CSIE5410 Optimization algorithms

Lecture 5: Mirror descent & subdifferential

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

11.10.2018

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

This lecture addresses the following questions.

- Why is the mirror descent called the mirror descent?
- Why do we consider relative smoothness instead of merely the standard smoothness?
- What if the objective function is not differentiable?

Recommended reading

- A. Beck and M. Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization.
- A. Juditsky and A. Nemirovski. 2010. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods.
- R. T. Rockafellar. 1970. *Convex Analysis*. (Chapter 23).
- *A. S. Nemirovsky and D. B. Yudin. 1983. *Problem Complexity and Method Efficiency in Optimization*. (Chapter 3).

Table of contents

Arkadi Nemirovski

Arkadi Nemirovski
(1947–present)

- *Problem Complexity and Method Efficiency in Optimization* (1983).
    - Black-box approach & complexity measures.
    - Complexity lower bounds.
    - *Mirror descent*.
    - Many other gems!
- *Interior-Point Polynomial Algorithms in Convex Programming* (1994).
- Aggregation of estimates (2000).
- Mirror-prox (2004).
- *Robust Optimization* (2009).
- ...

Why is the mirror descent called the mirror descent?

Original problem set-up of Nemirovski & Yudin

Let $(E, \|\cdot\|)$ be a *Banach space*. Consider the problem

$$f^\star = \min_x \{\, f(x) \mid x \in \mathcal{X} \,\},$$

where $\mathcal{X}$ is a bounded closed convex set in E, and f is an convex L-Lipschitz continuous function on $\mathcal{X}$.

Definition. We say that f is L-Lipschitz continuous on $\mathcal{X}$ for some $L > 0$, if and only if

$$|f(y) - f(x)| \leq L\|y - x\|, \quad \forall x, y \in \mathcal{X}.$$

A. Nemirovsky and D. B. Yudin. 1983. *Problem Complexity and Method Efficiency in Optimization.*

Naïve introduction to Banach spaces (1/2)

Definition. A Banach space $(E, \|\cdot\|)$ is a vector space E (over $\mathbb{R}$ or $\mathbb{C}$) with a norm $\|\cdot\|$, on which each Cauchy sequence converges to an element of E.

Remark. There is not any inner product in a Banach space!

Remark. If there is an inner product $\langle \cdot, \cdot \rangle$, such that

$$\langle x, x \rangle = \|x\|^2, \quad \forall x \in E,$$

then $(E, \langle \cdot, \cdot \rangle)$ is called a *Hilbert space*.

Example. The space $(\mathbb{R}^p, \|\cdot\|_q)$ is a Banach space for any $q \geq 1$, and is also a Hilbert space for $q = 2$.

Naïve introduction to Banach spaces (2/2)

Definition. We say that f is (Fréchet) differentiable at $x \in E$, if there exists a linear function $Df(x)$, such that

$$f(x+h) = f(x) + Df(x)[h] + o(\|h\|).$$

Theorem. For any linear function T on $\mathbb{R}^p$, there exists a unique $v_T \in \mathbb{R}^p$, such that

$$Th = \langle v_T, h \rangle_{\ell_2}, \quad \forall h \in \mathbb{R}^p.$$

Proof. Let $\{ e_1, \ldots, e_p \}$ be the canonical basis of $\mathbb{R}^p$. Set $v_T = \sum_{i=1}^p T(e_i) e_i$.

Remark. Hence, we could write $\langle \nabla f(x), h \rangle$ instead of $Df(x)[h]$ in previous lectures.

Deviating from the ℓ_2 paradigm

Let us consider $\mathbb{R}^p$ as a Banach space with some norm $\|\cdot\|$, without any inner product.

Question. Why?

Answer. Adopting the Banach space perspective, we can get rid of the ℓ_2-norm, and hope to benefit from *significantly smaller* Lipschitz parameters and initial distances (e.g., $\|x_0 - x^\star\|$).

Observation. Then the projected gradient descent is not valid!

Naïve introduction to duality

Definition. Let $(E, \|\cdot\|)$ be a Banach space over $\mathbb{R}$ or $\mathbb{C}$. Then the dual space $(E^*, \|\cdot\|_*)$ is defined as the space of (continuous) linear functions on E, with the dual norm

$$\|\varphi\|_* := \sup_x \{\, |\varphi(x)| \mid x \in E, \|x\| = 1 \,\}\,.$$

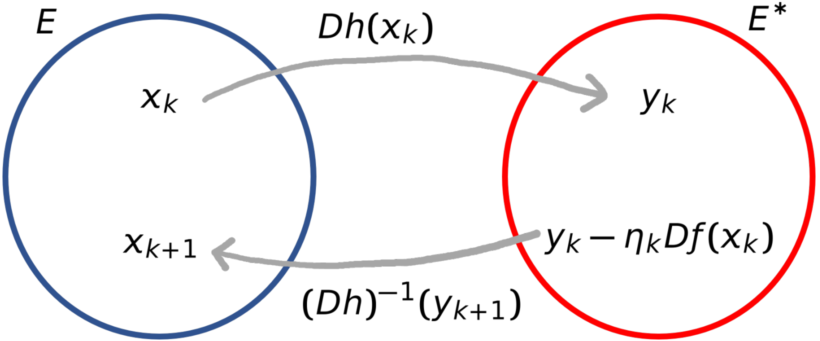
Remark. The space $(E^*, \|\cdot\|_*)$ is also Banach.

Example. Consider the Banach space $(\mathbb{R}^p, \|\cdot\|)$. Let f be a function differentiable at $x \in \mathbb{R}^p$. Then $Df(x)$ is an element in the dual space, and

$$\|Df(x)\|_* = \|\nabla f(x)\|_* := \sup_h \left\{\, |\langle \nabla f(x), h\rangle_{\ell_2}| \;\middle|\; h \in \mathbb{R}^p, \|h\| = 1 \,\right\}.$$

Illustration of the idea

Choose an *appropriate* function h.



Two versions

Algorithm Mirror Descent (ver. 1)

Set $x_0 \in \mathcal{X}$ and $T \in \mathbb{N}$.
for $t = 1, \ldots, T$ **do**
  $x_t \in \arg\min_x \{\eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_h(x, x_{t-1}) |$
  $x \in \mathcal{X}\}$
end for

Algorithm Mirror Descent (ver. 2α)

Set $x_0 \in \mathcal{X}$ and $T \in \mathbb{N}$.
for $t = 1, \ldots, T$ **do**
  $x_t \leftarrow (Dh)^{-1} (Dh(x_{t-1}) - \eta_{t-1} Df(x_{t-1}))$
end for

Question. Where is the effect of $\mathcal{X}$ in version 2α?

Mirror descent ver. 2

Algorithm Mirror Descent (ver. 2)

1: Set $x_0 \in \mathcal{X}$ and $T \in \mathbb{N}$.
2: **for** $t = 1, \dots, T$ **do**
3: $\quad \tilde{x}_t \leftarrow (Dh)^{-1} \left(Dh(x_{t-1}) - \eta_{t-1} Df(x_{t-1}) \right)$
4: $\quad x_t \leftarrow \arg\min_x \{ D_h(x, \tilde{x}_t) \mid x \in \mathcal{X} \}$ $\triangleright$ "projection"
5: **end for**

Assumption. Everything is *well-defined*. In particular, $\tilde{x}_t$ and x_t exist and are uniquely defined.

Remark. See the reference below for some sufficient conditions.

H. Bauschke *et al.* 2001. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces.

Equivalence

Theorem. The two versions are equivalent.

Proof. Notice that there is a one-to-one correspondence between Df and ∇f, and also Dh and ∇h. By definition, we write

$$\nabla h(x_{t-1}) - \eta_{t-1} \nabla f(x_{t-1}) = \nabla h(\tilde{x}_t).$$

The optimality condition of x_t says that

$$\langle \nabla h(x_t) - \nabla h(\tilde{x}_t), x - x_t \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Combining the two, by the optimality condition, we obtain that

$$x_t \in \arg\min_x \left\{ \eta_{t-1} \left\langle \nabla f(x_{t-1}), x - x_{t-1} \right\rangle + D_h(x, x_{t-1}) \mid x \in \mathcal{X} \right\}.$$

A. Beck and M. Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization.

Discussion: Understanding the # mapping

Recall from the last lecture:

Proposition For any $y \in \mathbb{R}^p$, define $y^{\#}$ to be any vector satisfying

$$y^{\#} \in \arg\max_{u} \left\{ \langle y, u \rangle - \frac{1}{2}\|u\|^2 \;\middle|\; u \in \mathbb{R}^p \right\}.$$

Then we have

$$f(x - \frac{1}{L}\,[\nabla f(x)]^{\#}) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_*^2, \quad \forall x \in \mathbb{R}^p.$$

Remark. The # mapping maps an element in the dual space to one in the primal space; moreover, $y^{\#} = y$ when the norm is ℓ_2.

Closely related notion: Duality map

Definition. Let $(E, \|\cdot\|)$ be a Banach space. The duality map is defined as

$$J(x) = \left\{ \varphi \in E^* \;\middle|\; \|\varphi\|_* = \|x\|, \varphi(x) = \|x\|^2 \right\}.$$

Proposition. The duality map is non-empty, closed, and convex.

Proposition. The set of all possible $y^{\#}$'s is equal to $J(y)$.

Proof. Exercise.

H. Brezis. 2011. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.*

Example: Entropic mirror descent

Issue of implementing the mirror descent

While there is a great flexibility in choosing the function h, there are very few *practically useful* choices.

Question. Why?

Answer. The *optimization sub-problem*

$$x_t \in \arg\min_x \{ \eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_h(x, x_{t-1}) \mid x \in \mathcal{X} \}$$

should be easily solvable, which is in general difficult to achieve.

Remark. The entropic mirror descent is perhaps the most successful instance.

Problem set-up

Consider the optimization problem

$$f^\star = \min_x \{ f(x) \mid x \in \Delta \},$$

where f is a differentiable convex function, and Δ denotes the probability simplex

$$\Delta := \{ x \in \mathbb{R}^p \mid x \geq 0, \|x\|_1 = 1 \}.$$

Remark. Each element in Δ defines a probability distribution.

Entropy & relative entropy (1/2)

Let $x \coloneqq (x_1, \ldots, x_p)$ and $y \coloneqq (y_1, \ldots, y_p)$ in $\Delta \subset \mathbb{R}^p$.

Definition. The entropy function is given by

$$S(x) \coloneqq -\sum_{i=1}^{p} x_i \log x_i - \sum_{i=1}^{p} x_i.$$

Definition. The relative entropy is given by

$$D_S(x, y) \coloneqq \begin{cases} +\infty, & \text{if } y_i = 0 \text{ while } x_i \neq 0 \text{ for some } i, \\ \sum_{i=1}^{p} x_i \log \frac{x_i}{y_i}, & \text{otherwise.} \end{cases}$$

The convention $0 \log 0 \coloneqq 0$ applies.

Entropy & relative entropy (2/2)

Proposition. The entropy function is concave.

Proof. A direct calculation gives

$$\nabla^2 S(x) = \mathrm{diag}\left(-\frac{1}{x_1}, \cdots, -\frac{1}{x_p}\right) \leq 0.$$

Proposition. The relative entropy D_S is the Bregman divergence induced by the negative entropy $-S$.

Proof. Exercise.

Entropic mirror descent

Proposition. For any $x_{t-1} \in \Delta$, $x_{t-1} > 0$, define

$$x_t \in \arg\min_x \left\{ \eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_S(x, x_{t-1}) \mid x \in \Delta \right\}.$$

Then x_t is unique and given by

$$x_t = \frac{x_{t-1} \circ \exp\left(-\eta_{t-1} \nabla f(x_{t-1})\right)}{c_{t-1}},$$

where $\circ$ denotes element-wise multiplication, $\exp$ denotes element-wise exponential, and c_{t-1} normalizes $\|x_t\|_1$. Moreover, $x_t > 0$.

Proof. Apply the optimality condition.

Comparison to projected gradient descent

Projected gradient descent

$$x_t \leftarrow \mathrm{proj}_\Delta \left(x_{t-1} - \eta_{t-1} \nabla f(x_{t-1})\right).$$

Entropic mirror descent

$$x_t \leftarrow \frac{x_{t-1} \circ \exp\left(-\eta_{t-1} \nabla f(x_{t-1})\right)}{c_{t-1}}$$

Remark. Much lower per-iteration computational complexity ($O(p)$) in comparison to the projected gradient descent ($O(p^2)$, see the reference below).

L. Condat. 2016. Fast projection onto the simplex and the ℓ_1 ball.

Smoothness relative to the negative entropy

Theorem. (Pinsker's inequality) It holds that

$$D_S(x, y) \geq \frac{1}{2}\|x - y\|_1^2, \quad \forall x, y \in \Delta.$$

Proposition. If f is L-smooth with respect to the ℓ_1-norm on Δ, i.e.,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_1^2, \quad \forall x, y \in \Delta,$$

then f is L-smooth relative to the negative entropy.

I. Csiszár and J. Körner. 2011. *Information Theory: Coding Theorems for Discrete Memoryless Systems*.

An example of true relative smoothness

Positron emission tomography

Recall that positron emission tomography (PET) corresponds to solving the problem

$$x^\star \in \arg\min_x \{ f(x) \mid x \in \Delta \},$$

where

$$f(x) := \sum_{i=1}^n \langle a_i, x\rangle - y_i \log \langle a_i, x\rangle,$$

for some $a_1, \ldots, a_n \in \mathbb{R}^p_{++}$ and $y_1, \ldots, y_n \in \mathbb{N}$.

Proposition. The function f is not smooth, nor smooth relative to the negative entropy.

Y.-H. Li and V. Cevher. 2017. Convergence of the exponentiated gradient method with Armijo line search.

Detour: Equivalent definition of relative smoothness

Definition. We say that a function f is L-smooth relative to a differentiable convex function h on a set $\mathcal{X}$, if and only if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x), \quad \forall x, y \in \mathcal{X}.$$

Theorem. A function f is L-smooth relative to a differentiable convex function h on a set $\mathcal{X}$, if and only if $Lh - f$ is convex on $\mathcal{X}$.

Proof. Plug in the definition of $D_h(y, x)$ in the inequality above.

H. Bauschke *et al.* 2017. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications.

H. Lu *et al.* 2018. Relatively smooth convex optimization by first-order methods, and applications.

Proof of non-smoothness

Proof. A direct calculation gives

$$\nabla^2 f(x) = \sum_{i=1}^{n} \frac{y_i}{\langle a_i, x \rangle^2} a_i \otimes a_i,$$

which is unbounded as $\langle a_i, x \rangle$ can be arbitrarily close to zero.

Consider the special case

$$f(x_1, x_2) := (x_1 + x_2) - \log x_1 - \log x_2, \quad \forall (x_1, x_2) \in \Delta \subset \mathbb{R}^2.$$

Then

$$\nabla^2(-LS - f)(x_1, x_2) = \operatorname{diag}\left(\frac{L}{x_1} - \frac{1}{x_1^2}, \frac{L}{x_2} - \frac{1}{x_2^2} \right),$$

which cannot be positive semi-definite for any fixed $L > 0$.

"True" relative smoothness

Theorem. The function f (for PET) is L-smooth relative to the Burg entropy

$$h(x) := -\sum_{i=1}^{p} \log x_i, \quad \forall x \in \mathbb{R}^p, x \geq 0,$$

for $L := \sum_{i=1}^{n} y_i$, where x_i denotes the i-th element of x.

H. Bauschke *et al.* 2017. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications.

Proof of relative smoothness

Proof. By convexity of the function $u \mapsto u^2$, we have for every $v \in \mathbb{R}^p$ and $i \in \{ 1, \ldots, n \}$,

$$\frac{\langle a_i, v \rangle^2}{\langle a_i, x \rangle^2} = \frac{1}{\langle a_i, x \rangle^2} \left(\sum_{j=1}^{p} (a_i)_j x_j \frac{v_j}{x_j} \right)^2$$

$$\leq \frac{1}{\langle a_i, x \rangle} \sum_{j=1}^{p} (a_i)_j x_j \left(\frac{v_j}{x_j} \right)^2 \leq \sum_{j=1}^{p} \left(\frac{v_j}{x_j} \right)^2 .$$

Therefore, for any $v \in \mathbb{R}^p$, we have

$$\langle v, \nabla^2 (Lh - f)(x) v \rangle = \sum_{i=1}^{n} y_i \sum_{j=1}^{p} \frac{v_j^2}{x_j^2} - \sum_{i=1}^{n} \frac{y_i \langle a_i, v \rangle^2}{\langle a_i, x \rangle^2} \geq 0.$$

H. Bauschke *et al.* 2017. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications.

Implementation issue

Question. Choose h as the Burg entropy. How does one compute

$$x_t \in \arg\min_x \{\, \eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_h(x, x_{t-1}) \mid x \in \Delta \,\}?$$

Exercise. There is not any closed-form solution, but one does not need to solve the whole optimization problem.

Subgradient & subdifferential

Back to the original problem set-up of Nemirovski & Yudin

Consider the Banach space $(\mathbb{R}^p, \|\cdot\|)$. Consider the problem

$$f^\star = \min_x \{ f(x) \mid x \in \mathcal{X} \},$$

where $\mathcal{X}$ is a bounded closed convex set in $\mathbb{R}^p$, and f is an convex L-Lipschitz continuous function on $\mathcal{X}$, *possibly non-differentiable*.

Question. How does one set up the mirror descent without a gradient?

Subgradient & subdifferential

Definition. Let $f : \mathcal{X} \subseteq \mathbb{R}^p \to]-\infty, +\infty]$ be convex. We say that a vector $g_x \in \mathbb{R}^p$ is a *subgradient* of f at a point $x \in \mathcal{X}$, if and only if

$$f(y) \geq f(x) + \langle g_x, y - x \rangle, \quad \forall y \in \mathcal{X}.$$

The set of all such g_x's is called the *subdifferential* of f at x, and is written as $\partial f(x)$.

Notation. We also write a subgradient in $\partial f(x)$ as $\nabla f(x)$.

R. T. Rockafellar. 1970. *Convex Analysis.*

Properties

Proposition. The set $\partial f(x)$ is convex.

Proof. Exercise.

Proposition. In general, $\partial f(x)$ can be empty.

Proof. For example, set $f(x) := -\sqrt{x}$; then $\partial f(x)$ is empty at $x = 0$. Notice that f is convex.

Theorem. Let f be a convex function differentiable at $x \in \mathbb{R}^p$. Then $\partial f(x) = \{ \nabla f(x) \}$.

Proof for the subdifferential of a differentiable function

Proof. It holds that $\nabla f(x) \in \partial f(x)$, as by convexity,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle , \quad \forall y \in \mathbb{R}^p.$$

Suppose that there exists some $g \neq \nabla f(x)$ in $\partial f(x)$. Then we have

$$f(x + \lambda y) \geq f(x) + \lambda \langle g, y \rangle , \quad \forall \lambda > 0, y \in \mathbb{R}^p.$$

Letting $\lambda \downarrow 0$, we get

$$\langle \nabla f(x), y \rangle \geq \langle g, y \rangle , \quad \forall y \in \mathbb{R}^p,$$

a contradiction.

Examples

Let $f : \mathbb{R} \to \mathbb{R}$.

Example. Define $f(x) := x^2$. Then $\partial f(x) = \{\, 2x \,\}$ for every $x \in \mathbb{R}$.

Example. Define $f(x) := |x|$. Then $\partial f(x) = \{\, \operatorname{sign} x \,\}$ if $x \neq 0$, and $\partial f(x) = [-1, 1]$ otherwise.

Let $f : \mathbb{R}^p \to [-\infty, +\infty]$.

Example. Define $f(x) := \chi_{\mathcal{X}}(x)$ for some closed convex set $\mathcal{X} \subseteq \mathbb{R}^p$. Then

$$\partial f(x) = \{\, g \mid g \in \mathbb{R}^p, \langle g, y - x \rangle \leq 0 \ \forall y \in \mathcal{X} \,\},$$

for every $x \in \mathcal{X}$. The set is also called the *normal cone of $\mathcal{X}$ at x*.

Terminologies

Let $f : \mathbb{R}^p \to [-\infty, +\infty]$ be a convex function.

Definition. The domain of f is given by

$$\operatorname{dom} f \coloneqq \{\, x \mid x \in \mathbb{R}^p, f(x) < +\infty \,\} .$$

Definition. We say that f is proper, if and only if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^p$.

Definition. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be convex. The interior and relative interior of $\mathcal{X}$ are given by, respectively,

$$\begin{aligned} \operatorname{int} \mathcal{X} &\coloneqq \{\, x \in \mathcal{X} \mid \exists \varepsilon > 0 : x + B_\varepsilon \subset \mathcal{X} \,\} , \\ \operatorname{ri} \mathcal{X} &\coloneqq \{\, x \in \operatorname{aff} \mathcal{X} \mid \exists \varepsilon > 0 : (x + B_\varepsilon) \cap \operatorname{aff} \mathcal{X} \subset \mathcal{X} \,\} , \end{aligned}$$

where B_ε denotes the unit ℓ_2-norm ball of radius ε, and $\operatorname{aff} \mathcal{X}$ denotes the affine hull of $\mathcal{X}$.

Subdifferential calculus

Theorem. Let f be a proper convex function.

- If $x \notin \operatorname{dom} f$, then $\partial f(x) = \emptyset$.
- If $x \in \operatorname{ri}(\operatorname{dom} f)$, then $\partial f(x)$ is non-empty.

Theorem. Let f_1 and f_2 be proper convex functions on $\mathbb{R}^p$. If $\operatorname{ri}(\operatorname{dom} f_1) \cap \operatorname{ri}(\operatorname{dom} f_2) \neq \emptyset$, then

$$\begin{aligned}\partial(f_1 + f_2)(x) &= \partial f_1(x) + \partial f_2(x) \\ &:= \{\, g_1 + g_2 \mid g_1 \in \partial f_1(x), g_2 \in \partial f_2(x) \,\}.\end{aligned}$$

R. T. Rockafellar. 1970. *Convex Analysis*.

Fermat's rule

Theorem. Let $f : \mathcal{X} \to]-\infty, +\infty]$ be convex. Then $x^\star$ is a minimizer of f, if and only if

$$0 \in \partial f(x^\star).$$

Proof. We have $0 \in \partial f(x^\star)$, if and only if

$$f(y) \geq f(x^\star) + \langle 0, y - x^\star \rangle = f(x^\star), \quad \forall y \in \mathcal{X}.$$

Remark. Therefore, the problem of minimizing a convex function f is equivalent to the inclusion problem:

$$0 \in \partial f(x).$$

Mirror descent

Closedness

Definition. Let $f : \mathbb{R}^p \to [-\infty, +\infty]$ be convex. We say that f is closed, if and only if $\operatorname{epi} f$ is closed.

Remark. Without closedness, a minimizer of f may not exist.

Example. Consider the function

$$f(x) := \begin{cases} +\infty, & x \leq 0, \\ x^2, & x > 0. \end{cases}$$

There does not exist a minimizer of f on $\mathbb{R}$.

Problem set-up

Consider the Banach space $(\mathbb{R}^p, \|\cdot\|)$. Consider the problem

$$f^\star = \min_x \{\, f(x) \mid x \in \mathcal{X} \,\},$$

where $\mathcal{X}$ is a bounded closed convex set in $\mathbb{R}^p$, and f is a proper closed convex L-Lipschitz continuous function on $\mathcal{X}$.

Recall the definition.

Definition. We say that f is L-Lipschitz continuous function on $\mathcal{X}$, if and only if

$$|f(y) - f(x)| \leq L\|y - x\|, \quad \forall x, y \in \mathcal{X}.$$

Mirror descent

Algorithm Mirror Descent

Set $x_1 \in \mathcal{X}$ and $T \in \mathbb{N}$.
for $t = 2, \dots, T$ **do**
    $x_t \in \arg\min_x \{\eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_h(x, x_{t-1}) \mid x \in \mathcal{X}\}$
end for

Theorem. Suppose that the function h is 1-strongly convex with respect to $\| \cdot \|$ on $\mathcal{X}$, i.e., for all $x, y \in \mathcal{X}$,

$$D_h(y, x) \coloneqq h(y) - (h(x) + \langle \nabla h(x), y - x \rangle) \geq \frac{1}{2} \|y - x\|^2.$$

Then it holds that for all $x \in \mathcal{X}$,

$$\min \{ f(x_1), \dots, f(x_t) \} - f(x) \leq \frac{D_h(x, x_1) + \frac{L^2}{2} \sum_{\tau=1}^{t} \eta_\tau^2}{\sum_{\tau=1}^{t} \eta_\tau}.$$

Proof of the convergence guarantee (1/3)

Lemma. It holds that $\|\nabla f(x)\|_* \le L$ for all $x \in \operatorname{int} \mathcal{X}$.

Proof. For any x, y such that $x \in \mathcal{X}$ and $x + y \in \mathcal{X}$, it holds that

$$f(x) + L\|y\| \ge f(x+y) \ge f(x) + \langle \nabla f(x), y \rangle .$$

Therefore,

$$\|\nabla f(x)\|_* := \sup_y \left\{ \frac{\langle \nabla f(x), y \rangle}{\|y\|} \;\middle|\; y \in \mathbb{R}^p \right\} \le L.$$

Lemma. It holds that, for all $x \in \mathcal{X}$,

$$\langle \nabla f(x_t), x_t - x \rangle$$
$$\le \langle \nabla f(x_t), x_t - x_{t+1} \rangle + \frac{1}{\eta_t} \left(D_h(x, x_t) - D_h(x, x_{t+1}) - D(x_{t+1}, x_t) \right) .$$

Proof. The key lemma in the last lecture.

Proof of the convergence guarantee (2/3)

Proof. For any $x \in \mathcal{X}$, we write

$$\begin{aligned}\eta_t \left(f(x_t) - f(x)\right) &\le \eta_t \left\langle \nabla f(x_t), x_t - x \right\rangle \\ &\le \left(D_h(x, x_t) - D_h(x, x_{t+1})\right) + \\ &\quad\ \eta_t \left\langle \nabla f(x_t), x_t - x_{t+1} \right\rangle - D_h(x_{t+1}, x_t).\end{aligned}$$

By the strong convexity of h, we have

$$\begin{aligned}&\eta_t \left\langle \nabla f(x_t), x_t - x_{t+1} \right\rangle - D_h(x_{t+1}, x_t) \\ &\quad \le \eta_t \left\langle \nabla f(x_t), x_t - x_{t+1} \right\rangle - \frac{1}{2}\|x_t - x_{t+1}\|^2 \\ &\quad \le \eta_t \|\nabla f(x_t)\|_* \|x_t - x_{t+1}\| - \frac{1}{2}\|x_t - x_{t+1}\|^2 \\ &\quad \le \frac{\eta_t^2}{2}\|\nabla f(x_t)\|_*^2.\end{aligned}$$

Proof of the convergence guarantee (3/3)

Proof continued. Then we obtain

$$\eta_t \left(f(x_t) - f(x) \right) \leq \frac{\eta_t^2}{2} \|\nabla f(x_t)\|_*^2 + \left(D_h(x, x_t) - D_h(x, x_{t+1}) \right).$$

Summing over all t, we get

$$\sum_{\tau=1}^{t} \eta_\tau \left(f(x_\tau) - f(x) \right) \leq D_h(x, x_1) + \frac{L^2}{2} \sum_{\tau=1}^{t} \eta_\tau^2.$$

Therefore, we get

$$\min \{ f(x_1), \ldots, f(x_t) \} - f(x) \leq \frac{D_h(x, x_1) + \frac{L^2}{2} \sum_{\tau=1}^{t} \eta_\tau^2}{\sum_{\tau=1}^{t} \eta_\tau}.$$

Standard form

Define $R := \max_{x,y} \{\, D_h(x,y) \mid x, y \in \mathcal{X} \,\}$.

Exercise. Check that R is well-defined.

Corollary. Fix $T \in \mathbb{N}$. Set $\eta_t = \frac{\sqrt{2R}}{L\sqrt{T}}$. Then it holds that

$$\min \{\, f(x_1), \ldots, f(x_T) \,\} - f^\star \leq \frac{L\sqrt{2R}}{\sqrt{T}} = O\left(\frac{L\sqrt{R}}{\sqrt{T}}\right).$$

When is a non-ℓ_2-norm preferred? (1/4)

Consider the problem of minimizing a proper closed convex function on the probability simplex $\Delta \subset \mathbb{R}^p$.

Suppose that the function is L_1-Lipshitz w.r.t the ℓ_1-norm, and L_2-Lipschitz w.r.t. the ℓ_2-norm on Δ.

Projected subgradient method: $\mathrm{error}_2 = O\left(\frac{L_2 1}{\sqrt{T}}\right)$.

Entropic mirror descent: $\mathrm{error}_1 = O\left(\frac{L_1\sqrt{\log p}}{\sqrt{T}}\right)$.

Question. Which one is better?

A. Ben-Tal *et al.* 2001. The ordered subsets mirror descent optimization method with applications to tomography.

A. Beck and M. Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization.

When is a non-ℓ_2-norm preferred? (2/4)

Proposition. Choose h to be the negative entropy, and $x_1 = (1/p, \ldots, 1/p)$. Then

$$D_h(x, x_1) \leq \log p.$$

Proof. We write

$$\begin{aligned} D_h(x, x_1) &= \sum_{i=1}^{p} x_i \log \frac{x_i}{(x_1)_i} \\ &= \sum_{i=1}^{p} x_i \log x_i + \log p \leq \log p. \end{aligned}$$

A. Beck and M. Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization.

When is a non-ℓ_2-norm preferred? (3/4)

Notice that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{p}\|x\|_2$. Then we have

$$\frac{1}{\log p} \leq \frac{\mathrm{err}_2}{\mathrm{err}_1} = \frac{L_2}{L_1 \log p} \leq \frac{\sqrt{p}}{\log p}.$$

Observation. Choosing the entropic mirror descent can have a significant gain or an "negligible" loss.

A. Beck and M. Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization.

When is a non-ℓ_2-norm preferred? (4/4)

Observation. The gain is significant when, for example, the objective function is given by $f(x) := \langle a, x \rangle$, where the entries of the vector a have similar absolute values.

Proof. Notice that

$$
\begin{aligned}
|\, \langle a, x \rangle - \langle a, y \rangle \,| &\leq \|a\|_2 \|x - y\|_2, \\
|\, \langle a, x \rangle - \langle a, y \rangle \,| &\leq \|a\|_\infty \|x - y\|_1.
\end{aligned}
$$

Therefore, we obtain that $L_2 = \|a\|_2$ and $L_1 = \|a\|_\infty$. It remains to notice that $\|a\|_2 \approx \sqrt{p}\|a\|_\infty$, when the entries of a have similar absolute values.

Conclusions

Summary (1/4)

Algorithm Mirror Descent (ver. 1)

Set $x_0 \in \mathcal{X}$ and $T \in \mathbb{N}$.
for $t = 1, \dots, T$ **do**
  $x_t \in \arg\min_x \{\eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_h(x, x_{t-1}) | x \in \mathcal{X}\}$
end for

Algorithm Mirror Descent (ver. 2)

Set $x_0 \in \mathcal{X}$ and $T \in \mathbb{N}$.
for $t = 1, \dots, T$ **do**
  $\tilde{x}_t \leftarrow (Dh)^{-1} (Dh(x_{t-1}) - \eta_{t-1} Df(x_{t-1}))$
  $x_t \leftarrow \arg\min_x \{ D_h(x, \tilde{x}_t) \mid x \in \mathcal{X} \}$ ▷ "projection"
end for

Summary (2/4)

Subdifferential & subgradient The subdifferential of a proper convex function $f : \mathbb{R}^p \to]-\infty, +\infty]$ at x is given by

$$\partial f(x) \coloneqq \{\, g \in \mathbb{R}^p \mid f(y) \geq f(x) + \langle g, y - x \rangle \, \forall x, y \in \mathbb{R}^p \,\}\,.$$

An element of $\partial f(x)$ is called a subgradient of f at x, and is denoted by $\nabla f(x)$.

Theorem. Let f_1 and f_2 be proper convex functions, satisfying that $\mathrm{ri}(\mathrm{dom}\, f_1) \cap \mathrm{ri}(\mathrm{dom}\, f_2) \neq \emptyset$. Then

$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x).$$

Summary (3/4)

Algorithm Mirror Descent

Set $x_1 \in \mathcal{X}$ and $T \in \mathbb{N}$.
for $t = 2, \ldots, T$ **do**
    $x_t \in \arg\min_x \{\eta_{t-1} \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + D_h(x, x_{t-1}) |$
        $x \in \mathcal{X}\}$
end for

Theorem. Suppose that f is L-Lipschitz on $\mathcal{X}$. Then the mirror descent achieves

$$\min \{ f(x_1), \ldots, f(x_t) \} - f^\star = O\left(\frac{L\sqrt{R}}{\sqrt{t}} \right).$$

Summary (4/4)

Entropic mirror descent. Minimize an L-Lipschitz convex function f on the probability simplex $\Delta \subset \mathbb{R}^p$, by the iteration

$$x_t = \frac{x_{t-1} \circ \exp\left(-\eta_{t-1} \nabla f(x_{t-1})\right)}{c_{t-1}}, \quad t = 2, 3, \ldots, T.$$

Set $x_1 = (1/p, \ldots, 1/p)$ and $\eta_t = \frac{\sqrt{2 \log p}}{L\sqrt{T}}$. The convergence rate is

$$\min\{\, f(x_1), \ldots, f(x_T) \,\} - f^\star = O\left(\frac{L\sqrt{\log p}}{\sqrt{T}}\right),$$

which is *almost dimension-independent*.

Question. What is the norm?

Next lecture

- Composite minimization.
- Proximal gradient method.