# Lecture 4. Gradient Descent Method II

Advanced Optimization (Fall 2024)

**Peng Zhao**

zhaop@lamda.nju.edu.cn

Nanjing University

# Outline

- GD for Smooth Optimization
  - Smooth and Convex Functions
  - Smooth and Strongly Convex Functions
- Momentum and Acceleration
  - Polyak's Momentum
  - Nesterov's Accelerated GD
- Extension to Composite Optimization
  - Proximal Gradient and Accelerated One

# 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:*

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

**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:***

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

□

# Refined Result for Smooth Optimization

***Proof:***

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

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

*only exploited convexity, but haven't used smoothness*

# Refined Result for Smooth Optimization

- Recall the first-order characterization of smooth functions

## Smoothness

**Theorem 2** (*First-order* Characterizations of $L$-smoothness). *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function, differentiable over $\mathcal{X}$. Then the following claims are equivalent:*

*(i)* *$f$ is $L$-smooth.*

*(ii)* $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

*(iii)* $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

*(iv)* $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

*(v)* $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{L}{2}\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2$ *for any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ *and* $\lambda \in [0, 1]$.

*Proofs can be found below Theorem 5.8 of Amir Beck's book.*

Advanced Optimization (Fall 2024) Lecture 2. Convex Optimization Basics 57

*co-coerciveity*

# 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***.

# Refined Result for Smooth Optimization

***Proof:***

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

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

*only exploited convexity, but haven't used smoothness*

# Refined Result for Smooth Optimization

***Proof:***

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

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

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

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

$$\le \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \tfrac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 \qquad \text{(by picking } \eta_t=\eta=\tfrac{1}{L} \text{ to minimize the r.h.s)}$$

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

# Smooth and Convex

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

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

$$\begin{aligned}
& f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \qquad \textbf{one-step improvement} \\
&= f(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)) - f(\mathbf{x}_t) \quad \text{(utilize unconstrained update)} \\
&\le \langle \nabla f(\mathbf{x}_t), -\eta_t \nabla f(\mathbf{x}_t) \rangle + \frac{L}{2}\eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \quad \text{(smoothness)} \\
&= \Big( -\eta_t + \frac{L}{2}\eta_t^2 \Big) \|\nabla f(\mathbf{x}_t)\|^2 \\
&= -\frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 \quad \text{(recall that we have picked } \eta_t = \eta = \tfrac{1}{L}\text{)}
\end{aligned}$$

*Cautious: This derivation even doesn't require convexity!!*

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

# Smooth and Convex

***Proof:***

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

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

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

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

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

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

# 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 $\beta \triangleq \frac{1}{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}$.

$$\Rightarrow \quad \delta_{t+1} \leq \delta_t - \beta\delta_t^2$$

$$\Rightarrow \quad \frac{1}{\delta_t} \leq \frac{1}{\delta_{t+1}} - \beta \quad \text{(noting that } \delta_t^2 \geq \delta_t \cdot \delta_{t+1} \text{ and then dividing } \delta_t\delta_{t+1} \text{ from both sides)}$$

$$\Rightarrow \quad \sum_{t=1}^{T-1} \beta \leq \frac{1}{\delta_T} - \frac{1}{\delta_1} \leq \frac{1}{\delta_T}$$

$$\Rightarrow \quad \delta_T \triangleq f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{\beta(T-1)} = \frac{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T-1}.$$

□

# Key Lemma for Smooth GD

- During the proof, we have obtained an important lemma for ***smooth*** optimization, that is, ***one-step improvement***

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \left(-\eta_t + \frac{L}{2}\eta_t^2\right) \|\nabla f(\mathbf{x}_t)\|^2 \quad \Longrightarrow \quad f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \mathcal{O}\left(\frac{1}{T}\right).$$

***last-iterated*** convergence

- 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*

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

This lemma usually implies convergence like $f(\bar{\mathbf{x}}_T) - f^\star \leq \ldots$ with $\bar{\mathbf{x}}_T \triangleq \sum_{t=1}^T \frac{\eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$ (or other average).

***average-iterated*** convergence

# One-Step Improvement Lemma for Smooth GD

**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*

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

Function progress is proportional to the square of gradient magnitude (*consider due reasons*).

# 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 $\mathbf{g} \in \partial f(\mathbf{x})$,

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

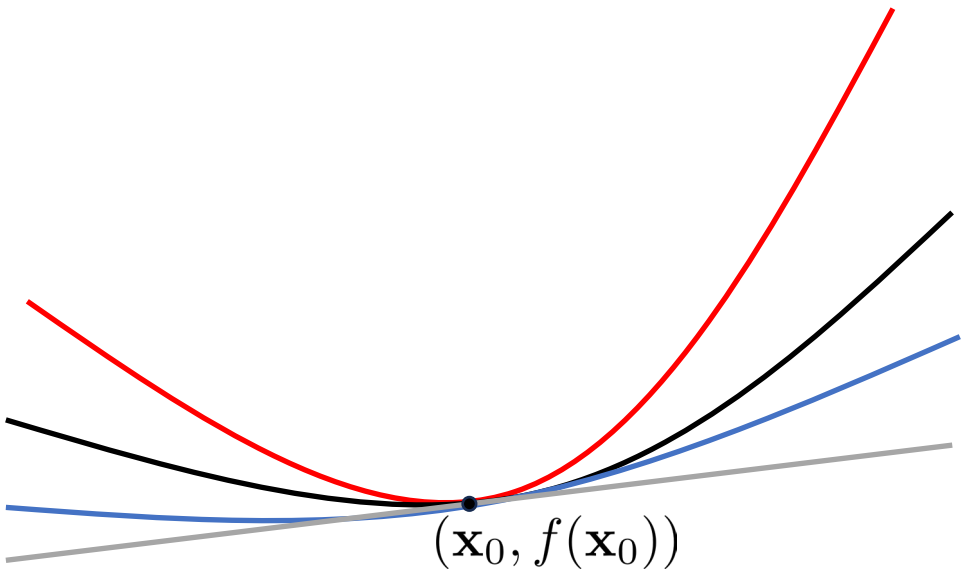
# Smooth and Strongly Convex

$f$ is $\sigma$-strongly convex

$f$ is $L$-smooth

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

$$h(\mathbf{x}) \triangleq f(\mathbf{x}_0)+\langle\nabla f(\mathbf{x}_0),\mathbf{x}-\mathbf{x}_0\rangle+\tfrac{L}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2$$




$$s(\mathbf{x}) \triangleq f(\mathbf{x}_0)+\langle\nabla f(\mathbf{x}_0),\mathbf{x}-\mathbf{x}_0\rangle+\tfrac{\sigma}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2$$

$$\ell(\mathbf{x}) \triangleq f(\mathbf{x}_0)+\langle\nabla f(\mathbf{x}_0),\mathbf{x}-\mathbf{x}_0\rangle$$

# 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. 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:***

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

*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:***

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

*exploiting co-coercivity of smooth and strongly convex function*

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

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

*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.

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

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

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

# 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, $\nabla f(\mathbf{x}^\star) = \mathbf{0}$)

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

□

# Constrained Optimization

- For unconstrained optimization, the key technical lemma is

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

  where $\nabla f(\mathbf{x})$ is used to measure the function progress.

- For constrained optimization, a ***generalized*** one-step improvement:

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

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

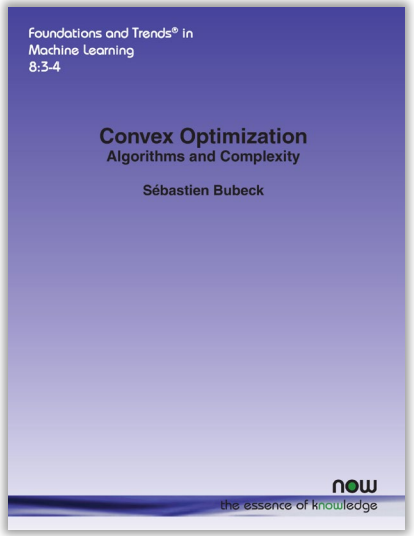
  - $g(\mathbf{x}_t)$ is used to qualify the progress; and in the unconstrained case, $g(\mathbf{x}_t)=\nabla f(\mathbf{x}_t)$.
  - comparator $\mathbf{u}$ is introduced because (projected) GD is not necessary "descent".

# Constrained Optimization

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

Detailed proofs for the constrained optimization will not be presented. The proof follows the same vein yet requires some additional 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 | $\mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right)$ | $\Omega\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right)\right)$ | ✗ |

⇨ **GD is suboptimal in *smooth* convex optimization!**

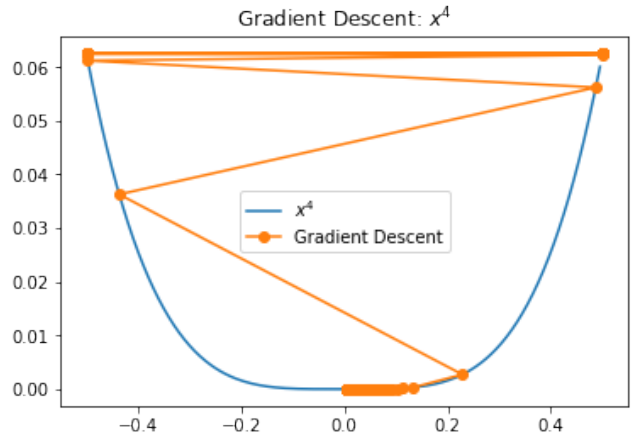
# Part 2. Momentum and Acceleration

- Polyak's Momentum
- Nesterov's Accelerated GD
- Smooth and Convex
- Smooth and Strongly Convex

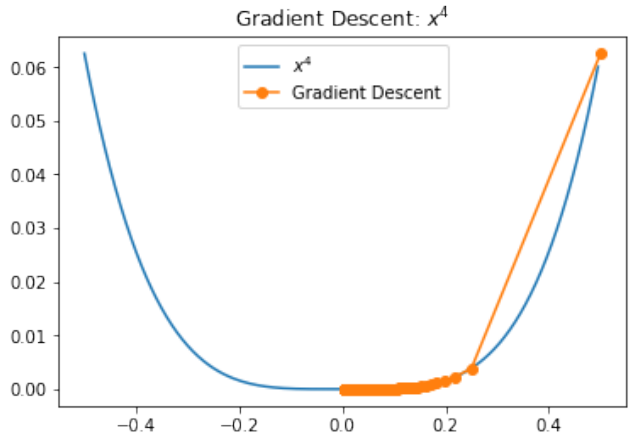
# Polyak's Momentum

- GD method (with a fixed step size): $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$, e.g., $\eta = \dfrac{1}{L}$
- The problem: ***pathological curvature***

Considera deploying GD on a quartic function $f(x) = x^4$.




(a) large step size




(b) large step size

**Motivation:**

- ❑ Ensure smaller steps in regions of high curvature to dampen oscillations.
- ❑ Ensure larger steps and accelerate in regions of low curvature.

Source: https://boostedml.com/2020/07/gradient-descent-and-momentum-the-heavy-ball-method.html

# Polyak's Momentum

- GD with momentum:

$$\mathbf{x}_{t+1} = \underbrace{\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)}_{\texttt{GD Update}} + \underbrace{\beta(\mathbf{x}_t - \mathbf{x}_{t-1})}_{\texttt{momentum}}$$

where $\beta$ is a hyperparameter (usually $\beta \in [0, 1]$ though not limited to it), which scales down the previous step adaptively.

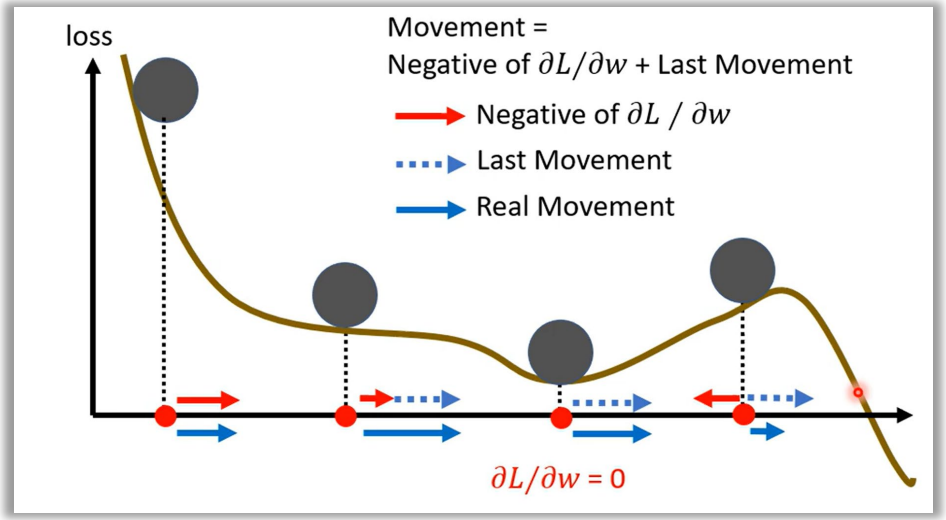
- ❑ If the current gradient step is in the same direction as the previous step (e.g., in the region of low curvature), then move a little further in that direction;
- ❑ If it's in the opposite direction (e.g., in the region of high curvature), move less far.

- Also known as the "**heavy ball method**" (think of the physical intuition).

# Polyak's Momentum

- Provable benefit: can achieve ***accelerated rate*** for optimizing the ***quadratic functions*** (but fail for more general cases like smooth and convex/strongly convex functions). Details are omitted [more details].

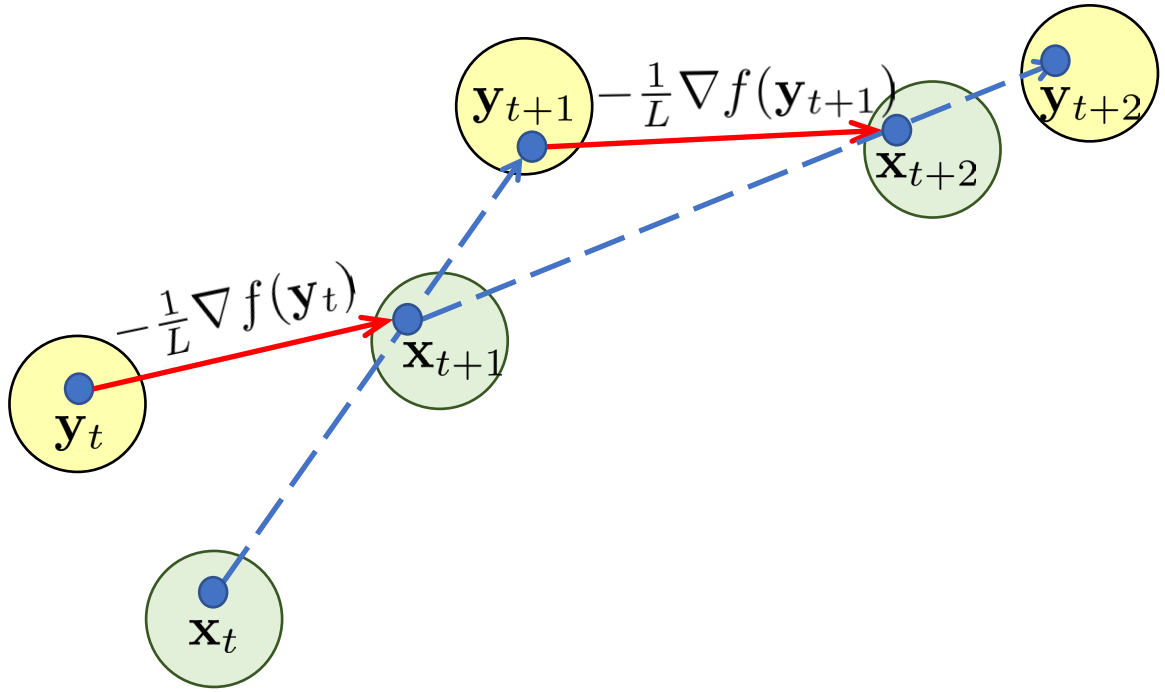
- Other benefit: help jump out of the local region (can be useful for non-convex opt)




Source: Hung-yi Lee ML 2021 Spring course Lecture on batch and momentum

# Nesterov's Accelerated GD

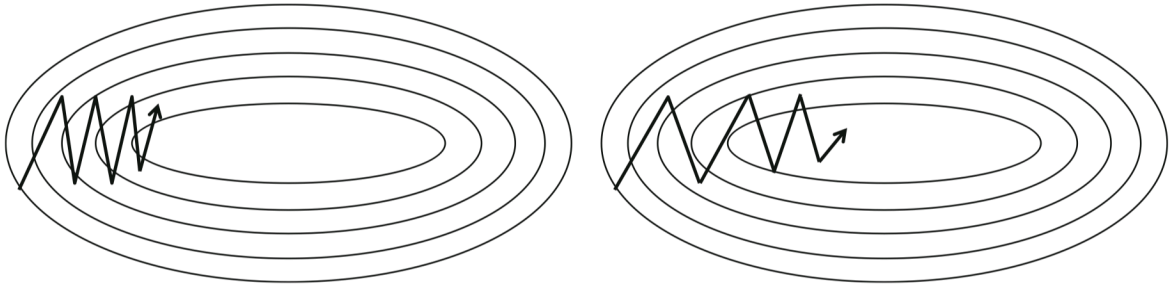
$$\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t)$$
$$\mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t)$$




- Define $\mathbf{x}_1 = \mathbf{y}_1$.
- $\beta_t > 0$ is a ***time-varying*** mixing rate of $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$; $\beta_t = 0$ recovers vanilla GD.
- AGD can be also thought a version of GD with ***momentum***.

# Nesterov's Accelerated GD

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

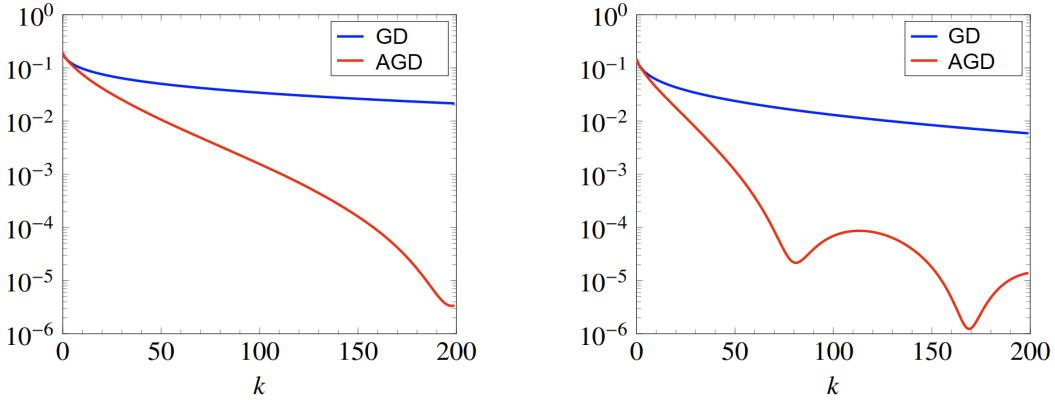


GD Accelarated GD

**Example**

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

- 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

# Polyak's Momentum v.s. Nesterov's AGD

- Polyak's Momentum:

$$\mathbf{x}_{t+1} = \underbrace{\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)}_{\texttt{GD Update}} + \underbrace{\beta(\mathbf{x}_t - \mathbf{x}_{t-1})}_{\texttt{momentum}}$$

- Nesterov's AGD:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{y}_t - \eta \nabla f(\mathbf{y}_t) \\ \mathbf{y}_{t+1} &= \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t) \end{aligned}$$

$$\mathbf{x}_{t+1} = \underbrace{\mathbf{x}_t - \eta \nabla f\big(\mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1})\big)}_{\texttt{GD Update}} + \underbrace{\beta(\mathbf{x}_t - \mathbf{x}_{t-1})}_{\texttt{momentum}}$$

**Main difference**: separate *the gradient calculation state* and *the momentum state*.

# Convergence of Nesterov's Accelerated GD

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

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

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

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

**Note**: for simplicity, we are working on ***unconstrained***.

# 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:*

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

a comparator variable $\mathbf{u}$ is introduced here,
because now AGD is not necessary "descent" due to the momentum

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

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$ *GD for smooth and convex functions*

# 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:***

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

□

# Proof of AGD Convergence

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

***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:*

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

(i) Plugging in $\mathbf{u} = \mathbf{x}_t$:

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

(ii) Plugging in $\mathbf{u} = \mathbf{x}^\star$:

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

# Proof of AGD Convergence

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

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

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

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

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

$$\begin{aligned} &(\lambda_t - 1)\big(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)\big) + f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \\ =\ &\lambda_t\big(f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star)\big) - (\lambda_t - 1)\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \end{aligned}$$

Define $\delta_t \triangleq f(\mathbf{x}_t) - f(\mathbf{x}^\star)$, then we have

$$\text{LHS} = \lambda_t \delta_{t+1} - (\lambda_t - 1)\delta_t$$

*Goal: design a telescoping series*

# Proof of AGD Convergence

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

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

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

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

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

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

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

That is

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

# Proof of AGD Convergence

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

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

Cautious: many terms of interest have already appeared in the following inequality.

*gradient inner product* *optimal point*

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

*optimality gap*
*telescoping structure*

*linear combination*
*related to momentum*

*gradient norm*

# Proof of AGD Convergence

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

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

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

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

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

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

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

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

# Proof of AGD Convergence

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

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

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

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

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

# Proof of AGD Convergence

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

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

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

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

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

***telescope***

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

$$\lambda_t^2\delta_{t+1} - \lambda_{t-1}^2\delta_t \le \frac{L}{2}(\|\mathbf{z}_t\|^2 - \|\mathbf{z}_{t+1}\|^2)$$

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

# Proof of AGD Convergence

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

***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):*** $\lambda_0 = 0$

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

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

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

# Proof

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

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

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

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

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

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

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

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

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

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

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

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

# Smooth and Strongly Convex

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

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

*satisfies*

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

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

***core technique:*** ***estimate sequence*** *(developed by Yurii Nesterov)*

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

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

# Smooth and Strongly Convex

- Proof sketch

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

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

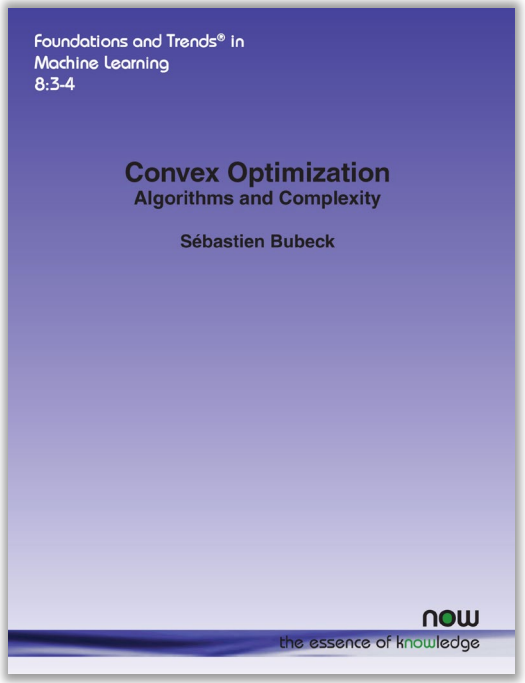
$$\Phi_{t+1}(\mathbf{x}) \triangleq (1-\theta)\Phi_t(\mathbf{x}) + \theta\left(f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}-\mathbf{x}_t\rangle + \frac{\sigma}{2}\|\mathbf{x}-\mathbf{x}_t\|^2\right)$$

$$\begin{aligned}
f(\mathbf{x}_t) - f(\mathbf{x}^\star) &\overset{(ii)}{\leq} \min_{\mathbf{x}\in\mathbb{R}^d} \Phi_t(\mathbf{x}) - f(\mathbf{x}^\star) \leq \Phi_t(\mathbf{x}^\star) - f(\mathbf{x}^\star) && \text{(by property } (ii)\text{)} \\
&\overset{(i)}{\leq} (1-\theta)^t(\Phi_1(\mathbf{x}^\star) - f(\mathbf{x}^\star)) && \text{(by property } (i)\text{)} \\
&= (1-\theta)^t\left(f(\mathbf{x}_1) + \frac{\sigma}{2}\|\mathbf{x}^\star-\mathbf{x}_1\|^2 - f(\mathbf{x}^\star)\right) && \text{(definition of } \Phi_1\text{)} \\
&\lesssim (\sigma+L)\|\mathbf{x}^\star-\mathbf{x}_1\|^2\exp(-\theta t) && \text{(smoothness)}
\end{aligned}$$

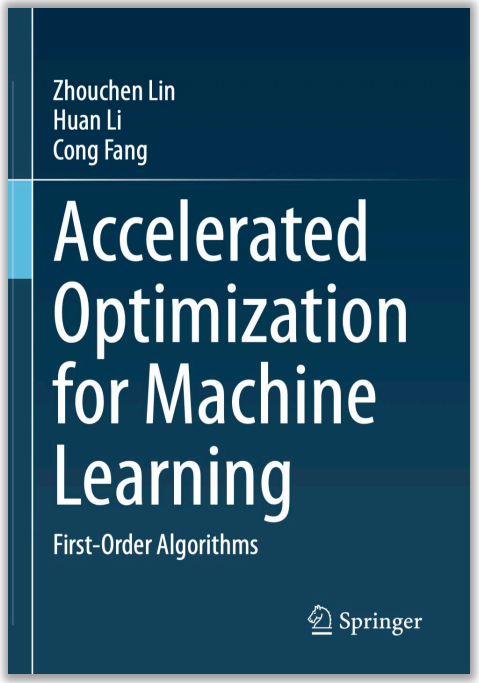
# Estimate Sequence

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

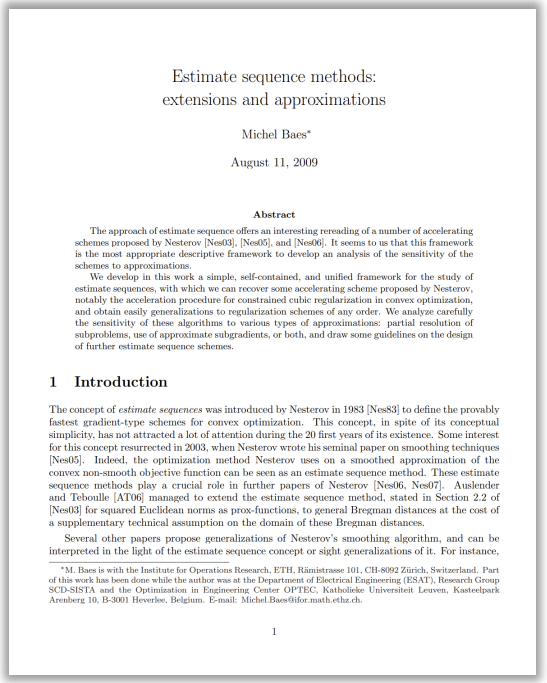
***References:***




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)

# More Explanations for Nesterov's AGD

- Ordinary Differentiable Equations
  - Su, W., Boyd, S., & Candes, E. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *NIPS 2014.*
  - Even, M., Berthier, R., Bach, F., Flammarion, N., Gaillard, P., Hendrikx, H., Taylor, A. A continuized view on nesterov acceleration for stochastic gradient descent and randomized gossip. *NeurIPS 2021*.

- Variational Analysis
  - Wibisono, A., Wilson, A. C., & Jordan, M. I. A variational perspective on accelerated methods in optimization. *PNAS 2016*, 113(47), E7351-E7358.

# More Explanations for Acceleration

- Linear Coupling of GD and MD
  - Allen-Zhu, Z., & Orecchia, L. Linear coupling: An ultimate unification of gradient and mirror descent. *ITCS 2017*.
  - Cutkosky A. Chapter 14 Momentum & Chapter 15 Acceleration. *Lecture Notes for EC525: Optimization for Machine Learning, 2022.*
- Online Learning with Suitable Optimism
  - Kavis, A., Levy, K. Y., Bach, F., & Cevher, V. UnixGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *NeurIPS 2019.*
  - Kreisler, I., Ivgi, M., Hinder, O., & Carmon, Y. Accelerated Parameter-Free Stochastic Optimization. *COLT 2024*.

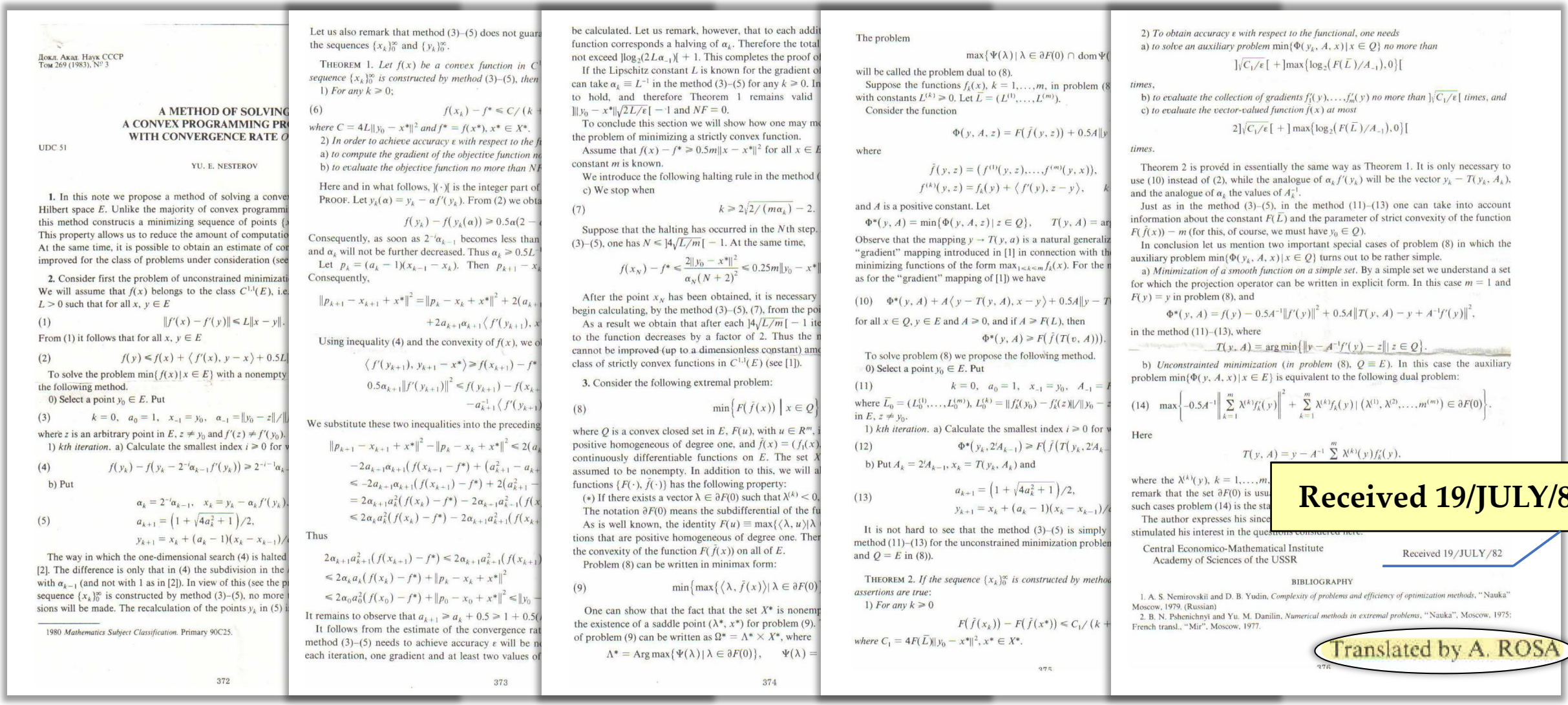
# History Bits

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.



Докл. Акад. Наук СССР
Том 269 (1983), № 3

**A METHOD OF SOLVING A CONVEX PROGRAMMING PROBLEM WITH CONVERGENCE RATE $O$**

UDC 51

YU. E. NESTEROV

**1.** In this note we propose a method of solving a convex ... Hilbert space $E$. Unlike the majority of convex programming ... this method constructs a minimizing sequence of points ... This property allows us to reduce the amount of computation ... At the same time, it is possible to obtain an estimate of ... improved for the class of problems under consideration (see ...

**2.** Consider first the problem of unconstrained minimization ... We will assume that $f(x)$ belongs to the class $C^{1,1}(E)$, i.e. $L > 0$ such that for all $x, y \in E$

$$\|f'(x) - f'(y)\| \leqslant L\|x - y\|. \tag{1}$$

From (1) it follows that for all $x, y \in E$

$$f(y) \leqslant f(x) + \langle f'(x), y - x\rangle + 0.5L\| \tag{2}$$

To solve the problem $\min\{f(x) \mid x \in E\}$ with a nonempty ... the following method.

0) Select a point $y_0 \in E$. Put

$$k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad \alpha_{-1} = \|y_0 - z\|/\| \tag{3}$$

where $z$ is an arbitrary point in $E$, $z \neq y_0$ and $f'(z) \neq f'(y_0)$.

1) *kth iteration*. a) Calculate the smallest index $i \geqslant 0$ for ...

$$f(y_k) - f(y_k - 2^{-i}\alpha_{k-1}f'(y_k)) \geqslant 2^{-i-1}\alpha_k \tag{4}$$

b) Put

$$\alpha_k = 2^{-i}\alpha_{k-1}, \quad x_k = y_k - \alpha_k f'(y_k), \quad a_{k+1} = \left(1 + \sqrt{4a_k^2 + 1}\right)/2, \quad y_{k+1} = x_k + (a_k - 1)(x_k - x_{k-1})/ \tag{5}$$

The way in which the one-dimensional search (4) is halted ... [2]. The difference is only that in (4) the subdivision in the ... with $\alpha_{k-1}$ (and not with 1 as in [2]). In view of this (see the p... sequence $\{x_k\}_0^\infty$ is constructed by method (3)–(5), no more ... sions will be made. The recalculation of the points $y_k$ in (5) i...

1980 *Mathematics Subject Classification.* Primary 90C25.

372

Let us also remark that method (3)–(5) does not guara... the sequences $\{x_k\}_0^\infty$ and $\{y_k\}_0^\infty$.

THEOREM 1. *Let $f(x)$ be a convex function in $C^1$... sequence $\{x_k\}_0^\infty$ is constructed by method (3)–(5), then*
1) *For any $k \geqslant 0$;*

$$f(x_k) - f^* \leqslant C/(k + \tag{6}$$

*where $C = 4L\|y_0 - x^*\|^2$ and $f^* = f(x^*)$, $x^* \in X^*$.*
2) *In order to achieve accuracy $\varepsilon$ with respect to the f...*
a) *to compute the gradient of the objective function n...*
b) *to evaluate the objective function no more than $NF$...*

Here and in what follows, $]\cdot[$ is the integer part of ...

PROOF. Let $y_k(\alpha) = y_k - \alpha f'(y_k)$. From (2) we obta...

$$f(y_k) - f(y_k(\alpha)) \geqslant 0.5\alpha(2 - \alpha$$

Consequently, as soon as $2^{-i}\alpha_{k-1}$ becomes less than ... and $\alpha_k$ will not be further decreased. Thus $\alpha_k \geqslant 0.5L^{-1}$.

Let $p_k = (a_k - 1)(x_{k-1} - x_k)$. Then $p_{k+1} - x_k$... Consequently,

$$\|p_{k+1} - x_{k+1} + x^*\|^2 = \|p_k - x_k + x^*\|^2 + 2(a_{k+1}$$
$$+ 2a_{k+1}\alpha_{k+1}\langle f'(y_{k+1}), x$$

Using inequality (4) and the convexity of $f(x)$, we ob...

$$\langle f'(y_{k+1}), y_{k+1} - x^*\rangle \geqslant f(x_{k+1}) - f^*$$
$$0.5\alpha_{k+1}\|f'(y_{k+1})\|^2 \leqslant f(y_{k+1}) - f(x_{k+$$
$$-a_{k+1}^{-1}\langle f'(y_{k+1})$$

We substitute these two inequalities into the preceding ...

$$\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \leqslant 2(a_k$$
$$-2a_{k+1}\alpha_{k+1}(f(x_{k+1} - f^*) + (a_{k+1}^2 - a_{k+$$
$$\leqslant -2a_{k+1}\alpha_{k+1}(f(x_{k+1}) - f^*) + 2(a_{k+1}^2 -$$
$$= 2a_{k+1}a_k^2(f(x_k) - f^*) - 2a_{k+1}a_{k+1}^2(f(x$$
$$\leqslant 2a_k a_k^2(f(x_k) - f^*) - 2a_{k+1}a_{k+1}^2(f(x_{k+$$

Thus

$$2a_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \leqslant 2a_{k+1}a_{k+1}^2(f(x_{k+1})$$
$$\leqslant 2a_k a_k(f(x_k) - f^*) + \|p_k - x_k + x^*\|^2$$
$$\leqslant 2a_0 a_0^2(f(x_0) - f^*) + \|p_0 - x_0 + x^*\|^2 \leqslant \|y_0 -$$

It remains to observe that $a_{k+1} \geqslant a_k + 0.5 \geqslant 1 + 0.5($

It follows from the estimate of the convergence rat... method (3)–(5) needs to achieve accuracy $\varepsilon$ will be n... each iteration, one gradient and at least two values of ...

373

be calculated. Let us remark, however, that to each addit... function corresponds a halving of $\alpha_k$. Therefore the total ... not exceed $]\log_2(2L\alpha_{-1})[ + 1$. This completes the proof o...

If the Lipschitz constant $L$ is known for the gradient o... can take $\alpha_k \equiv L^{-1}$ in the method (3)–(5) for any $k \geqslant 0$. In ... to hold, and therefore Theorem 1 remains valid ... $\|y_0 - x^*\|\sqrt{2L/\varepsilon}[ - 1$ and $NF = 0$.

To conclude this section we will show how one may m... the problem of minimizing a strictly convex function.

Assume that $f(x) - f^* \geqslant 0.5m\|x - x^*\|^2$ for all $x \in E$... constant $m$ is known.

We introduce the following halting rule in the method ...
c) We stop when

$$k \geqslant 2\sqrt{2/(m\alpha_k)} - 2. \tag{7}$$

Suppose that the halting has occurred in the $N$th step. ... (3)–(5), one has $N \leqslant ]4\sqrt{L/m}[ - 1$. At the same time,

$$f(x_N) - f^* \leqslant \frac{2\|y_0 - x^*\|^2}{\alpha_N(N+2)^2} \leqslant 0.25m\|y_0 - x^*\|$$

After the point $x_N$ has been obtained, it is necessary ... begin calculating, by the method (3)–(5), (7), from the poi...

As a result we obtain that after each $]4\sqrt{L/m}[ - 1$ ite... to the function decreases by a factor of 2. Thus the ... cannot be improved (up to a dimensionless constant) amo... class of strictly convex functions in $C^{1,1}(E)$ (see [1]).

**3.** Consider the following extremal problem:

$$\min\left\{F(\tilde f(x)) \;\middle|\; x \in Q\right\} \tag{8}$$

where $Q$ is a convex closed set in $E$, $F(u)$, with $u \in R^m$, i... positive homogeneous of degree one, and $\tilde f(x) = (f_1(x)$... continuously differentiable functions on $E$. The set $X$... assumed to be nonempty. In addition to this, we will a... functions $\{F(\cdot), \tilde f(\cdot)\}$ has the following property:

(*) If there exists a vector $\lambda \in \partial F(0)$ such that $\lambda^{(k)} < 0$...

The notation $\partial F(0)$ means the subdifferential of the fu...

As is well known, the identity $F(u) \equiv \max\{\langle \lambda, u\rangle \mid \lambda$... tions that are positive homogeneous of degree one. Ther... the convexity of the function $F(\tilde f(x))$ on all of $E$.

Problem (8) can be written in minimax form:

$$\min\left\{\max\left\{\langle \lambda, \tilde f(x)\rangle \mid \lambda \in \partial F(0)\right.\right. \tag{9}$$

One can show that the fact that the set $X^*$ is nonemp... the existence of a saddle point $(\lambda^*, x^*)$ for problem (9). ... of problem (9) can be written as $\Omega^* = \Lambda^* \times X^*$, where

$$\Lambda^* = \operatorname{Arg}\max\{\Psi(\lambda) \mid \lambda \in \partial F(0)\}, \qquad \Psi(\lambda) =$$

374

The problem

$$\max\{\Psi(\lambda) \mid \lambda \in \partial F(0) \cap \operatorname{dom}\Psi($$

will be called the problem dual to (8).

Suppose the functions $f_k(x)$, $k = 1, \ldots, m$, in problem (8... with constants $L^{(k)} \geqslant 0$. Let $\bar L = (L^{(1)}, \ldots, L^{(m)})$.

Consider the function

$$\Phi(y, A, z) = F(\tilde f(y, z)) + 0.5A\|y$$

where

$$\tilde f(y, z) = (f^{(1)}(y, z), \ldots, f^{(m)}(y, x)),$$
$$f^{(k)}(y, z) = f_k(y) + \langle f'(y), z - y\rangle, \qquad k$$

and $A$ is a positive constant. Let

$$\Phi^*(y, A) = \min\{\Phi(y, A, z) \mid z \in Q\}, \qquad T(y, A) = \arg$$

Observe that the mapping $y \to T(y, a)$ is a natural generaliz... "gradient" mapping introduced in [1] in connection with th... minimizing functions of the form $\max_{1 \leqslant k \leqslant m} f_k(x)$. For the ... as for the "gradient" mapping of [1]) we have

$$\Phi^*(y, A) + A\langle y - T(y, A), x - y\rangle + 0.5A\|y - T \tag{10}$$

for all $x \in Q$, $y \in E$ and $A \geqslant 0$, and if $A \geqslant F(L)$, then

$$\Phi^*(y, A) \geqslant F(\tilde f(T(v, A))).$$

To solve problem (8) we propose the following method.

0) Select a point $y_0 \in E$. Put

$$k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad A_{-1} = F \tag{11}$$

where $\bar L_0 = (L_0^{(1)}, \ldots, L_0^{(m)})$, $L_0^{(k)} = \|f_k'(y_0) - f_k'(z)\|/\|y_0 - z$... in $E$, $z \neq y_0$.

1) *kth iteration*. a) Calculate the smallest index $i \geqslant 0$ for ...

$$\Phi^*(y_k, 2^i A_{k-1}) \geqslant F(\tilde f(T(y_k, 2^i A_{k- \tag{12}$$

b) Put $A_k = 2^i A_{k-1}$, $x_k = T(y_k, A_k)$ and

$$a_{k+1} = \left(1 + \sqrt{4a_k^2 + 1}\right)/2, \qquad y_{k+1} = x_k + (a_k - 1)(x_k - x_{k-1})/ \tag{13}$$

It is not hard to see that the method (3)–(5) is simply ... method (11)–(13) for the unconstrained minimization problem ... and $Q = E$ in (8)).

THEOREM 2. *If the sequence $\{x_k\}_0^\infty$ is constructed by metho... assertions are true:*
1) *For any $k \geqslant 0$*

$$F(\tilde f(x_k)) - F(\tilde f(x^*)) \leqslant C_1/(k + \tag{}$$

*where $C_1 = 4F(\bar L)\|y_0 - x^*\|^2$, $x^* \in X^*$.*

375

2) *To obtain accuracy $\varepsilon$ with respect to the functional, one needs*
a) *to solve an auxiliary problem $\min\{\Phi(y_k, A, x) \mid x \in Q\}$ no more than*

$$]\sqrt{C_1/\varepsilon}[ + ]\max\{\log_2(F(\bar L)/A_{-1}), 0\}[$$

*times,*
b) *to evaluate the collection of gradients $f_1'(y), \ldots, f_m'(y)$ no more than $]\sqrt{C_1/\varepsilon}[$ times, and*
c) *to evaluate the vector-valued function $\tilde f(x)$ at most*

$$2]\sqrt{C_1/\varepsilon}[ + ]\max\{\log_2(F(\bar L)/A_{-1}), 0\}[$$

*times.*

Theorem 2 is proved in essentially the same way as Theorem 1. It is only necessary to use (10) instead of (2), while the analogue of $\alpha_k f'(y_k)$ will be the vector $y_k - T(y_k, A_k)$, and the analogue of $\alpha_k$ the values of $A_k^{-1}$.

Just as in the method (3)–(5), in the method (11)–(13) one can take into account information about the constant $F(\bar L)$ and the parameter of strict convexity of the function $F(\tilde f(x)) - m$ (for this, of course, we must have $y_0 \in Q$).

In conclusion let us mention two important special cases of problem (8) in which the auxiliary problem $\min\{\Phi(y_k, A, x) \mid x \in Q\}$ turns out to be rather simple.

a) *Minimization of a smooth function on a simple set.* By a simple set we understand a set for which the projection operator can be written in explicit form. In this case $m = 1$ and $F(y) = y$ in problem (8), and

$$\Phi^*(y, A) = f(y) - 0.5A^{-1}\|f'(y)\|^2 + 0.5A\|T(y, A) - y + A^{-1}f'(y)\|^2,$$

in the method (11)–(13), where

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

b) *Unconstrained minimization (in problem (8), $Q \equiv E$).* In this case the auxiliary problem $\min\{\Phi(y, A, x) \mid x \in E\}$ is equivalent to the following dual problem:

$$\max\left\{-0.5A^{-1}\left\|\sum_{k=1}^m \lambda^{(k)}f_k'(y)\right\|^2 + \sum_{k=1}^m \lambda^{(k)}f_k(y) \;\middle|\; (\lambda^{(1)}, \lambda^{(2)}, \ldots, m^{(m)}) \in \partial F(0)\right\}. \tag{14}$$

Here

$$T(y, A) = y - A^{-1}\sum_{k=1}^m \lambda^{(k)}(y)f_k'(y),$$

where the $\lambda^{(k)}(y)$, $k = 1, \ldots, m$, ... remark that the set $\partial F(0)$ is usu... such cases problem (14) is the sta...

The author expresses his since... stimulated his interest in the questions considered here.

Central Economico-Mathematical Institute
Academy of Sciences of the USSR

Received 19/JULY/82

BIBLIOGRAPHY

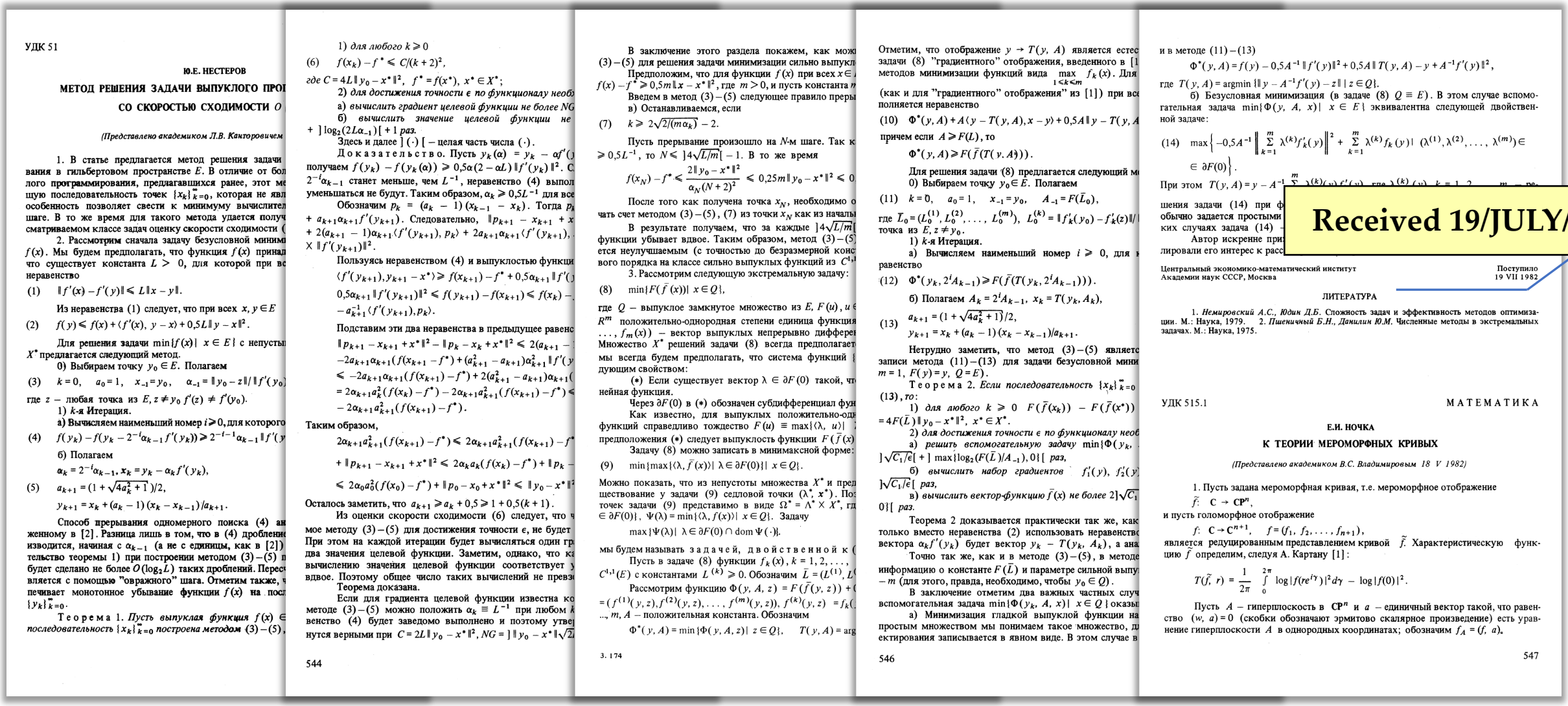
1. A. S. Nemirovskiĭ and D. B. Yudin, *Complexity of problems and efficiency of optimization methods*, "Nauka" Moscow, 1979. (Russian)
2. B. N. Pshenichnyĭ and Yu. M. Danilin, *Numerical methods in extremal problems*, "Nauka", Moscow, 1975; French transl., "Mir", Moscow, 1977.

376

**Received 19/JULY/82**

**Translated by A. ROSA**

# 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 ПРОГ

СО СКОРОСТЬЮ СХОДИМОСТИ $O$

*(Представлено академиком Л.В. Канторовичем*

1. В статье предлагается метод решения задачи вания в гильбертовом пространстве $E$. В отличие от бол лого программирования, предлагавшихся ранее, этот ме щую последовательность точек $\{x_k\}_{k=0}^{\infty}$, которая не явл особенность позволяет свести к минимуму вычислител шаге. В то же время для такого метода удается получ сматриваемом классе задач оценку скорости сходимости

2. Рассмотрим сначала задачу безусловной минимиз $f(x)$. Мы будем предполагать, что функция $f(x)$ принад что существует константа $L > 0$, для которой при вс неравенство

(1) $\|f'(x) - f'(y)\| \leqslant L\|x - y\|$.

Из неравенства (1) следует, что при всех $x, y \in E$

(2) $f(y) \leqslant f(x) + \langle f'(x),\ y - x\rangle + 0{,}5L\|y - x\|^2$.

Для решения задачи $\min\{f(x) \mid x \in E\}$ с непусты $X^*$ предлагается следующий метод.

0) Выбираем точку $y_0 \in E$. Полагаем

(3) $k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad \alpha_{-1} = \|y_0 - z\| / \|f'(y_0)$

где $z$ – любая точка из $E$, $z \neq y_0$ $f'(z) \neq f'(y_0)$.

1) $k$-я Итерация.

а) Вычисляем наименьший номер $i \geqslant 0$, для которого

(4) $f(y_k) - f(y_k - 2^{-i}\alpha_{k-1} f'(y_k)) \geqslant 2^{-i-1}\alpha_{k-1}\|f'(y$

б) Полагаем

$$\alpha_k = 2^{-i}\alpha_{k-1}, \quad x_k = y_k - \alpha_k f'(y_k),$$

(5) $$a_{k+1} = (1 + \sqrt{4a_k^2 + 1})/2,$$

$$y_{k+1} = x_k + (a_k - 1)(x_k - x_{k-1})/a_{k+1}.$$

Способ прерывания одномерного поиска (4) ан женному в [2]. Разница лишь в том, что в (4) дробление изводится, начиная с $\alpha_{k-1}$ (а не с единицы, как в [2]) тельство теоремы 1) при построении методом (3)–(5) будет сделано не более $O(\log_2 L)$ таких дроблений. Перес вляется с помощью "овражного" шага. Отметим также, печивает монотонное убывание функции $f(x)$ на посл $\{y_k\}_{k=0}^{\infty}$.

Теорема 1. *Пусть выпуклая функция* $f(x) \in$ *последовательность* $\{x_k\}_{k=0}^{\infty}$ *построена методом* (3)–(5),

1) *для любого* $k \geqslant 0$

(6) $f(x_k) - f^* \leqslant C/(k+2)^2$,

*где* $C = 4L\|y_0 - x^*\|^2$, $f^* = f(x^*)$, $x^* \in X^*$;

2) *для достижения точности* $\epsilon$ *по функционалу необх*

а) *вычислить градиент целевой функции не более NG*

б) *вычислить значение целевой функции не* + $]\log_2(2L\alpha_{-1})[ + 1$ *раз.*

Здесь и далее $]\cdot)[$ – целая часть числа $(\cdot)$.

Доказательство. Пусть $y_k(\alpha) = y_k - \alpha f'($ получаем $f(y_k) - f(y_k(\alpha)) \geqslant 0{,}5\alpha(2 - \alpha L)\|f'(y_k)\|^2$. С $2^{-i}\alpha_{k-1}$ станет меньше, чем $L^{-1}$, неравенство (4) выпол уменьшаться не будут. Таким образом, $\alpha_k \geqslant 0{,}5L^{-1}$ для все

Обозначим $p_k = (a_k - 1)(x_{k-1} - x_k)$. Тогда $p_k$ + $a_{k+1}\alpha_{k+1}f'(y_{k+1})$. Следовательно, $\|p_{k+1} - x_{k+1} + x$ + $2(a_{k+1} - 1)\alpha_{k+1}\langle f'(y_{k+1}), p_k\rangle + 2a_{k+1}\alpha_{k+1}\langle f'(y_{k+1}),$ $\times \|f'(y_{k+1})\|^2$.

Пользуясь неравенством (4) и выпуклостью функци

$$\langle f'(y_{k+1}), y_{k+1} - x^*\rangle \geqslant f(x_{k+1}) - f^* + 0{,}5\alpha_{k+1}\|f'(y$$

$$0{,}5\alpha_{k+1}\|f'(y_{k+1})\|^2 \leqslant f(y_{k+1}) - f(x_{k+1}) \leqslant f(x_k) -$$

$$- a_{k+1}^{-1}\langle f'(y_{k+1}), p_k\rangle.$$

Подставим эти два неравенства в предыдущее равенс

$$\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \leqslant 2(a_{k+1} -$$

$$-2a_{k+1}\alpha_{k+1}(f(x_{k+1} - f^*) + (a_{k+1}^2 - a_{k+1})\alpha_{k+1}^2\|f'(y$$

$$\leqslant -2a_{k+1}\alpha_{k+1}(f(x_{k+1}) - f^*) + 2(a_{k+1}^2 - a_{k+1})\alpha_{k+1}($$

$$= 2\alpha_{k+1}a_k^2(f(x_k) - f^*) - 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \leqslant$$

$$- 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*).$$

Таким образом,

$$2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \leqslant 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*$$

$$+ \|p_{k+1} - x_{k+1} + x^*\|^2 \leqslant 2\alpha_k a_k(f(x_k) - f^*) + \|p_k -$$

$$\leqslant 2\alpha_0 a_0^2(f(x_0) - f^*) + \|p_0 - x_0 + x^*\|^2 \leqslant \|y_0 - x^*\|^2$$

Осталось заметить, что $a_{k+1} \geqslant a_k + 0{,}5 \geqslant 1 + 0{,}5(k+1)$.

Из оценки скорости сходимости (6) следует, что мое методу (3)–(5) для достижения точности $\epsilon$, не будет При этом на каждой итерации будет вычисляться один гр два значения целевой функции. Заметим, однако, что к вычислению значения целевой функции соответствует вдвое. Поэтому общее число таких вычислений не превз Теорема доказана.

Если для градиента целевой функции известна ко методе (3)–(5) можно положить $\alpha_k \equiv L^{-1}$ при любом $k$ венство (4) будет заведомо выполнено и поэтому утве нутся верными при $C = 2L\|y_0 - x^*\|^2$, $NG = ]\|y_0 - x^*\|\sqrt{2}[$

544

В заключение этого раздела покажем, как мож (3)–(5) для решения задачи минимизации сильно выпукл

Предположим, что для функции $f(x)$ при всех $x \in$ $f(x) - f^* \geqslant 0{,}5m\|x - x^*\|^2$, где $m > 0$, и пусть константа $m$

Введем в метод (3)–(5) следующее правило преры

в) Останавливаемся, если

(7) $k \geqslant 2\sqrt{2/(m\alpha_k)} - 2$.

Пусть прерывание произошло на $N$-м шаге. Так к $\geqslant 0{,}5L^{-1}$, то $N \leqslant ]4\sqrt{L/m}[ - 1$. В то же время

$$f(x_N) - f^* \leqslant \frac{2\|y_0 - x^*\|^2}{\alpha_N(N+2)^2} \leqslant 0{,}25m\|y_0 - x^*\|^2 \leqslant 0$$

После того как получена точка $x_N$, необходимо о чать счет методом (3)–(5), (7) из точки $x_N$ как из начал

В результате получаем, что за каждые $]4\sqrt{L/m}[$ функции убывает вдвое. Таким образом, метод (3)–(5) ется неулучшаемым (с точностью до безразмерной конс вого порядка на классе сильно выпуклых функций из $C^{1,1}$

3. Рассмотрим следующую экстремальную задачу:

(8) $\min\{F(\bar{f}(x)) \mid x \in Q\}$,

где $Q$ – выпуклое замкнутое множество из $E$, $F(u)$, $u \in$ $R^m$ положительно-однородная степени единица функция $\ldots, f_m(x))$ – вектор выпуклых непрерывно дифферен Множество $X^*$ решений задачи (8) всегда предполагает мы всегда будем предполагать, что система функций дующим свойством:

(*) Если существует вектор $\lambda \in \partial F(0)$ такой, чт нейная функция.

Через $\partial F(0)$ в (*) обозначен субдифференциал фун

Как известно, для выпуклых положительно-одн функций справедливо тождество $F(u) \equiv \max\{\langle\lambda, u\rangle \mid$ предположения (*) следует выпуклость функции $F(\bar{f}(x)$

Задачу (8) можно записать в минимаксной форме:

(9) $\min\{\max\{\langle\lambda, \bar{f}(x)\rangle \mid \lambda \in \partial F(0)\} \mid x \in Q\}$.

Можно показать, что из непустоты множества $X^*$ и пред ществование у задачи (9) седловой точки $(\lambda^*, x^*)$. По точек задачи (9) представимо в виде $\Omega^* = \Lambda^* \times X^*$, гд $\in \partial F(0)\}$, $\Psi(\lambda) = \min\{\langle\lambda, f(x)\rangle \mid x \in Q\}$. Задачу

$$\max\{\Psi(\lambda) \mid \lambda \in \partial F(0) \cap \operatorname{dom}\Psi(\cdot)\}.$$

мы будем называть задачей, двойственной к

Пусть в задаче (8) функции $f_k(x)$, $k = 1, 2, \ldots,$ $C^{1,1}(E)$ с константами $L^{(k)} \geqslant 0$. Обозначим $\bar{L} = (L^{(1)}, L$

Рассмотрим функцию $\Phi(y, A, z) = F(\bar{f}(y, z)) +$ $= (f^{(1)}(y, z), f^{(2)}(y, z), \ldots, f^{(m)}(y, z))$, $f^{(k)}(y, z) = f_k($ $\ldots, m$, $A$ – положительная константа. Обозначим

$$\Phi^*(y, A) = \min\{\Phi(y, A, z) \mid z \in Q\}, \qquad T(y, A) = \arg$$

3. 174

Отметим, что отображение $y \to T(y, A)$ является естес задачи (8) "градиентного" отображения, введенного в [1 методов минимизации функций вида $\max_{1 \leqslant k \leqslant m} f_k(x)$. Для (как и для "градиентного" отображения" из [1]) при вс полняется неравенство

(10) $\Phi^*(y, A) + A\langle y - T(y, A), x - y\rangle + 0{,}5A\|y - T(y, A$

причем если $A \geqslant F(L)$, то

$$\Phi^*(y, A) \geqslant F(\bar{f}(T(y, A))).$$

Для решения задачи (8) предлагается следующий м

0) Выбираем точку $y_0 \in E$. Полагаем

(11) $k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad A_{-1} = F(\bar{L}_0)$,

где $\bar{L}_0 = (L_0^{(1)}, L_0^{(2)}, \ldots, L_0^{(m)})$, $L_0^{(k)} = \|f_k'(y_0) - f_k'(z)\|/$ точка из $E$, $z \neq y_0$.

1) $k$-я Итерация.

а) Вычисляем наименьший номер $i \geqslant 0$, для равенство

(12) $\Phi^*(y_k, 2^i A_{k-1}) \geqslant F(\bar{f}(T(y_k, 2^i A_{k-1})))$.

б) Полагаем $A_k = 2^i A_{k-1}$, $x_k = T(y_k, A_k)$,

(13) $$a_{k+1} = (1 + \sqrt{4a_k^2 + 1})/2,$$

$$y_{k+1} = x_k + (a_k - 1)(x_k - x_{k-1})/a_{k+1}.$$

Нетрудно заметить, что метод (3)–(5) являетс записи метода (11)–(13) для задачи безусловной мини $m = 1$, $F(y) = y$, $Q = E$).

Теорема 2. *Если последовательность* $\{x_k\}_{k=0}^{\infty}$ (13), *то*:

1) *для любого* $k \geqslant 0$ $F(\bar{f}(x_k)) - F(\bar{f}(x^*)) = 4F(\bar{L})\|y_0 - x^*\|^2$, $x^* \in X^*$.

2) *для достижения точности* $\epsilon$ *по функционалу необ*

а) *решить вспомогательную задачу* $\min\{\Phi(y_k,$ $]\sqrt{C_1/\epsilon}[ + ] \max\{\log_2(F(\bar{L})/A_{-1}), 0\}[$ *раз,*

б) *вычислить набор градиентов* $f_1'(y), f_2'(y$ $]\sqrt{C_1/\epsilon}[$ *раз,*

в) *вычислить вектор-функцию* $\bar{f}(x)$ *не более* $2]\sqrt{C_1}$ $0\}[$ *раз.*

Теорема 2 доказывается практически так же, как только вместо неравенства (2) использовать неравенство вектора $\alpha_k f'(y_k)$ будет вектор $y_k - T(y_k, A_k)$, а ана

Точно так же, как и в методе (3)–(5), в методе информацию о константе $F(\bar{L})$ и параметре сильной выпу $- m$ (для этого, правда, необходимо, чтобы $y_0 \in Q$).

В заключение отметим два важных частных случ вспомогательная задача $\min\{\Phi(y_k, A, x) \mid x \in Q\}$ оказы

а) Минимизация гладкой выпуклой функции на простым множеством мы понимаем такое множество, д ектирования записывается в явном виде. В этом случае в

546

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

$$\left. \in \partial F(0)\right\}.$$

При этом $T(y, A) = y - A^{-1}\sum_{k=1}^{m}\lambda^{(k)}(y)f'(y)$, где $\lambda^{(k)}(y)$, $k = 1, 2, \ldots, m$ – ре

шения задачи (14) при ф обычно задается простыми ких случаях задача (14)

Автор искренне при лировали его интерес к расс

Received 19/JULY/82

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

$\bar{f}$: $\mathbf{C} \to \mathbf{CP}^n$,

и пусть голоморфное отображение

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

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

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

Пусть $A$ – гиперплоскость в $\mathbf{CP}^n$ 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

# History Bits

- Polyak's Momentum, credit goes to Polyak, date back to 1960s

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, 1964.

**Boris T. Polyak**
1935-2023

Math. Program., Ser. B 91: 401–416 (2002)
Digital Object Identifier (DOI) 10.1007/s101070100258

B.T. Polyak

**History of mathematical programming in the USSR: analyzing the phenomenon***

Received: January 29, 2001 / Accepted: May 17, 2001
Published online October 2, 2001 – © Springer-Verlag 2001

**Abstract**. I am not a historian; these are just reminiscences of a person involved in the development of optimization theory and methods in the former USSR. I realize that my point of view may be very personal; however, I am trying to present as broad and unbiased picture as I can.

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

***How to effectively leverage the (partial) smoothness to improve convergence?***

# Recall Non-composite Optimization

Recall how we ***invent*** GD for unconstrained non-composite optimization.

- **Idea: surrogate optimization**

  We aim to find a sequence of ***local upper bounds*** $U_1, \cdots, U_T$, where the surrogate function $U_t : \mathbb{R}^d \mapsto \mathbb{R}$ may depend on $\mathbf{x}_t$ such that

  (i) $f(\mathbf{x}_t) = U_t(\mathbf{x}_t)$;

  (ii) $f(\mathbf{x}) \leq U_t(\mathbf{x})$ holds for all $\mathbf{x} \in \mathbb{R}^d$;

  (iii) $U_t(\mathbf{x})$ should be simple enough to minimize.

  $\Longrightarrow$ Then, our proposed algorithm would be $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} U_t(\mathbf{x})$

# Recall Non-composite Optimization

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

$$\text{By smoothness: } f(\mathbf{x}) \le \underbrace{f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2}_{\triangleq U_t(\mathbf{x}) \quad \textit{surrogate objective}}$$

$\Rightarrow$ to minimize $f(\mathbf{x})$, it suffices to minimize the ***surrogate*** sequence $\{U_t(\mathbf{x})\}_{t=1}^T$.

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

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

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

# Recall Non-composite Optimization

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

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

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

***Proof:***

$$\begin{aligned}
\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} U_t(\mathbf{x}) &= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\langle \nabla f(\mathbf{x}_t), \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{x}\|^2 - L\langle \mathbf{x}, \mathbf{x}_t\rangle\right\} && \text{(remove irrelative terms)}\\
&= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\frac{L}{2}\left(-2\left\langle \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \mathbf{x}\right\rangle + \|\mathbf{x}\|^2\right)\right\} && \text{(rearrange)}\\
&= \arg\min_{\mathbf{x}\in\mathcal{X}} \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\|^2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)\right\| = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right]
\end{aligned}$$

□

# 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***.

**A natural idea for surrogate objective:**

Following previous argument (for non-composite optimization), to minimize $F \triangleq f + h$, it's natural to optimize surrogate sequence $\{U_t(\mathbf{x})\}_{t=1}^T$ defined as

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

# Composite Optimization

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

*surrogate objective*

$\Rightarrow$ to minimize $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$, it suffices to minimize $\boxed{U_t(\mathbf{x}) \triangleq u_t(\mathbf{x}) + h(\mathbf{x}).}$

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

# Composite Optimization

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

*surrogate objective*

$\Rightarrow$ to minimize $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$, it suffices to minimize $\boxed{U_t(\mathbf{x}) \triangleq u_t(\mathbf{x}) + h(\mathbf{x}).}$

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

*this will be abstracted as an operator, a subproblem to optimize*

# Composite Optimization

- Iteratively solve the surrogate optimization problem.

Deploying the following update rule:

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

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

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

# Proximal Gradient

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

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

**Proximal Gradient Method**

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

*An equivalent notation:* $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t) \triangleq \mathbf{prox}_{\frac{1}{L}h}\left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right).$

# Proximal Gradient

**Proximal Gradient Method**

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

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

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*

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

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

***Proof:***

**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 $\mathbf{u} \in \mathcal{X}$,*

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

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

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

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

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

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

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

# Proof of PG Convergence

***Proof:***

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

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: $F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \leq -\frac{1}{2L}\|g(\mathbf{x}_t)\|^2 \leq 0$.

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

# 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:***

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

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

□

# 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:* $F(\mathbf{x}) \leq U_t(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X} \Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq U_t(\mathbf{x}_{t+1}) - F(\mathbf{u})$

*analyzing this quantity*

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

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

*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:*** $F(\mathbf{x}) \le U_t(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X} \Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \le U_t(\mathbf{x}_{t+1}) - F(\mathbf{u})$

*analyzing this quantity*

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

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

$\Downarrow$ *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:*** $F(\mathbf{x}) \leq U_t(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X} \Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq U_t(\mathbf{x}_{t+1}) - F(\mathbf{u})$

*analyzing this quantity*

$$\begin{cases} U_t(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t) + \nabla h(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \dfrac{1}{2L} \|g(\mathbf{x}_t)\|^2 \\ \text{and the fact that } \nabla f(\mathbf{x}_t) + \nabla h(\mathbf{x}_{t+1}) = -L(\mathbf{x}_{t+1} - \mathbf{x}_t) = -g(\mathbf{x}_t) \end{cases}$$

$$\begin{aligned} \Rightarrow \quad U_t(\mathbf{x}_{t+1}) - F(\mathbf{u}) &\leq \langle g(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \frac{1}{2L} \|g(\mathbf{x}_t)\|^2 \\ &= \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L} \|g(\mathbf{x}_t)\|^2 \end{aligned}$$

□

# One-Step Improvement Lemma

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

| | |
|---|---|
| ***unconstrained, GD*** | $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$ |
| ***unconstrained, AGD*** | $f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$ |
| ***constrained, GD*** | $f(\mathbf{x}_{t+1}) - f(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$ |
| ***composite, GD/AGD*** | $F(\mathbf{x}_{t+1}) - F(\mathbf{u}) \leq \langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle - \frac{1}{2L}\|g(\mathbf{x}_t)\|^2$ |

*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 achieve AGD in composite optimization?

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

**Nesterov's Accelerated GD**

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

**Accelerated Proximal Gradient**

$$\mathbf{x}_{t+1} = \mathbf{prox}_{\frac{1}{L}h}\left(\mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t)\right), \quad \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t)$$

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***.

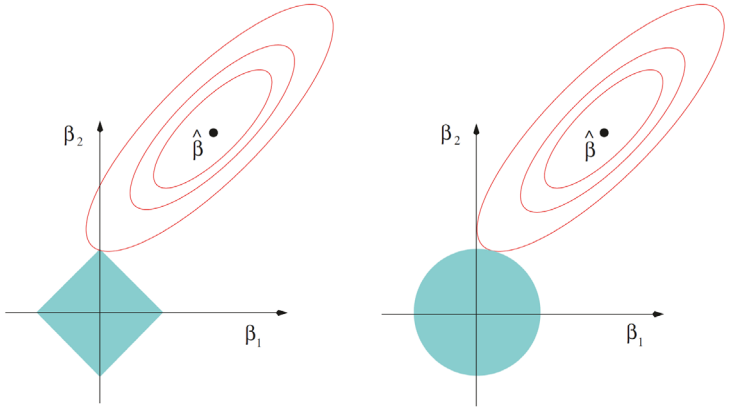
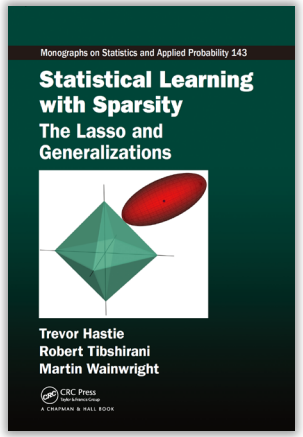
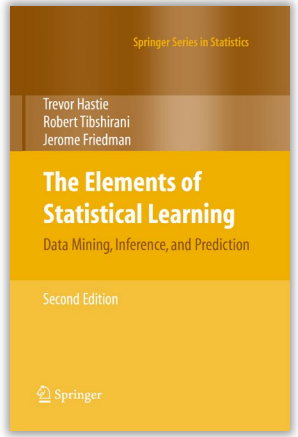
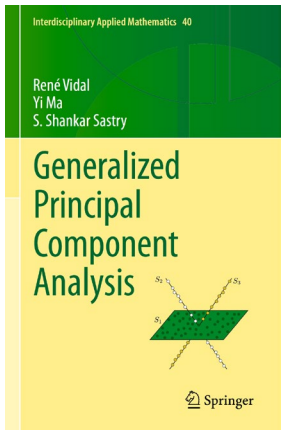



FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$*, respectively, while the red ellipses are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




Regression shrinkage and selection via the lasso 61368 1996
R Tibshirani
Journal of the Royal Statistical Society. Series B (Methodological), 267-288

# Application to LASSO

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

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

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

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

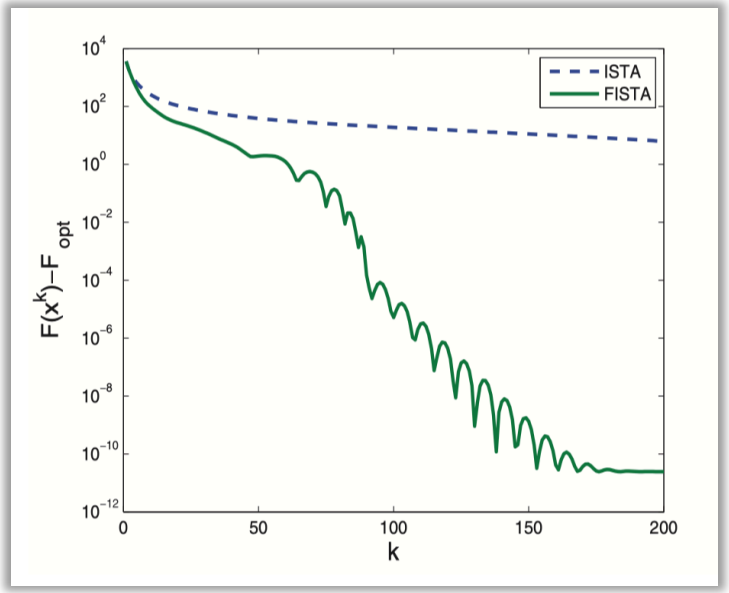
- **ISTA** (**I**terative **S**hrinkage-**T**hresholding **A**lgorithm): **PG** for LASSO
- **FISTA** (**F**ast **ISTA**): **APG** for LASSO

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

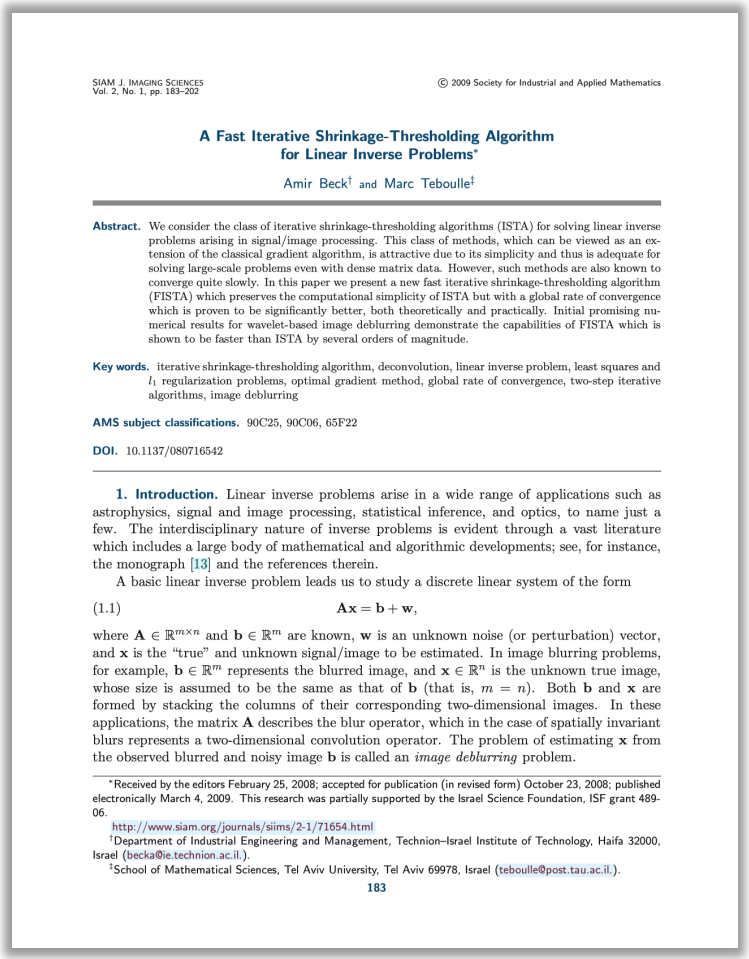
$$[\mathcal{P}_L^h(\mathbf{w}_t)]_i = \operatorname{sign}\left(\left[\mathbf{w}_t - \frac{1}{L}\nabla f(\mathbf{w}_t)\right]_i\right)\left(\left|\left[\mathbf{w}_t - \frac{1}{L}\nabla f(\mathbf{w}_t)\right]_i\right| - \frac{\lambda}{L}\right)_+$$

# 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.).

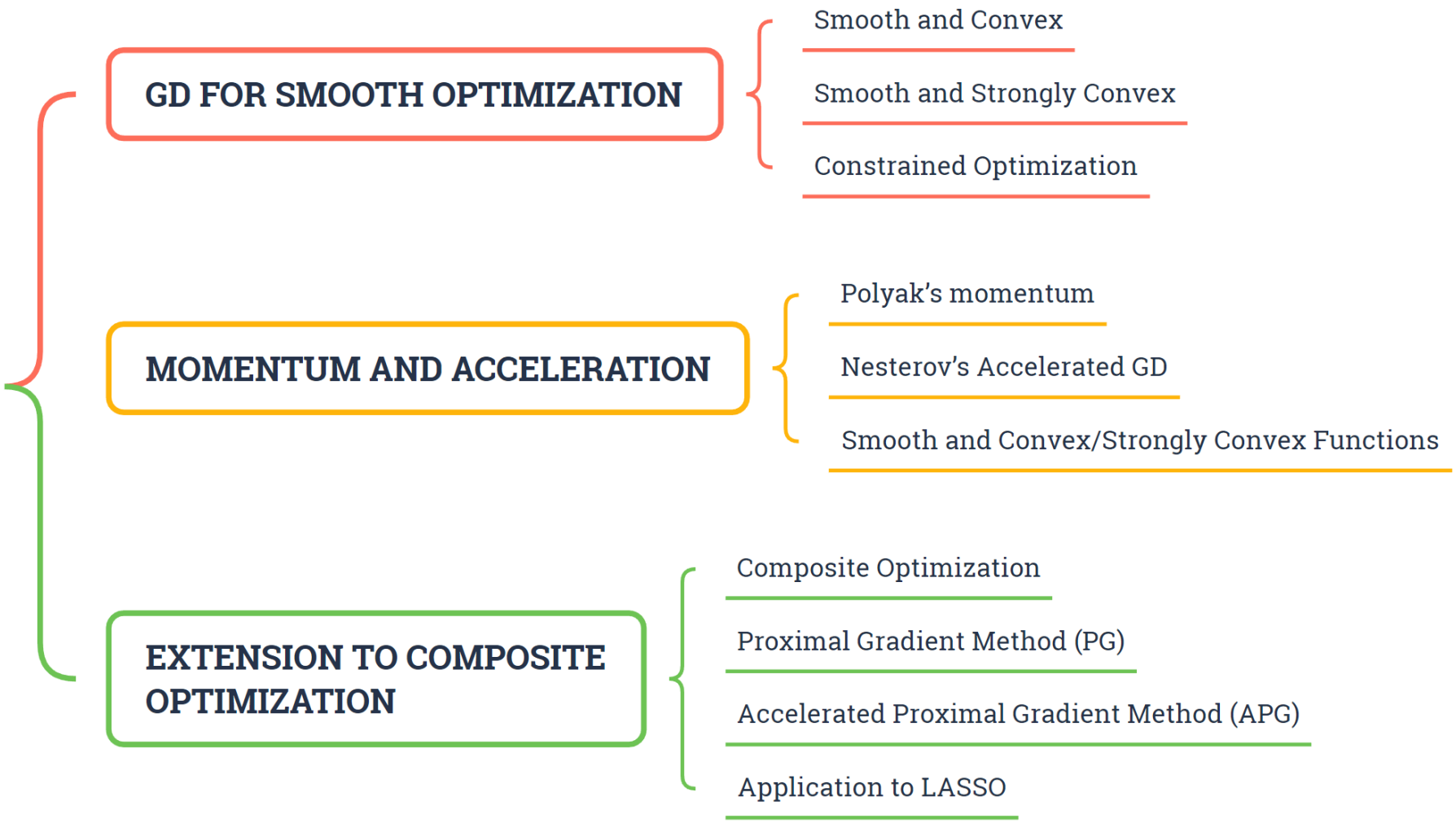

A fast iterative shrinkage-thresholding algorithm for linear inverse problems 13994 2009
A Beck, M Teboulle
SIAM journal on imaging sciences 2 (1), 183-202

# Summary

Table 1: A summary of convergence rates of GD method for smooth optimization.

| Algorithm | Function Family | Step Size | Output Sequence | Convergence Rate | Remark |
|---|---|---|---|---|---|
| GD | $L$-smooth and convex | $\eta = \frac{1}{L}$ | $\bar{\mathbf{x}}_T \triangleq \mathbf{x}_T$ | $\mathcal{O}(1/T)$ | suboptimal |
| | $L$-smooth and $\sigma$-strongly convex | $\eta = \frac{2}{\sigma+L}$ | $\bar{\mathbf{x}}_T \triangleq \mathbf{x}_T$ | $\mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right)$ | suboptimal |
| AGD | $L$-smooth and convex | $\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t),\ \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t)$ | $\bar{\mathbf{x}}_T \triangleq \mathbf{x}_T$ | $\mathcal{O}(1/T^2)$ | optimal |
| | $L$-smooth and $\sigma$-strongly convex | $\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t),\ \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}(\mathbf{x}_{t+1} - \mathbf{x}_t)$ | $\bar{\mathbf{x}}_T \triangleq \mathbf{x}_T$ | $\mathcal{O}\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right)\right)$ | optimal |
| PG | $F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{x})$<br>$f$ and $h$ are convex<br>$f$ is $L$-smooth but $h$ is not smooth | $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{x}_t) \triangleq \mathbf{prox}_{\frac{1}{L}h}\left(\mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)\right)$ | $\bar{\mathbf{x}}_T \triangleq \mathbf{x}_T$ | $\mathcal{O}(1/T)$ | suboptimal |
| APG | | $\mathbf{x}_{t+1} = \mathcal{P}_L^h(\mathbf{y}_t),\ \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t)$ | $\bar{\mathbf{x}}_T \triangleq \mathbf{x}_T$ | $\mathcal{O}(1/T^2)$ | optimal |

# Summary


GD FOR SMOOTH OPTIMIZATION
Smooth and Convex
Smooth and Strongly Convex
Constrained Optimization
MOMENTUM AND ACCELERATION
Polyak's momentum
Nesterov's Accelerated GD
Smooth and Convex/Strongly Convex Functions
EXTENSION TO COMPOSITE OPTIMIZATION
Composite Optimization
Proximal Gradient Method (PG)
Accelerated Proximal Gradient Method (APG)
Application to LASSO