Unconstrained Optimization III: some convergence results of line search methods

Note Title 4/10/2022

In the analysis of line search methods, we shall assume that the objective functions $f:\mathbb{R}^n \to \mathbb{R}$ is C^1, and that its gradient is Lipschitz, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\| \quad \forall x,y \in \text{either the whole or part of } \mathbb{R}^n$$

choice of norm cannot affect the existence of L, but affects its numerical value.

Typically, L depends on how the domain of f is restricted (and hence on x_0).

E.g. $f(x) = x^\beta \quad \beta > 2$

$$|f'(x) - f'(y)| = |f''(\xi)|\,|x-y| \quad \text{for some } \xi \in [x,y].$$

$$|f''(\xi)| = |\beta(\beta-1)x^{\beta-2}| \leq \beta(\beta-1)|x_0|^{\beta-2} \quad \text{for } |x| \leq |x_0|.$$

So L can be chosen to be $\beta(\beta-1)|x_0|^{\beta-2}$.

Also, $\nexists L$ st $|f'(x) - f'(y)| \leq L|x-y| \ \forall x,y \in \mathbb{R}$.

Write $C^{1,1}_L(\mathcal{U}) := \{ f \in C^1(\mathbb{R}^n) \mid \|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \ \forall x, y \in \mathcal{U} \}$

E.g. Linear functions $f(x) = a^T x + b$ is in $C^{1,1}_0(\mathbb{R}^n)$

Quadratics $f(x) = \frac{1}{2}x^T A x + bx + c$ is C^∞, does its gradient have a uniform Lipshitz bound over the whole $\mathbb{R}^n$?

Yes it does: $\nabla f(x) = Ax + b$

$$\|\nabla f(x) - \nabla f(y)\| = \|A(x-y)\| \le \|A\| \|x-y\|$$

the matrix norm induced by whichever vector norm used to define Lipshitz continuity i.e. $\|A\| := \max_{\|x\|=1} \|Ax\|$

Recall: $\|A\|_2 = \sigma_{max}(A)$

if $A \succeq 0$, $\|A\|_2 = \lambda_{max}(A)$.

It's easy to show that, using the fundamental theorem of calculus, that if $f \in C^2(\mathbb{R}^n)$,

$$f \in C^{1,1}_L(\mathcal{U}) \iff \|\nabla^2 f(x)\| \le L, \ \forall x \in \text{Convex Hull}(\mathcal{U}).$$

Two global convergence results

Theorem (Zoutendijk) Let $f \in C^1(\mathbb{R}^n)$ be bounded below.
Consider any line search method $x_{k+1} = x_k + \alpha_k p_k$, where

- p_k is a descent direction $(\nabla f(x_k)^T p_k < 0)$
- α_k is a step size that satisfies the Wolfe conditions
  $\phi(\alpha) := \phi(x_k + \alpha p_k)$, $\phi(\alpha_k) \leq \phi(0) + c_1 \alpha_k \phi'(0)$, $\phi'(\alpha_k) \geq c_2 \phi'(0)$, $0 < c_1 < c_2 < 1$

If $f \in C_L^{1,1}(\mathcal{L})$ for $\mathcal{L} := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$.
Then

$$\sum_{k \geq 0} \cos^2\theta_k \, \|\nabla f_k\|_2^2 < \infty, \quad \text{where } \cos\theta_k = -\nabla f_k^T p_k / \|\nabla f_k\|_2 \cdot \|p_k\|_2.$$

Proof: The 2nd Wolfe condition gives

$$(\nabla f_{k+1} - \nabla f_k)^T p_k = \phi'(\alpha_k) - \phi'(0) \geq (c_2 - 1)\phi'(0) = -(1-c_2)\phi'(0)$$

while Lipschitz condition gives $(\nabla f_{k+1} - \nabla f_k)^T p_k \leq L \|\alpha_k p_k\|_2 \|p_k\|_2 = \alpha_k L \|p_k\|_2^2$.

The two inequalities imply : $\alpha_k \geq -\dfrac{1-c_2}{L} \dfrac{\phi'(0)}{\|p_k\|_2^2}$

Substitute this lower bound of α_k into the 1st Wolfe condition,

$$f_{k+1} \leq f_k - c_1 \frac{1-c_2}{L} \frac{\overbrace{\phi'(0)^2}^{=(\nabla f_k^T p_k)^2}}{\|p_k\|_2^2}$$

$$= f_k - \underbrace{c_1 \frac{1-c_2}{L}}_{=:c} \cos^2\theta_k \|\nabla f_k\|_2^2$$

So,

$$f_{k+1} \leq f_0 - c \sum_{j=0}^{k} \cos^2\theta_j \|\nabla f_j\|^2$$

But f is bounded below, so $f_{k+1} > -\infty$ and $\underbrace{\sum_{j=0}^{k} \cos^2\theta_j \|\nabla f_j\|^2 < \infty}_{\text{called the Zoutendijk condition (Z)}}$. Q.E.D.

Condition (Z) $\Rightarrow \lim_{k\to\infty} \cos^2\theta_k \|\nabla f_k\|^2 = 0$

Under the extra condition that $\cos\theta_k \geq \delta > 0 \ \forall k$, $\lim_{k\to\infty} \|\nabla f_k\| = 0$.

For GD, $\cos\theta_k = 1$, $\forall k$, so we can conclude that any cluster point of x_k is a stationary point of f.

For GD at least, we can dispense with the (full) Wolfe condition and use backtracking to attain the first Wolfe (aka Armijo) condition.

Recall:

Armijo condition	satisfied by any small enough α	backtracking alg. easy to implement
Wolfe condition	not satisfied by arb. small α	alg. complicated to implement

Theorem (convergence of GD) Let $f \in C_L^{1,1}(\mathbb{R}^n)$ be bounded below.

Let $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$, $k \geq 0$, with α_k chosen with one of the following stepsize strategies:

- constant step size $\alpha_k = \bar{\alpha} \in (0, 2/L)$ (choice requires knowing L)
- exact line search (too expensive in general)
- backtracking with parameters $\bar{\alpha}$, $c \in (0,1)$, $\rho \in (0,1)$. (perhaps most practical/robust)

~~Assume $f \in C_L^{1,1}(\mathcal{L})$, $\mathcal{L} := \{x : f(x) \leq f(x_0)\}$.~~

Then $f(x_k)$ is non-increasing and convergent, and $\lim_{k\to\infty} \|\nabla f_k\| = 0$.

Ex: Compare the two theorems. Make sure you understand the similarities and differences of what the two results say.

Under the setting of either theorem,

$$f(x_{k+1}) < f(x_k), \text{ unless } \nabla f(x_k) = 0 \quad \text{(in which case the iteration terminates.)}$$

Neither theorem (in the way it is stated) says anything about rate of convergence. The analysis of rate of convergence is very subtle in this setting. Recall from last week that it is impossible to get linear convergence for GD.

Proof

① Similar to the previous proof, we get the desired result if we can show that

$$f_{k+1} \leq f_k - M \cancel{\cos^2\theta_k} \|\nabla f_k\|^2 \quad \text{for some constant } M > 0. \quad \text{— (M)}$$

[Since f_k is $\downarrow$ and bounded below, $\{f_k\}$ must be convergent. But then $M\|\nabla f_k\|^2 \leq f_k - f_{k+1} \to 0$, which also implies $\|\nabla f_k\| \to 0$.]

② Descent lemma: $\forall x, y$, $|f(y) - f(x) - \nabla f(x)^T(y-x)| \leq \frac{L}{2}\|x-y\|_2^2$

Proof: By the fundamental thm of calculus,

$$f(y)-f(x)=\int_0^1 \overbrace{\langle \nabla f(x+t(y-x)), y-x\rangle}^{=\frac{d}{dt}f(x+t(y-x))}\,dt$$

$$= \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x+t(y-x)), y-x\rangle - \langle \nabla f(x), y-x\rangle\,dt$$

Thus,

$$\left| f(y)-f(x)-\langle \nabla f(x), y-x\rangle \right| = \left| \int_0^1 \langle \nabla f(x+t(y-x)), y-x\rangle - \langle \nabla f(x), y-x\rangle\,dt \right|$$

$$\leq \int_0^1 \left| \langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle \right| dt$$

Cauchy-Schwartz

$$\leq \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|\,\|y-x\|\,dt$$

$$\leq \int_0^1 tL\|y-x\|^2 dt = L/2\,\|y-x\|^2.$$

③ Sufficient decrease lemma: For any $x\in\mathbb{R}^n$, $\alpha>0$,

(specific for GD)

$$f(x)-f(x-\alpha\nabla f(x)) \geq \alpha\left(1-\frac{L\alpha}{2}\right)\|\nabla f(x)\|^2.$$

Proof: By ②, $f(x-\alpha\nabla f(x)) \leq f(x) - \alpha\|\nabla f(x)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x)\|^2$

$$= f(x) - \alpha\left(1-\frac{L\alpha}{2}\right)\|\nabla f(x)\|^2.$$

rearrange

④ By ③, we see that a constant step size $\bar{\alpha} \in (0, 2/L)$ guarantees

$$f(x_k) - f(\underbrace{x_{k+1}}_{= x_k - \bar{\alpha}\nabla f_k}) \geq \bar{\alpha}\underbrace{\left(1 - \frac{L\bar{\alpha}}{2}\right)}_{>0}\|\nabla f_k\|^2$$

So in this case, the constant M desired in ① can be set to $M = \bar{\alpha}\left(1 - \frac{L\bar{\alpha}}{2}\right)$.

Note : $\bar{\alpha} = 1/L$ maximizes $\alpha(1 - \frac{L}{2}\alpha)$ over $(0, 2/L)$, and thus a popular choice of step size. In this case

$$f_k - f_{k+1} \geq \frac{1}{2L}\|\nabla f_k\|^2.$$

Exact line search $\alpha_k \in \operatorname{argmin}_{\alpha \geq 0} f(x_k - \alpha \nabla f_k)$ guarantees

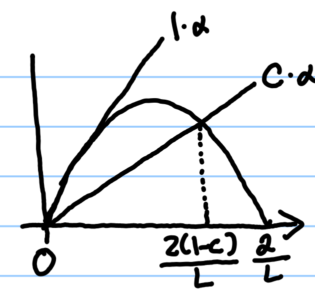
$$f_k - f_{k+1} \geq f_k - f\left(x_k - \frac{1}{L}\nabla f_k\right) \geq \frac{1}{2L}\|\nabla f_k\|^2.$$

So in this case, the constant M desired in ① can be set to $M = 1/2L$.

In backtracking, we seek a small enough α_k for which

$$f(x_k) - f(x_k - \alpha_k \nabla f_k) \geq c\,\alpha_k \|\nabla f_k\|^2, \quad c \in (0,1). \quad \text{— (SD)}$$

By ③, $\quad f(x) - f(x - \alpha \nabla f(x)) \geq \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x)\|^2.$



So (SD) is satisfied as long as $\alpha \in \left[0, \frac{2(1-c)}{L}\right]$.

Backtracking guarantees that α_k is either $\bar{\alpha}$ (when no backtracking is needed),

OR

$$\alpha_k/\rho > 2(1-c)/L, \text{ i.e. } \alpha_k > \frac{2(1-c)\rho}{L}.$$

So $\alpha_k \geq \min\left\{\bar{\alpha}, \frac{2(1-c)\rho}{L}\right\}$.

By (SD), the constant M desired in ① can be set to

$$M = c \min\left\{\bar{\alpha}, \frac{2(1-c)\rho}{L}\right\}.$$

Q.E.D.

Note: For the backtracking case, it is easy to see that the proof goes through if we weaken the "$f \in C_L^{1,1}(\mathbb{R}^n)$" assumption to

$$f \in C_L^{1,1}(\mathcal{L}),\ \mathcal{L} := \{x : f(x) \leq f(x_0)\} \text{ (as in Zoutendijk's theorem).}$$

But for the constant step size case, the weaker condition does not seem to be sufficient.

Thm (ROC of gradient norms) Under the setting of the previous theorem, write $f^* = \lim_{k\to\infty} f(x_k)$. Then $\forall n \geq 0$, $\min_{0\leq k\leq n} \|\nabla f(x_k)\| \leq \sqrt{\dfrac{f(x_0)-f^*}{M(n+1)}}$,

where
$$M = \begin{cases} \bar{\alpha}\left(1-\frac{\bar{\alpha}L}{2}\right) & \text{constant step size} \\ 1/2L & \text{exact line search} \\ c\min\left\{\bar{\alpha}, \frac{2(1-c)\rho}{L}\right\} & \text{backtracking} \end{cases}$$

<u>Proof</u>: We proved that $f_{k+1} \leq f_k - M\|\nabla f_k\|^2$

$$\left.\begin{aligned} f_0 &\geq f_1 + M\|\nabla f_0\|^2 \\ f_1 &\geq f_2 + M\|\nabla f_1\|^2 \\ &\vdots \\ f_n &\geq f_{n+1} + M\|\nabla f_n\|^2 \end{aligned}\right\} \Rightarrow f_0 - f_{n+1} \geq M\sum_{k=0}^{n}\|\nabla f_k\|^2$$

$$\Downarrow$$

$$f_0 - f^* \geq M \underbrace{\sum_{k=0}^{n}\|\nabla f_k\|^2}_{\geq (n+1)\min_{0\leq k\leq n}\|\nabla f_k\|^2}$$

$$\Downarrow$$

the desired result.

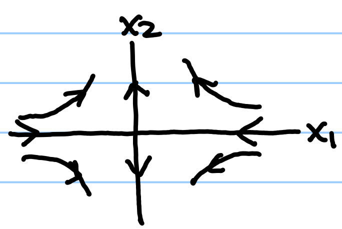
Q.E.D.

Note: This $O(\frac{1}{\sqrt{n}})$ bound is probably not tight. In the convex case (see below), it can be shown that $\|\nabla f_n\| = O(1/n)$.

The following example tells you something subtle about GD applied to a non-convex objective.

> It is possible that GD converges to a saddle point.
> (Recall that our theorem only promises GD, when converges, converges to a critical point. It does not promise the limit must be a local minimizer.

Consider $f(x) = \underset{\overset{\vee}{0}}{a} x_1^2 - \underset{\overset{\vee}{0}}{b} x_2^2$ The origin is a saddle point.



Any initial point $x^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \end{bmatrix} \neq 0$ would be sent to ∞ by GD.

But any initial point $x^0 = \begin{bmatrix} x_1^0 \\ 0 \end{bmatrix}$ stays on the x_1-axis and is attracted to the saddle point.

Note : $\nabla f(x) = \begin{bmatrix} 2ax_1 \\ -2bx_2 \end{bmatrix} \underset{\text{if } x_2 = 0}{\overset{\uparrow}{=}} \begin{bmatrix} 2ax_1 \\ 0 \end{bmatrix}$

However, as this and the next example suggest, for most initial points GD converges (when it converges) to a local minimizer. This has been proved rigorously.

A slightly more interesting example

$$f(x_1, x_2) = \frac{1}{2}x_1^2 - \frac{1}{2}x_2^2 + \frac{1}{4}x_2^4$$

Ex: (i) Prove that the only critical points are $\begin{bmatrix}0\\0\end{bmatrix}$, $\begin{bmatrix}0\\-1\end{bmatrix}$, $\begin{bmatrix}0\\1\end{bmatrix}$.

(ii) Are these critical points local min, local max, or saddle points?

(iii) What happens if GD (with any step size selection method) is applied to f if x^0 is on the x_1-axis?

(iv) Is $f \in C_L^{1,1}(\mathbb{R}^2)$ for some $L>0$?

Is f bounded below?

Is f coercive?

(v) With any step size selection method, is it possible to get global convergence for GD applied to this objective?

Two rate of convergence results for GD

(I) Suppose that $f \in C^2(\mathbb{R}^n)$ and that the iterates generated by the GD method with exact line searches converge to a point x^* at which the Hessian matrix $\nabla^2 f(x^*)$ is positive definite. Let

$$r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1 \right),$$

where $0 < \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\nabla^2 f(x^*)$. The for all k sufficiently large, we have

$$f(x_{k+1}) - f(x^*) \leqslant r^2 [f(x_k) - f(x^*)].$$

(II) Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and convex, and admit a minimizer x^*. Then the GD method with constant step size $\bar{\alpha} = 1/L$ generates a sequence x_k such that

$$\|\nabla f(x_k)\|_2 \leq \frac{L^2}{\sqrt{k(k+1)}} \|x_0 - x^*\|_2$$

and

$$f(x_k) - f(x^*) \leqslant \frac{L}{2(k+1)} \|x_0 - x^*\|_2.$$

recall the computation example at the end of wo-2.

Two implications of the ROC results

1. In finding critical point of a function $f: \mathbb{R}^n \to \mathbb{R}$, instead of solving $\nabla f(x) = 0$, one might consider solving

$$\min_x \tfrac{1}{2}\|\nabla f(x)\|^2 = h(x)$$

An appealing feature of this approach is that as long as x is a critical point of f, be it a local min, local max, a saddle point etc, x is always a global minimizer of h, with minimum value 0.

A problem of this method is that the conditioning of the original problem is squared, making it susceptible to very slow convergence if GD is used.

If $\bar{x}$ is a critical point of f (i.e. $\nabla f(\bar{x}) = 0$), then

$$f(x) = f(\bar{x}) + \cancel{\nabla f(\bar{x})^T (x-\bar{x})}^{\,0} + \tfrac{1}{2}(x-\bar{x})^T \nabla^2 f(\bar{x})(x-\bar{x}) + \cdots$$

$$\nabla f(x) = \nabla^2 f(\bar{x})(x-\bar{x}) + (\text{higher order terms in } (x-\bar{x}))$$

$$h(x) = \tfrac{1}{2}\nabla f(x)^T \nabla f(x) = \tfrac{1}{2}(x-\bar{x})^T [\nabla^2 f(\bar{x})]^2 (x-\bar{x}) + \text{h.o.t}$$

assume $\nabla^2 f(\bar{x}) \succ 0$ for simplicity

$$\nabla^2 h(\bar{x}) = [\nabla^2 f(\bar{x})]^2 \Rightarrow \kappa(\nabla^2 h(\bar{x})) = \frac{\lambda_{max}(\nabla^2 h(\bar{x}))}{\lambda_{min}(\nabla^2 h(\bar{x}))} = \left[\frac{\lambda_{max}(\nabla^2 f(\bar{x}))}{\lambda_{min}(\nabla^2 f(\bar{x}))}\right]^2 = \kappa(\nabla^2 f(\bar{x}))^2.$$

2. Later in the course we shall study methods for constrained optimization:

Consider (P) $\quad \min f(x) \quad$ s.t. $h(x)=0 \qquad f, h : \mathbb{R}^n \to \mathbb{R}$

We may turn it into an unconstraint optimization problem by solving

$(P_\mu) \quad \min f(x) + \frac{\mu}{2} h(x)^2$ for a large $\mu > 0$. (called a penalty method)

The larger the μ the better. (In fact, if x_μ^* is a solution of (P_μ), then we expect that $\lim_{\mu\to\infty} x_\mu^*$, if exists, is a solution of (P).)

Problem is that the bigger the μ, the more ill-conditioned (P_μ) is at a minimizer.

For example, consider a quadratic program (QP) $\quad \min \frac{1}{2}x^T A x + b^T x \quad$ s.t. $c^T x = c_0$

(P_μ) is $\quad \min \underbrace{\frac{1}{2}x^T A x + b^T x + \mu (c^T x - c_0)^2}_{} \qquad$ (assume $A \succ 0$ for simplicity)

$$= \frac{1}{2}x^T (A + \mu c c^T) x + (b - 2\mu c_0 c)^T x + \mu c_0^2$$

The Hessian of the objective of (P_μ) is $A+\mu cc^T$.

Some juicy linear algebra:

the eigenvalues of A and $A+(\sqrt{\mu}c)(\sqrt{\mu}c)^T$ are interlaced:
$\lambda_1\le\cdots\le\lambda_n$ $\quad$ $\lambda_1^\mu\le\cdots\le\lambda_n^\mu$

$$\lambda_1 \le \lambda_1^\mu \le \lambda_2 \le \lambda_2^\mu \le \cdots\cdots \le \lambda_n \le \lambda_n^\mu .$$

moreover, (for most vectors c) $\lambda_n^\mu \to \infty$ as $\mu \uparrow \infty$.

so $\kappa(A+\mu cc^T) = \lambda_{max}(A+\mu cc^T)/\lambda_{min}(A+\mu cc^T) \uparrow \infty$ as $\mu\uparrow\infty$.

This means, to the very least, it is unsuitable to apply GD to solve (P_μ) for a large penalty parameter μ.