Exercises in Convex Optimization
Lecture 10

Michel Baes, Patrick Cheridito

November 27, 2018

Infimum convolution of convex functions 1. Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be two convex functions, and assume that for every $z \in \mathbb{R}^n$, the infimum $\inf_{x\in\mathbb{R}^n}\{f(x) + g(z - x)\}$ is attained. Define the function

$$h(z) := \min_{x\in\mathbb{R}^n}\{f(x) + g(z - x)\},$$

for every $z \in \mathbb{R}^n$. The function h is the *infimum convolution* of f and g and is sometimes denoted as $f \,\square\, g$.

Prove that h is convex and that $f \,\square\, g = g \,\square\, f$.

2. Such a construction can be useful to *smooth* nondifferentiable convex functions. Consider for instance the function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|$, and the function $g : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2/(2\epsilon)$, for $\epsilon > 0$. What is the function

$$h(z) := \inf_{x\in\mathbb{R}^n}\{f(x) + g(z - x)\}?$$

Note: this function is called the *Moreau envelope* of f.

`Lecture_11/InfConvolution`

Minimizing over the simplex In this exercise, we will specify a Mirror Descent method for problems of the form:

$$\min\{f(x) : x \in \Delta_n\},$$

where f is a Lipschitz continuous convex function and $\Delta_n := \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$ is the standard simplex, called also probability simplex. We assume that we have at our disposal a first-order oracle, which returns $f(x)$ and $g(x) \in \partial f(x)$ for every $x \in \Delta_n$.

We define the function $\psi : \mathbb{R}_+ \to \mathbb{R}$ by $\psi(t) := t\ln(t)$ for $t > 0$ and $\psi(0) := 0$. Consider the following function $V : \mathbb{R}^n \to \mathbb{R}$ defined by

$$x \mapsto V(x) = \begin{cases} \sum_{i=1}^n \psi(x_i), & \text{for } x \in \Delta_n \\ +\infty, & \text{otherwise.} \end{cases}$$

1. Show that the function V is convex on $\mathbb{R}^n$, continuous on Δ_n, and differentiable on relint(Δ_n).
2. What is the minimum of V and where is it attained? What is its maximum on Δ_n?
3. Compute the conjugate V_* of V and its gradient.
4. Determine the complexity (in terms of number of iterations) of the Mirror Descent method that uses the above function V, with an initial point $x_1 \in \text{relint}(\Delta_n)$. You can use the fact that $s(x) - V_*(s) \leq V(x)$ (why?). Explain in particular why we can use the same algorithm as for unconstrained problems.
   Note that V_* has a Lipschitz continuous gradient with Lipschitz constant 1 with respect to the norm $||\cdot||_1$, a result that you can use without a proof.

Notes: 1. This algorithm happens to be much faster than the standard gradient method for very high dimensional convex problems over the simplex, provided that the required accuracy is moderate.
2. In practice, to avoid possible numerical issues when x_k goes too close to the boundary of Δ_n, the function V is replaced by $\sum_{i=1}^n \psi(x_i + \delta)$ for a small $\delta > 0$ (typically δ is in the ranges of 10^{-12}.)

`Lecture_11/MirrorSimplex`

An optimal gradient method for smooth convex problems *(bonus exercise)* This method is due to Arkadi Nemirovski (private communication). It might have been published in Russian in the eighties.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function with a Lipschitz continuous gradient, i.e. there exists a constant $L > 0$ for which every point $x, y \in \mathbb{R}^n$ satisfies

$$||f'(x) - f'(y)||_2 \leq L||x - y||_2.$$

We would like to minimize the function f over $\mathbb{R}^n$. We assume that $f^* := \min\{f(x) : x \in \mathbb{R}^n\}$ is finite and that $R := \min\{||x||_2 : f(x) = f^*\}$ is finite as well, or in other words that there exists a minimizer in the ball $B[0, R]$.

A minimizing sequence full of properties Let $x_0 = 0, x_1, x_2, \ldots$ be a sequence of points of $\mathbb{R}^n$ that satisfy the following three properties. Here, we abbreviate $f'(x_t)$ by g_t.

1. $\langle g_t, x_t \rangle = 0$ for every $t \geq 0$.
2. There exists a sequence $\lambda_0, \lambda_1, \lambda_2, \ldots$ of positive numbers (still to be determined) for which $\langle \sum_{k=0}^{t-1} \lambda_k g_k, g_t \rangle = 0$ for every $t \geq 0$.
3. $f(x_{t+1}) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$ for every $t \geq 0$.

Let $\epsilon_t := f(x_t) - f^*$.

1. Show that $\{\epsilon_t : t \geq 0\}$ is a decreasing sequence.
2. Show that

$$\sum_{k=0}^{t} \lambda_k \epsilon_k \leq R\sqrt{\sum_{k=0}^{t} \lambda_k^2 ||g_k||_2^2} \leq R\sqrt{2L \sum_{k=0}^{t} \lambda_k^2 (\epsilon_k - \epsilon_{k+1})}.$$

Hint: we suggest you to use successively the convexity of f, and the three properties of the sequence $\{x_t : t \geq 0\}$ in the given order.

Defining a sequence λ_t that works

Observe that the sum on the right-hand side is:

$$S_t := \lambda_0^2\epsilon_0 + (\lambda_1^2 - \lambda_0^2)\epsilon_1 + (\lambda_2^2 - \lambda_1^2)\epsilon_2 + \cdots + (\lambda_t^2 - \lambda_{t-1}^2)\epsilon_t - \lambda_t^2\epsilon_{t+1},$$

and that this number is nonnegative (why?). We denote the left-hand side by $V_t := \sum_{k=0}^{t} \lambda_k\epsilon_k$. There is a natural choice for the constants $\lambda_0, \lambda_1, \ldots$ so that $V_t - \lambda_t^2\epsilon_{t+1} = S_t$: We can take $\lambda_0 = 1$, $\lambda_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_k^2})$ (why?). Show that $\lambda_k \geq \frac{k+1}{2}$ for every $k \geq 0$.

Prove that with this choice of λ_k, we have $\sqrt{V_t} \leq R\sqrt{2L}$ and thus:

$$\epsilon_{t+1} = f(x_{t+1}) - f^* \leq \frac{V_t}{\lambda_t^2} \leq \frac{2LR^2}{\lambda_t^2} \leq \frac{8LR^2}{(t+1)^2}.$$

Observe that this result reaches the optimal complexity for L-smooth convex problems with a first-order oracle.

We can construct such a minimizing sequence

Show that if $\hat{x}_t := x_t - \frac{g_t}{L}$, we have $f(\hat{x}_t) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$.

Define now $x_{t+1} := \arg\min \left\{ f(x) : x \in \text{span}\{\hat{x}_t, \sum_{k=0}^{t} \lambda_k g_k\} \right\}$.

Prove that:

1. $\langle g_{t+1}, x_{t+1} \rangle = 0$.
2. $\langle \sum_{k=0}^{t} \lambda_k g_k, g_{t+1} \rangle = 0$.
3. $f(x_{t+1}) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$.

Note: the drawback of this method is that one needs to solve at every step a 2D convex optimization problem exactly. Nesterov's accelerated method, which is based on entirely different principles, does not have this unpleasant feature.

`Lecture_11/OptimalGradientMethodForSmoothConvexProblems`