Convergence speed with strong convexity

Huanle Xu

July 7, 2020

1 Strongly convex functions

Today we will talk about another property of convex functions that can significantly speed-up the convergence of first-order methods: strong convexity. We say that $f : \mathbb{R}^n \to \mathbb{R}$ is α–strongly convex if it satisfies

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

Of course this definition does not require differentiability of the function f, and one can replace $\nabla f(x)$ in the inequality above by $g \in \partial f(x)$. It is immediate to verify that a function f is α-strongly convex if and only if $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.

Note that (1) can be interpreted as follows: at any point x one can find a (convex) quadratic lower bound $q_x^-(y) = f(x) + \nabla f(x)^\top (y-x) + \frac{\alpha}{2}\|x-y\|^2$ to the function f, i.e. $\mathrm{q}_x^-(y) \leq f(y), \forall y \in \mathbb{R}^n$ (and $q_x^-(x) = f(x)$). Thus in some sense strong convexity is a dual assumption to the smoothness assumption from previous lectures. Indeed recall that smoothness can be defined via the inequality:

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

which implies that at any point y one can find a (convex) quadratic upper bound $q_y^+(x) = f(y) + \nabla f(y)^\top (x-y) + \frac{\beta}{2}\|x-y\|^2$ to the function f, i.e. $q_y^+(x) \geq f(x), \forall x \in \mathbb{R}^n$ (and $q_y^+(y) = f(y)$). In fact we will see later a precise sense in which smoothness and strong convexity are dual notions (via Fenchel duality). Remark also that clearly one always has$\beta \geq \alpha$.

2 Projected Subgradient Descent for strongly convex and Lipschitz functions

In this section we investigate the setting where f is strongly convex but potentially non-smooth. As we have already seen in a previous lecture, in the case of non-smooth functions we have to project back on the set where we control the norm of the gradients. Precisely let us assume that $\mathcal{X}$ is a compact and convex set such that $\forall x \in \mathcal{X}, \forall g \in \partial f(x), \|g\| \leq L$. We consider the Projected Subgradient Descent algorithm with time-varying step size, that is

$$\begin{aligned} y_{t+1} &= x_t - \eta_t g_t, \text{ where } g_t \in \partial f(x_t) \\ x_{t+1} &= \mathrm{argmin}_{x \in \mathcal{X}} \|x - y_{t+1}\|. \end{aligned}$$

The following result is extracted from a recent paper of Simon Lacoste-Julien, Mark Schmidt, and Francis Bach.

Theorem 1. *Let $\eta_s = \frac{2}{\alpha(s+1)}$, then Projected Subgradient Descent satisfies for* $\bar{x}_t \in \left\{ \mathrm{argmin}_{1 \le s \le t} f(x_s); \sum_{s=1}^t \frac{2s}{t(t-1)} x_s \right\}$,

$$f(\bar{x}_t) - \min_{x \in \mathcal{X}} f(x) \le \frac{2L^2}{\alpha(t+1)}.$$

Note that one can immediately see from the analysis in this lecture that the rate is optimal. Indeed, one can always find a function f and set $\mathcal{X}$ (an ℓ_2 ball) that satisfies the above assumptions and such that no black-box procedure can go at a rate faster than $\frac{L^2}{8\alpha t}$ for $t \le n$ (in fact the constant $1/8$ can be improved to $1/2$).

Proof. Let $x^* \in \mathrm{argmin}_{x \in \mathcal{X}} f(x)$. Coming back to our original analysis of Projected Subgradient Descent and using the strong convexity assumption one immediately obtains

$$f(x_s) - f(x^*) \le \frac{\eta_s}{2} L^2 + \left(\frac{1}{2\eta_s} - \frac{\alpha}{2} \right) \|x_s - x^*\|^2 - \frac{1}{2\eta_s} \|x_{s+1} - x^*\|^2.$$

Multiplying this inequality by s yields

$$s(f(x_s) - f(x^*)) \le \frac{L^2}{\alpha} + \frac{\alpha}{4} \Big(s(s-1)\|x_s - x^*\|^2 - s(s+1)\|x_{s+1} - x^*\|^2 \Big).$$

Now sum the resulting inequality over $s = 1$ to $s = t$, and apply Jensen's inequality to obtain the claimed statement. □

3 Gradient Descent for strongly convex and smooth functions

As will see now, having both strong convexity and smoothness allows for a drastic improvement in the convergence rate. The key observation is the following lemma.

Lemma 1. *Let f be β-smooth and α-strongly convex. Then for all $x, y \in \mathbb{R}^n$, one has*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \ge \frac{\alpha\beta}{\beta + \alpha} \|x - y\|^2 + \frac{1}{\beta + \alpha} \|\nabla f(x) - \nabla f(y)\|^2.$$

Proof. Using the definitions it is easy to prove that $\phi(x) = f(x) - \frac{\alpha}{2}\|x\|^2$is convex and $(\beta - \alpha)$-smooth, and thus using a result from the previous lecture one has

$$(\nabla \phi(x) - \nabla \phi(y))^\top (x - y) \ge \frac{1}{\beta - \alpha} \|\nabla \phi(x) - \nabla \phi(y)\|^2,$$

which gives the claimed result with straightforward computations. (Note that if $\alpha = \beta$ then one just has to apply directly the above inequality to f.) □

Theorem 2. *Let f be β-smooth and α-strongly convex, and let $Q = \frac{\beta}{\alpha}$ be the condition number of f. Then Gradient Descent with $\eta = \frac{2}{\alpha+\beta}$satisfies*

$$f(x_t) - f(x^*) \le \frac{\beta}{2} \left(\frac{Q-1}{Q+1} \right)^{2(t-1)} \|x_1 - x^*\|^2.$$

Proof. First note that by β-smoothness one has

$$f(x_t) - f(x^*) \leq \frac{\beta}{2}\|x_t - x^*\|^2.$$

Now using the previous lemma one obtains

$$\begin{aligned}
\|x_t - x^*\|^2 &= \|x_{t-1} - \eta\nabla f(x_{t-1}) - x^*\|^2 \\
&= \|x_{t-1} - x^*\|^2 - 2\eta\nabla f(x_{t-1})^\top(x_{t-1} - x^*) + \eta^2\|\nabla f(x_{t-1})\|^2 \\
&\leq \left(1 - 2\frac{\eta\alpha\beta}{\beta+\alpha}\right)\|x_{t-1} - x^*\|^2 + \left(\eta^2 - 2\frac{\eta}{\beta+\alpha}\right)\|\nabla f(x_{t-1})\|^2 \\
&= \left(\frac{Q-1}{Q+1}\right)^2\|x_{t-1} - x^*\|^2,
\end{aligned}$$

which concludes the proof. $\square$