CSIE5410 Optimization algorithms

Lecture 4: Accelerated gradient descent & Mirror descent*

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

04.10.2018

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

The first part of this lecture addresses the optimality of gradient descent for minimizing an L-smooth convex function.

The second part introduces a class of algorithms for solving a constrained convex optimization problem

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

for some L-smooth convex function f and closed convex set $\mathcal{X}$.

Recommended reading

- Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*. (Chapter 2).
- S. Bubeck. 2015. *Convex Optimization: Algorithms and Complexity*. (Chapter 3).
- H. Lu *et al.* 2018. Relatively smooth convex optimization by first-order methods, and applications.
- H. H. Bauschke *et al.* 2017. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications.

Table of contents

Accelerated gradient descent

Yurii Nesterov

Yurii Nesterov
(1956–present)

Known for:

- *Accelerated gradient descent* (1983, 1988, 2005).
- Interior point method (1994).
- Pioneer of sum-of-squares (2000).
- Smoothing (2005).
- Cubic regularization (2006).
- Dual averaging (2009).
- ...

Recap

Consider the optimization problem:

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

for some convex L-smooth function f.

Recall the gradient descent algorithm:

$$x_t \leftarrow x_{t-1} - \frac{1}{L}\nabla f(x_{t-1}), \quad t \in \mathbb{N},$$

for some $x_0 \in \mathbb{R}^p$. Then we have

$$f(x_t) - f^\star = O\left(\frac{L\|x_0 - x^\star\|_2^2}{t}\right).$$

Question. Is the gradient descent optimal?

Minmax optimality

Ideally, the sequence of iterates $(x_{t,\mathcal{A}})_{t\in\mathbb{N}}$ generated by *the best* algorithm $\mathcal{A}$ should achieve

$$\min_{\mathcal{A}} \max_{f} \left\{ f(x_{t,\mathcal{A}}) - f^\star \mid f \text{ } L\text{-smooth}, \mathcal{A} \in ? \right\}, \quad \forall t \in \mathbb{N}.$$

Let us specify a class of algorithms.

Definition. We say that an iterative algorithm $\mathcal{A}$ is in the class $\mathcal{M}$, if and only if its iterates satisfies

$$x_{t,\mathcal{A}} \in x_0 + \mathrm{span} \left\{ \nabla f(x_0), \nabla f(x_{1,\mathcal{A}}), \ldots, \nabla f(x_{t-1,\mathcal{A}}) \right\}, \quad \forall t \in \mathbb{N}.$$

Lemma. The gradient descent is in $\mathcal{M}$.

Lower bound

Theorem. For any initial iterate $x_0 \in \mathbb{R}^p$, and any $t \in \mathbb{N}$ such that

$$t \leq \frac{1}{2}(p-1),$$

there exists an L-smooth function f such that for any algorithm $\mathcal{A} \in \mathcal{M}$,

$$f(x_{t,\mathcal{A}}) - f^\star \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}, \quad \|x_{t,\mathcal{A}} - x^\star\|_2^2 \geq \frac{1}{8}\|x_0 - x^\star\|_2^2.$$

Question. How do we interpret?

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization.*

Nesterov's accelerated gradient descent

Algorithm Accelerated Gradient Descent

```
Set $x_{-1} = y_0 \in \mathbb{R}^p$, $\gamma_0 = 1$, and $T_\varepsilon \in \mathbb{N}$.
for $t = 0, 1, \ldots, T_\varepsilon$ do
    $x_t \leftarrow y_t - \frac{1}{L}\nabla f(y_t)$
    $\gamma_{t+1} \leftarrow \frac{1+\sqrt{4\gamma_t^2+1}}{2}$
    $y_{t+1} \leftarrow x_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(x_t - x_{t-1})$
end for
```

Theorem. It holds that $f(x_t) - f^\star \leq \frac{4L\|y_0 - x^\star\|_2^2}{(t+2)^2}$.

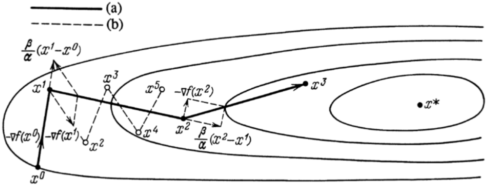
Yu. E. Nesterov. 1983. A method of solving a convex programming problem with convergence rate $O(1/k^2)$.

Detour: Heavy-ball method (1/3)

A closely related algorithm is the *heavy-ball method*.

Algorithm Heavy-Ball Method

```
Set $x_{-1}, x_0 \in \mathbb{R}^p$ and $T_\varepsilon \in \mathbb{N}$.
for $t = 1, \dots, T_\varepsilon$ do
    $x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1}) + \theta(x_{t-1} - x_{t-2})$.
end for
```



B. T. Polyak. 1964. Some methods of speeding up the convergence of iterative methods.

Detour: Heavy-ball method (2/3)

Theorem. Suppose that f is a quadratic function; that is, for some $A \in \mathbb{R}^{p \times p}$, $b \in \mathbb{R}^p$, and $c \in \mathbb{R}$,

$$f(x) := \frac{1}{2} \langle x, Ax \rangle + \langle b, x \rangle + c$$

Suppose that f is L-smooth and μ-strongly convex. Then, with properly chosen η and θ, for some $\delta_t \to 0$,

$$\begin{aligned} &\|x_t - x^\star\|_2 + \|x_{t-1} - x^\star\|_2 \\ &\quad \le \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} + \delta_t \right)^t \left(\|x_{-1} - x^\star\|_2 + \|x_0 - x^\star\|_2 \right). \end{aligned}$$

L. Lessard *et al.* 2016. Analysis and design of optimization algorithms via integral quadratic constraints.

Detour: Heavy-ball method (3/3)

Theorem. For any initial iterate $x_0 \in \mathbb{R}^p$, there exists an L-smooth and μ-strongly convex function f, such that

$$\|x_{t,\mathcal{A}} - x^\star\|_2 \geq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^t \|x_0 - x^\star\|_2, \quad \forall \mathcal{A} \in \mathcal{M}.$$

The iteration complexity is then $\Omega(\sqrt{(L/\mu)}\log(1/\varepsilon))$.

Remark. The gradient descent achieves a worse complexity $O((L/\mu)\log(1/\varepsilon))$.

Remark. However, convergence of the heavy-ball method in general is unclear.

Yu. E. Nesterov. 1983. A method of solving a convex programming problem with convergence rate $O(1/k^2)$.

Proof of convergence of the accelerated gradient method (1/2)

Proof. Recall that

$$
\begin{gathered}
f^\star \geq f(y_{k+1}) + \langle \nabla f(y_{k+1}), x^\star - y_{k+1} \rangle + \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2, \\
f(x_{k+1}) \leq f(y_{k+1}) - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2, \\
f(x_k) \geq f(y_{k+1}) + \langle \nabla f(y_{k+1}), x_k - y_{k+1} \rangle .
\end{gathered}
$$

Define $q_t := (\gamma_t - 1)(x_{t-1} - x_t)$. By the inequalities above, one can show that

$$
\begin{aligned}
&\frac{2}{L} \gamma_{t+1}^2 \left(f(x_{t+1}) - f^\star \right) + \|q_{t+1} - x_{t+1} + x^\star\|^2 \\
&\qquad \leq \frac{2}{L} \gamma_t^2 \left(f(x_t) - f^\star \right) + \|q_t - x_t + x^\star\|^2.
\end{aligned}
$$

Proof of convergence of the accelerated gradient method (2/2)

Proof continued. Therefore,

$$\begin{aligned}
&\frac{2}{L}\gamma_{t+1}^2\left(f(x_{t+1})-f^\star\right)+\|q_{t+1}-x_{t+1}+x^\star\|^2\\
&\quad\le \frac{2}{L}\gamma_t^2\left(f(x_t)-f^\star\right)+\|q_t-x_t+x^\star\|^2\\
&\quad\le \cdots\\
&\quad\le \frac{2}{L}\gamma_0^2\left(f(x_0)-f^\star\right)+\|q_0-x_0+x^\star\|^2\\
&\quad\le \frac{2}{L}\left(f(y_0)-f^\star\right)\\
&\quad\le \|y_0-x^\star\|^2.
\end{aligned}$$

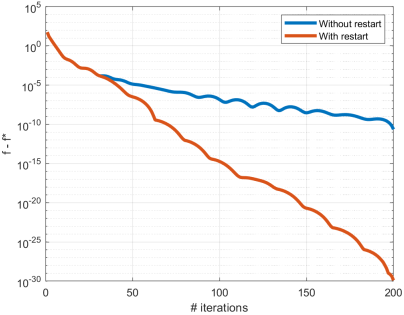
It remains to notice that

$$\gamma_{t+1}\ge\gamma_t+0.5\ge\cdots\ge 1+0.5(t+1).$$

Remarks

Remark. The accelerated gradient descent is not monotone.

Remark. One may make it monotone and faster by *restart*.



B. O'Donoghue and E. Candès. 2015. Adaptive restart for accelerated gradient schemes.

Remarks

Remark. There are also accelerated gradient methods that achieves the optimal $O(\sqrt{(L/\mu)}\log(1/\varepsilon))$ iteration complexity for L-smooth and μ-strongly convex functions, e.g.:

$$\begin{aligned} x_t &\leftarrow y_{t-1} - \frac{1}{L}\nabla f(y_{t-1}), \\ y_t &\leftarrow x_t + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\left(x_t - x_{t-1}\right). \end{aligned}$$

Remark. There are two more variants of the accelerated gradient methods, both by Nesterov.

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

P. Tseng. 2008. On accelerated proximal gradient methods for convex-concave optimization.

Remarks

Remark. Why acceleration can be achieved is one of the most mysterious topics in optimization.

- Estimate sequence (by Nesterov himeself).
- Coupling of gradient descent and mirror descent.
- Differential equation/dynamical systems (abundant!).
- Online learning.
- ...

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

Z. Allen-Zhu and L. Orecchia. 2014. Linear coupling: An ultimate unification of gradient and mirror descent.

W. Su *et al.* 2016. A differential equation for modeling Nesterov's accelerated gradient method: Theory and practice.

J.-K. Wang and J. Abernethy. 2018. Acceleration through optimistic no-regret dynamics.

Differential equation approach

Consider a slightly different accelerated gradient method.

$$x_t \leftarrow y_{t-1} - \eta \nabla f(y_{t-1}),$$
$$y_t \leftarrow x_t + \frac{t-1}{t+2}\left(x_t - x_{t-1}\right).$$

Theorem. A continuous limit is the differential equation

$$\ddot{x}(\tau) + \frac{3}{\tau}\dot{x}(\tau) + \nabla f(x(\tau)) = 0, \quad \tau \in \ ]0, +\infty[,$$
$$x(0) = x_0,$$
$$\dot{x}(0) = 0.$$

Furthermore,

$$f(x(\tau)) - f^\star = O(\tau^{-2}).$$

W. Su *et al.* 2016. A differential equation for modeling Nesterov's accelerated gradient method: Theory and practice.

Mirror descent & relative smoothness

Caveat

The algorithm to be introduced is not exactly the “mirror descent” in literature, which we will see soon.

However, it takes basically the same form as the mirror descent, and currently it lacks a name, so I will also call it “mirror descent” in this course.

Recall: Smoothness

Definition. We say that a function f is L-smooth on $\mathbb{R}^p$, if and only if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^p.$$

Question. Is it necessary to work with the Euclidean (ℓ_2) norm?
It there a gradient descent with some other norm?
What is the benefit of working with non-Euclidean norms?

Norm & dual norm

Definition. A norm on $\mathbb{R}^p$ is a non-negative function $\|\cdot\|$ that satisfies, for any $x, y \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$:

1. $\|x + y\| \leq \|x\| + \|y\|$.
2. $\|\alpha x\| = |\alpha| \|x\|$.
3. $\|x\| = 0$ if and only if $x = 0$.

Example.

- The q-norm for $q \geq 1$, defined as $\|x\|_q := \left(\sum_{i=1}^p |x^{(i)}|^q\right)^{1/q}$.
- The Schatten q-norm of a matrix, defined as the q-norm of the vector of singular values.
- Let $A \in \mathbb{R}^{p \times p}$ be positive definite. Then $\|x\|_A := \sqrt{\langle x, Ax \rangle}$ is a norm.

Norm & dual norm

Definition. Let $\| \cdot \|$ be a norm on $\mathbb{R}^p$. The associated dual norm is given by

$$\|y\|_* := \max_x \left\{ \langle y, x \rangle \mid x \in \mathbb{R}^p, \|x\| \leq 1 \right\}.$$

Theorem. For any $x, y \in \mathbb{R}^p$, $\langle y, x \rangle \leq \|y\|_* \|x\|$.

Example. Fix $q \in [1, +\infty]$.

- The dual norm of the q-norm is the q'-norm, such that $1/q + 1/q' = 1$.
- The dual norm of the Schatten q-norm is the Schatten q'-norm, $1/q + 1/q' = 1$.
- The dual norm of $\| \cdot \|_A$ is $\| \cdot \|_{A^{-1}}$.

Smoothness in general norms

Definition. We say that f is L-smooth on $\mathbb{R}^p$ with respect to a norm $\|\cdot\|$, if and only if

$$\|\nabla f(y) - \nabla f(x)\|_* \leq L\|y - x\|, \quad \forall x, y \in \mathbb{R}^p.$$

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

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

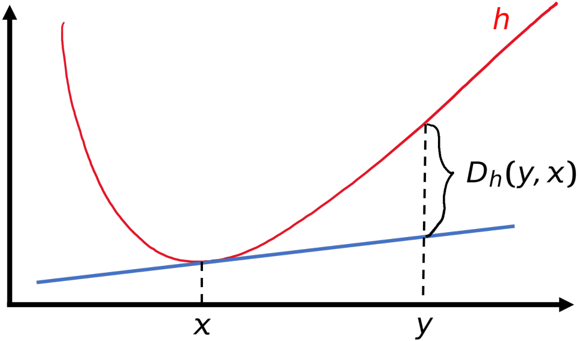
Yu. Nesterov. 2012. Efficiency of coordinate descent methods on huge-scale optimization problems.

J. A. Kelner *et al.* 2014. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations.

Relative smoothness (1/3)

Definition. Let h be a differentiable convex function on $\mathcal{X} \subseteq \mathbb{R}^p$. The associated Bregman divergence is given by

$$D_h(y,x) := h(y) - (h(x) + \langle \nabla h(x), y - x \rangle)\,.$$



Relative smoothness (2/3)

Example. If we set $h(x) := (1/2)\|x\|_2^2$, then

$$D_h(y,x) = (1/2)\|y-x\|_2^2.$$

Example. If we set $h(x) := \sum_{i=1}^{p} x^{(i)} \log x^{(i)} - \sum_{i=1}^{p} x^{(i)} \quad x \geq 0,$ (the negative Shannon entropy), then D_h is the relative entropy

$$D_h(y,x) = \sum_{i=1}^{p} y^{(i)} \log \frac{y^{(i)}}{x^{(i)}} - \sum_{i=1}^{p} y^{(i)} + \sum_{i=1}^{p} x^{(i)}.$$

Example. If we set $h(x) := -\sum_{i=1}^{p} \log x^{(i)}$ (Burg entropy), then D_h is given by

$$D_h(y,x) = \sum_{i=1}^{p} \frac{y^{(i)}}{x^{(i)}} - \log \prod_{i=1}^{p} \frac{y^{(i)}}{x^{(i)}} - p.$$

Relative smoothness (3/3)

Definition. We say that a function f is L-smooth relative to h on $\mathcal{X} \subseteq \mathbb{R}^p$, if and only if

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

Remark. If $h(x) = (1/2)\|x\|_2^2$, then the L-relative smoothness becomes the standard L-smoothness.

H. 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.

Mirror descent (1/3)

Recall the gradient descent method:

$$x_t \leftarrow x_{t-1} - \eta_{t-1} \nabla f(x_{t-1}).$$

Proposition. The expression above is equivalent to

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

Proof. Notice that the function to be minimized is strongly convex, so x_t is uniquely defined. By Fermat's rule, we have

$$\eta_{t-1} \nabla f(x_{t-1}) + x_t - x_{t-1} = 0.$$

Mirror descent (2/3)

Consider a slightly more sophisticated problem:

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

for some closed convex set $\mathcal{X} \subseteq \mathbb{R}^p$, and convex function f L-smooth relatively to a differentiable convex function h on $\mathcal{X}$.

A natural extension of the gradient descent is

$$x_t \leftarrow \arg\min_x \left\{ \, \eta_{t-1} \left\langle \nabla f(x_{t-1}), x - x_{t-1} \right\rangle + \frac{1}{2} \|x - x_{t-1}\|_2^2 \; \middle| \; x \in \mathcal{X} \, \right\},$$

which in fact corresponds to the *projected gradient descent*

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

Mirror descent (3/3)

Algorithm Mirror Descent

```
Set $x_0 \in \operatorname{ri} \mathcal{X}$ and $T_\varepsilon \in \mathbb{N}$.
for $t = 1, \dots, T_\varepsilon$ 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. Set $\eta_t = 1/L$ for all t. It holds that

$$f(x_t) - f(x) \le \frac{L D_h(x, x_0)}{t}, \quad \forall x \in \mathcal{X}.$$

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

Pre-proof: Three-point equality

Theorem. (Three-point equality) Suppose that h is a convex function taking a finite value at u, and differentiable at v and w. Then it holds that

$$D_h(u,v) + D_h(v,w) = D_h(u,w) + \langle \nabla h(w) - \nabla h(v), u - v \rangle \,.$$

Proof. Plug in the definitions.

Remark. This can be viewed as a generalization of the Pythagorean theorem.

G. Chen and M. Teboulle. 1993. Convergence analysis of a proximal-like minimization algorithm using Bregman functions.

Pre-proof: Key lemma

Lemma. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be convex and closed, and φ be differentiable and convex on $\mathcal{X}$. Suppose that h is differentiable at z. Define

$$z_+ \in \arg\min_x \left\{ \varphi(x) + D_h(x, z) \mid x \in \mathcal{X} \right\}.$$

Then, for any $x \in \mathcal{X}$,

$$D(x, z_+) + D(z_+, z) \le D(x, z) + \varphi(x) - \varphi(z_+).$$

G. Chen and M. Teboulle. 1993. Convergence analysis of a proximal-like minimization algorithm using Bregman functions.

P. Tseng. 2008. On accelerated proximal gradient methods for convex-concave optimization.

Pre-proof: Proof of the key lemma

Proof. By the three-point equality, we write

$$D_h(x, z_+) + D_h(z_+, z) - D(x, z) = \langle x - z_+, \nabla h(z) - \nabla h(z_+) \rangle .$$

By the optimality condition for z_+, we have

$$\langle \nabla \varphi(z_+) + \nabla h(z_+) - \nabla h(z), x - z_+ \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Then we have

$$D_h(x, z_+) + D_h(z_+, z) - D(x, z) \leq \langle \nabla \varphi(z_+), x - z_+ \rangle , \quad \forall x \in \mathcal{X}.$$

The theorem follows from the convexity of φ:

$$\varphi(x) \geq \varphi(z_+) + \langle \nabla \varphi(z_+), x - z_+ \rangle .$$

Proof of convergence of mirror descent (1/2)

Proof. We write, by L-relative smoothness,

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

By the key lemma, we write, for any $x \in \mathcal{X}$,

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

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

Proof of convergence of mirror descent (2/2)

Proof continued. By convexity of f, we obtain

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

Summing over all t, we obtain

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

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

$$t(f(x_t) - f(x)) \leq LD_h(x, x_0).$$

Special case: Projected gradient descent

Consider the problem of minimizing an L-smooth convex function f on a closed convex set $\mathcal{X}$, by the algorithm

$$x_t \in \arg\min_x \left\{ \langle \nabla f(x_{t-1}), x - x_{t-1} \rangle + \frac{L}{2} \|x - x_{t-1}\|_2^2 \;\middle|\; x \in \mathcal{X} \right\}.$$

Proposition. The algorithm is *projected gradient descent*:

$$x_t \leftarrow \operatorname{proj}_{\mathcal{X}} \left(x_{t-1} - \frac{1}{L} \nabla f(x_{t-1}) \right).$$

Theorem. It holds that $f(x_t) - f^\star = O(L\|x_0 - x^\star\|_2^2 t^{-1})$.

Proof of the proposition

Proof. Notice that x_t is uniquely defined (Why?). The optimality condition of x_t in the first formulation is

$$\langle \nabla f(x_{t-1}) + L(x_t - x_{t-1}), x - x_t \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

The second formulation is equivalent to

$$x_t \in \arg\min_x \left\{ \frac{1}{2} \left\| x - \left(x_{t-1} - \frac{1}{L} \nabla f(x_{t-1}) \right) \right\|_2^2 \;\middle|\; x \in \mathcal{X} \right\},$$

which has the same optimality condition.

Can the algorithm be accelerated?

Theorem. Suppose that f is L-smooth with respect to a norm $\|\cdot\|$, and that h is μ-strongly convex with respect to the same norm on $\mathcal{X}$, i.e.,

$$D_h(y,x) \geq \frac{\mu}{2}\|y-x\|^2, \quad \forall x,y \in \mathcal{X}.$$

Then there exists an iterative algorithm, whose t-th iterate x_t satisfies

$$f(x_t) - f^\star = O\left(\frac{L(D_h(x^\star, x_0) + f(x_0) - f^\star)}{\mu t^2}\right).$$

Remark. Possibility unclear for general relative smooth functions.

A. Auslender and M. Teboulle. 2006. Interior gradient and proximal methods for convex and conic optimization.

Conclusions

Summary (1/2)

Algorithm Accelerated Gradient Descent

```
Set $x_{-1} = y_0 \in \mathbb{R}^p$, $\gamma_0 = 1$, and $T_\varepsilon \in \mathbb{N}$.
for $t = 0, 1, \ldots, T_\varepsilon$ do
    $x_t \leftarrow y_t - \frac{1}{L}\nabla f(y_t)$
    $\gamma_{t+1} \leftarrow \frac{1+\sqrt{4\gamma_t^2+1}}{2}$
    $y_{t+1} \leftarrow x_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(x_t - x_{t-1})$
end for
```

Theorem. It holds that $f(x_t) - f^\star = O(L\|x_0 - x^\star\|_2^2/t^2)$.

Summary (2/2)

Algorithm Mirror Descent

```
Set $x_0 \in \operatorname{ri} \mathcal{X}$ and $T_\varepsilon \in \mathbb{N}$.
for $t = 1, \ldots, T_\varepsilon$ 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
```

Definition. Relative smoothness ⇔

$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x)$.

Theorem. Mirror descent achieves

$f(x_t) - f^\star = O(L D(x^\star, x_0) t^{-1})$.

Remark. Now we know how to do *constrained convex optimization*!

Next lecture

- Examples of relative smoothness.
- Subgradient & subdifferential.
- Mirror descent.