Math of ML : Exercises 8 *

November 10, 2025

In this exercise sheet, we will derive the discrete-time mirror descent algorithm and prove its convergence to a minimizer. The family of mirror descent algorithms generalize the gradient descent algorithm. The key notion allowing this generalization is Bregman divergence – a distance-like function that, as we shall see, in various ways behaves similarly to the squared Euclidean distance. Before we proceed, we introduce the following definitions.

Definition 1 (Closed and Convex Functions)**.** *A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called convex if its epigraph* $\mathrm{epi} f = \{(x,t) : f(x) \le t\} \subseteq \mathbb{R}^{d+1}$ *is a convex set. If in addition* epi$f$ *is closed, we say that the function f is closed. The set of values $\{x : f(x) < \infty\}$ is denoted by* $\mathrm{dom}(f)$*.*

Definition 2 (Legendre-Fenchel Transform)**.** *For any function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ define $f^*(x^*) = \sup_x \{\langle x, x^*\rangle - f(x)\}$. Restricted to the set of closed and convex functions, the mapping $f \mapsto f^*$ is one-to-one. In particular, for closed and convex f we have $f = f^{**}$.*

Exercise 1 (Bregman Divergences)**.** *Let ψ be a function with* $\mathrm{dom}\,\psi = \mathcal{X} \subseteq \mathbb{R}^d$ *and suppose that ψ is differentiable on the interior of its domain denoted by* $\mathrm{int}(\mathcal{X})$*. The Bregman divergence of ψ is a function $D_\psi : \mathcal{X} \times \mathrm{int}(\mathcal{X}) \to \mathbb{R}$ defined by*

$$D_\psi(x, x') = \psi(x) - \psi(x') - \langle \nabla\psi(x'), x - x'\rangle.$$

Compute ψ^*, $\mathrm{dom}(\psi^*)$ *and $D_\psi(x, x')$ when:*

1. *$\mathcal{X} = \mathbb{R}^d$ and $\psi(x) = \frac{1}{2}\|x\|_2^2$;*
2. *$\mathcal{X} = \{x \in \mathbb{R}^d : x_1, \dots, x_d \ge 0\}$ and $\psi(x) = \sum_{i=1}^d x_i \log(x_i) - x_i$ (defining $0 \log 0 = 0$).*
3. *$\mathcal{X} = \{x \in \mathbb{R}^d : x_1, \dots, x_d > 0\}$ and $\psi(x) = \sum_{i=1}^d -\log(x)$.*

Exercise 2 (Mirror Descent: Derivation of the Algorithm and the Key Identity)**.** *When minimizing a differentiable convex function $f : \mathcal{X} \to \mathbb{R}$ via the gradient descent updates $x_{t+1} = x_t - \eta\nabla f(x_t)$, the key to proving upper bounds on convergence rates (cf. Lecture 8) is the following identity, which holds for any $x^* \in \mathbb{R}^d$ (typically taken to be a minimizer of f):*

$$\underbrace{\frac{1}{2}\|x_t - x^*\|_2^2 - \frac{1}{2}\|x_t - x^*\|_2^2}_{\textit{change in potential}} = \langle -\eta\nabla f(x_t), x^* - x_t\rangle - \underbrace{\frac{1}{2}\|\eta\nabla f(x_t)\|_2^2}_{\textit{discretization error}}. \tag{1}$$

The purpose of this exercise is to generalize the identity (1) to Bregman divergences, which will allow us to repeat (nearly verbatim) the gradient decent convergence proof scheme shown in Lecture 8 to a general family of mirror descent algorithms (which includes gradient descent as a special case).

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

1. *Let $\psi : \mathcal{X} \to \mathbb{R}$ be differentiable on the interior of its domain. Prove the following identity, which holds for any $x, x', x^* \in \mathrm{int}(\mathcal{X})$:*

$$D_\psi(x^*, x) - D_\psi(x^*, x') = \langle \nabla\psi(x') - \nabla\psi(x), x^* - x\rangle - D_\psi(x, x'). \tag{2}$$

2. *Let us now relabel $x = x_t$ and $x' = x_{t+1}$. We wish to put the identity (2) closer in its form to the identity (1). This suggests defining the* mirror descent updates *as follows:*

$$\nabla\psi(x_{t+1}) = \nabla\psi(x_t) - \eta\nabla f(x_t). \tag{3}$$

Note that for $\psi(x) = \frac{1}{2}\|x\|_2^2$ the above updates coincide with the gradient descent updates. Suppose that the sequence $(x_t)_{t\geq 0}$ satisfies (3). Further, suppose that ψ is α-strongly convex[1] with respect to some norm $\|\cdot\|$. Let $\|\cdot\|_*$ denote the dual norm[2] of $\|\cdot\|$. Prove that for any $t = 0, 1, \ldots$ and any x^* it holds that

$$\underbrace{D_\psi(x^*, x_t) - D_\psi(x^*, x_{t+1})}_{\textit{change in potential}} \geq \langle -\eta\nabla f(x_t), x^* - x_t\rangle - \underbrace{\frac{1}{2\alpha}\|\eta\nabla f(x_t)\|_*^2}_{\textit{discretization error}}. \tag{4}$$

Remarks:

- *In (1) and (4) the term "discretization error" is not present in continuous time versions of gradient/mirror descent. This is because the inner product term is linear in the step size η while the discretization error term is quadratic in η.*
- *Note that while gradient descent pays for the ℓ_2 norms of the gradient in terms of the discretization error, with a suitable choice of the "mirror map" ψ we can instead pay for the size of gradients in a different norm. This is especially relevant, for example, if gradients live in a very high dimensional space with each coordinate bounded by a constant in absolute value. Then, the discretization error term in squared ℓ_2 norm scales linearly with the ambient dimension, while it scales only as a constant when measured in squared ℓ_∞ norm.*
- *In this exercise, we derived the mirror descent algorithm (3) by attempting to design update rules for which the gradient descent convergence proof would work with a different choice of a potential function than the squared ℓ_2 norm. This is very similar in spirit to how the mirror descent algorithm was originally designed. See [Nemirovskii and Yudin, 1983, Chapter 3] (the method of mirror descent and its convergence analysis dates back at least to late 70s; the above cited reference is an English edition of a book previously published in 1978 in Russian).*
- *Observe that the inequality (4) is actually an equality when $\psi(x) = \frac{1}{2}\|x\|_2^2$ and it recovers exactly the identity (1).*

To go further, we shall introduce additional regularity properties on the function ψ in (3). Below, we define the class of Legendre functions, a class of sufficiently regular convex functions (note that the functions considered in Exercise 1 are Legendre).

Definition 3 (Subgradients)**.** *A vector $z \in \mathbb{R}^d$ is a sub-gradient of a convex function ψ at a point x if and only if for any y we have $\psi(y) \geq \psi(x) + \langle z, y - x\rangle$. The set of sub-gradients of ψ at a point x is denoted by $\partial\psi(x)$. A convex function ψ is differentiable at a point x if and only if $\partial\psi(x) = \{\nabla\psi(x)\}$ (see [Rockafellar, 1970, Section 25] for a proof).*

[1] A function f is α-strongly convex with respect to the norm $\|\cdot\|$ if for any x, y it holds that $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|x - y\|^2$.

[2] The dual norm is defined by $\|x\|_* = \sup_{y:\|y\|\leq 1}\langle x, y\rangle$.

Definition 4 (Legendre Functions). *We say that a convex function ψ is Legendre (or of Legendre type) if (i) ψ is closed and convex; (ii)* $\mathrm{int}(\mathrm{dom}(\psi)) \neq 0$*; (iii) ψ is strictly convex and continuously differentiable on the interior of its domain; (iv) for any sequence $x_k \in \mathrm{int}(\mathrm{dom}(\psi))$ converging to a boundary point of* $\mathrm{int}(\mathrm{dom}(\psi))$ *we have $\|\nabla\psi(x_k)\|_2 \to \infty$.*

The class of Legendre functions is closed under convex conjugation; that is, whenever ψ is Legendre, so is ψ^*; see [Rockafellar, 1970, Section 26].

Exercise 3 (Inverting Gradient Map of Legendre Functions). *In order to implement the mirror descent updates (3) we need to compute an inverse of the gradient map $\nabla\psi$. Let ψ be a Legendre function with $\mathrm{dom}(\psi) = \mathcal{X}$. In this exercise, we show that the inverse of the gradient mapping $\nabla\psi$ exists and can be computed via $\nabla\psi^*$.*

1. *Show that for any closed and convex function $f : \mathrm{int}(\mathcal{X}) \to \mathbb{R}$ we have $x^* \in \partial f(x)$ if and only if $x \in \partial f^*(x^*)$.*

2. *For a Legendre function ψ, using the fact that ψ^* is differentiable on the interior of its domain, deduce that the inverse gradient map $(\nabla\psi)^{-1}$ exists and equals $\nabla\psi^*$.*

Exercise 4 (Bregman Projections and the Generalized Pythagorean Theorem). *Let ψ be a Legendre function with $\mathrm{dom}(\psi) = \mathcal{X}$. Let $\mathcal{C} \subseteq \mathcal{X}$ be a closed and convex set such that $\mathcal{C} \cap \mathrm{int}(\mathcal{X}) \neq \emptyset$. Define the Bregman projection operator $\Pi_\psi^{\mathcal{C}} : \mathrm{int}(\mathcal{X}) \to \mathcal{C} \cap \mathrm{int}(\mathcal{X})$ by*

$$\Pi_\psi^{\mathcal{C}}(x) = \mathrm{argmin}_{y \in \mathcal{C}} D_\psi(y, x).$$

When ψ is a Legendre function, the above map is well-defined: there exists a unique minimizer in the above minimization problem and the mininimizer lies in the set $\mathcal{C} \cap \mathrm{int}(\mathcal{X})$ (for a proof of this fact see, e.g., [Bauschke, Borwein, et al., 1997]). Prove that for any $x \in \mathcal{X}$ and any $x^* \in \mathcal{C}$ it holds that

$$D_\psi(x^*, x) \geq D_\psi(\Pi_\psi^{\mathcal{C}}(x), x) + D_\psi(x^*, \Pi_\psi^{\mathcal{C}}(x)).$$

Exercise 5 (Convergence Rate of Projected Sub-Gradient Mirror Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. We are interested in minimizing f over a closed and convex set $\mathcal{C} \subseteq \mathbb{R}^d$. Let ψ be a Legendre function with $\mathrm{dom}(\psi) = \mathcal{X} \subseteq \mathbb{R}^d$ and let $\Pi_\psi^{\mathcal{C}} : \mathrm{int}(\mathcal{X}) \to \mathcal{C} \cap \mathrm{int}(\mathcal{X})$ be the Bregman projection mapping onto the set $\mathcal{C} \subseteq \mathcal{X}$ (cf. Exercise 4), where we also assume that $\mathcal{C} \cap \mathrm{int}(\mathcal{X}) \neq \emptyset$.*

Suppose that f is α-strongly convex with respect to some norm $\|\cdot\|$ on the set $\mathcal{C} \cap \mathrm{int}(\mathcal{X})$. Also, suppose that there exists some constant L such that $\sup_{x \in \mathcal{C} \cap \mathrm{int}(\mathcal{X})} \sup_{g \in \partial f(x)} \|g\|_* \leq L$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ [3]*. Consider the projected gradient descent scheme:*

$$\begin{aligned} x_0 &\in \mathrm{int}(\mathcal{X}) \cap \mathcal{C}, \\ \nabla\psi(y_{t+1}) &= \nabla\psi(x_t) - \eta g_t, \ \text{where } g_t \in \partial f(x_t) \text{ for } t = 0, 1, 2, \ldots, \\ x_{t+1} &= \Pi_{\mathcal{C}}^{\psi}(y_{t+1}) \text{ for } t = 0, 1, 2, \ldots \end{aligned}$$

Let $\bar{x}_t = \frac{1}{t}\sum_{s=0}^{t-1} x_s$. Then, for any $x^* \in \mathcal{C}$ and any T we have

$$f(\bar{x}_T) - f(x^*) \leq \frac{D_\psi(x^*, x_0)}{\eta T} + \frac{L^2 \eta}{2\alpha}.$$

[3] In other words, f is L-Lipschitz with respect to the norm $\|\cdot\|$.

Suppose that we have the upper bound $D_\psi(x^*, x_0) \leq R^2$ for some constant $R > 0$. Deduce that with the step size choice $\eta = \frac{\sqrt{2\alpha}}{\sqrt{T}} \frac{R}{L}$[4] we have

$$f(\bar{x}_T) - f(x^*) \leq \frac{\sqrt{2}RL}{\sqrt{\alpha T}}.$$

References

Heinz H Bauschke, Jonathan M Borwein, et al. Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67, 1997.

Arkadij Semenovič Nemirovskii and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983. URL https://www2.isye.gatech.edu/~nemirovs/Nemirovskii_Yudin_1983.pdf.

R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

[4]This choice of η depends on knowing the number of iterations in advance; see Lecture 8 for a decreasing step size choice which circumvents this limitation but pays an extra logarithmic factor in the convergence rate bound.