Exercices
First-order optimization
Université Paris–Saclay

Exercice 1 (*Smooth and strongly convex functions*). — Let $L > 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ a L-smooth (for $\|\cdot\|_2$) differentiable function that admits a global minimizer $x_* \in \mathbb{R}^d$.

1) Prove that for all $x, x' \in \mathbb{R}^d$,

$$f(x') \geqslant f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2L} \|\nabla f(x') - \nabla f(x)\|_2^2 .$$

Indication: For each $x \in \mathbb{R}^d$, consider function $g_x : x' \mapsto f(x') - \langle \nabla f(x), x' \rangle$ and use Lemma 7.4.1 from the lecture notes.

2) Deduce that for all $x, x' \in \mathbb{R}^d$,

$$\langle \nabla f(x') - \nabla f(x), x' - x \rangle \geqslant \frac{1}{L} \|\nabla f(x') - \nabla f(x)\|_2^2 .$$

Let $K > 0$. We now further assume that f is also K-strongly convex for $\|\cdot\|_2$.

4) Prove that $f - \frac{K}{2} \|\cdot\|_2^2$ is $(L - K)$-smooth for $\|\cdot\|_2$.

5) Deduce that for all $x, x' \in \mathbb{R}^d$,

$$D_f(x', x) \geqslant \frac{1}{2(L-K)} \|\nabla f(x') - \nabla f(x)\|_2^2 + \frac{KL}{2(L-K)} \|x' - x\|_2^2 - \frac{K}{L-K} \langle \nabla f(x') - \nabla f(x), x' - x \rangle .$$

6) Deduce that for all $x, x' \in \mathbb{R}^d$,

$$\langle \nabla f(x') - \nabla f(x), x' - x \rangle \geqslant \frac{\mathrm{KL}}{\mathrm{K}+\mathrm{L}} \|x' - x\|_2^2 + \frac{1}{\mathrm{K}+\mathrm{L}} \|\nabla f(x') - \nabla f(x)\|_2^2 .$$

Exercice 2 (*Smooth and strongly convex optimization with Gradient Descent*). — Let $\mathrm{L}, \mathrm{K} > 0$, $f : \mathbb{R}^d \to \mathbb{R}$ a function that we assume differentiable, L-smooth and K-strongly convex for $\|\cdot\|_2$. We assume that f admits a global minimizer $x_* \in \mathbb{R}^d$. Let $x_1 \in \mathbb{R}^d$, $(\gamma_t)_{t \geqslant 1}$ a positive sequence and for $t \geqslant 1$, consider

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t).$$

1) Asssume that $\gamma_t = 1/\mathrm{L}$, and for all $t \geqslant 1$.

   a) Prove that for all $t \geqslant 1$,

   $$\|x_{t+1} - x_*\|^2 \leqslant \left(1 - \frac{\mathrm{K}}{\mathrm{L}}\right) \|x_t - x_*\|^2 .$$

   b) For $\mathrm{T} \geqslant 1$, deduce an upper bound on $f(x_{\mathrm{T}+1}) - f(x_*)$.

2) Assume that $\gamma_t = 2/(\mathrm{K}+\mathrm{L})$ for all $t \geqslant 1$. Let $t \geqslant 1$.

   a) Using the previous exercice, prove that

   $$\frac{1}{\mathrm{L}+\mathrm{K}} \|\nabla f(x_t)\|_2^2 + \frac{\mathrm{KL}}{\mathrm{L}+\mathrm{K}} \|x_t - x_*\|_2^2 \leqslant \langle \nabla f(x_t), x_t - x_* \rangle .$$

   b) Deduce that

   $$\|x_{t+1} - x_*\|_2^2 \leqslant \left(1 - \frac{2}{\mathrm{L}/\mathrm{K}+1}\right)^2 \|x_t - x_*\|_2^2 .$$

   c) Deduce, for $\mathrm{T} \geqslant 1$, an upper bound on $f(x_{\mathrm{T}+1}) - f(x_*)$.

Exercice 3 (*Smooth nonconvex optimization*). — Let $\mathrm{L} > 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ a L-smooth (for $\|\cdot\|_2$) differentiable function that admits a global minimizer $x_* \in \mathbb{R}^d$. Let $x_1 \in \mathbb{R}^d$ and for $t \geqslant 1$, consider

$$x_{t+1} = x_t - \frac{1}{\mathrm{L}} \nabla f(x_t).$$

1) Using the fact that for all $t \geqslant 1$, $D_f(x_{t+1}, x_t) \leqslant \frac{L}{2} \|x_{t+1} - x_t\|_2^2$, prove that for all $T \geqslant 1$

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leqslant \frac{2L(f(x_1) - f(x_*))}{T}.$$

2) Let $\mathscr{X} \subset \mathbb{R}^d$ be a closed convex set, and assume that f admits a minimizer $\tilde{x}_* \in \mathscr{X}$ on $\mathscr{X}$. Let $\tilde{x}_1 \in \mathbb{R}^d$ and for $t \geqslant 1$,

$$\tilde{x}_{t+1} = \Pi_{\mathscr{X}} \left(\tilde{x}_t - \frac{1}{L} \nabla f(\tilde{x}_t) \right).$$

For $x \in \mathbb{R}^d$, define

$$G(x) = L \left(x - \Pi_{\mathscr{X}}(x - \frac{1}{L} \nabla f(x)) \right).$$

Generalize the above analysis and establish for $T \geqslant 1$ an upper bound on

$$\frac{1}{T} \sum_{t=1}^{T} \|G(\tilde{x}_t)\|_2^2.$$

Exercice 4 (*Dual averaging for stochastic nonsmooth convex optimization*). — In the context of stochastic nonsmooth convex optimization from Section 6.4, define Dual Averaging iterates with time-dependent parameters and derive guarantees that get rid of the $\log T$ factor.