OPTIMIZATION FOR MACHINE LEARNING

September 26, 2024

Today: Lab session
Jupyter notebook available online
(see Google doc on my webpage)

These notes:
A proof for the convergence rate of gradient descent on strongly convex functions with Lipschitz continuous gradients

Setup

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2} x^T C x \qquad \mu I \preceq C \preceq L I$$

$$C = C^T \in \mathbb{R}^{d \times d}$$

$L > 0$
$\mu > 0$

$\longrightarrow$ f is $C^{1,1}_L$ and μ-strongly convex

$\longrightarrow$ $x^* := \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$, $\quad f^* = f(x^*)$

Lemma: For any $(x,y) \in (\mathbb{R}^d)^2$, we have

$$[\nabla f(x) - \nabla f(y)]^T (x-y) \geq \frac{1}{L+\mu} \|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{L+\mu} \|x-y\|^2$$

NB: This result holds for any $C^{1,1}_L$, μ-strongly convex function (not necessarily quadratic)

Proof: Define $q : x \longmapsto q(x) = f(x) - \frac{\mu}{2}\|x\|^2$.

For any $(x,y) \in (\mathbb{R}^d)^2$, $\langle \nabla q(x) - \nabla q(y), x-y \rangle = \langle \nabla f(x) - \nabla f(y), x-y \rangle - \mu \|x-y\|^2$

Therefore,

$$\langle \nabla q(x) - \nabla q(y), x-y \rangle \geq \mu \|x-y\|^2 - \mu \|x-y\|^2 = 0$$

$$\langle \nabla q(x) - \nabla q(y), x-y \rangle \leq L \|x-y\|^2 - \mu \|x-y\|^2 = (L-\mu) \|x-y\|^2$$

As a result, q is convex and $C^{1,1}_{L-\mu}$

$g : \mathbb{R}^d \to \mathbb{R}$ convex and $C^{1,1}_L$ $\iff$ $0 \leq \langle \nabla g(x) - \nabla g(y), x-y \rangle \leq L\|x-y\|^2$ $\forall (x,y) \in (\mathbb{R}^d)^2$

Apply co-coercivity (1) to q yields

$$\forall (x,y) \in (\mathbb{R}^d)^2, \quad \left[\nabla q(x) - \nabla q(y)\right]^T (x-y) \geq \frac{1}{L-\mu} \|\nabla q(x) - \nabla q(y)\|^2$$

$$\left[\nabla q(x) - \nabla q(x)\right]^T (x-y) = \left[\nabla f(x) - \nabla f(y)\right]^T (x-y) - \mu \|x-y\|^2$$

$$\frac{1}{L-\mu} \|\nabla q(x) - \nabla q(y)\|^2 = \frac{1}{L-\mu} \|\nabla f(x) - \nabla f(y)\|^2 - \frac{2\mu}{L-\mu} \left[\nabla f(x) - \nabla f(y)\right]^T (x-y) + \frac{\mu^2}{L-\mu} \|x-y\|^2$$

Thus,

$$\left[\nabla f(x) - \nabla f(y)\right]^T (x-y) - \mu \|x-y\|^2 \geq \frac{1}{L-\mu} \|\nabla f(x) - \nabla f(y)\|^2 - \frac{2\mu}{L-\mu} \left[\nabla f(x) - \nabla f(y)\right]^T (x-y) + \frac{\mu^2}{L-\mu} \|x-y\|^2$$

$$\left(1 + \frac{2\mu}{L-\mu}\right) \left[\nabla f(x) - \nabla f(y)\right]^T (x-y) \geq \frac{1}{L-\mu} \|\nabla f(x) - \nabla f(y)\|^2 + \left[\mu + \frac{\mu^2}{L-\mu}\right] \|x-y\|^2$$

$$\frac{L+\mu}{L-\mu} \left[\nabla f(x) - \nabla f(y)\right]^T (x-y) \geq \frac{1}{L-\mu} \|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{L-\mu} \|x-y\|^2$$

$$\left[\nabla f(x) - \nabla f(y)\right]^T (x-y) \geq \frac{1}{L+\mu} \|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{L+\mu} \|x-y\|^2$$

$\square$

(1) Co-coercivity for g convex $C_L^{1,1}$ ($g: \mathbb{R}^d \to \mathbb{R}$) : $\forall (x,y) \in \mathbb{R}^d, \ (\nabla g(x) - \nabla g(y))^T (x-y) \geq \frac{1}{L} \|\nabla g(x) - \nabla g(y)\|^2$

Theorem: Apply gradient descent with stepsize $\frac{1}{L+\mu}$ to $\min_{x \in \mathbb{R}^d} f(x)$ where f $C_L^{1,1}$, μ-strongly convex with unique minimum x^*.

Then, for any $k \in \mathbb{N}$,

$$\|x_k - x^*\| \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - x^*\|$$

Preuve For any $j = 0 \dots k-1$,

$$\|x_{j+1} - x^*\|^2 = \|x_j - \frac{2}{L+\mu}\nabla f(x_j) - x^*\|^2$$

$$= \|x_j - x^* - \left(\frac{2}{L+\mu}\nabla f(x_j) - \frac{2}{L+\mu}\nabla f(x^*)\right)\|^2$$

using that $x^* = \operatorname{argmin}_x f(x)$ and $\nabla f(x^*) = 0_{\mathbb{R}^d}$

Expanding the square gives

$$\|x_{j+1} - x^*\|^2 = \|x_j - x^*\|^2 - \frac{4}{L+\mu}\langle \nabla f(x_j) - \nabla f(x^*), x_j - x^*\rangle + \frac{4}{(L+\mu)^2}\|\nabla f(x_j) - \nabla f(x^*)\|^2$$

Applying the lemma with $x = x_j$ and $y = x^*$, we obtain

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

Thus,

$$\|x_{j+1} - x^*\|^2 \leq \|x_j - x^*\|^2 - \frac{4}{(L+\mu)^2}\|\nabla f(x_j) - \nabla f(x^*)\|^2 - \frac{4\mu L}{(L+\mu)^2}\|x_j - x^*\|^2 + \frac{4}{(L+\mu)^2}\|\nabla f(x_j) - \nabla f(x^*)\|^2$$

$$= \left(1 - \frac{4\mu L}{(L+\mu)^2}\right)\|x_j - x^*\|^2 = \left(\frac{L-\mu}{L+\mu}\right)^2 \|x_j - x^*\|^2$$

We have thus established

$$\|x_{j+1} - x^*\|^2 \leq \left(\frac{L-\mu}{L+\mu}\right)^2 \|x_j - x^*\|^2$$

from which it follows that

$$\|x_k - x^*\| \leq \left(\frac{L-\mu}{L+\mu}\right)\|x_{k-1} - x^*\| \leq \dots \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - x^*\|$$

$\square$