Sub-optimality bounds

Stephen Becker
Applied Math, U. Colorado Boulder stephen.becker@colorado.edu

January 12, 2021

Lipschitz continuity of derivative and/or strong convexity of f

The definition of Lipschitz continuity of ∇f (with constant L) is

$$\forall x, y \quad \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \tag{1}$$

and the definition of f being μ strongly convex means that the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex[1]. In the lines below, if L or μ appears, then we are assuming the gradient is Lipschitz with constant L or f is strongly convex with constant μ, respectively. Most references to Nesterov's book are to his first edition [Nes04], not the recent 2018 edition [Nes18].

These two inequalities are very helpful; see, e.g., Thm 2.1.5 and Thm 2.1.10 from [Nes04].

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 \tag{2}$$

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2 \tag{3}$$

If we drop convexity but keep Lipschitz continuity of the gradient, then the first equation is still true, but the second equation is not true with $\mu = 0$, but it is true with $\mu = -L$. This is often written as $|f(y) - (f(x) + \langle \nabla f(x), y - x\rangle)| \le \frac{L}{2}\|x - y\|^2$.

The main inequalities can be summarized by:

$$\left.\begin{array}{rl} L^{-1}\|\nabla f(x) - \nabla f(y)\|^2 & \text{(a)} \\ \mu\|x - y\|^2 & \text{(b)} \\ \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2 & \text{(c)} \end{array}\right\} \le \langle \nabla f(x) - \nabla f(y), x - y\rangle \le \begin{cases} \text{(d)} & L\|x - y\|^2 \\ \text{(e)} & \mu^{-1}\|\nabla f(x) - \nabla f(y)\|^2 \end{cases} \tag{4}$$

The inequality (a) really follows from the co-coercivity of gradients; this result is actually surprisingly strong, since it makes implicit use of the Baillon-Haddad theorem. The result (e) for μ also requires f be continuously differentiable. The (c) inequality assumes both strong convexity and Lipschitz continuity of the gradient; see [Nes04, Thm. 2.1.12] for a derivation.

Sub-optimality bounds

For unconstrained smooth optimization, if $x^\star$ is a minimizer, then $\nabla f(x^\star) = 0$. Note there are 3 equivalent definitions of optimality: x is optimal if

$$\|x - x^\star\| = 0, \quad f(x) - f^\star = 0, \quad \|\nabla f(x)\| = 0 \tag{5}$$

[1] See Thm. 5.17 and Remark 5.18 in [Bec17] — this is actually only true if $\|\cdot\|$ is the induced norm from the inner product. However, most other properties hold for a general norm.

and this would be "iff" if we assume the optimal solution is unique. Now, given a Lipschitz continuous derivative, we can bound

$$\|\nabla f(x)\| = \|\nabla f(x) - \nabla f(x^\star)\| \le L\|x - x^\star\| \quad \text{by (1)} \tag{6}$$

$$f(x) - f^\star \le \frac{L}{2}\|x - x^\star\|^2 \quad \text{by (2)} \tag{7}$$

$$\|\nabla f(x)\|^2 \le 2L\,(f(x) - f^\star) \quad \text{by Eq. (9.14) in [BV04]} \tag{8}$$

and given μ strong convexity, we can bound in the other direction:

$$\|x - x^\star\|^2 \le \frac{1}{\mu^2}\|\nabla f(x)\|^2 \quad \text{by (4) (b) and (e). This is basically EB.} \tag{9}$$

$$\|x - x^\star\|^2 \le \frac{2}{\mu}\,(f(x) - f^\star) \quad \text{by (3), with } x = x^\star,\ y = x. \text{ This is basically QG.} \tag{10}$$

$$f(x) - f^\star \le \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \text{by Eq. (9.9) in [BV04]. This is PL} \tag{11}$$

Note: at least Eq. (10) holds for any norm [Bec17, Thm. 5.25]. Given both L and μ, we can combine the bounds, and bound any one of the 3 error metrics in terms of another, i.e., $\|\nabla f(x)\|^2 \le \frac{2L^2}{\mu}\,(f(x) - f^\star)$ and $f(x) - f^\star \le \frac{L}{2\mu^2}\|\nabla f(x)\|^2$. But these are not good bounds; the bounds in Eq (8) and (11) are better. Note: (11) is the Polyak-Lojasiewicz (PL) inequality, see Karimi, Nutini, Schmidt for details. Eq. (8) as derived in [BV04] requires f must be twice-continuously differentiable, but there are other derivations that do not require twice-continuously differentiable, e.g., [Nes18, Thm. 2.1.5, Eq. 2.1.10], and also a simple proof in section 12.1.3 of Shalev-Shwartz' book).

The reason we say "basically EB/QG" above is that EB (Error Bound condition)/QG (Quadratic Growth condition) (see later notes) apply when $x^\star$ is the *closest* optimal point to x. Under strong convexity (hence strict convexity), there's a unique optimal point, so then there's no need to specify. PL implies every stationary point is a global solution, but doesn't prove uniqueness.

Note that since the gradient is in the subdifferential, combined with Hölder's inequality, we also have (see [Nes18, §2.2.2])

$$f(x) - f^\star \le \|\nabla f(x)\|_p \|x - x^\star\|_{p'} \quad 1/p + 1/p' = 1 \tag{12}$$

which doesn't require Lipshitz continuity or strong convexity. This can be useful if it is known x lies in a bounded set, since then $\|x - x^\star\|$ can be bounded.

Another way to think of these quantities is as **Lyapunov functions**. See arXiv:1906.10053 for some ideas, e.g., a Lyapunov function can be the cost/objective, or Bregman or Euclidean distance, or with strong convexity, distance to unique solution. An alternative framework is via quasi-Fejér monotonicity.

References

[BC11] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces,* 1st edition, Springer, 2011.

[BC17] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces,* 2nd edition, Springer, 2017.

[Bec17] A. Beck, *First-Order Methods in Optimization*, SIAM, 2017.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Nes04] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer, Boston, 2004.

[Nes18] Yu. Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, 2018.