A Second-Order Path-Following Algorithm for Unconstrained Convex Optimization

Yinyu Ye
Department is Management Science & Engineering
and Institute of Computational & Mathematical Engineering
Stanford University

May 31, 2017

Abstract

We present more details of the minimal-norm path following algorithm for unconstrained smooth convex optimization described in the lecture note of CME307 and MS&E311 [10].

Iterative optimization algorithms have been modeled as following certain paths to the optimal solution set, such as the central path of linear programming (e.g., [1, 3]) and, more recently, the trajectory of accelerated gradients of Nesterov ([8, 9]). Moreover, there is an interest in accelerating and globalizing Newton's or second-order methods for unconstrained smooth convex optimization, e.g., [5].

Let $f(x),\ x \in R^n$ be any smooth convex function with continuous second-order derivatives, and it meets a local Lipschitz condition: for any point $x \neq 0$ and a constant $\beta \geq 1$

$$\|\nabla f(x+d) - \nabla f(x) - \nabla^2(x)d\| \leq \beta d^T \nabla^2 f(x) d, \text{ whenever } \|d\| \leq O(1). \tag{1}$$

and $x + d$ is in the function domain. Here, $\|\cdot\|$ represents the L_2 norm, and it resembles the self-concordant condition of [6]. Note that all convex power, logarithmic, barrier, and exponential functions meet this condition, and the function does not need to be strictly or strongly convex nor has a bounded solution set. Furthermore, we assume that $x = 0$ is not a minimizer or $\nabla f(0) \neq 0$.

We consider the path constructed from the strictly convex minimization problem

$$\min_x \ f(x) + \frac{\mu}{2}\|x\|^2 \tag{2}$$

where μ is any positive parameter, and the minimizer, denoted by $x(\mu)$, satisfies the necessary and sufficient condition:

$$\nabla f(x) + \mu x = 0. \tag{3}$$

We now prove a theorem on the path convergence similar to the one in [2]:

Theorem 1. *The following properties on the minimizer of (2) hold.*

i). The minimizer $x(\mu)$ of (2) is unique and continuous with μ.

ii). The function $f(x(\mu))$ is strictly increasing and $\|x(\mu)\|$ is strictly decreasing function of μ.

iii). $\lim_{\mu \to 0^+} x(\mu)$ *converges to the minimal norm solution of $f(x)$.*

Proof Property i) is based on the fact that $f(x) + \frac{\mu}{2}\|x\|^2$ is a strictly convex function for any $\mu > 0$ and its Hessian is positive definite.

We prove ii). Let $0 < \mu' < \mu$. Then

$$f(x(\mu')) + \frac{\mu'}{2}\|x(\mu')\|^2 < f(x(\mu)) + \frac{\mu'}{2}\|x(\mu)\|^2$$

and

$$f(x(\mu)) + \frac{\mu}{2}\|x(\mu)\|^2 < f(x(\mu')) + \frac{\mu}{2}\|x(\mu')\|^2.$$

Add the two inequalities on both sides and rearrange them, we have

$$\frac{\mu - \mu'}{2}\|x(\mu')\|^2 > \frac{\mu - \mu'}{2}\|x(\mu)\|^2.$$

Since $\mu - \mu' > 0$, we have $\|x(\mu')\|^2 > \|x(\mu)\|^2$, that is, $\|x(\mu)\|$ is strictly decreasing function of μ. Then, using any one of the original two inequalities, we have $f(x(\mu')) < f(x(\mu))$.

Finally, we prove iii). Let $\bar{x}$ be an optimizer with the the minimum L_2 norm, then $\nabla f(\bar{x}) = 0$, which, together with (3), indicate

$$\nabla f(x(\mu)) - \nabla f(\bar{x}) + \mu x(\mu) = 0.$$

Pre-multiplying $x(\mu) - \bar{x}$ to both sides, and using the convexity of f,

$$-\mu(x(\mu) - \bar{x})^T x(\mu) = (x(\mu) - \bar{x})^T(\nabla f(x(\mu)) - \nabla f(\bar{x})) \geq 0.$$

Thus, we have $\|x(\mu)\|^2 \leq \bar{x}^T x(\mu) \leq \|\bar{x}\|\|x(\mu)\|$, that is, $\|x(\mu)\| \leq \|\bar{x})\|$ for any $\mu > 0$. If the accumulating limit point $x(0) \neq \bar{x}$, f must have two different minimum L_2 norm solutions in the convex optimal solution set of f. Then $\frac{1}{2}(x(0) + \bar{x})$ would remain an optimal solution and it has a norm strictly less than $\|\bar{x}\|$. Thus, $\bar{x}$ is unique and every accumulating limit point $x(0) = \bar{x}$, which completes the proof.

Let us call the path minimum-norm path and let x^k be an approximate path solution for $\mu = \mu^k$ and the path error be

$$\|\nabla f(x^k) + \mu^k x^k\| \leq \frac{1}{2\beta}\mu^k,$$

which defines a neighborhood of the path. Then, we like to compute a new iterate x^{k+1} remains in the neighborhood of the path, similar to the interior-point path-following algorithms (e.g., [7]), that is,

$$\|\nabla f(x^{k+1}) + \mu^{k+1} x^{k+1}\| \leq \frac{1}{2\beta}\mu^{k+1}, \quad \text{where } 0 \leq \mu^{k+1} < \mu^k. \tag{4}$$

Note that the neighborhood become smaller and smaller as the iterates go.

When μ^k is replaced by μ^{k+1}, say $(1-\eta)\mu^k$ for some number $\eta \in (0,\ 1]$, we aim to find the solution x such that

$$\nabla f(x) + (1-\eta)\mu^k x = 0.$$

To proceed, we use x^k as the initial solution and apply the Newton iteration:

$$\begin{array}{c} \nabla f(x^k) + \nabla^2 f(x^k)d + (1-\eta)\mu^k(x^k+d) = 0, \quad \text{or} \\ \nabla^2 f(x^k)d + (1-\eta)\mu^k d = -\nabla f(x^k) - (1-\eta)\mu^k x^k, \end{array} \tag{5}$$

and let the new iterate

$$x^{k+1} = x^k + d.$$

From the second expression, we have

$$\begin{array}{rcl} \|\nabla^2 f(x^k)d + (1-\eta)\mu^k d\| &=& \|-\nabla f(x^k) - (1-\eta)\mu^k x^k\| \\ &=& \|-\nabla f(x^k) - \mu^k x^k + \eta\mu^k x^k\| \\ &\le& \|-\nabla f(x^k) - \mu^k x^k\| + \eta\mu^k\|x^k\| \\ &\le& (\frac{1}{2\beta} + \eta\|x^k\|)\mu^k. \end{array} \tag{6}$$

On the other hand

$$\|\nabla^2 f(x^k)d + (1-\eta)\mu^k d\|^2 = \|\nabla^2 f(x^k)d\|^2 + 2(1-\eta)\mu^k d^T\nabla^2 f(x^k)d + ((1-\eta)\mu^k)^2\|d\|^2.$$

From convexity of f, $d^T\nabla^2 f(x^k)d \ge 0$, together using (6), we have

$$\begin{array}{rcl} ((1-\eta)\mu^k)^2\|d\|^2 &\le& (\frac{1}{2\beta} + \eta\|x^k\|)^2(\mu^k)^2 \quad \text{and} \\ 2(1-\eta)\mu^k d^T\nabla^2 f(x^k)d &\le& (\frac{1}{2\beta} + \eta\|x^k\|)^2(\mu^k)^2. \end{array} \tag{7}$$

The first inequality of (7) implies

$$\|d\|^2 \le \left(\frac{1}{2\beta(1-\eta)} + \frac{\eta\|x^k\|}{1-\eta}\right)^2.$$

The second inequality of (7) implies

$$\begin{array}{rl} & \|\nabla f(x^+) + (1-\eta)\mu^k x^+\| \\ = & \|\nabla f(x^+) - (\nabla f(x^k) + \nabla^2 f(x^k)d) + (\nabla f(x^k) + \nabla^2 f(x^k)d) + (1-\eta)\mu^k(x^k+d)\| \\ = & \|\nabla f(x^+) - \nabla f(x^k) + \nabla^2 f(x^k)d\| \\ \le & \beta d^T\nabla^2 f(x^k)d \le \frac{\beta}{2(1-\eta)}(\frac{1}{2\beta} + \eta\|x^k\|)^2\mu^k. \end{array}$$

We now just need to choose $\eta \in (0,\ 1)$ such that

$$\begin{array}{rcl} \left(\frac{1}{2\beta(1-\eta)} + \frac{\eta\|x^k\|}{1-\eta}\right)^2 &\le& 1 \quad \text{and} \\ \frac{\beta}{2(1-\eta)}(\frac{1}{2\beta} + \eta\|x^k\|)^2 &\le& \frac{1}{2\beta}(1-\eta). \end{array}$$

to satisfy (4), due to $(1-\eta)\mu^k = \mu^{k+1}$. Since $\beta \ge 1$, set

$$\eta = \frac{1}{2\beta(1+\|x^k\|)}$$

would suffice. This would give a linear convergence of μ down to zero,

$$\mu^{k+1} \leq \left(1 - \frac{1}{2\beta(1 + \|x^k\|)}\right)\mu^k$$

and x^k follows the path to the optimality. From Theorem 1, the size $\|x^k\|$ is bounded above by the size of $\|x^*\|$ where x^* is the minimum-norm optimal solution of $f(x)$ that is fixed.

Theorem 2. *There is a linearly convergent second-order or Newton method in minimizing any smooth convex function that satisfies the local Lipschitz condition. More precisely, the convergence rate is* $\left(1 - \frac{1}{2\beta(1+\|x^*\|)}\right)$ *where* x^* *is the minimum-norm optimal solution of* $f(x)$.

Practically, one can implement the algorithm in a predictor-corrector fashion (e.g., [4]) to explore wide neighborhoods and without knowing Lipschitz constant β. One can also scale variable x such that the norm of the minimum-norm solution x^* is about 1.

References

[1] D. A. Bayer and J. C. Lagarias (1989), The nonlinear geometry of linear programming, Part I: Affine and projective scaling trajectories. *Transactions of the American Mathematical Society*, 314(2):499–526.

[2] O. Güler and Y. Ye (1993), Convergence behavior of interior point algorithms. *Math. Programming*, 60:215–228.

[3] N. Megiddo (1989), Pathways to the optimal set in linear programming. In N. Megiddo, editor, *Progress in Mathematical Programming : Interior Point and Related Methods*, pages 131–158. Springer Verlag, New York.

[4] S. Mizuno, M. J. Todd, and Y. Ye (1993), On adaptive step primal–dual interior–point algorithms for linear programming. *Mathematics of Operations Research*, 18:964–981.

[5] Yurii Nesterov (2017), Accelerating the Universal Newton Methods. In Nemfest, George Tech, Atlanta.

[6] Yu. E. Nesterov and A. S. Nemirovskii (1993), *Interior Point Polynomial Methods in Convex Programming: Theory and Algorithms.* SIAM Publications. SIAM, Philadelphia.

[7] J. Renegar (1988), A polynomial–time algorithm, based on Newton's method, for linear programming. *Math. Programming*, 40:59–93.

[8] Weijie Su, Stephen Boyd and Emmanuel J. Candés (2014), A Differential Equation for Modeling Nesterovs Accelerated Gradient Method: Theory and Insights. In Advances in Neural Information Processing Systems (NIPS) 27.

[9] Andre Wibisono, Ashia C. Wilson, Michael I. Jordan (2016), A Variational Perspective on Accelerated Methods in Optimization. https://arxiv.org/abs/1603.04245

[10] Y. Ye (2017), Lecture Note 13: Second-Order Optimization Algorithms II. http://web.stanford.edu/class/msande311/lecture13.pdf