Ch 12: Convexity
APPM 7400 Theory of Machine Learning
Spring 2020

Stephen Becker

University of Colorado Boulder

March 16 2020

Smoothness and Strong Convexity

The definition of "**smoothness**" in some books (or "strong smoothness") of f means Lipschitz continuity of ∇f (with constant L):

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

The definition of f being $\mu > 0$ **strongly convex** means that the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex[1].

In the slides 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].

[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.

Under- and over-approximations

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

Related, [Nes18, Thm. 2.1.5, Eq. 2.1.10] gives

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$$

Inequalities

Some nice inequalities can be summarized by:

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

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.

We can actually get a tighter lower bound if we assume *both* strong convexity and Lipschitz continuity of the gradient; see [Nes04, Thm. 2.1.12] for a derivation. That result is:

$$\frac{\mu L}{\mu+L}\|x-y\|^2 + \frac{1}{\mu+L}\|\nabla f(x)-\nabla f(y)\|^2 \le \langle \nabla f(x)-\nabla f(y), x-y\rangle$$

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

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

If we change all the zeros above to $\epsilon > 0$, are these conditions equivalent? On the next slides, we'll investigate this.

To start with, here's a first result: note that since the gradient is in the subdifferential, combined with Hölder's inequality, then ([Nes18, §2.2.2])

$$f(x) - f^\star \le \|\nabla f(x)\|_p \|x - x^\star\|_{p'} \quad (\forall p, p' \text{ s.t. } 1/p + 1/p' = 1) \tag{6}$$

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.

Sub-optimality bounds: assuming strong smoothness

If f has a L-Lipschitz continuous derivative, we can bound

$$\|\nabla f(x)\| = \|\nabla f(x) - \nabla f(x^\star)\| \le L\|x - x^\star\| \quad \text{by (1)} \tag{7}$$

$$f(x) - f^\star \le \frac{L}{2}\|x - x^\star\|^2 \quad \text{by (2)} \tag{8}$$

$$\|\nabla f(x)\|^2 \le 2L\,(f(x) - f^\star) \quad \text{by Eq. (9.14) in [BV04]} \tag{9}$$

Note further that f must be twice-continuously differentiable to apply (9) is proved in [BV04] assuming f is twice-differentiable, but without assuming twice differentiability it can be proved using [Nes18, Thm. 2.1.5, Eq. 2.1.10].

Sub-optimality bounds: assuming strong convexity

Assuming f is $\mu > 0$ strong convexity, we can bound in the other direction:

$$\|x - x^\star\|^2 \leq \frac{1}{\mu^2}\|\nabla f(x)\|^2 \quad \text{by (4) (b) and (e)} \tag{10}$$

$$\|x - x^\star\|^2 \leq \frac{2}{\mu}(f(x) - f^\star) \quad \text{by (3), with } x = x^\star,\ y = x \tag{11}$$

$$f(x) - f^\star \leq \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \text{by Eq. (9.9) in [BV04]. This is PL} \tag{12}$$

Note: at least Eq. (11) holds for any norm [Bec17, Thm. 5.25].
Note: (12) is the Polyak-Lojasiewicz (PL) inequality, see Karimi, Nutini, Schmidt for details.

References

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

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

A. Beck, First-Order Methods in Optimization, SIAM, 2017.

S. Boyd and L. Vandenberghe.
Convex Optimization.
Cambridge University Press, 2004.

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

Yu. Nesterov.
Lectures on Convex Optimization.
Springer International Publishing, 2018.