Gradient descent with momentum: exercises

Irène Waldspurger[*]

December 5, 2022

Exercise 1

Let μ, L be real numbers such that $0 < \mu < L$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth and μ-strongly convex function.
We consider Nesterov's algorithm applied to f: for any $x_0 \in \mathbb{R}$, we define

$$x_{t+1} = x_t - \frac{1}{L}\nabla f\left(x_t + \beta(x_t - x_{t-1})\right) + \beta(x_t - x_{t-1}),$$

where we have set $x_{-1} = x_0$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.
We assume that f has a minimizer. The goal of this exercise is to prove the theorem seen during the class on the convergence speed of Nesterov's algorithm to the minimizer.
Without loss of generality, we assume that the minimizer is 0, and $f(0) = 0$.
For any $t \in \mathbb{N}$, we define

$$y_t = x_t + \beta(x_t - x_{t-1}),$$
$$V_t = f(x_t) + \frac{L}{2}\left|\left|x_t - \left(1 - \sqrt{\frac{\mu}{L}}\right)x_{t-1}\right|\right|^2.$$

1. (a) Show that, for any $a, b \in \mathbb{R}^n$,

$$f(a) = f(b) + \int_0^1 \langle \nabla f(b + t(a-b)), a - b\rangle\, dt.$$

   (b) Deduce that

$$f(a) \leq f(b) + \langle \nabla f(b), a - b\rangle + \frac{L}{2}||a-b||^2.$$

[*] waldspurger@ceremade.dauphine.fr

2. (a) Show that, for all t,

$$f(x_{t+1}) \leq f(y_t) - \frac{L}{2}||x_{t+1} - y_t||^2.$$

(b) Deduce that

$$V_{t+1} \leq f(y_t) + L\left\langle x_{t+1}, y_t - \left(1 - \sqrt{\frac{\mu}{L}}\right) x_t \right\rangle + \frac{L}{2}\left(1 - \sqrt{\frac{\mu}{L}}\right)^2 ||x_t||^2 - \frac{L}{2}||y_t||^2.$$

3. (a) Show that, for all t,

$$f(y_t) \leq L\left\langle y_t - x_{t+1}, y_t \right\rangle - \frac{\mu}{2}||y_t||^2.$$

[Hint: apply the strong convexity assumption at points 0 and y_t.]

(b) Deduce that

$$V_{t+1} \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(f(y_t) + L\left\langle x_{t+1}, y_t - x_t \right\rangle + \frac{L}{2}\left(1 - \sqrt{\frac{\mu}{L}}\right)||x_t||^2 - \frac{L}{2}\left(1 - \sqrt{\frac{\mu}{L}} - \frac{\mu}{L}\right)||y_t||^2\right).$$

4. (a) Show that, for all t,

$$f(y_t) \leq f(x_t) + L\left\langle y_t - x_{t+1}, y_t - x_t \right\rangle - \frac{\mu}{2}||y_t - x_t||^2.$$

[Hint: apply the strong convexity assumption at points x_t and y_t.]

(b) Deduce that

$$V_{t+1} \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(f(x_t) + \frac{L}{2}\left|\left|\left(1 + \sqrt{\frac{\mu}{L}}\right) y_t - x_t\right|\right|^2 - \frac{1}{2}\left(\sqrt{\mu L} + \mu\right)||x_t - y_t||^2\right).$$

5. (a) Show that, for all t,

$$x_t - \left(1 - \sqrt{\frac{\mu}{L}}\right) x_{t-1} = \left(1 + \sqrt{\frac{\mu}{L}}\right) y_t - x_t$$

   (b) Deduce that

$$V_{t+1} \leq \left(1 - \sqrt{\frac{\mu}{L}}\right) V_t.$$

6. Show that, for all t, $f(x_t) \leq 2\left(1 - \sqrt{\frac{\mu}{L}}\right)^t f(x_0)$.

Exercise 2

Let $L > 0$ be fixed.
The goal of this exercise is to show that, up to a multiplicative constant, no first-order algorithm has a better convergence rate than Nesterov's method on the class of L-smooth convex functions.
Let Alg be a first-order algorithm. For any $n \in \mathbb{N}$, L-smooth convex function $f : \mathbb{R}^n \to \mathbb{R}$ and starting point x_0, we denote $(x_k^f)_{k\in\mathbb{N}}$ the sequence generated by Alg. We make a simplifying assumption on Alg:

(H): For any n, f, x_0, for all $k \in \mathbb{N}$, $x_k^f - x_0 \in \text{Vect}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\}$.

[Remark: standard gradient descent and heavy ball, for instance, satisfy this assumption. Nesterov's method does too, after slight reformulations.]
Let k be fixed. We define

$$\begin{array}{cccc} f : & \mathbb{R}^{2k+1} & \to & \mathbb{R} \\ & (s_1, \dots, s_{2k+1}) & \to & \frac{L}{4}\left(\frac{s_1^2}{2} + \frac{1}{2}\sum_{i=1}^{2k}(s_{i+1} - s_i)^2 + \frac{s_{2k+1}^2}{2} - s_1\right). \end{array}$$

1. Show that f is convex.

2. Compute ∇f and show that it is L-Lipschitz.

3. Show that the only minimizer of f is

$$x_* = \frac{1}{2(k+1)}(2k+1, 2k, ..., 1)$$

   and that

$$f(x_*) = -\frac{L}{8}\left(1 - \frac{1}{2(k+1)}\right).$$

4. We set $x_0 = 0$. Show that, for any $t = 1, \dots, 2k$,

$$x_t^f \in \{(s_1, \dots, s_t, 0, \dots, 0) | s_1, \dots, s_t \in \mathbb{R}\}.$$

5. Compute

$$\min\{f(s_1, \dots, s_k, 0, \dots, 0) | s_1, \dots, s_k \in \mathbb{R}\}.$$

6. Show that

$$f(x_k^f) - f(x_*) \geq \frac{L}{16(k+1)}.$$

7. Deduce that

$$f(x_k^f) - f(x_*) \geq \frac{3L}{32(k+1)^2} ||x_0 - x_*||^2.$$

[On rappelle que, pour tout $\ell \in \mathbb{N}$, $\sum_{r=1}^{\ell} r^2 = \frac{\ell(\ell+1)(2\ell+1)}{6}$.]

8. (Difficult) Show the same result without Assumption (H).