M2A 2023-2024
M2 Statistique & Learning

Exercises 2: Gradient descent

Exercise 1 (Optimal step-size for GD with L-smooth and μ-strongly convex functions)**.** *Assume the function $F : \mathbb{R}^d \to \mathbb{R}$ to be L-smooth and μ-strongly convex.*

1. *Show that, by μ-strong convexity, for all θ and $\theta_0 \in \mathbb{R}^d$*

$$\langle \nabla F(\theta) - \nabla F(\theta_0), \theta - \theta_0 \rangle \geq \mu \left\| \theta - \theta_0 \right\|_2^2 .$$

2. *Show that when F is L-smooth and μ-strongly convex, then*

$$\langle \nabla F(\theta) - \nabla F(\theta_0), \theta - \theta_0 \rangle \geq \frac{\mu L}{\mu + L} \left\| \theta - \theta_0 \right\|_2^2 + \frac{1}{\mu + L} \left\| \nabla F(\theta) - \nabla F(\theta_0) \right\|_2^2$$

3. *Using the previous inequalities, establish the following convergence rate for GD iterates*

$$\|\theta_{k+1} - \theta^\star\|_2^2 \leq \exp\left(-\frac{4k}{\mu/L + 1} \right) \|\theta_k - \theta^\star\|_2^2$$

   when choosing the step size as $\gamma = \frac{2}{\mu+L}$.

   NB: note that in such a case, we are performing bolder jumps, since $\frac{2}{\mu+L} > \frac{1}{L}$.

4. *Deduce a convergence rate on the objective function.*

Exercise 2 (Nesterov's Acceleration)**.** *The goal of this exercise is to derive a convergence rate for Nesterov's acceleration when minimizing a function F, assumed to be convex and L-smooth.*

Algorithm 1: Nesterov's acceleration

$d_0 \leftarrow 0$;
$\lambda_0 \leftarrow 1$;
$\gamma = 1/L$;
for $t = 0, \ldots, T-1$ **do**
  $\beta_t = \theta_t + d_t$;
  $\theta_{t+1} = \beta_t - \gamma \nabla F(\beta_t)$;
  $\lambda_{t+1} =$ largest solution of $\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$;
  $d_{t+1} = \frac{\lambda_t - 1}{\lambda_{t+1}}(\theta_{t+1} - \theta_t)$ (Momentum term)
end
return θ_T

Define $\delta_t := F(\theta_t) - F^\star$.

1. *Show that*

$$\delta_{t+1} - \delta_t \leq \left(\frac{L\gamma^2}{2} - \gamma \right) \|\nabla F(\theta_t + d_t)\|^2 + F(\theta_t + d_t) - F(\theta_t).$$

2. *Deduce that*

$$\delta_{t+1} - \delta_t \leq -\frac{L}{2} \left(\|g_t\|_2^2 - 2\langle g_t, d_t \rangle \right)$$

   with $g_t := \frac{1}{L} \nabla F(\theta_t + d_t)$.

3. By doing similar computations, deduce that

$$\delta_{t+1} \leq -\frac{L}{2}\left(\|g_t\|_2^2 - 2\langle g_t, \theta_t + d_t - \theta^\star\rangle\right).$$

4. By remarking that $\theta_t - \lambda_t g_t + \lambda_t d_t = \theta_{t+1} + \lambda_{t+1} d_{t+1}$, derive the bounds

$$(\lambda_t - 1)(\delta_{t+1} - \delta_t) + \delta_{t+1} \leq -\frac{L}{2\lambda_t}\left(\|\theta_{t+1} + \lambda_{t+1} d_{t+1} - \theta^\star\|_2^2 - \|\theta_t + \lambda_t d_t - \theta^\star\|_2^2\right)$$

and

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq -\frac{L}{2}\left(\|\theta_{t+1} + \lambda_{t+1} d_{t+1} - \theta^\star\|_2^2 - \|\theta_t + \lambda_t d_t - \theta^\star\|_2^2\right).$$

5. Note that $\lambda_t \geq (t+1)/2$ for all t, and conclude that Nesterov's acceleration gives

$$F(\theta_T) - F^\star \leq 2L\frac{\|\theta_0 - \theta^\star\|_2^2}{T^2}.$$