# CSIE5410 Optimization algorithms

Lecture 7: Convergence of the proximal gradient method

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

08.11.2018

Department of Computer Science and Information Engineering
National Taiwan University

## Abstract

We have seen the proximal gradient method. This lecture addresses the following questions.

- Under what conditions does the algorithm provably converge?
- How fast does it converge?

# Recommended reading

- A. Beck and M. Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
- Yu. Nesterov and A. Nemirovski. 2013. On first-order algorithms for $\ell_1$/nuclear norm minimization.
- H. Lu *et al.* 2018. Relatively smooth convex optimization by first-order methods, and applications.
- M. Teboulle. 2018. A simplified view of first order methods for optimization.

# Table of contents

# Smooth case

# Recap: Proximal gradient method

Consider the optimization problem

$$\varphi^\star = \min_x \{ f(x) + g(x) \mid x \in \mathbb{R}^p \},$$

for some $L$-smooth convex function $f$ and proper closed convex function $g$.

**Proximal gradient method**

$$\begin{aligned} x_{t+1} &\leftarrow (I + \eta_t \partial g)^{-1}(I - \eta_t \nabla f)x_t \\ &= \operatorname{prox}_{\eta_t g}(x_t - \eta_t \nabla f(x_t)) \end{aligned}$$

## Equivalent formulation

$$x_{t+1} \leftarrow \text{prox}_{\eta_t g}\left(x_t - \eta_t \nabla f(x_t)\right).$$

**Proposition.** Equivalently, we write

$$x_{t+1} = \arg\min_x \left\{ \langle \nabla f(x_t), x - x_t \rangle + g(x) + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \;\middle|\; x \in \mathbb{R}^p \right\}.$$

*Proof.* Plug in the definition of a proximal mapping.

**Remark.** Recall that when $g$ is the indicator function of a closed convex set $\mathcal{X}$, the iteration rule of the projected gradient method becomes a special case.

# Proximal set-up (1/3)

Consider the problem

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

for some convex function $f$ that is $L$-smooth on $\mathcal{X}$, proper closed convex function $g$, and closed convex set $\mathcal{X} \subseteq \mathbb{R}^p$.

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

# Proximal set-up (2/3)

**Definition.** We say that a function $h : \mathcal{X} \to \mathbb{R}$ is a *distance generating function (DGF)*, if and only if the following are satisfied.

- The function $h$ is continuous and convex on $\mathcal{X}$.
- There exists a function $h'$ that is continuous on $\mathcal{X}^\circ := \mathcal{X} \cap \mathrm{dom}(\partial h)$ and satisfies

$$h'(x) \in \partial h(x), \quad \forall x \in \mathcal{X}^\circ.$$

- The function $h$ is $1$-strongly convex on $\mathcal{X}^\circ$, i.e.,

$$\langle h'(x) - h'(y), x - y \rangle \geq \|x - y\|^2, \quad \forall x, y \in \mathcal{X}^\circ.$$

# Proximal set-up (3/3)

**Definition.** The *Bregman divergence* is given by

$$D_h(y,x) \coloneqq h(y) - h(x) - \langle h'(x), y - x \rangle \,, \quad \forall x \in \mathcal{X}^\circ, y \in \mathcal{X}.$$

**Definition.** The *(composite) prox-mapping* is given by

$$T_L(x) \coloneqq \operatorname*{arg\,min}_{y \in \mathcal{X}} \{ \, \langle \nabla f(x), y - x \rangle + g(y) + L D_h(y,x) \mid y \in \mathcal{X} \, \} \,.$$

**Bregman proximal gradient method.**

$$x_t \leftarrow T_L(x_{t-1}).$$

# Interpretation of the Bregman proximal gradient method (1/3)

Recall the derivation of the proximal point method:

$$\begin{aligned} 0 \in \partial f(x^\star) &\Leftrightarrow x^\star \in (I + \partial f)x^\star \\ &\Leftrightarrow x^\star = (I + \partial f)^{-1} x^\star \end{aligned}$$

How about the following derivation? Let $h$ be a differentiable $\mu$-strongly convex function (w.r.t. the $2$-norm).

$$\begin{aligned} 0 \in \partial f(x^\star) &\Leftrightarrow \nabla h(x^\star) \in (\nabla h + \partial f)x^\star \\ &\Leftrightarrow x^\star = (\nabla h + \partial f)^{-1}(\nabla h(x^\star)) \end{aligned}$$

---

J. Eckstein. 1993. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming.
H. H. Bauschke *et al.* 2003. Bregman monotone optimization algorithms.

# Interpretation of the Bregman proximal gradient method (2/3)

Consider the algorithm:

$$x_{t+1} \leftarrow (\nabla h + \partial f)^{-1}(\nabla h(x_t)).$$

Then, we write

$$\nabla h(x_t) \in (\nabla h + \partial f)x_{t+1} \Leftrightarrow 0 \in \partial f(x_{t+1}) + \nabla h(x_{t+1}) - \nabla h(x_t),$$

which is equivalent to

$$x_{t+1} \in \arg\min_x \left\{ f(x) + D_h(x, x_t) \mid x \in \mathbb{R}^p \right\}.$$

---

J. Eckstein. 1993. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming.
H. H. Bauschke *et al.* 2003. Bregman monotone optimization algorithms.

# Interpretation of the Bregman proximal gradient method (3/3)

Similarly, consider the monotone inclusion problem

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

We may write

$$\begin{aligned}(\nabla h - \nabla f)(x^\star) &\in (\nabla h + \partial g)(x^\star)\\ \Leftrightarrow x^\star &= (\nabla h + \partial g)^{-1}(\nabla h - \nabla f)(x^\star),\end{aligned}$$

which motivates the algorithm:

$$\begin{aligned}x_{t+1} &\leftarrow (\nabla h + \partial g)^{-1}(\nabla h - \nabla f)(x_t)\\ &= \arg\min_x \left\{ \langle \nabla f(x_t), x - x_t \rangle + g(x) + D_h(x, x_t) \mid x \in \mathbb{R}^p \right\}.\end{aligned}$$

## Convergence of the Bregman proximal gradient method

**Algorithm** Bregman proximal gradient method

Set $x_0 \in \mathcal{X}^\circ$.
**for** $t = 0, 1, \ldots, T$ **do**
    $x_{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \{ \langle \nabla f(x_t), x - x_t \rangle + LD(x, x_t) + g(x) \}$
**end for**

**Theorem.** Suppose that $g$ is finite on $\operatorname{ri} \mathcal{X}$. Then, for all $t \geq 1$, it holds that $x_t \in \mathcal{X}^\circ$, and

$$\varphi(x_t) - \varphi(x) \leq \frac{LD_h(x, x_0)}{t}, \quad \forall x \in \mathcal{X}.$$

Yu. Nesterov and A. Nemirovski. 2013. On first-order algorithms for $\ell_1$/nuclear norm minimization.

# Proof of the theorem (1/5)

**Lemma.** Let $\psi$ be a proper closed convex function that is finite on $\operatorname{ri}\mathcal{X}$. Define

$$x^\star_\psi \in \arg\min_x \{\, \psi(x) + h(x) \mid x \in \mathcal{X} \,\}.$$

Then, $x^\star_\psi \in \mathcal{X}^\circ$, and there exists some $\xi \in \partial\psi(x^\star_\psi)$, such that

$$\langle \xi + h'(x^\star_\psi), x - x^\star_\psi \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

*Proof.* See the reference.

---

Yu. Nesterov and A. Nemirovski. 2013. On first-order algorithms for $\ell_1$/nuclear norm minimization.

# Proof of the theorem (2/5)

**Lemma.** Suppose that $\tilde{\psi} \coloneqq \psi - Lh$ is convex. Let $x_\psi^\star \in \arg\min_{x \in \mathcal{X}} \psi(x)$. Then $x_\psi^\star \in \mathcal{X}^\circ$, and

$$\psi(x) \geq \psi(x_\psi^\star) + LD_h(x, x_\psi^\star), \quad \forall x \in \mathcal{X}.$$

*Proof.* By the previous lemma, $x_\psi^\star \in \mathcal{X}^\circ$, and there exists some $\eta \in \partial\tilde{\psi}(x_\psi^\star)$ such that

$$\langle \eta + Lh'(x_\psi^\star), x - x_\psi^\star \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

## Proof of the theorem (3/5)

*Proof (of the lemma) continued.* We write

$$\begin{aligned}\psi(x) &= \tilde{\psi}(x) + Lh(x) \\ &\geq \tilde{\psi}(x^\star_\psi) + \langle \eta, x - x^\star_\psi \rangle + Lh(x) \\ &\geq \tilde{\psi}(x^\star_\psi) - \langle Lh'(x^\star_\psi), x - x^\star_\psi \rangle + Lh(x) \\ &= \tilde{\psi}(x^\star_\psi) + Lh(x^\star_\psi) + Lh(x) - \big(Lh(x^\star_\psi) + \langle Lh'(x^\star_\psi), x - x^\star_\psi \rangle\big) \\ &= \psi(x^\star_\psi) + LD_h(x, x^\star_\psi).\end{aligned}$$

# Proof of the theorem (4/5)

*Proof of the theorem.* Define

$$m_L(x) := f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + g(x) + LD_h(x, x_t).$$

By smoothness of $f$ and strong convexity of $h$,

$$m_L(x) \geq \varphi(x) := f(x) + g(x).$$

By the previous lemma, $x_{t+1} \in \mathcal{X}^\circ$, and

$$m_L(x) \geq m_L(x_{t+1}) + LD_h(x, x_{t+1}), \quad \forall x \in \mathcal{X}.$$

Then, by smoothness of $f$,

$$\begin{aligned} \varphi(x_t) = m_L(x_t) &\geq m_L(x_{t+1}) + LD_h(x_t, x_{t+1}) \\ &\geq \varphi(x_{t+1}), \end{aligned}$$

showing that the sequence $(\varphi(x_t))_{t \in \mathbb{N}}$ is non-increasing.

# Proof of the theorem (5/5)

*Proof of the theorem continued.*

Moreover, we have

$$\begin{aligned}\varphi(x) + LD_h(x, x_t) \geq m_L(x) &\geq m_L(x_{t+1}) + LD_h(x, x_{t+1}) \\ &\geq \varphi(x_{t+1}) + LD_h(x, x_{t+1}).\end{aligned}$$

Summing up the inequalities over $t$, we obtain

$$\begin{aligned}t(\varphi(x_t) - \varphi(x)) &\leq \sum_{\tau=0}^{t-1} (\varphi(x_\tau) - \varphi(x)) \\ &\leq L\left(D_h(x, x_0) - D_h(x, x_t)\right).\end{aligned}$$

# Acceleration for the smoothness case

# Estimate sequence approach (1/2)

**Estimate sequence approach.** Consider an iterative algorithm that maintains three sequences:

1. A sequence of iterates $(y_t)_{t\geq 0}$.
2. A sequence of increasing numbers $(A_t)_{t\geq 0}$ such that

$$A_0 = 0, \quad A_t := A_{t-1} + a_t, \quad \forall t \in \mathbb{N}.$$

3. A sequence of *estimate functions*

$$\psi_t := \langle \xi_k, \cdot \rangle + A_k g(x) + \frac{1}{2}\| \cdot - x_0\|^2, \quad \forall t \geq 0.$$

# Estimate sequence approach (2/2)

**Conditions.** The algorithm has to satisfy for all $t$:

1. $A_t\varphi(y_t) \leq \psi_t^\star := \min_x \psi_t(x)$.
2. $\psi_t(x) \leq A_t\varphi(x) + LD_h(x, y_0)$ for all $x \in \mathbb{R}^p$.

**Theorem.** If the conditions are satisfied, then it holds that

$$\varphi(y_t) - \varphi(x) \leq \frac{LD_h(x, y_0)}{A_t}, \quad \forall x \in \mathcal{X}.$$

---

Yu. Nesterov. 2008. Accelerating the cubic regularization of Newton's method on convex problems.

# Accelerated Bregman proximal gradient method

**Algorithm** Accelerated Bregman proximal gradient method

Set $y_0 = \arg\min_{x \in \mathcal{X}} h(x)$ and $\psi_0(x) = L D_h(x, y_0)$.
**for** $t = 0, 1, \ldots, T$ **do**
  $z_t \leftarrow \arg\min_{x \in \mathcal{X}} \psi_t(x)$.
  $\gamma_t \leftarrow \frac{2(t+2)}{(t+1)(t+4)}$.
  $x_{t+1} \leftarrow \gamma_t z_t + (1 - \gamma_t) y_t$.
  $\hat{x}_{t+1} \leftarrow T_{\frac{2L}{t+2}}(x_{t+1})$.
  $y_{t+1} \leftarrow \gamma_t \hat{x}_{t+1} + (1 - \gamma_t) y_t$.
  $\psi_{t+1} \leftarrow \psi_t + \frac{t+2}{2} \left[ f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + g(x) \right]$
**end for**

---

Yu. Nesterov. 2013. Gradient methods for minimizing composite functions.

Yu. Nesterov. 2013. On first-order algorithms for $\ell_1$/nuclear norm minimization.

# Iteration complexity

**Theorem.** It holds that for all $t$, $y_t \in \mathcal{X}$, and

$$\varphi(y_t) - \varphi^\star \leq \frac{4LD_h(x^\star, y_0)}{t(t+3)}.$$

**Remark.** The algorithm requires evaluating two prox-mappings per iteration.

**Remark.** A similar algorithm for the case where $g \equiv 0$ can be found in the first reference below.

---

Yu. Nesterov. 2005. Smooth minimization of non-smooth functions.

Yu. Nesterov. 2013. Gradient methods for minimizing composite functions.

Yu. Nesterov. 2013. On first-order algorithms for $\ell_1$/nuclear norm minimization.

# Proof of the theorem (1/5)

**Lemma.** It holds that $y_t \in \mathcal{X}$ for all $t$.

*Proof.* Notice that $\hat{x}_{t+1} \in \mathcal{X}$ as the output of a prox-mapping. The lemma follows by induction.

Define $a_{t+1} \coloneqq \frac{t+2}{2}$. Then, $A_t = \frac{t(t+3)}{4}$.

**Lemma.** It holds that

$$\psi_t(x) \leq A_t \varphi(x) + LD(x, y_0), \quad \forall x \in \mathcal{X}.$$

*Proof.* The lemma follows by induction with the observation that

$$\begin{aligned} \psi_{t+1} &= \psi_t + a_{t+1}\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + g(x)\right] \\ &\leq \psi_t + a_{t+1}\left[f(x) + g(x)\right]. \end{aligned}$$

# Proof of the theorem (2/5)

**Lemma.** It holds that

$$A_t\varphi(y_t) \le \psi_t^\star := \min_x \psi_t(x).$$

*Proof.* We prove by induction. The inequality obvious holds for $t = 0$. Suppose now that the inequality holds for some $t \ge 0$.

Since $\psi_t - Lh$ is obviously convex, we have

$$\psi_t(x) \ge \psi_t^\star + LD_h(x, z_t).$$

By the convexity of $f$, we write

$$\begin{aligned}\psi_t(x) &\ge A_t\varphi(y_t) + LD_h(x, z_t)\\ &\ge A_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_t - x_{t+1}\rangle + g(y_t)\right] + LD_h(x, z_t).\end{aligned}$$

# Proof of the theorem (3/5)

*Proof continued.* Then, by the algorithm, we obtain for every $x \in \mathcal{X}$,

$$\psi_{t+1}(x) \geq A_t \left[ f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_t - x_{t+1} \rangle + g(y_t) \right] + LD_h(x, z_t) + a_{t+1} \left[ f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + g(x) \right].$$

Also by the algorithm, we have

$$A_t(y_t - x_{t+1}) - a_{t+1}x_{t+1} = -a_{t+1}z_t.$$

Therefore, we obtain

$$\psi_{t+1}(x) \geq a_{t+1} \left[ \langle \nabla f(x_{t+1}), x - z_t \rangle + g(x) \right] + A_{t+1} f(x_{t+1}) + A_t g(y_t) + LD_h(x, z_t).$$

*Proof continued.* Then, we write

$$
\begin{aligned}
\psi^\star_{t+1} &\geq \min_{x \in \mathcal{X}} \{a_{t+1} \left[\langle \nabla f(x_{t+1}), x - z_t \rangle + g(x)\right] + \\
&\qquad\qquad A_{t+1} f(x_{t+1}) + A_t g(y_t) + L D_h(x, z_t)\} \\
&\geq a_{t+1} \left[\langle \nabla f(x_{t+1}), \hat{x}_{t+1} - z_t \rangle + g(\hat{x}_{t+1})\right] + \\
&\qquad\qquad A_{t+1} f(x_{t+1}) + A_t g(y_t) + L D_h(\hat{x}_{t+1}, z_t) \\
&\geq a_{t+1} \langle \nabla f(x_{t+1}), \hat{x}_{t+1} - z_t \rangle + A_{t+1} g(y_{t+1}) + A_{t+1} f(x_{t+1}) + \\
&\qquad\qquad \frac{L}{2} \|\hat{x}_{t+1} - z_t\|^2,
\end{aligned}
$$

where we have exploited the definition of $\hat{x}_{t+1}$, strong convexity of $h$, convexity of $g$, and the definition of $y_{t+1}$.

# Proof of the theorem (5/5)

*Proof continued.* Notice that

$$\hat{x}_{t+1} - z_t = \frac{y_{t+1} - x_{t+1}}{\gamma_t} = \frac{A_{t+1}}{a_{t+1}} \left( y_{t+1} - x_{t+1} \right).$$

Then, we obtain

$$\begin{aligned} \psi_{t+1}^{\star} &\geq A_{t+1} \left\langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1} \right\rangle + A_{t+1} g(y_{t+1}) + \\ &\qquad A_{t+1} f(x_{t+1}) + A_{t+1} \frac{L}{2} \left\| y_{t+1} - x_{t+1} \right\|^2 \\ &\geq A_{t+1} \left[ f(y_{t+1}) + g(y_{t+1}) \right]. \end{aligned}$$

The lemma follows.

*Proof of the theorem.* Directly apply the estimation sequence result.

# FISTA

**Remark.** Arguably, FISTA looks more natural, but its proof is even less principled.

**Assumption.** Set $\mathcal{X} = \mathbb{R}^p$. The function $f$ is $L$-smooth (w.r.t the $2$-norm) on $\mathbb{R}^p$.

---

**Algorithm** Fast iterative shrinkage thresholding algorithm (FISTA)

---

Set $y_1 = x_0 \in \mathbb{R}^p$, $\gamma_1 = 1$, and $\eta = 1/L$.
**for** $t = 1, \ldots, T$ **do**
  $x_t \leftarrow (I + \eta \partial g)^{-1}(I - \eta \nabla f) y_t$.
  $\gamma_{t+1} \leftarrow \frac{1+\sqrt{1+4\gamma_t^2}}{2}$.
  $y_{t+1} = x_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(x_t - x_{t-1})$.
**end for**

---

A. Beck and M. Teboulle. 2009. A fast iterative shrinkage thresholding algorithm for linear inverse problems.

# Relatively smooth case

# Problem formulation

Consider the optimization problem

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

where $f$ is convex and $L$-smooth relative to a proper closed convex function $h$ on $\operatorname{int}(\operatorname{dom} h)$, $g$ is a proper closed convex function, and $\mathcal{X}$ is a closed convex set.

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

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

# Another prox set-up (1/3)

**Definition.** We say that $h$ is *essentially smooth*, if and only if $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$, $h$ is differentiable on $\operatorname{int}(\operatorname{dom} h)$, and $\|\nabla h(x_k)\| \to +\infty$ for any sequence $(x_k)_{k\in\mathbb{N}}$ in $\operatorname{int}(\operatorname{dom} h)$ converging to a point on the boundary of $\operatorname{dom} h$.

**Theorem.** The function $h$ is essentially smooth, if and only if $\operatorname{dom} \partial h = \operatorname{int}(\operatorname{dom} h) \neq \emptyset$.

**Definition.** We say that $h$ is *strictly convex*, if and only if

$$h(\alpha x + (1-\alpha)y) < \alpha h(x) + (1-\alpha)h(y), \quad \forall \alpha \in ]0,1[, x, y.$$

**Definition.** We say that $h$ is *Legendre*, if and only if it is essentially smooth and strictly convex.

# Another prox set-up (2/3)

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

**Assumptions.**

1. $h$ is Legendre with $\mathrm{cl}(\mathrm{dom}\, h) = \mathcal{X}$.
2. $\mathrm{dom}\, f \supset \mathrm{dom}\, h$ and $f$ is differentiable on $\mathrm{int}(\mathrm{dom}\, h)$.
3. $\mathrm{dom}\, g \cap \mathrm{int}(\mathrm{dom}\, h) \neq \emptyset$.

**Question.** How do we specialize the set-up for entropic mirror descent?

# Another prox set-up (3/3)

**Theorem.** Let $x \in \operatorname{int}(\operatorname{dom} h)$. Let $\psi$ be a proper closed convex function. Consider the optimization problem

$$x_+ = \arg\min_y \{\, \psi(y) + D_h(y, x) \mid y \in \mathcal{X} \,\}\,.$$

1. If $\psi(x) > -\infty$ on $\mathcal{X}$, then $x_+$ exists and is unique.
2. If in addition, $\operatorname{ri}(\operatorname{dom} \psi \cap \mathcal{X}) \subset \operatorname{int}(\operatorname{dom} h)$, then $x^+ \in \operatorname{dom} \psi \cap \operatorname{int}(\operatorname{dom} h)$, and

$$0 \in \partial\psi(x_+) + \nabla h(x_+) - \nabla h(x).$$

---

M. Teboulle. 2018. A simplified view of first order methods for optimization.

## Bregman proximal inequality

This is indeed the "key lemma" in Lecture 4.

**Theorem.** Let $x \in \operatorname{int}(\operatorname{dom} h)$. Let $\psi$ be proper closed convex. Define

$$x_+ = \arg\min_y \{ \psi(y) + D_h(y, x) \mid y \in \mathcal{X} \}.$$

Then $x_+ \in \operatorname{dom} \psi \cap \operatorname{int}(\operatorname{dom} h)$, and

$$\psi(x_+) - \psi(u) \leq D_h(u, x) - D_h(u, x_+) - D_h(x_+, x), \quad \forall u \in \operatorname{dom} h.$$

# Convergence guarantee

**Algorithm** Bregman proximal gradient method

```
Set $x_0 \in \mathcal{X}^\circ$.
for $t = 0, 1, \ldots, T$ do
    $x_{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \{ \langle \nabla f(x_t), x - x_t \rangle + LD(x, x_t) + g(x) \}$
end for
```

**Theorem.** Define $\varphi := f + g$. For all $t \geq 1$, it holds that $x_t \in \operatorname{dom} g \cap \operatorname{int}(\operatorname{dom} h)$, and

$$\varphi(x_t) - \varphi(u) \leq \frac{LD_h(u, x_0)}{t}, \quad \forall u \in \operatorname{dom} h.$$

---

M. Teboulle. 2018. A simplified view of first order methods for optimization.

# Proof of the theorem (1/2)

*Proof.* We write, by $L$-relative smoothness,

$$f(x_t) \leq f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1} \rangle + LD_h(x_t, x_{t-1}).$$

By the Bregman proximal inequality, we write, for any $x \in \operatorname{dom} h$,

$$\begin{aligned} LD_h(x_t, x_{t-1}) \leq \langle \nabla f(x_{t-1}), x - x_t \rangle + g(x) - g(x_t) + \\ L(D_h(x, x_{t-1}) - D_h(x, x_t)). \end{aligned}$$

Then, we obtain

$$\begin{aligned} \varphi(x_t) &\leq f(x_{t-1}) + \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + g(x) + \\ &\quad L(D_h(x, x_{t-1}) - D_h(x, x_t)) \\ &\leq \varphi(x) + L(D_h(x, x_{t-1}) - D_h(x, x_t)). \end{aligned}$$

# Proof of the theorem (2/2)

*Proof continued.* Notice that then the sequence $(\varphi(x_t))_{t\in\mathbb{N}}$ is non-increasing.

Summing over all $t$, we obtain

$$\sum_{\tau=1}^{t} \varphi(x_\tau) \le t\varphi(x) + LD_h(x, x_0) - LD_h(x, x_t).$$

Since $(\varphi(x_t))_{t\in\mathbb{N}}$ is non-increasing, we have

$$t(\varphi(x_t) - \varphi(x)) \le LD_h(x, x_0).$$

**Remark.** The proof strategy is exact the same as that for the "mirror descent".

# Is the proof easy?

Below is a slightly earlier convergence guarantee due to Nesterov.

**Theorem.** Suppose that $f$ is $L$-smooth on $\mathcal{X}$. Suppose that

$$\|x - x^\star\| \le R, \quad \forall x \in \mathcal{X} \text{ such that } \varphi(x) \le \varphi(x_0).$$

If $\varphi(x_0) - \varphi(x^\star) \ge LR^2$, then

$$\varphi(x_1) - \varphi^\star \le \frac{LR^2}{2}.$$

Otherwise, for every $t \in \mathbb{N}$,

$$\varphi(x_t) - \varphi^\star \le \frac{2LR^2}{t+2}.$$

Yu. Nesterov. 2013. Gradient methods for minimizing composite functions.

# Conclusions

# Summary

Two new notions:

1. Prox-mapping.
2. Estimate sequence approach.

Two algorithms:

1. Bregman proximal gradient algorithm.
2. Accelerated Bregman proximal gradient algorithm.

Two things to notice:

1. Subtle difference between the two proximal set-ups.
2. Similarity of the proof strategies.

# Next lecture

- Frank-Wolfe algorithm.