2 \\ &\leq f(x) - \nabla f(x^\ell)^\top (x - x^\ell) + \nabla f(x^\ell)^\top (x^{\ell+1} - x^\ell) + \frac{L}{2}\|x^{\ell+1} - x^\ell\|^2 \\ &= f(x) + \nabla f(x^\ell)^\top (x^{\ell+1} - x) + \frac{L}{2}\|x^{\ell+1} - x^\ell\|^2. \end{aligned}$$

Denoting $F(x) = f(x) + h(x)$, letting $x = x^*$ and adding up the previous two inequalities we have

$$\begin{aligned} F(x^{\ell+1}) &\leq F(x^*) + \frac{L}{2}\left[2(x^{\ell+1} - x^\ell)^\top (x^* - x^{\ell+1}) + \|x^{\ell+1} - x^\ell\|^2\right] \\ &= F(x^*) + \frac{L}{2}\left[\|x^\ell - x^*\|^2 - \|x^{\ell+1} - x^*\|^2\right], \end{aligned}$$

where we used the identity

$$(w_1 - w_2)^\top (w_3 - w_1) = \frac{1}{2}\left(\|w_2 - w_3\|^2 - \|w_1 - w_2\|^2 - \|w_1 - w_3\|^2\right).$$

Similar as in the proximal point algorithm case, one observes that the iterates are monotone. Therefore, summing from $\ell = 0$ to $\ell = k-1$, we have

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

5 Proof of Theorem 6

If Φ is strongly convex, then there is $\sigma > 0$ such that $B(y,x) \geq \sigma\|y - x\|^2$. Moreover,

$$B(w,u) + B(u,v) - B(w,v) = \left[\nabla\Phi(v) - \nabla\Phi(u)\right]^\top (w - u).$$

Consider the ℓth iteration. The optimality condition leads to

$$\left\{t_\ell d^\ell - \left[\nabla\Phi(x^\ell) - \nabla\Phi(x^{\ell+1})\right]\right\}^\top (x - x^{\ell+1}) \geq 0, \ \ \forall x \in \mathcal{X}.$$

Therefore,

$$\begin{aligned} t_\ell (d^\ell)^\top (x^* - x^{\ell+1}) &\geq \left[\nabla\Phi(x^\ell) - \nabla\Phi(x^{\ell+1})\right]^\top (x^* - x^{\ell+1}) \\ &= B(x^*, x^{\ell+1}) + B(x^{\ell+1}, x^\ell) - B(x^*, x^\ell) \\ &\geq B(x^*, x^{\ell+1}) - B(x^*, x^\ell) + \sigma\|x^\ell - x^{\ell+1}\|^2. \end{aligned}$$

Now we have

$$\begin{aligned} t_\ell(f(x^*) - f(x^\ell)) &\geq t_\ell (d^\ell)^\top (x^* - x^\ell) \\ &= t_\ell (d^\ell)^\top (x^* - x^{\ell+1}) + t_\ell (d^\ell)^\top (x^{\ell+1} - x^\ell) \\ &\geq B(x^*, x^{\ell+1}) - B(x^*, x^\ell) + \sigma\|x^\ell - x^{\ell+1}\|^2 \\ &\quad - \left(\frac{\tau}{2}\|x^\ell - x^{\ell+1}\|^2 + \frac{t_\ell^2}{2\tau}\|d^\ell\|^2\right), \end{aligned}$$

for any $\tau > 0$. By the Lipschitz continuity of f, we have $\|d^\ell\| \leq L$. Choosing $\tau = 2\sigma$ we have

$$t_\ell(f(x^*) - f(x^\ell)) \geq B(x^*, x^{\ell+1}) - B(x^*, x^\ell) - \frac{L^2 t_\ell^2}{4\sigma}.$$

Letting $t_\ell = 1/\sqrt{k}$ for all $0 \leq \ell \leq k-1$, and summing up, we have

$$\min_{0\leq\ell\leq k-1} \left(f(x^\ell) - f(x^*)\right) \leq \frac{B(x^*, x^0) + \frac{L^2}{4\sigma}}{\sqrt{k}}.$$

6 Proof of Theorem 7

If additionally, ∇f is Lipschitz, say $(\nabla f(x) - \nabla f(y))^\top (x - y) \leq M\|x - y\|^2$, then

$$\begin{aligned} \nabla f(x^\ell)^\top (x^{\ell+1} - x^\ell) &\geq \nabla f(x^{\ell+1})^\top (x^{\ell+1} - x^\ell) - M\|x^\ell - x^{\ell+1}\|^2 \\ &\geq f(x^{\ell+1}) - f(x^\ell) - M\|x^\ell - x^{\ell+1}\|^2. \end{aligned}$$

Therefore,

$$\begin{aligned} t_\ell(f(x^*) - f(x^\ell)) &\geq B(x^*, x^{\ell+1}) - B(x^*, x^\ell) + (\sigma - M t_\ell)\|x^\ell - x^{\ell+1}\|^2 \\ &\quad + t_\ell(f(x^{\ell+1}) - f(x^\ell)). \end{aligned}$$

Taking $t_\ell = \sigma/M$ and summing up, the theorem is proven.

7 Proof of Theorem 8

Let us analyze what happens after this modification. Let ℓ be an iteration.

By the optimality condition:

$$\left[\nabla f(y^{\ell+1}) + L(x^{\ell+1} - y^{\ell+1})\right]^\top (x - x^{\ell+1}) \geq 0, \ \forall x \in \mathcal{X}.$$

Now, since

$$f(x^{\ell+1}) \leq f(y^{\ell+1}) + \nabla f(y^{\ell+1})^\top (x^{\ell+1} - y^{\ell+1}) + \frac{L}{2}\|x^{\ell+1} - y^{\ell+1}\|^2$$

and by the subgradient inequality

$$f(y^{\ell+1}) \leq f(x) - \nabla f(y^{\ell+1})^\top (x - y^{\ell+1})$$

we have

$$f(x^{\ell+1}) \leq f(x) + \nabla f(y^{\ell+1})^\top (x^{\ell+1} - x) + \frac{L}{2}\|x^{\ell+1} - y^{\ell+1}\|^2.$$

Combining with the optimality condition we have

$$f(x^{\ell+1}) \leq f(x) + L(x^{\ell+1} - y^{\ell+1})^\top (x - x^{\ell+1}) + \frac{L}{2}\|x^{\ell+1} - y^{\ell+1}\|^2, \ \forall x \in \mathcal{X}.$$

Taking $x = x^*$ and $x = x^\ell$ respectively:

$$\begin{aligned} f(x^{\ell+1}) &\leq f(x^*) + L(x^{\ell+1} - y^{\ell+1})^\top (x^* - x^{\ell+1}) + \frac{L}{2}\|x^{\ell+1} - y^{\ell+1}\|^2 \\ f(x^{\ell+1}) &\leq f(x^\ell) + L(x^{\ell+1} - y^{\ell+1})^\top (x^\ell - x^{\ell+1}) + \frac{L}{2}\|x^{\ell+1} - y^{\ell+1}\|^2. \end{aligned}$$

Multiplying the first inequality by $1/t_{\ell+1}$ and multiplying the second inequality by $1 - 1/t_{\ell+1}$, and then adding up:

$$\begin{aligned} & \frac{1}{t_{\ell+1}}\left(f(x^{\ell+1}) - f(x^*)\right) + \left(1 - \frac{1}{t_{\ell+1}}\right)\left(f(x^{\ell+1}) - f(x^\ell)\right) \\ \leq \;& L(x^{\ell+1} - y^{\ell+1})^\top \left(\frac{1}{t_{\ell+1}} x^* + \left(1 - \frac{1}{t_{\ell+1}}\right) x^\ell - x^{\ell+1}\right) + \frac{L}{2}\|x^{\ell+1} - y^{\ell+1}\|^2. \end{aligned}$$

The left hand side of the above inequality is

$$f(x^{\ell+1}) - f(x^*) - \left(1 - \frac{1}{t_{\ell+1}}\right)\left(f(x^\ell) - f(x^*)\right).$$

Using the identity

$$(w_1 - w_2)^\top (w_3 - w_1) = \frac{1}{2}\left(\|w_2 - w_3\|^2 - \|w_1 - w_2\|^2 - \|w_1 - w_3\|^2\right)$$

on the right hand side, it becomes

$$\begin{aligned} & \frac{L}{2}\left(\left\|y^{\ell+1}-\left(1-\frac{1}{t_{\ell+1}}\right)x^{\ell}-\frac{1}{t_{\ell+1}}x^*\right\|^2-\left\|x^{\ell+1}-\left(1-\frac{1}{t_{\ell+1}}\right)x^{\ell}-\frac{1}{t_{\ell+1}}x^*\right\|^2\right) \\ = \; & \frac{L}{2t_{\ell+1}^2}\left(\left\|t_{\ell+1}y^{\ell+1}-(t_{\ell+1}-1)x^{\ell}-x^*\right\|^2-\left\|t_{\ell+1}x^{\ell+1}-(t_{\ell+1}-1)x^{\ell}-x^*\right\|^2\right). \end{aligned}$$

The formula $y^{\ell+1}=x^{\ell}+\frac{t_{\ell}-1}{t_{\ell+1}}\left(x^{\ell}-x^{\ell-1}\right)$ gives us

$$t_{\ell+1}y^{\ell+1}-(t_{\ell+1}-1)x^{\ell}=t_{\ell}x^{\ell}-(t_{\ell}-1)x^{\ell-1}.$$

Rearranging and using $t_{\ell+1}(t_{\ell+1}-1)=t_{\ell}^2$, we have

$$\begin{aligned} & t_{\ell+1}^2\left(f(x^{\ell+1})-f(x^*)\right)-t_{\ell}^2\left(f(x^{\ell})-f(x^*)\right) \\ \leq \; & \frac{L}{2}\left(\left\|t_{\ell}x^{\ell}-(t_{\ell}-1)x^{\ell-1}-x^*\right\|^2-\left\|t_{\ell+1}x^{\ell+1}-(t_{\ell+1}-1)x^{\ell}-x^*\right\|^2\right). \end{aligned}$$

Summing up $\ell=1$ to $\ell=k-1$, the theorem follows.

8 Proof of Theorem 10

Consider ℓ-th iteration. The optimality condition is:

$$\left[\nabla f(x^{\ell+1})-A^\top\lambda^{\ell}+\gamma A^\top(Ax^{\ell+1}-b)\right]^\top(x-x^{\ell+1})\geq 0,\ \forall x\in\mathcal{X}.$$

Now take any fixed $\lambda\in\mathbf{R}^m$. We have

$$\begin{aligned} 0 \; & \leq \; \left[\nabla f(x^{\ell+1})-A^\top\lambda^{\ell}+\gamma A^\top(Ax^{\ell+1}-b)\right]^\top(x^*-x^{\ell+1}) \\ & \leq \; f(x^*)-f(x^{\ell+1})-(\lambda^{\ell+1})^\top A(x^*-x^{\ell+1}) \\ & = \; f(x^*)-f(x^{\ell+1})+(\lambda^{\ell+1})^\top(Ax^{\ell+1}-b) \\ & = \; f(x^*)-f(x^{\ell+1})+\lambda^\top(Ax^{\ell+1}-b)+(\lambda^{\ell+1}-\lambda)^\top\left[\frac{1}{\gamma}\left(\lambda^{\ell}-\lambda^{\ell+1}\right)\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} & f(x^{\ell+1})-f(x^*)-\lambda^\top(Ax^{\ell+1}-b) \\ \leq \; & \frac{1}{\gamma}(\lambda^{\ell+1}-\lambda)^\top(\lambda^{\ell}-\lambda^{\ell+1}) \\ = \; & \frac{1}{2\gamma}\left[\|\lambda-\lambda^{\ell}\|^2-\|\lambda-\lambda^{\ell+1}\|^2-\|\lambda^{\ell+1}-\lambda^{\ell}\|^2\right] \\ \leq \; & \frac{1}{2\gamma}\left[\|\lambda-\lambda^{\ell}\|^2-\|\lambda-\lambda^{\ell+1}\|^2\right]. \end{aligned}$$

Summing from $\ell = 0$ to $k-1$, and noting the convexity of f (Jensen's inequality) we have

$$f(\bar{x}^k) - f(x^*) - \lambda^\top (A\bar{x}^k - b) \leq \frac{\|\lambda^0 - \lambda\|^2}{2k\gamma}.$$

The desired result follows by taking

$$\lambda = -\rho \cdot \frac{A\bar{x}^k - b}{\|A\bar{x}^k - b\|}.$$

In fact, we still need to establish the feasibility violation of $\bar{x}^k$.

Let $v(z) := \min\{f(x) \mid x \in \mathcal{X},\ Ax - b = z\}$, which is a convex function.

Take any $\mu \in \partial v(0)$. By the subgradient and Cauchy-Schwartz inequalities

$$v(z) \geq v(0) + \mu^\top z \geq f(x^*) - \|\mu\| \cdot \|z\|.$$

Let $z = A\bar{x}^k - b$. We have

$$f(\bar{x}^k) \geq f(x^*) - \|\mu\| \cdot \|A\bar{x}^k - b\|.$$

Choose $\rho > \|\mu\|$. It follows that

$$(\rho - \|\mu\|)\|A\bar{x}^k - b\| \leq f(\bar{x}^k) - f(x^*) + \rho\|A\bar{x}^k - b\| \leq \frac{\rho^2/\gamma + \|\lambda^0\|^2/\gamma}{k}.$$

Hence,

$$\begin{cases} \|A\bar{x}^k - b\| \leq \frac{\rho^2 + \|\lambda^0\|^2}{\gamma(\rho - \|\mu\|)k} = O(1/k); \\ f(\bar{x}^k) - f(x^*) \leq \frac{\rho^2 + \|\lambda^0\|^2}{\gamma k} = O(1/k). \end{cases}$$

This shows that in k iterations the multiplier method does produce a solution that has $O(1/k)$ error in terms of constraint violation and the objective value.

References

[1] Amir Beck and Marc Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, 31, 167–175, 2003.

[2] Amir Beck and Marc Teboulle, *A Fast Iterative Shrinkage-Threshold Algorithm Linear Inverse Problems*, SIAM Journal on Imaging Sciences, 2, 183–202, 2009.

[3] Dimitri Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[4] Dimitri Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, 2015.

[5] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato and Jonathan Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Foundations and Trends in Machine Learning, 3 (1), 1–122, 2010.

[7] Sébastien Bubeck, *The Complexities of Optimization*, Lecture Notes, 2013.

[8] Tianyi Lin, Shiqian Ma and Shuzhong Zhang, *On the Convergence Rate of Multi-Block ADMM*, `http://arxiv.org/pdf/1408.4265v2.pdf`.

[9] Tianyi Lin, Shiqian Ma and Shuzhong Zhang, *On the Global Linear Convergence of the ADMM with Multi-Block Variables*, `http://arxiv.org/abs/1408.4266`.

[10] David Luenberger and Yinyu Ye, *Linear and Nonlinear Programming*, Springer, 2007.

[11] Yurii Nesterov, *Introductory Lectures on Convex Optimization: a Basic Course*, Kluwer Academic Publisher, 2004.

[12] Lieven Vandenberghe's lecture notes: `http://www.seas.ucla.edu/~vandenbe/ee236c.html`