# Lecture 5. Gradient Descent Method II

Advanced Optimization (Fall 2022)

**Peng Zhao**

zhaop@lamda.nju.edu.cn

Nanjing University

# Outline

- GD for Smooth Optimization
  - Smooth and Convex Functions
  - Smooth and Strongly Convex Functions
- Nesterov's Accelerated GD
- Extension to Composite Optimization

# Part 1. GD for Smooth Optimization

- Smooth and Convex

- Smooth and Strongly Convex

- Extension to Constrained Case

# Overview

Table 1: A summary of convergence rates of GD for different function families, where we use $\kappa \triangleq L/\sigma$ to denote the condition number.

| Function Family | | Step Size | Output Sequence | Convergence Rate | |
|---|---|---|---|---|---|
| $G$-Lipschitz | convex | $\eta = \frac{D}{G\sqrt{T}}$ | $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$ | $\mathcal{O}(1/\sqrt{T})$ | *last lecture* |
| | $\sigma$-strongly convex | $\eta_t = \frac{2}{\sigma(t+1)}$ | $\bar{\mathbf{x}}_T = \sum_{t=1}^{T}\frac{2t}{T(T+1)}\mathbf{x}_t$ | $\mathcal{O}(1/T)$ | |
| $L$-smooth | convex | $\eta = \frac{1}{L}$ | $\bar{\mathbf{x}}_T = \mathbf{x}_T$ | $\mathcal{O}(1/T)$ | *this lecture* |
| | $\sigma$-strongly convex | $\eta = \frac{2}{\sigma+L}$ | $\bar{\mathbf{x}}_T = \mathbf{x}_T$ | $\mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right)$ | |

For simplicity, we mostly focus on ***unconstrained*** domain, i.e., $\mathcal{X} = \mathbb{R}^d$.

# Convex and Smooth

**Theorem 1.** *Suppose the function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex and differentiable, and also L-smooth. GD updates by $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$ with step size $\eta_t = \frac{1}{L}$, and then GD enjoys the following convergence guarantee:*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{2L \|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T-1} = \mathcal{O}\left(\frac{1}{T}\right).$$

**Note**: we are working on ***unconstrained*** setting and using a ***fixed*** step size tuning.

# The First Gradient Descent Lemma

**Lemma 1.** *Suppose that $f$ is proper, closed and convex; the feasible domain $\mathcal{X}$ is nonempty, closed and convex. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by the gradient descent method, $\mathcal{X}^\star$ be the optimal set of the optimization problem and $f^\star$ be the optimal value. Then for any $\mathbf{x}^\star \in \mathcal{X}^\star$ and $t \geq 0$,*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t(f(\mathbf{x}_t) - f^\star) + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2.$$

***Proof:***

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &= \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^\star\|^2 \quad \text{(GD)} \\
&\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^\star\|^2 \quad \text{(Pythagoras Theorem)} \\
&= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t(f(\mathbf{x}_t) - f^\star) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}$$

(convexity: $f(\mathbf{x}_t) - f^\star = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle$) $\square$

# Refined Result for Smooth Optimization

***Proof:***

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 = \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^\star\|^2 \quad \text{(GD)}$$

$$\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^\star\|^2 \quad \text{(Pythagoras Theorem)}$$

$$= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$$

$$\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t (f(\mathbf{x}_t) - f^\star) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$$

(convexity: $f(\mathbf{x}_t) - f^\star = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle$)

*haven't used smoothness*

**Lemma 2** (co-coercivity). *Let $f$ be convex and L-smooth over $\mathbb{R}^d$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, one has*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

# Co-coercive Operator

**Lemma 2** (co-coercivity). *Let $f$ be convex and L-smooth over $\mathbb{R}^d$. Then for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, *one has*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

**Definition 1** (co-coercive operator). An operator $C$ is called *$\beta$-co-coercive* (or *$\beta$-inverse-strongly monotone*, for $\beta > 0$, if for any $x, y \in \mathcal{H}$,

$$\langle Cx - Cy, x - y \rangle \geq \beta \|Cx - Cy\|^2.$$

The co-coercive condition is relatively standard in *operator splitting* literature and *variational inequalities*.

# Smooth and Convex

***Proof:***

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 = \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^\star\|^2 \quad \text{(GD)}$$

$$\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^\star\|^2 \quad \text{(Pythagoras Theorem)}$$

$$= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$$

$$\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \left(\eta_t^2 - \frac{2\eta_t}{L}\right) \|\nabla f(\mathbf{x}_t)\|^2$$

*exploiting coercivity of smoothness and unconstrained first-order optimality*

$$\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle = \langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^\star), \mathbf{x}_t - \mathbf{x}^\star \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^\star)\|^2 = \frac{1}{L} \|\nabla f(\mathbf{x}_t)\|^2$$

$$\Rightarrow \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \left(\eta_t^2 - \tfrac{2\eta_t}{L}\right) \|\nabla f(\mathbf{x}_t)\|^2$$

by picking $\eta_t = \eta = \frac{1}{L}$ to minimize the r.h.s

$$\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \tfrac{1}{L^2} \|\nabla f(\mathbf{x}_t)\|^2$$

$$\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 \leq \ldots \leq \|\mathbf{x}_1 - \mathbf{x}^\star\|^2$$

*which already implies the convergence*

# Smooth and Convex

***Proof:*** Now, we consider the function-value level,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) = f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) + f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

$$\begin{aligned} & f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \\ &= f(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)) - f(\mathbf{x}_t) \\ &\leq \langle \nabla f(\mathbf{x}_t), -\eta_t \nabla f(\mathbf{x}_t) \rangle + \frac{L}{2}\eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \quad \text{(smoothness)} \\ &= \left(-\eta_t + \frac{L}{2}\eta_t^2\right) \|\nabla f(\mathbf{x}_t)\|^2 \\ &= -\frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 \quad \text{(recall that we have picked } \eta_t = \eta = \tfrac{1}{L}\text{)} \end{aligned}$$

***one-step improvement***

$$\Longrightarrow \quad f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq -\frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 + f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

# Smooth and Convex

***Proof:***

$$\Longrightarrow \quad f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \le -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

***Next step:*** relating $\|\nabla f(\mathbf{x}_t)\|$ to function-value gap to form a telescoping structure.

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle \le \|\nabla f(\mathbf{x}_t)\| \|\mathbf{x}_t - \mathbf{x}^\star\| \quad \Rightarrow \|\nabla f(\mathbf{x}_t)\|^2 \ge \frac{(f(\mathbf{x}_t) - f(\mathbf{x}^\star))^2}{\|\mathbf{x}_t - \mathbf{x}^\star\|^2}$$

$$\Longrightarrow \quad f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \le -\frac{(f(\mathbf{x}_t) - f(\mathbf{x}^\star))^2}{2L\|\mathbf{x}_t - \mathbf{x}^\star\|^2} + f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

$$\le -\frac{(f(\mathbf{x}_t) - f(\mathbf{x}^\star))^2}{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2} + f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

(by optimizer's convergence, i.e., $\|\mathbf{x}_t - \mathbf{x}^\star\| \le \|\mathbf{x}_1 - \mathbf{x}^\star\|$)

# Smooth and Convex

***Proof:*** $$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq -\frac{1}{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}(f(\mathbf{x}_t) - f(\mathbf{x}^\star))^2 + f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

Define $\Delta_t \triangleq f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ and $\beta_t \triangleq \frac{1}{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}$.

$$\Longrightarrow \quad \Delta_{t+1} \leq \Delta_t - \beta_t \Delta_t^2 \qquad \Rightarrow \frac{\beta_t \Delta_t}{\Delta_{t+1}} \leq \frac{1}{\Delta_{t+1}} - \frac{1}{\Delta_t}$$

$$\Longrightarrow \quad \beta_t \leq \frac{1}{\Delta_{t+1}} - \frac{1}{\Delta_t} \qquad \Rightarrow \sum_{t=1}^{T-1} \beta_t \leq \frac{1}{\Delta_T} - \frac{1}{\Delta_1} \leq \frac{1}{\Delta_T}$$

$$\Longrightarrow \quad \Delta_T \triangleq f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{\sum_{t=1}^{T-1} \beta_t} = \frac{2L\left\|\mathbf{x}_1 - \mathbf{x}^\star\right\|^2}{T-1}. \qquad \square$$

# Key Lemma for Smooth GD

- During the proof, we have obtained an important lemma for smooth optimization.

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \left(-\eta_t + \frac{L}{2}\eta_t^2\right) \|\nabla f(\mathbf{x}_t)\|^2$$ *one-step improvement*

Compare a similar result that holds for convex and Lipschitz functions.

**Lemma 2.** *Under the same assumptions as Theorem 1. Let $\{\mathbf{x}_t\}_{t=1}^T$ be the sequence generated by GD. Then we have*

$$\sum_{t=1}^{T} \eta_t (f(\mathbf{x}_t) - f^\star) \leq \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}^\star\|^2 + \frac{1}{2}\sum_{t=1}^{T} \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2.$$

***average-iterated*** convergence vs ***last-iterated*** convergence

# Key Lemma for Smooth GD

- One-step improvement for ***smooth*** GD under ***unconstrained*** setting.

**Lemma 3** (one-step improvement)**.** *Suppose the function* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *is convex and differentiable, and also* $L$*-smooth. Consider the following unconstrained GD update:* $\mathbf{x}' = \mathbf{x} - \eta \nabla f(\mathbf{x})$*. Then,*

$$f(\mathbf{x}') - f(\mathbf{x}) \leq \left( -\eta + \frac{L}{2}\eta^2 \right) \|\nabla f(\mathbf{x})\|^2 .$$

*In particular, when choosing* $\eta = \frac{1}{L}$*, we have*

$$f\left( \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x}) \right) - f(\mathbf{x}) \leq -\frac{1}{2L} \|\nabla f(\mathbf{x})\|^2 .$$

# Smooth and Strongly Convex

- Recall the definition of strongly convex functions ( *first-order* version).

**Definition 5** (Strong Convexity). A function $f$ is $\sigma$-strongly convex if, for any $\mathbf{x} \in \mathrm{dom}(\partial f), \mathbf{y} \in \mathrm{dom}(f)$ and $\mathbf{g} \in \partial f(\mathbf{x})$,

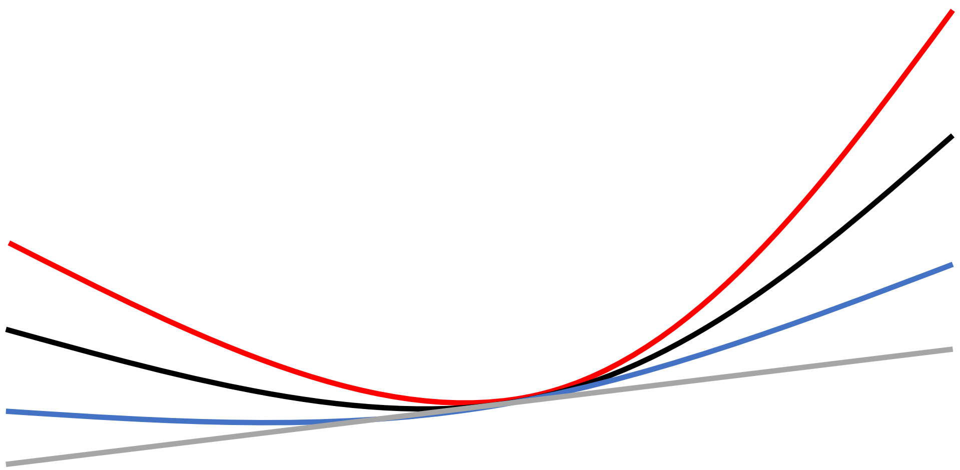
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

# Smooth and Strongly Convex

$f$ is $\sigma$-strongly convex

$f$ is $L$-smooth

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \le f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$



# Smooth and Strongly Convex

**Theorem 2.** *Suppose the function* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *is* $\sigma$*-strongly-convex and differentiable, and also L-smooth; and the feasible domain* $\mathcal{X} \subseteq \mathbb{R}^d$ *is compact and convex with a diameter* $D > 0$*. Then, setting* $\eta_t = \frac{2}{\sigma+L}$*, GD satisfies*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2} \exp\left(-\frac{4(T-1)}{\kappa+1}\right) \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 = \mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right),$$

*where* $\kappa \triangleq L/\sigma$ *denotes the condition number of* $f$*.*

**Note**: we are working on ***unconstrained*** setting and using a ***fixed*** step size tuning.

# Smooth and Strongly Convex

***Proof:***

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 = \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^\star\|^2 \text{ (GD)}$$

$$\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^\star\|^2 \text{ (Pythagoras Theorem)}$$

$$= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2$$

*how to exploiting the **strong convexity** and **smoothness** simultaneously*

**Lemma 4** (co-coercivity of smooth and strongly convex function). *Let $f$ be $L$-smooth and $\sigma$-strongly convex on $\mathbb{R}^d$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, one has*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\sigma L}{\sigma + L}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\sigma + L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

# Coercivity of Smooth and Strongly Convex Function

**Lemma 4** (co-coercivity of smooth and strongly convex function). *Let $f$ be $L$-smooth and $\sigma$-strongly convex on $\mathbb{R}^d$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, one has*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\sigma L}{\sigma + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\sigma + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

***Proof:*** Define $h(\mathbf{x}) \triangleq f(\mathbf{x}) - \frac{\sigma}{2}\|\mathbf{x}\|^2$. Then, $h$ enjoys the following properties:

- $h$ is convex: by $\sigma$-strong convexity (see previous lecture).
- $h$ is $(L - \sigma)$-smooth. $\nabla^2 h(\mathbf{x}) = \nabla^2 f(\mathbf{x}) - \sigma I \preceq (L - \sigma) I$.

$$\Longrightarrow \langle \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L - \sigma} \|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\|^2$$

*by co-coercivity of smooth and convex functions*

Then, rearranging the terms finishes the proof. □

# Smooth and Strongly Convex

***Proof:*** 
$$\begin{aligned}\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 &= \|\Pi_{\mathcal{X}}[\mathbf{x}_t-\eta_t\nabla f(\mathbf{x}_t)]-\mathbf{x}^\star\|^2 \quad \text{(GD)}\\ &\leq \|\mathbf{x}_t-\eta_t\nabla f(\mathbf{x}_t)-\mathbf{x}^\star\|^2 \quad \text{(Pythagoras Theorem)}\\ &= \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - 2\eta_t\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^\star\rangle + \eta_t^2\|\nabla f(\mathbf{x}_t)\|^2\\ &\leq \left(1-\frac{2\eta_t\sigma L}{L+\sigma}\right)\|\mathbf{x}_t-\mathbf{x}^\star\|^2 + \left(\eta_t^2-\frac{2\eta_t}{L+\sigma}\right)\|\nabla f(\mathbf{x}_t)\|^2\end{aligned}$$

*exploiting co-coercivity of smoothness and strong convexity*

$$\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^\star\rangle = \langle\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}^\star),\mathbf{x}_t-\mathbf{x}^\star\rangle \geq \frac{1}{L+\sigma}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L\sigma}{L+\sigma}\|\mathbf{x}_t-\mathbf{x}^\star\|^2$$

$$\Longrightarrow \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 \leq \left(1-\frac{2\eta_t\sigma L}{L+\sigma}\right)\|\mathbf{x}_t-\mathbf{x}^\star\|^2 + \left(\eta_t^2-\frac{2\eta_t}{L+\sigma}\right)\|\nabla f(\mathbf{x}_t)\|^2$$

*serving as the "one-step improvement" in the analysis*

# Smooth and Strongly Convex

***Proof:*** $$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{2\eta_t \sigma L}{L+\sigma}\right) \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \left(\eta_t^2 - \frac{2\eta_t}{L+\sigma}\right) \|\nabla f(\mathbf{x}_t)\|^2$$

The step size configuration:

(i) first, we need $1 - \frac{2\eta_t \sigma L}{L+\sigma} < 1$ to ensure the contraction property;

(ii) second, we hope $(\eta_t^2 - \frac{2\eta_t}{L+\sigma}) \leq 0$, or it becomes $0$ is enough.

$\Longrightarrow$ a feasible (simple) setting: $\eta_t = \eta = \dfrac{2}{L+\sigma}$

$$\Longrightarrow \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{4\sigma L}{(L+\sigma)^2}\right) \|\mathbf{x}_t - \mathbf{x}^\star\|^2 = \left(\frac{L-\sigma}{L+\sigma}\right)^2 \|\mathbf{x}_t - \mathbf{x}^\star\|^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^2 \|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

$$\Longrightarrow \|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2(T-1)} \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 \leq \exp\left(-\frac{4(T-1)}{\kappa+1}\right) \|\mathbf{x}_1 - \mathbf{x}^\star\|^2$$

# Smooth and Strongly Convex

***Proof:*** $\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \le \left(\frac{\kappa-1}{\kappa+1}\right)^{2(T-1)} \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 \le \exp\left(-\frac{4(T-1)}{\kappa+1}\right) \|\mathbf{x}_1 - \mathbf{x}^\star\|^2$

***Next step:*** relating $\|\mathbf{x}_T - \mathbf{x}^\star\|^2$ to $f(\mathbf{x}_T) - f(\mathbf{x}^\star)$.

$$f(\mathbf{x}_t) \le f(\mathbf{x}^\star) + \langle \nabla f(\mathbf{x}^\star), \mathbf{x}_t - \mathbf{x}^\star \rangle + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 = f(\mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

(in unconstrained case, $\nabla f(\mathbf{x}^\star) = \mathbf{0}$)

$$\Longrightarrow \quad f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2} \exp\left(-\frac{4(T-1)}{\kappa+1}\right) \|\mathbf{x}_1 - \mathbf{x}^\star\|^2 = \mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right).$$

$\square$

# Constrained Optimization

- A ***generalized*** one-step improvement lemma for smooth optimization.

**Lemma 5.** *Suppose $f$ is $L$-smooth. Let $\mathbf{x}, \mathbf{u} \in \mathcal{X}, \mathbf{x}_{t+1} = \Pi_{\mathcal{X}}[\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)]$, and $g(\mathbf{x}) = L(\mathbf{x} - \mathbf{x}_{t+1})$. Then the following holds true:*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2.$$

comparator $\mathbf{u}$ is introduced because now GD is not necessary "descent" due to the projection

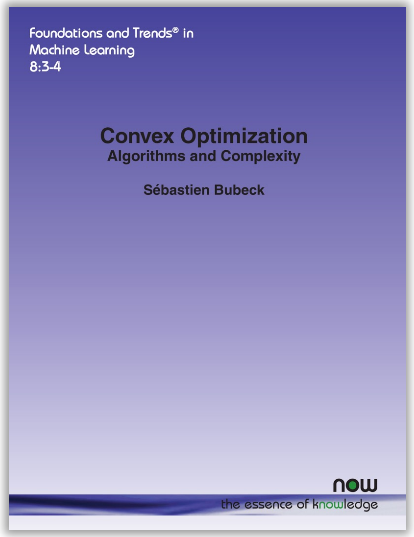
- In unconstrained case, $g(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)$.
- In unconstrained case, setting $\mathbf{u} = \mathbf{x}_t$ recovers the one-step improvement: $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$.

# Constrained Optimization

**Same** convergence rates as unconstrained case can be obtained in the constrained setting for smooth optimization.

Detailed proofs for constrained case are not presented in our course. The proof follows the same vein via some twists, we refer anyone interested to the following parts in **Bubeck's book**:

- *Constrained* + smooth + convex: **Section 3.2**
- *Constrained* + smooth + strongly convex: **Section 3.4.2**




**Convex Optimization: Algorithms and Complexity**
Sebastien Bubeck
Foundations and Trends in ML, 2015

# Lower Bound

Lower bounds reflect the difficulty of the problem, regardless of algorithms.

*notice: this lower bound only holds for first-order methods*

Table 1: A summary of convergence rates of GD for different function families.

| Function Family | | Convergence Rate | *Lower Bound* | *Optimal?* |
|---|---|---|---|---|
| $G$-Lipschitz | convex | $\mathcal{O}(1/\sqrt{T})$ | $\Omega(1/\sqrt{T})$ | ✓ |
| | $\sigma$-strongly convex | $\mathcal{O}(1/T)$ | $\Omega(1/T)$ | ✓ |
| $L$-smooth | convex | $\mathcal{O}(1/T)$ | $\Omega(1/T^2)$ | ✗ |
| | $\sigma$-strongly convex | $\mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right)$ | $\Omega\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right)\right)$ | ✗ |

$\Longrightarrow$ **GD is suboptimal in *smooth* convex optimization!**

# Part 2. Nesterov's Accelerated GD

- Algorithm

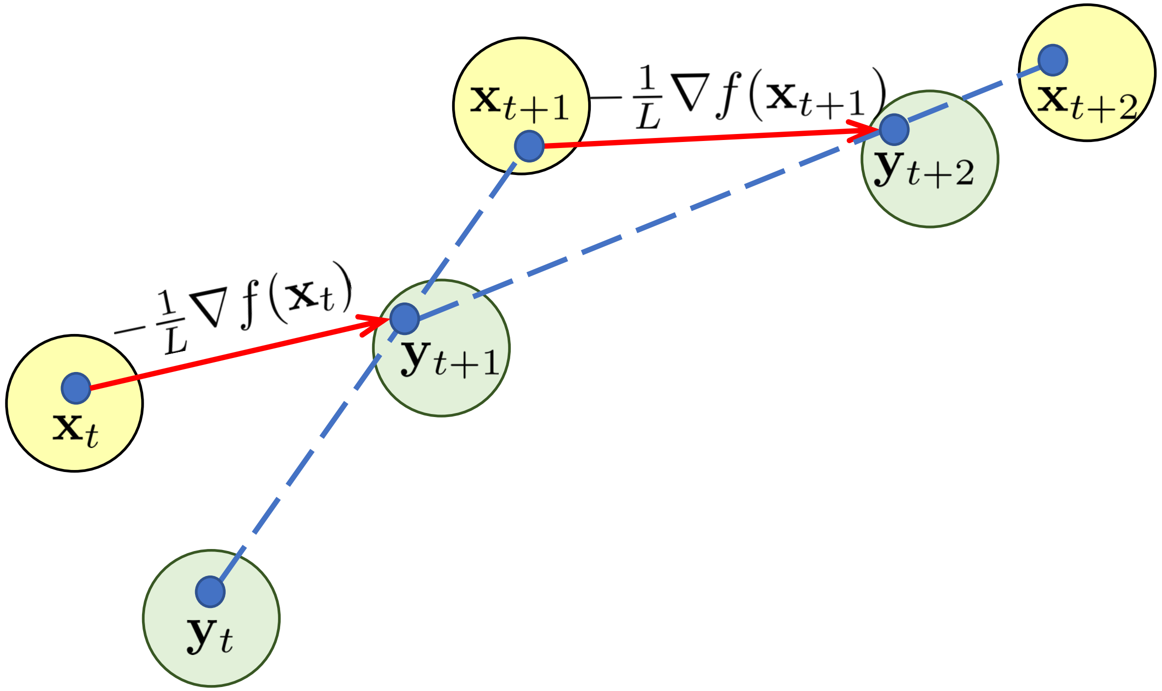
- Smooth and Convex

- Smooth and Strongly Convex

# Nesterov's Accelerated GD

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$

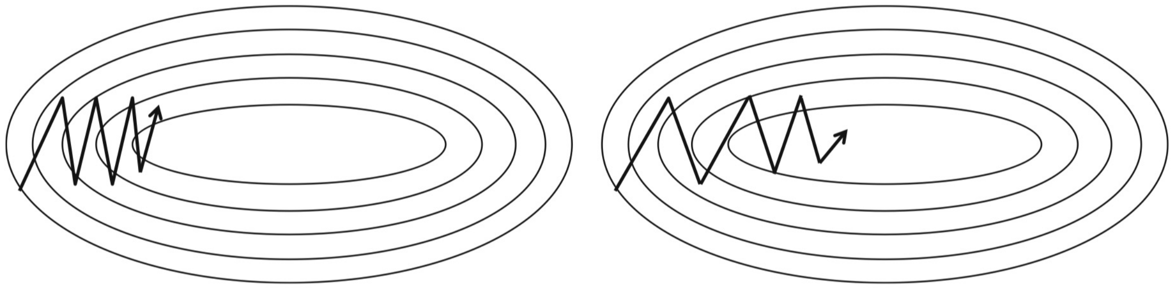
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$



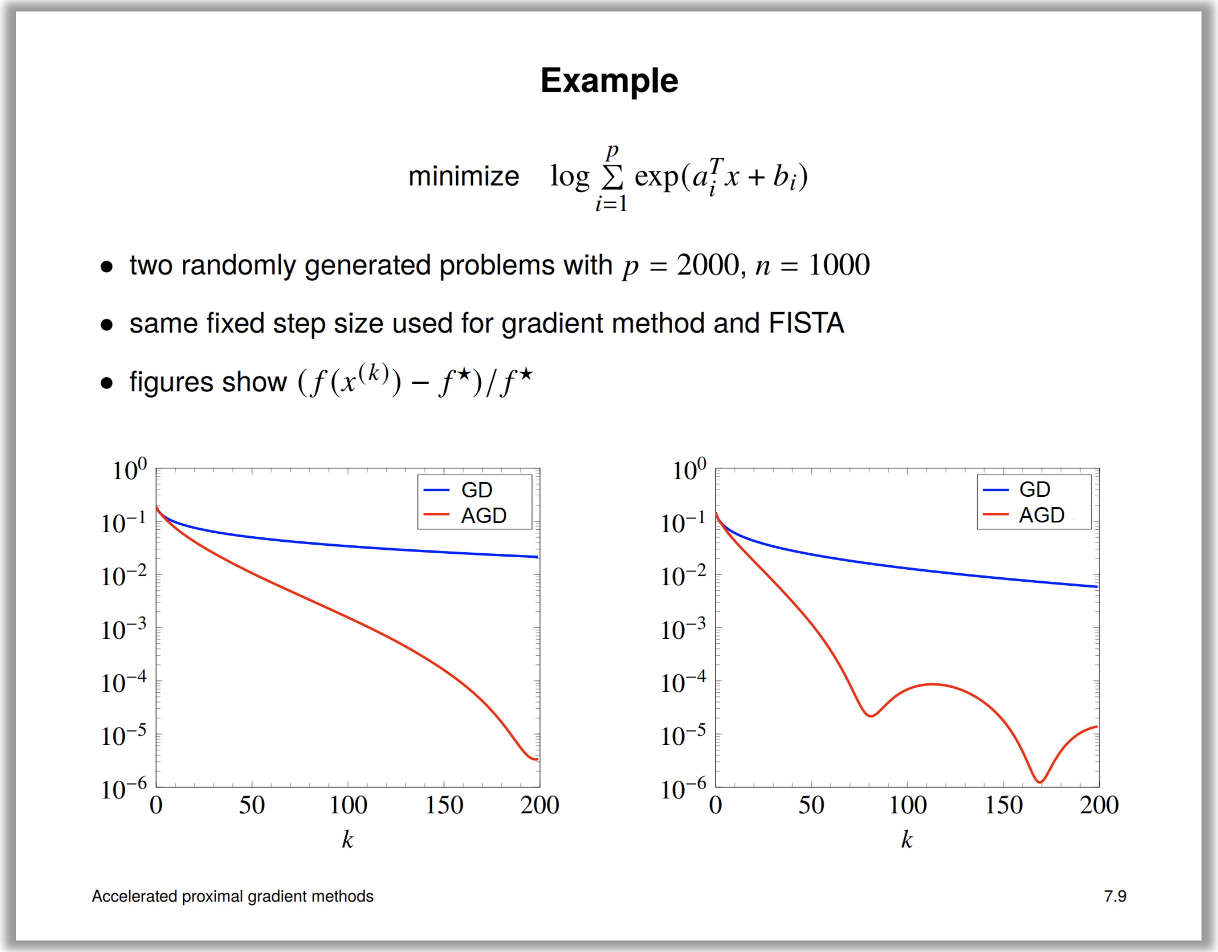

- Define $\mathbf{x}_1 = \mathbf{y}_1$.
- $\alpha_t < 0$ is a ***time-varying*** mixing rate of $\mathbf{y}_t$ and $\mathbf{y}_{t+1}$.
- $\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \alpha_t(\mathbf{y}_t - \mathbf{y}_{t+1})$ is an ***extrapolated*** point, i.e., with ***momentum***.

# Nesterov's Accelerated GD

- a momentum term is added to improve convergence
- the descent property is relaxed and not ensured now



GD                Accelarated GD



**Example**

minimize $\log \sum_{i=1}^{p} \exp(a_i^T x + b_i)$

- two randomly generated problems with $p = 2000$, $n = 1000$
- same fixed step size used for gradient method and FISTA
- figures show $(f(x^{(k)}) - f^\star)/f^\star$



Accelerated proximal gradient methods 7.9

https://www.seas.ucla.edu/~vandenbe/236C/lectures/fgrad.pdf

# Convergence of Nesterov's Accelerated GD

**Theorem 3.** *Let $f$ be convex and $L$-smooth. Nesterov's accelerated GD is configured as*

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t,$$

*where* $\lambda_0 = 0$, $\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$, *and* $\alpha_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. *Then, we have*

$$f(\mathbf{y}_T) - f(\mathbf{x}^\star) \le \frac{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right).$$

# Proof of AGD Convergence

***Proof:*** first we prove the following *generalized one-step improvement lemma*.

**Lemma 6.** *For any* $\mathbf{u} \in \mathcal{X}$, *if* $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$, *then the following holds true:*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2.$$

comparator $\mathbf{u}$ is introduced because now GD is not necessary "descent" due to the momentum

Setting $\mathbf{u} = \mathbf{x}_t$ recovers the one-step improvement used in earlier analysis.

# Generalized One-Step Improvement

**Lemma 6.** *For any* $\mathbf{u} \in \mathcal{X}$, *if* $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$, *then the following holds true:*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2.$$

Setting $\mathbf{u} = \mathbf{x}_t$ recovers the one-step improvement used in earlier analysis.

***Proof:***

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{u}) &= f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) + f(\mathbf{x}_t) - f(\mathbf{u}) \\
&\leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle \quad \text{(smoothness and convexity)} \\
&= \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \quad (\mathbf{x}_{t+1} = \mathbf{x}_t - \tfrac{1}{L}\nabla f(\mathbf{x}_t)) \\
&= \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}
$$

$\square$

# Proof of AGD Convergence

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

***Proof:*** (continued proving Theorem 3)

**Lemma 6.** *For any $\mathbf{u} \in \mathcal{X}$, if $\mathbf{x}' = \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})$, then the following holds true:*

$$f(\mathbf{x}') - f(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{u}\rangle - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2.$$

(i) Plugging in $\mathbf{u} = \mathbf{y}_t$: $f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y}_t\rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$.

(ii) Plugging in $\mathbf{u} = \mathbf{x}^\star$: $f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$.

LHS of $(\lambda_t - 1)$(i) + (ii) equals:

$$(\lambda_t - 1)(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star) = \lambda_t(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)) - (\lambda_t - 1)(f(\mathbf{y}_t) - f(\mathbf{x}^\star))$$

Define $\delta_t \triangleq f(\mathbf{y}_t) - f(\mathbf{x}^\star)$, LHS = $\lambda_t \delta_{t+1} - (\lambda_t - 1)\delta_t$

***Goal: design a telescoping series***

# Proof of AGD Convergence

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

***Proof:*** (continued proving Theorem 3)

(i) Plugging in $\mathbf{u} = \mathbf{y}_t$: $f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t) \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$.

(ii) Plugging in $\mathbf{u} = \mathbf{x}^\star$: $f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star) \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$.

RHS of $(\lambda_t - 1)$(i) + (i) equals:

$$(\lambda_t - 1)\left(\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2\right) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

$$= \langle \nabla f(\mathbf{x}_t), \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star \rangle - \frac{\lambda_t}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

That is

$$\lambda_t \delta_{t+1} - (\lambda_t - 1)\delta_t \le \langle \nabla f(\mathbf{x}_t), \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star \rangle - \frac{\lambda_t}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

# Proof of AGD Convergence

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

***Proof:*** (continued proving Theorem 3)

$$\lambda_t \delta_{t+1} - (\lambda_t - 1)\delta_t \leq \langle \nabla f(\mathbf{x}_t), \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star \rangle - \frac{\lambda_t}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

$$\Rightarrow \lambda_t^2 \delta_{t+1} - \lambda_t(\lambda_t - 1)\delta_t \leq \frac{1}{2L}\left(2\langle \lambda_t \nabla f(\mathbf{x}_t), L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star)\rangle - \|\lambda_t \nabla f(\mathbf{x}_t)\|^2\right)$$

***Requirement (1):*** $\lambda_t(\lambda_t - 1) = \lambda_{t-1}^2$

$$\Rightarrow \lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{1}{2L}\left(2\langle \lambda_t \nabla f(\mathbf{x}_t), L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star)\rangle - \|\lambda_t \nabla f(\mathbf{x}_t)\|^2\right)$$

Denote by $\boldsymbol{a} \triangleq \lambda_t \nabla f(\mathbf{x}_t), \boldsymbol{b} \triangleq L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star)$.

$$\Rightarrow \lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{1}{2L}(2\langle \boldsymbol{a}, \boldsymbol{b}\rangle - \|\boldsymbol{b}\|^2) \leq \frac{1}{2L}(\|\boldsymbol{b}\|^2 - \|\boldsymbol{b} - \boldsymbol{a}\|^2)$$

# Proof of AGD Convergence

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

***Proof:*** (continued proving Theorem 3)

Denote by $\boldsymbol{a} \triangleq \lambda_t \nabla f(\mathbf{x}_t), \boldsymbol{b} \triangleq L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star)$.

$$\begin{aligned}
&\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \\
&\leq \frac{1}{2L}\left(L^2\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2 - \|L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star) - \lambda_t \nabla f(\mathbf{x}_t)\|^2\right) \\
&= \frac{L}{2}\left(\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2 - \left\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star - \lambda_t \frac{\nabla f(\mathbf{x}_t)}{L}\right\|^2\right) \\
&= \frac{L}{2}\left(\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2 - \|\lambda_t \mathbf{y}_{t+1} - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2\right)
\end{aligned}$$

***Goal: design a telescoping series***

# Proof of AGD Convergence

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

***Proof:*** (continued proving Theorem 3)

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{L}{2}\left(\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2 - \|\lambda_t \mathbf{y}_{t+1} - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2\right)$$

***Requirement (2):*** $\lambda_t \mathbf{y}_{t+1} - (\lambda_t - 1)\mathbf{y}_t = \lambda_{t+1}\mathbf{x}_{t+1} - (\lambda_{t+1} - 1)\mathbf{y}_{t+1}$

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{L}{2}\left(\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star\|^2 - \|\lambda_{t+1}\mathbf{x}_{t+1} - (\lambda_{t+1} - 1)\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2\right)$$

***telescope***

Define $\mathbf{z}_t \triangleq \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^\star$

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{L}{2}\left(\|\mathbf{z}_t\|^2 - \|\mathbf{z}_{t+1}\|^2\right) \Rightarrow \lambda_{T-1}^2 \delta_T - \lambda_0^2 \delta_1 = \frac{L}{2}\left(\|\mathbf{z}_1\|^2 - \|\mathbf{z}_T\|^2\right)$$

# Proof of AGD Convergence

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

***Proof:*** (continued proving Theorem 3)

$$\lambda_{T-1}^2 \delta_T - \lambda_0^2 \delta_1 = \frac{L}{2}(\|\mathbf{z}_1\|^2 - \|\mathbf{z}_T\|^2)$$

***Requirement (3):*** $\lambda_0 = 0$

$$\lambda_{T-1}^2 \delta_T \leq \frac{L}{2}\|\mathbf{z}_1\|^2 \Rightarrow \delta_T \leq \frac{L\|\mathbf{z}_1\|^2}{2\lambda_{T-1}^2} = \frac{L\|\lambda_1 \mathbf{x}_1 - (\lambda_1 - 1)\mathbf{y}_1 - \mathbf{x}^\star\|^2}{2\lambda_{T-1}^2}$$

***Requirement (4):*** $\mathbf{x}_1 = \mathbf{y}_1$

$$\lambda_{T-1}^2 \delta_T \leq \frac{L}{2}\|\mathbf{z}_1\|^2 \Rightarrow \delta_T \leq \frac{L\|\mathbf{z}_1\|^2}{2\lambda_{T-1}^2} = \frac{L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\lambda_{T-1}^2}$$

# Proof

**Theorem 3.** *Let $f$ be convex and $L$-smooth. Nesterov's accelerated GD is configured as*

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t,$$

*where* $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$, *and* $\alpha_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. *Then, we have*

$$f(\mathbf{y}_T) - f(\mathbf{x}^\star) \le \frac{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right).$$

***Proof:*** (continued proving Theorem 3)

***Requirement (1):*** $\lambda_t(\lambda_t - 1) = \lambda_{t-1}^2$

$$\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$$

***Requirement (2):*** $\lambda_t \mathbf{y}_{t+1} - (\lambda_t - 1)\mathbf{y}_t = \lambda_{t+1}\mathbf{x}_{t+1} - (\lambda_{t+1} - 1)\mathbf{y}_{t+1}$

$$\mathbf{x}_{t+1} = \mathbf{y}_{t+1} - \frac{1-\lambda_t}{\lambda_{t+1}}(\mathbf{y}_t - \mathbf{y}_{t+1}) \Rightarrow \alpha_t = \frac{1-\lambda_t}{\lambda_{t+1}}$$

***Requirement (3):*** $\lambda_0 = 0$

***Requirement (4):*** $\mathbf{x}_1 = \mathbf{y}_1$

$$\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2} \Rightarrow \lambda_t \ge \frac{t+1}{2} \Rightarrow \delta_T \le \frac{L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{2\lambda_{T-1}^2} \le \frac{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right) \quad \square$$

# Smooth and Strongly Convex

**Theorem 4.** *Let $f$ be $\sigma$-strongly convex and $L$-smooth, then Nesterov's accelerated gradient descent:*

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}(\mathbf{y}_{t+1} - \mathbf{y}_t)$$

*satisfies*

$$f(\mathbf{y}_T) - f(\mathbf{x}^\star) \leq \frac{\sigma + L}{2}\|\mathbf{x}^\star - \mathbf{x}_1\|^2 \exp\left(-\frac{T}{\sqrt{\gamma}}\right),$$

*where $\gamma \triangleq L/\sigma$ denotes the condition number.*

***core technique: estimate sequence***

# Smooth and Strongly Convex

- Proof sketch

*Core technique:* construct an estimate sequence *(developed by Yurii Nesterov)*

$$\Phi_1(\mathbf{x}) \triangleq f(\mathbf{x}_1) + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2$$

$$\Phi_{t+1}(\mathbf{x}) \triangleq (1-\theta)\Phi_t(\mathbf{x}) + \theta\left(f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_t\|^2\right)$$

The estimate sequence $\{\Phi_t\}_{t=1}^T$ is required to satisify some nice properties:

*(i)* $\Phi_{t+1}(\mathbf{x}) - f(\mathbf{x}) \leq (1-\theta)^t(\Phi_1(\mathbf{x}) - f(\mathbf{x})) \Rightarrow$ approximate $f$ well.

*(ii)* $f(\mathbf{y}_t) \leq \min_{\mathbf{x}\in\mathbb{R}^d} \Phi_t(\mathbf{x}) \Rightarrow$ useful when giving the convergence rate.

*It can be proven that the above construction satisfies the two properties.*

# Smooth and Strongly Convex

- Proof sketch

*Core technique:* construct an estimate sequence *(developed by Yurii Nesterov)*

$$\Phi_1(\mathbf{x}) \triangleq f(\mathbf{x}_1) + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_1\|^2$$

$$\Phi_{t+1}(\mathbf{x}) \triangleq (1-\theta)\Phi_t(\mathbf{x}) + \theta\left(f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_t\|^2\right)$$

$$f(\mathbf{y}_t) - f(\mathbf{x}^\star) \overset{(ii)}{\leq} \min_{\mathbf{x}\in\mathbb{R}^d} \Phi_t(\mathbf{x}) - f(\mathbf{x}^\star) \leq \Phi_t(\mathbf{x}^\star) - f(\mathbf{x}^\star) \qquad \text{(by property } (ii))$$

$$\overset{(i)}{\leq} (1-\theta)^t(\Phi_1(\mathbf{x}^\star) - f(\mathbf{x}^\star)) \qquad \text{(by property } (i))$$

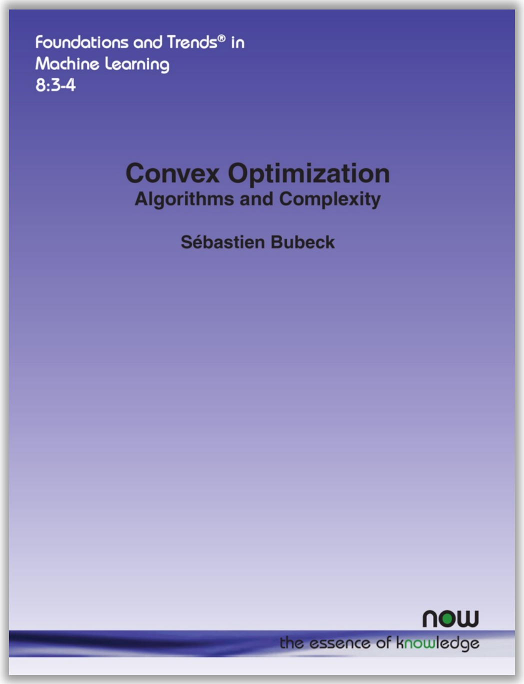
$$= (1-\theta)^t\left(f(\mathbf{x}_1) + \frac{\sigma}{2}\|\mathbf{x}^\star - \mathbf{x}_1\|^2 - f(\mathbf{x}^\star)\right) \qquad \text{(definition of } \Phi_1)$$

$$\lesssim (\sigma + L)\|\mathbf{x}^\star - \mathbf{x}_1\|^2 \exp(-\theta t) \qquad \text{(smoothness)}$$

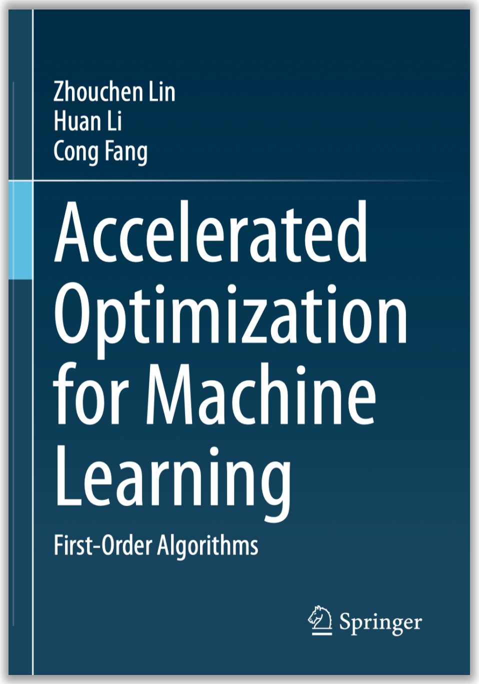
# Estimate Sequence

- Admittedly, how to construct estimate sequence is highly ***tricky***

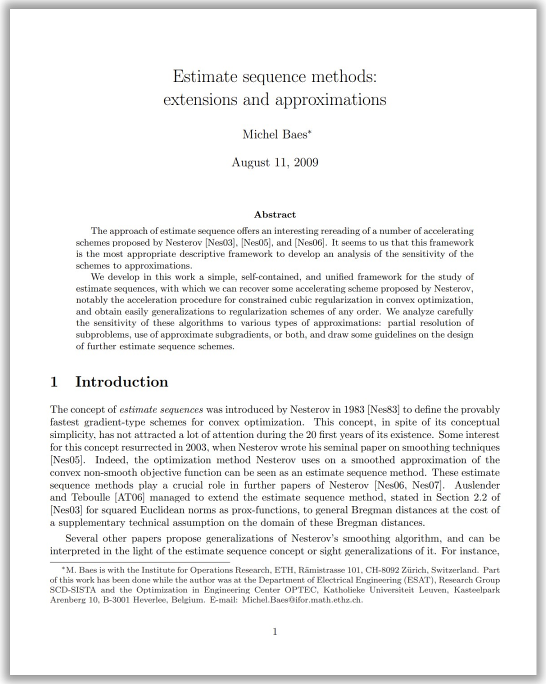
***References:***




Chapter 3.7




Chapter 2.1



Estimate sequence methods:
extensions and approximations

Michel Baes*

August 11, 2009

**Abstract**

The approach of estimate sequence offers an interesting rereading of a number of accelerating schemes proposed by Nesterov [Nes03], [Nes05], and [Nes06]. It seems to us that this framework is the most appropriate descriptive framework to develop an analysis of the sensitivity of the schemes to approximations.

We develop in this work a simple, self-contained, and unified framework for the study of estimate sequences, with which we can recover some accelerating scheme proposed by Nesterov, notably the acceleration procedure for constrained cubic regularization in convex optimization, and obtain easily generalizations to regularization schemes of any order. We analyze carefully the sensitivity of these algorithms to various types of approximations: partial resolution of subproblems, use of approximate subgradients, or both, and draw some guidelines on the design of further estimate sequence schemes.

**1 Introduction**

The concept of *estimate sequences* was introduced by Nesterov in 1983 [Nes83] to define the provably fastest gradient-type schemes for convex optimization. This concept, in spite of its conceptual simplicity, has not attracted a lot of attention during the 20 first years of its existence. Some interest for this concept resurrected in 2003, when Nesterov wrote his seminal paper on smoothing techniques [Nes05]. Indeed, the optimization method Nesterov uses on a smoothed approximation of the convex non-smooth objective function can be seen as an estimate sequence method. These estimate sequence methods play a crucial role in further papers of Nesterov [Nes06, Nes07]. Auslender and Teboulle [AT06] managed to extend the estimate sequence method, stated in Section 2.2 of [Nes03] for squared Euclidean norms as prox-functions, to general Bregman distances at the cost of a supplementary technical assumption on the domain of these Bregman distances.

Several other papers propose generalizations of Nesterov's smoothing algorithm, and can be interpreted in the light of the estimate sequence concept or sight generalizations of it. For instance,

*M. Baes is with the Institute for Operations Research, ETH, Rämistrasse 101, CH-8092 Zürich, Switzerland. Part of this work has been done while the author was at the Department of Electrical Engineering (ESAT), Research Group SCD-SISTA and the Optimization in Engineering Center OPTEC, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium. E-mail: Michel.Baes@ifor.math.ethz.ch.

1

M. Baes, Estimate sequence methods: extensions and approximations. Technical report, ETH, Zürich (2009)

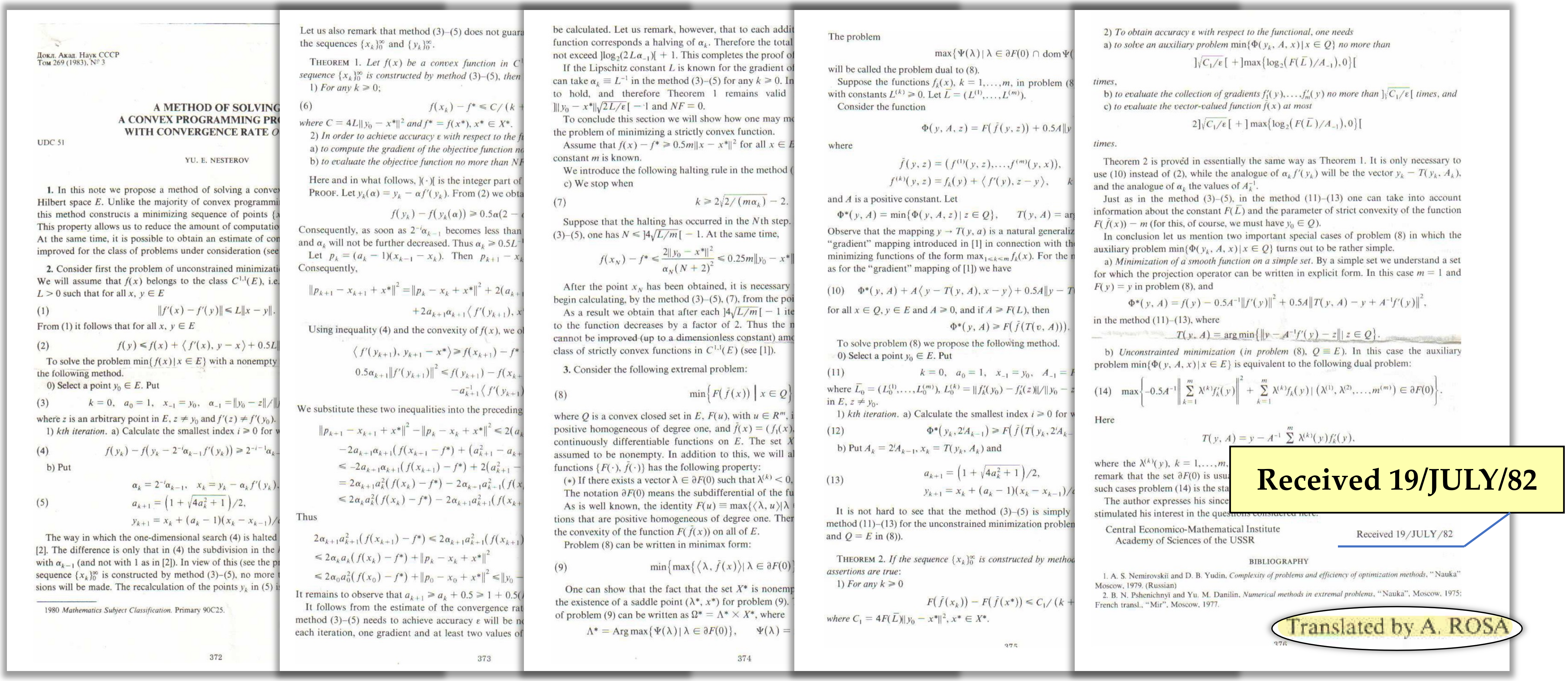
# References for Nesterov's Accelerated GD

Nesterov's four ideas (three acceleration methods):

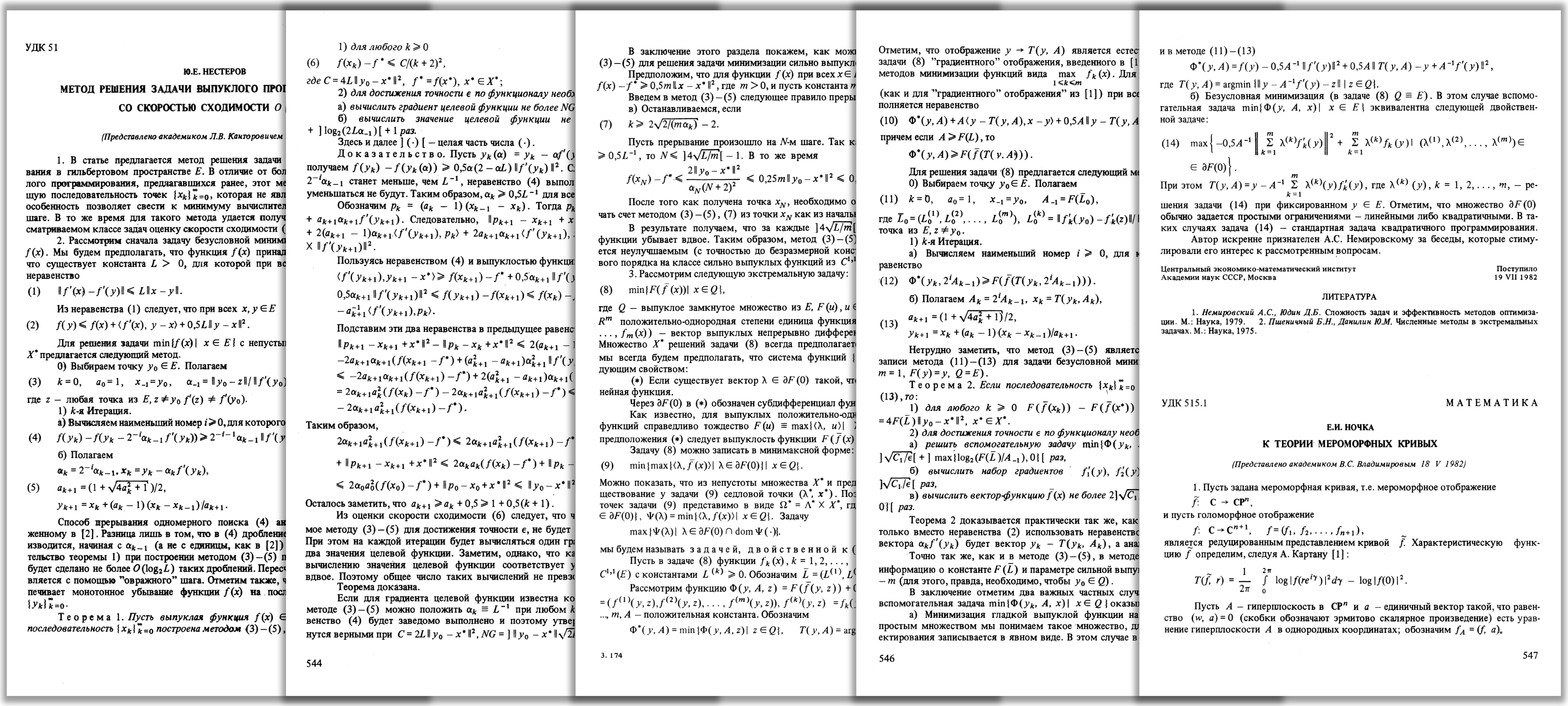
- Y. Nesterov (1983), A method for solving a convex programming problem with convergence rate $O(1/k^2)$
- Y. Nesterov (1988), On an approach to the construction of optimal methods of minimization of smooth convex functions
- Y. Nesterov (2005), Smooth minimization of non-smooth functions
- Y. Nesterov (2007), Gradient methods for minimizing composite objective function

**Yurii Nesterov**
1956 –
UCLouvain, Belgium

# Nesterov, Y. (1983), A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Soviet Mathematics Doklady 27(2), 372–376.



Nesterov, Y. (1983), A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Soviet Mathematics Doklady 27(2), 372–376.


УДК 51

Ю.Е. НЕСТЕРОВ

МЕТОД РЕШЕНИЯ ЗАДАЧИ ВЫПУКЛОГО ПРО[…]

СО СКОРОСТЬЮ СХОДИМОСТИ $O$[…]

(Представлено академиком Л.В. Канторовичем […])

[…]

и в методе (11) – (13)

$\Phi^*(y, A) = f(y) - 0{,}5A^{-1}\|f'(y)\|^2 + 0{,}5A\|T(y, A) - y + A^{-1}f'(y)\|^2$,

где $T(y, A) = \operatorname{argmin}\{\|y - A^{-1}f'(y) - z\| \mid z \in Q\}$.

б) Безусловная минимизация (в задаче (8) $Q \equiv E$). В этом случае вспомогательная задача $\min\{\Phi(y, A, x) \mid x \in E\}$ эквивалентна следующей двойственной задаче:

(14) $\max\left\{-0{,}5A^{-1}\left\|\sum_{k=1}^{m}\lambda^{(k)}f_k'(y)\right\|^2 + \sum_{k=1}^{m}\lambda^{(k)}f_k(y) \mid (\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(m)}) \in \partial F(0)\right\}$.

При этом $T(y, A) = y - A^{-1}\sum_{k=1}^{m}\lambda^{(k)}(y)f_k'(y)$, где $\lambda^{(k)}(y)$, $k = 1, 2, \ldots, m$, – решения задачи (14) при фиксированном $y \in E$. Отметим, что множество $\partial F(0)$ обычно задается простыми ограничениями – линейными либо квадратичными. В таких случаях задача (14) – стандартная задача квадратичного программирования.

Автор искренне признателен А.С. Немировскому за беседы, которые стимулировали его интерес к рассмотренным вопросам.

Центральный экономико-математический институт Академии наук СССР, Москва

Поступило 19 VII 1982

ЛИТЕРАТУРА

1. Немировский А.С., Юдин Д.Б. Сложность задач и эффективность методов оптимизации. М.: Наука, 1979. 2. Пшеничный Б.Н., Данилин Ю.М. Численные методы в экстремальных задачах. М.: Наука, 1975.

УДК 515.1 МАТЕМАТИКА

Е.И. НОЧКА

К ТЕОРИИ МЕРОМОРФНЫХ КРИВЫХ

(Представлено академиком В.С. Владимировым 18 V 1982)

1. Пусть задана мероморфная кривая, т.е. мероморфное отображение

$\tilde{f}$: C → CP$^n$,

и пусть голоморфное отображение

$f$: C → C$^{n+1}$, $f = (f_1, f_2, \ldots, f_{n+1})$,

является редуцированным представлением кривой $\tilde{f}$. Характеристическую функцию $\tilde{f}$ определим, следуя А. Картану [1]:

$T(\tilde{f}, r) = \frac{1}{2\pi}\int_0^{2\pi}\log|f(re^{i\gamma})|^2 d\gamma - \log|f(0)|^2$.

Пусть $A$ – гиперплоскость в CP$^n$ и $a$ – единичный вектор такой, что равенство $(w, a) = 0$ (скобки обозначают эрмитово скалярное произведение) есть уравнение гиперплоскости $A$ в однородных координатах; обозначим $f_A = (f, a)$.

547

# More Explanations for Nesterov's AGD

- Ordinary Differentiable Equations
  - Su, W., Boyd, S., & Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. Advances in neural information processing systems 27 (NIPS).
  - Berthier, R., Bach, F., Flammarion, N., Gaillard, P., & Taylor, A. (2021). A continuized view on Nesterov acceleration. arXiv preprint arXiv:2102.06035.
- Variational Analysis
  - Wibisono, A., Wilson, A. C., & Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences (PNAS), 113(47), E7351-E7358.
- Linear Coupling of GD and MD
  - Allen-Zhu, Z., & Orecchia, L. (2017). Linear coupling: An ultimate unification of gradient and mirror descent. The 8th Innovations in Theoretical Computer Science Conference (ITCS).

# Part 3. Extension to Composite Optimization

- Composite Optimization

- Proximal Gradient Method (PG)

- Accelerated Proximal Gradient Method (APG)

- Application to LASSO

# Composite Optimization

- Problem setup

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{x})$$

where $f$ is ***smooth*** (namely, gradient Lipschitz) while $h$ is ***not smooth***.

- The composite optimization problem is common in practice.

**Example 1.** The objective of ***LASSO***: $F(\mathbf{w}) = \frac{1}{2} \left\| \mathbf{w}^\top X - \mathbf{y} \right\|_2^2 + \lambda \left\| \mathbf{w} \right\|_1$, where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n], \mathbf{y} = [y_1, \ldots, y_n]^\top$.

# Recall Non-composite Optimization

- Consider $\min_{\mathbf{x}} f(\mathbf{x})$, and assume $f$ is $L$-smooth.

$$\text{By smoothness: } f(\mathbf{x}) \le \underbrace{f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2}_{\triangleq Q(\mathbf{x};\mathbf{y})}$$

*surrogate objective*

$\Longrightarrow$ to minimize $f(\mathbf{x})$, it suffices to minimize the ***surrogate*** objective $Q(\mathbf{x}; \mathbf{y})$.

**Claim.** GD for smooth functions can be equivalently represented by

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} Q(\mathbf{x}; \mathbf{x}_t) = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right],$$

where $Q(\mathbf{x}; \mathbf{x}_t) \triangleq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$ is a quadratic upper bound of $f$ at $\mathbf{x}_t$.

# Another View of GD Method

**Claim.** GD for smooth functions can be equivalently represented by

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} Q(\mathbf{x};\mathbf{x}_t) = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right],$$

where $Q(\mathbf{x};\mathbf{x}_t) \triangleq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}-\mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{x}_t\|^2$ is a quadratic upper bound of $f$ at $\mathbf{x}_t$.

***Proof:***

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} Q(\mathbf{x};\mathbf{x}_t) = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\langle \nabla f(\mathbf{x}_t), \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{x}\|^2 - L\langle \mathbf{x}, \mathbf{x}_t\rangle\right\} \text{(remove irrelative terms)}$$

$$= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\frac{L}{2}\left(-2\left\langle \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \mathbf{x}\right\rangle + \|\mathbf{x}\|^2\right)\right\} \quad \text{(rearrange)}$$

$$= \arg\min_{\mathbf{x}\in\mathcal{X}} \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\|^2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\| = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right]$$

□

# Another View of GD Method

**Claim.** GD for smooth functions can be equivalently represented by

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} Q(\mathbf{x};\mathbf{x}_t) = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right],$$

where $Q(\mathbf{x};\mathbf{x}_t) \triangleq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$ is a quadratic upper bound of $f$ at $\mathbf{x}_t$.

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} Q(\mathbf{x};\mathbf{x}_t) = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \boxed{f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle} + \boxed{\frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2} \right\}$$

*linear approximation of $f$ at $\mathbf{x}_t$*

*prevent $\mathbf{x}_t$ from getting too far*

# Composite Optimization

- Problem setup

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{x})$$

where $f$ is ***smooth*** (namely, gradient Lipschitz) while $h$ is ***not smooth***.

***An idea:***

Following previous argument (for non-composite optimization), to minimize $F \triangleq f + h$, it suffices to minimize

$$Q(\mathbf{x};\mathbf{x}_t) \triangleq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 + h(x)$$

# Formulation

By smoothness: $f(\mathbf{x}) \leq \underbrace{f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2}_{\triangleq q(\mathbf{x};\mathbf{y})}$

*surrogate objective*

$\Longrightarrow$ to minimize $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$, it suffices to minimize $Q(\mathbf{x};\mathbf{y}) \triangleq q(\mathbf{x};\mathbf{y}) + h(\mathbf{x})$.

$$\begin{aligned} \arg\min_{\mathbf{x}} Q(\mathbf{x};\mathbf{y}) &= \arg\min_{\mathbf{x}} \left\{ f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + h(\mathbf{x}) \right\} \\ &= \arg\min_{\mathbf{x}} \left\{ \langle \nabla f(\mathbf{y}), \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x}\|^2 - L\langle \mathbf{x}, \mathbf{y} \rangle + h(\mathbf{x}) \right\} \\ &= \arg\min_{\mathbf{x}} \left\{ \frac{L}{2}\left( -2\left\langle \mathbf{y} - \frac{\nabla f(\mathbf{y})}{L}, \mathbf{x} \right\rangle + \|\mathbf{x}\|^2 \right) + h(\mathbf{x}) \right\} \end{aligned}$$

# Composite Optimization

By smoothness: $f(\mathbf{x}) \leq \underbrace{f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2}_{\triangleq q(\mathbf{x};\mathbf{y})}$

*surrogate objective*

$\Rightarrow$ to minimize $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$, it suffices to minimize $Q(\mathbf{x};\mathbf{y}) \triangleq q(\mathbf{x};\mathbf{y}) + h(\mathbf{x})$.

$$\arg\min_{\mathbf{x}} Q(\mathbf{x};\mathbf{y}) = \arg\min_{\mathbf{x}} \left\{ \frac{L}{2} \left( -2 \left\langle \mathbf{y} - \frac{\nabla f(\mathbf{y})}{L}, \mathbf{x} \right\rangle + \|\mathbf{x}\|^2 \right) + h(\mathbf{x}) \right\}$$

$$= \arg\min_{\mathbf{x}} \left\{ \frac{L}{2} \left\| \mathbf{x} - \left( \mathbf{y} - \frac{\nabla f(\mathbf{y})}{L} \right) \right\|^2 + h(\mathbf{x}) \right\}$$

*an operator is defined for this (sub-)optimization problem*

# Composite Optimization

- Iteratively solve the surrogate optimization problem.

Deploying the following update rule:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x};\mathbf{x}_t) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{\frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\|^2 + h(\mathbf{x})\right\}$$

**Definition 2** (proximal mapping). Given a function $h : \mathbb{R}^d \mapsto \mathbb{R}$, the *proximal mapping* (or called *proximal operator*) of $h$ is the operator given by

$$\mathbf{prox}_h(\mathbf{x}) \triangleq \arg\min_{\mathbf{u}\in\mathbb{R}^d} \left\{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2\right\} \text{ for any } \mathbf{x} \in \mathbb{R}^d.$$

# Proximal Gradient

**Definition 2** (proximal mapping). Given a function $h : \mathbb{R}^d \mapsto \mathbb{R}$, the *proximal mapping* (or called *proximal operator*) of $h$ is the operator given by

$$\mathbf{prox}_h(\mathbf{x}) \triangleq \arg\min_{\mathbf{u}\in\mathbb{R}^d} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} \text{ for any } \mathbf{x} \in \mathbb{R}^d.$$

**Proximal Gradient Method.**

$$\begin{aligned}\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t) &\triangleq \mathbf{prox}_{\frac{1}{L}h}\left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right) \\ &= \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\|^2 + h(\mathbf{x}) \right\}.\end{aligned}$$

# Proximal Gradient

**Proximal Gradient Method.**

$$\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t) \triangleq \mathbf{prox}_{\frac{1}{L}h}\left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)$$
$$= \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min}\left\{\frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\|^2 + h(\mathbf{x})\right\}.$$

- In LASSO, where $h(\mathbf{x}) = \|\mathbf{x}\|_1$, $\mathcal{P}_L^h$ is easy to compute and has closed form solution.
- Algorithmically, PG induces famous algorithms for solving LASSO problem, which are called **ISTA** (GD-type) and **FISTA** (Nesterov's AGD-type).

# Convergence of Proximal Gradient

***Smooth Optimization***

problem: $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$

assumption: $f$ is $L$-smooth

GD: $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$

Convergence: $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \mathcal{O}\left(\frac{1}{T}\right)$

***Smooth Composite Optimization***

problem: $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{x})$

assumption: $f$ is $L$-smooth, $h$ not

PG: $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t)$

Convergence: $F(\mathbf{x}_T) - F(\mathbf{x}^\star) \leq ?$

# Convergence of Proximal Gradient

**Theorem 5.** *Suppose that $f$ and $h$ are convex and $f$ is $L$-smooth. Setting the parameters properly, Proximal Gradient (PG) enjoys*

$$F(\mathbf{x}_T) - F(\mathbf{x}^\star) \le \frac{L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2(T-1)} = \mathcal{O}\left(\frac{1}{T}\right)$$

Proximal gradient can also achieve an $\mathcal{O}(1/T)$ convergence rate, which is the ***same*** as the non-composite optimization counterpart.

The result can be further boosted to $\mathcal{O}(\exp(-T/\kappa))$ when the function $f$ is ***$\sigma$-strongly convex*** (where $\kappa = L/\sigma$ is the condition number).

# Convergence of Proximal Gradient

- Generalized one-step improvement lemma on $F \triangleq f + h$

**Lemma 7.** *Suppose that $f$ and $h$ are convex and $f$ is $L$-smooth. Let $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t)$ and $g(\mathbf{x}) \triangleq L(\mathbf{x} - \mathbf{x}_{t+1})$. Then for any $\mathbf{u} \in \mathcal{X}$,*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L} \|g(\mathbf{x}_t)\|^2.$$

Suppose the above lemma holds for a moment, we now prove the $\mathcal{O}(1/T)$ convergence rate of **PG**.

# Proof of PG Convergence

> **Lemma 7.** *Suppose that $f$ and $h$ are convex and $f$ is $L$-smooth. Let $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t)$ and $g(\mathbf{x}) \triangleq L(\mathbf{x} - \mathbf{x}_{t+1})$. Then for any $\mathbf{u} \in \mathcal{X}$,*
>
> $$F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \le \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2.$$

***Proof:***

Setting $\mathbf{u} = \mathbf{x}^\star$ in Lemma 7:

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^\star) \le \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$$

$$\Longrightarrow \quad F(\mathbf{x}_{t+1}) - F(\mathbf{x}^\star) \le L\langle \mathbf{x}_t - \mathbf{x}_{t+1}, \mathbf{x}_t - \mathbf{x}^\star \rangle - \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \quad (g(\mathbf{x}_t) \triangleq L(\mathbf{x}_t - \mathbf{x}_{t+1}))$$

$$= \frac{L}{2}\left(2\langle \mathbf{x}_t - \mathbf{x}_{t+1}, \mathbf{x}_t - \mathbf{x}^\star \rangle - \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\right)$$

$$= \frac{L}{2}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \quad (2\langle \boldsymbol{a}, \boldsymbol{b} \rangle - \|\boldsymbol{a}\|^2 = \|\boldsymbol{b}\|^2 - \|\boldsymbol{b} - \boldsymbol{a}\|^2)$$

$$\Longrightarrow \quad \sum_{t=1}^{T-1} F(\mathbf{x}_t) - (T-1)F(\mathbf{x}^\star) \le \tfrac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

# Proof of PG Convergence

***Proof:***

$$\Longrightarrow \quad \frac{1}{T-1}\sum_{t=1}^{T-1} F(\mathbf{x}_t) - F(\mathbf{x}^\star) \le \frac{L\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{2(T-1)}$$

which already gives an $\mathcal{O}(1/T)$ convergence rate of $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$.

*What we want:* $F(\mathbf{x}_T) - F(\mathbf{x}^\star)$

*Next step:* analyzing $F(\mathbf{x}_T) - \frac{1}{T-1}\sum_{t=1}^{T-1} F(\mathbf{x}_t)$.

Setting $\mathbf{u} = \mathbf{x}_t$ in Lemma 7: $F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \le -\frac{1}{2L}\|g(\mathbf{x}_t)\|^2 \le 0$.

$$\Longrightarrow \quad \sum_{t=1}^{T} t(F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)) \le 0$$

# Proof of PG Convergence

***Proof:***

***What we want:*** $F(\mathbf{x}_T) - F(\mathbf{x}^\star) \Rightarrow$ ***Next step:*** analyzing $F(\mathbf{x}_T) - \frac{1}{T-1}\sum_{t=1}^{T-1} F(\mathbf{x}_t)$.

$$\sum_{t=1}^{T-1} t(F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)) = \sum_{t=1}^{T-1} t(F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)) + F(\mathbf{x}_t) - F(\mathbf{x}_t)$$

$$= \sum_{t=1}^{T-1} \Big(tF(\mathbf{x}_{t+1}) - (t-1)F(\mathbf{x}_t)\Big) - \sum_{t=1}^{T-1} F(\mathbf{x}_t) = (T-1)F(\mathbf{x}_T) - \sum_{t=1}^{T-1} F(\mathbf{x}_t) \le 0$$

***What we have:***

- $F(\mathbf{x}_T) - \frac{1}{T-1}\sum_{t=1}^{T-1} F(\mathbf{x}_t) \le 0$
- $\frac{1}{T-1}\sum_{t=1}^{T-1} F(\mathbf{x}_t) - F(\mathbf{x}^\star) \le \frac{L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2(T-1)}$

$$\Longrightarrow \quad F(\mathbf{x}_T) - F(\mathbf{x}^\star) \le \frac{L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2(T-1)}$$

□

# Proof of One-Step Improvement Lemma

**Lemma 7.** *Suppose that $f$ and $h$ are convex and $f$ is $L$-smooth. Let $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t)$ and $g(\mathbf{x}_t) \triangleq L(\mathbf{x}_t - \mathbf{x}_{t+1})$. Then for any $\mathbf{u} \in \mathcal{X}$,*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2.$$

***Proof:*** ***What we have:*** $F(\mathbf{x}) \leq Q(\mathbf{x}; \mathbf{y})$ for any $\mathbf{y} \in \mathcal{X} \Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq Q(\mathbf{x}_{t+1}; \mathbf{x}_t) - F(\mathbf{u})$

*analyzing this quantity*

$$\begin{cases} F(\mathbf{u}) = f(\mathbf{u}) + h(\mathbf{u}) \geq \cancel{f(\mathbf{x}_t)} + \langle \nabla f(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_t \rangle + \cancel{h(\mathbf{x}_{t+1})} + \langle \nabla h(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1} \rangle & \text{(convexity)} \\ Q(\mathbf{x}_{t+1}; \mathbf{x}_t) = \cancel{f(\mathbf{x}_t)} + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 + \cancel{h(\mathbf{x}_{t+1})} \end{cases}$$

$$\Rightarrow Q(\mathbf{x}_{t+1}; \mathbf{x}_t) - F(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t) + \nabla h(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \underbrace{\frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2}_{= \frac{1}{2L}\|g(\mathbf{x}_t)\|^2 \ (g(\mathbf{x}_t) \triangleq L(\mathbf{x}_t - \mathbf{x}_{t+1}))}$$

***Next step:*** relate $\nabla f(\mathbf{x}_t) + \nabla h(\mathbf{x}_{t+1})$ to $g(\mathbf{x}_t)$.

# Proof of One-Step Improvement Lemma

***Proof:***

*What we have:* $F(\mathbf{x}) \leq Q(\mathbf{x};\mathbf{y})$ for any $\mathbf{y} \in \mathcal{X} \Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq Q(\mathbf{x}_{t+1};\mathbf{x}_t) - F(\mathbf{u})$

*analyzing this quantity*

$$\Longrightarrow Q(\mathbf{x}_{t+1};\mathbf{x}_t) - F(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t) + \nabla h(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{u}\rangle + \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$$

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\|^2 \right\}$$

$$\triangleq H(\mathbf{x})$$

$\Downarrow$ by *Fermat's optimality condition*

**Theorem 8** (Fermat's Optimality Condition). *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a proper convex function. Then*

$$\mathbf{x}^\star \in \operatorname{argmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^d\}$$

*if and only if $\mathbf{0} \in \partial f(\mathbf{x}^\star)$.*

$$\mathbf{0} = \nabla H(\mathbf{x}_{t+1}) = \nabla h(\mathbf{x}_{t+1}) + L(\mathbf{x}_{t+1} - \mathbf{x}_t) + \nabla f(\mathbf{x}_t)$$

# Proof of One-Step Improvement Lemma

***Proof:***

*What we have:* $F(\mathbf{x}) \leq Q(\mathbf{x};\mathbf{y})$ for any $\mathbf{y} \in \mathcal{X} \Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq Q(\mathbf{x}_{t+1};\mathbf{x}_t) - F(\mathbf{u})$

*analyzing this quantity*

$$\Longrightarrow Q(\mathbf{x}_{t+1};\mathbf{x}_t) - F(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t) + \nabla h(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$$

$$\Longrightarrow \mathbf{0} = \nabla H(\mathbf{x}_{t+1}) = \nabla h(\mathbf{x}_{t+1}) + L(\mathbf{x}_{t+1} - \mathbf{x}_t) + \nabla f(\mathbf{x}_t)$$

$$\Longrightarrow g(\mathbf{x}_t) = L(\mathbf{x}_t - \mathbf{x}_{t+1}) = \nabla h(\mathbf{x}_{t+1}) + \nabla f(\mathbf{x}_t)$$

Finally we have $Q(\mathbf{x}_{t+1};\mathbf{x}_t) - F(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$

$$= \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2 \quad \square$$

# One-Step Improvement Lemma

- A ***fundamental*** result for GD of smoothed optimization.

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

$$f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

$$f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$$

$$F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$$

*specialized* ↓ *general*

***Corollary:*** the proof of **PG** can also be used to prove the $\mathcal{O}(1/T)$ convergence rate of GD.

# Accelerated Proximal Gradient Method

- A natural idea

Can we extend the Nesterov's AGD to the composite optimization?

⇨ This induces the Accelerated Proximal Gradient (**APG**) method.

**Nesterov's Accelerated GD**

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

**Accelerated Proximal Gradient**

$$\mathbf{y}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1-\alpha_t)\mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

The covergence rates can be similarly obtained. *Proofs are omitted.*

# Accelerated Proximal Gradient Method

**Theorem 6.** *Suppose that $f$ and $h$ are convex and $f$ is $L$-smooth. Setting the parameters properly, APG enjoys*

$$F(\mathbf{x}_T) - F(\mathbf{x}^\star) \leq \frac{2L}{(T+1)^2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

*Suppose that $h$ is convex and $f$ is $\sigma$-strongly convex and $L$-smooth. Setting the parameters properly, APG enjoys*

$$F(\mathbf{x}_T) - F(\mathbf{x}^\star) \leq \exp\left(-\frac{T}{\sqrt{\kappa}}\right)\left(F(\mathbf{x}_0) - F(\mathbf{x}^\star) + \frac{\sigma}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\right),$$

*where $\kappa \triangleq L/\sigma$ denotes the condition number.*

***The convergence rates can be obtained same as those in non-composite optimization.***

# Application to LASSO

- **LASSO:** $\ell_1$-regularized least squares

$$F(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{w}^\top X - \mathbf{y}\right\|^2 + \lambda\left\|\mathbf{w}\right\|_1$$

commonly encountered in ***signal/image processing***.

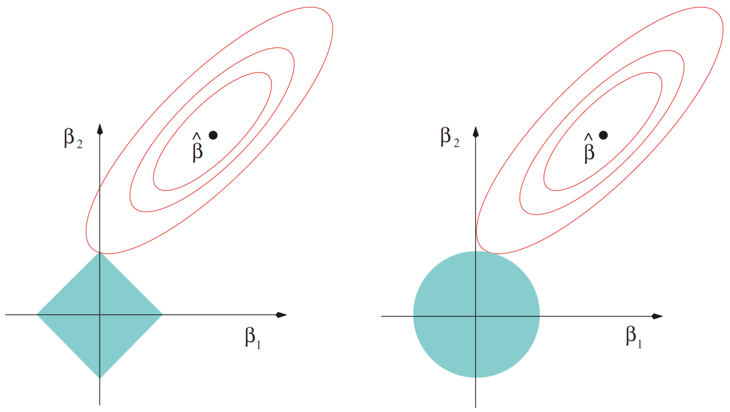
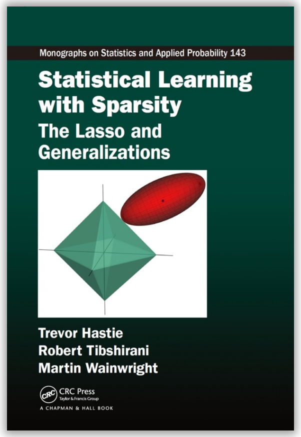
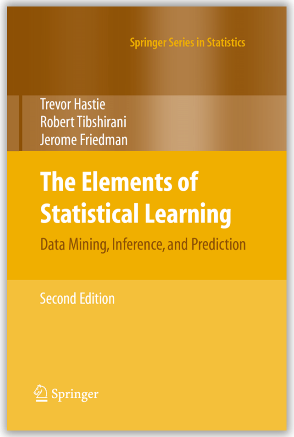



FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*





Regression shrinkage and selection via the lasso — 47812 — 1996
R Tibshirani
Journal of the Royal Statistical Society. Series B (Methodological), 267-288

# Application to LASSO

- **LASSO:** $\ell_1$-regularized least squares

$$F(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{w}^\top X - \mathbf{y}\right\|^2 + \lambda \left\|\mathbf{w}\right\|_1$$

commonly encountered in ***signal/image processing***.

$\Rightarrow$ **composite optimization**: first part is *smooth*, the other one is *non-smooth*

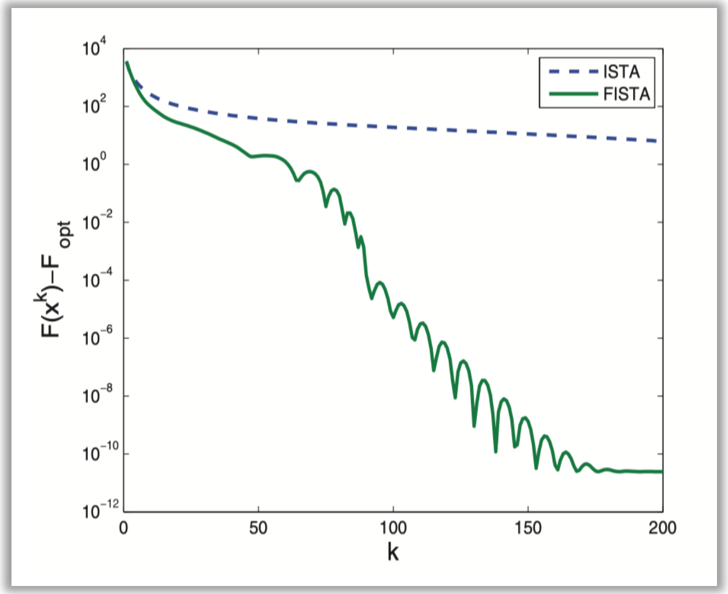
- **ISTA** (**I**terative **S**hrinkage-**T**hresholding **A**lgorithm): **PG** for LASSO
- **FISTA** (**F**ast **ISTA**): **APG** for LASSO

Closed-form solution: $(x_+ \triangleq \max\{x, 0\})$

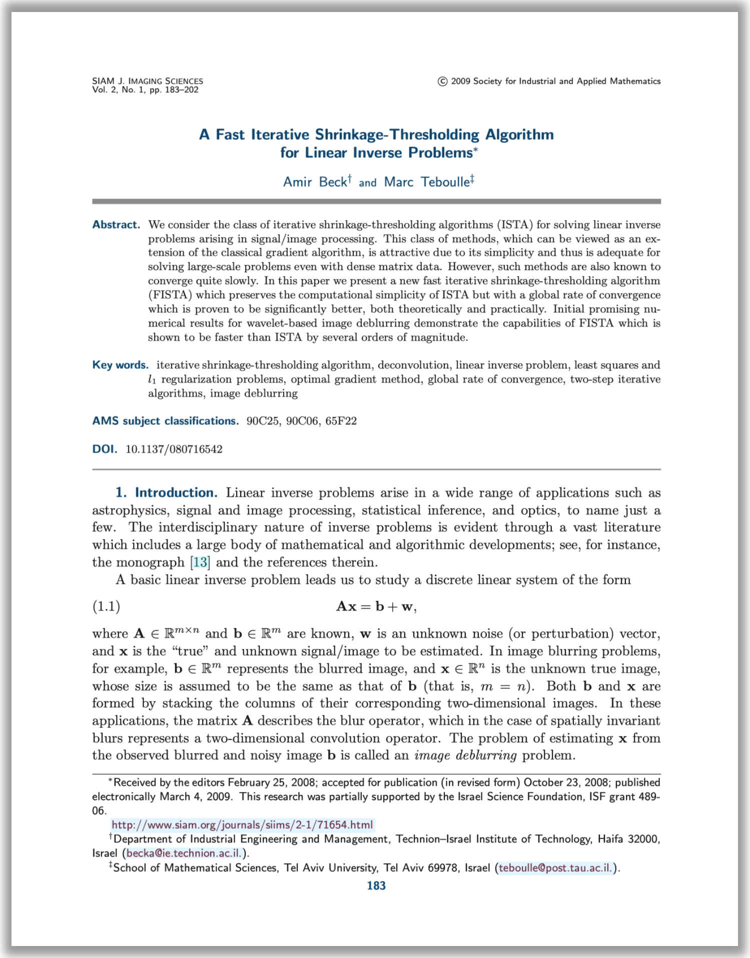
$$[\mathcal{P}_L^h(\mathbf{w}_t)]_i = \text{sign}\left(\left[\mathbf{w}_t - \frac{1}{L}\nabla f(\mathbf{w}_t)\right]_i\right)\left(\left|\left[\mathbf{w}_t - \frac{1}{L}\nabla f(\mathbf{w}_t)\right]_i\right| - \frac{\lambda}{L}\right)_+$$

# Application to LASSO

- Comparison of **ISTA** and **FISTA**




Comparison of **ISTA** and **FISTA**.



SIAM J. IMAGING SCIENCES
Vol. 2, No. 1, pp. 183–202

© 2009 Society for Industrial and Applied Mathematics

**A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems***

Amir Beck† and Marc Teboulle‡

**Abstract.** We consider the class of iterative shrinkage-thresholding algorithms (ISTA) for solving linear inverse problems arising in signal/image processing. This class of methods, which can be viewed as an extension of the classical gradient algorithm, is attractive due to its simplicity and thus is adequate for solving large-scale problems even with dense matrix data. However, such methods are also known to converge quite slowly. In this paper we present a new fast iterative shrinkage-thresholding algorithm (FISTA) which preserves the computational simplicity of ISTA but with a global rate of convergence which is proven to be significantly better, both theoretically and practically. Initial promising numerical results for wavelet-based image deblurring demonstrate the capabilities of FISTA which is shown to be faster than ISTA by several orders of magnitude.

**Key words.** iterative shrinkage-thresholding algorithm, deconvolution, linear inverse problem, least squares and $l_1$ regularization problems, optimal gradient method, global rate of convergence, two-step iterative algorithms, image deblurring

**AMS subject classifications.** 90C25, 90C06, 65F22

**DOI.** 10.1137/080716542

**1. Introduction.** Linear inverse problems arise in a wide range of applications such as astrophysics, signal and image processing, statistical inference, and optics, to name just a few. The interdisciplinary nature of inverse problems is evident through a vast literature which includes a large body of mathematical and algorithmic developments; see, for instance, the monograph [13] and the references therein.

A basic linear inverse problem leads us to study a discrete linear system of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{w}, \tag{1.1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are known, $\mathbf{w}$ is an unknown noise (or perturbation) vector, and $\mathbf{x}$ is the "true" and unknown signal/image to be estimated. In image blurring problems, for example, $\mathbf{b} \in \mathbb{R}^m$ represents the blurred image, and $\mathbf{x} \in \mathbb{R}^n$ is the unknown true image, whose size is assumed to be the same as that of $\mathbf{b}$ (that is, $m = n$). Both $\mathbf{b}$ and $\mathbf{x}$ are formed by stacking the columns of their corresponding two-dimensional images. In these applications, the matrix $\mathbf{A}$ describes the blur operator, which in the case of spatially invariant blurs represents a two-dimensional convolution operator. The problem of estimating $\mathbf{x}$ from the observed blurred and noisy image $\mathbf{b}$ is called an *image deblurring* problem.

*Received by the editors February 25, 2008; accepted for publication (in revised form) October 23, 2008; published electronically March 4, 2009. This research was partially supported by the Israel Science Foundation, ISF grant 489-06.
http://www.siam.org/journals/siims/2-1/71654.html
†Department of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel (becka@ie.technion.ac.il).
‡School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (teboulle@post.tau.ac.il).

183

| | | |
|---|---|---|
| A fast iterative shrinkage-thresholding algorithm for linear inverse problems<br>A Beck, M Teboulle<br>SIAM journal on imaging sciences 2 (1), 183-202 | 11526 | 2009 |

# Summary

- **GD FOR SMOOTH OPTIMIZATION**
  - Smooth and Convex
  - Smooth and Strongly Convex
  - Constrained Optimization
- **NESTEROV'S ACCELERATED GD**
  - Algorithm
  - Smooth and Convex
  - Smooth and Strongly Convex
- **EXTENSION TO COMPOSITE OPTIMIZATION**
  - Composite Optimization
  - Proximal Gradient Method (PG)
  - Accelerated Proximal Gradient Method (APG)
  - Application to LASSO