



Lecture 4. Gradient Descent Method II

Advanced Optimization (Fall 2023)

Peng Zhao zhaop@lamda.nju.edu.cn Nanjing University

Outline

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

Part 1. GD for Smooth Optimization

Smooth and Convex

• Smooth and Strongly Convex

• Extension to Constrained Case

Overview

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

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

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:

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

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}^* be the optimal set of the optimization problem and f^* be the optimal value. Then for any $\mathbf{x}^* \in \mathcal{X}^*$ and $t \ge 0$,

$$\|\mathbf{x}_{t+1} - \mathbf{x}^{\star}\|^{2} \le \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} - 2\eta_{t}(f(\mathbf{x}_{t}) - f^{\star}) + \eta_{t}^{2}\|\nabla f(\mathbf{x}_{t})\|^{2}$$

Advanced Optimization (Fall 2023)

Refined Result for Smooth Optimization

only exploited convexity, but haven't used smoothness

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

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

Advanced Optimization (Fall 2023)

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

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

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

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

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

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

exploiting coercivity of smoothness and unconstrained first-order optimality

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

$$\begin{split} \Longrightarrow \|\mathbf{x}_{t+1} - \mathbf{x}^{\star}\|^{2} &\leq \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} + \left(\eta_{t}^{2} - \frac{2\eta_{t}}{L}\right) \|\nabla f(\mathbf{x}_{t})\|^{2} \\ &\leq \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} - \frac{1}{L^{2}} \|\nabla f(\mathbf{x}_{t})\|^{2} \quad \text{(by picking } \eta_{t} = \eta = \frac{1}{L} \text{ to minimize the r.h.s)} \\ &\leq \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} \leq \ldots \leq \|\mathbf{x}_{1} - \mathbf{x}^{\star}\|^{2} \quad \text{which already implies the convergence} \end{split}$$

Advanced Optimization (Fall 2023)

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

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

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

one-step improvement

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

Advanced Optimization (Fall 2023)

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

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

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

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

Advanced Optimization (Fall 2023)

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

Advanced Optimization (Fall 2023)

Lecture 4. Gradient Descent Method II

Key Lemma for Smooth GD

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

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

$$last-iterated \text{ convergence}$$

• 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

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

This lemma usually implies convergence like $f(\bar{\mathbf{x}}_T) - f^* \leq \dots$ 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

Key Lemma for Smooth GD

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

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

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

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

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

Advanced Optimization (Fall 2023)

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

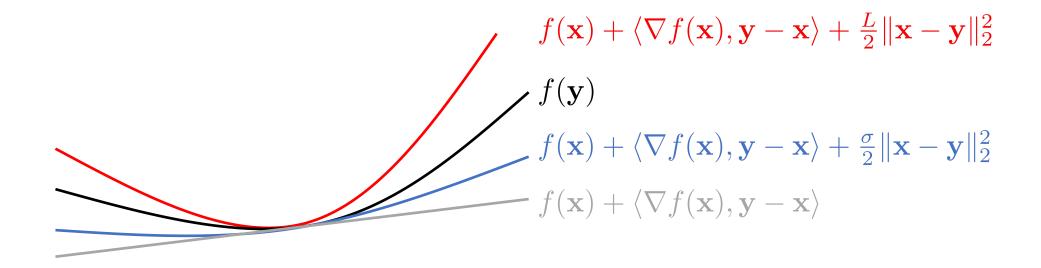
Definition 5 (Strong Convexity). A function f is σ -strongly convex if, for any $\mathbf{x} \in \operatorname{dom}(\partial f), \mathbf{y} \in \operatorname{dom}(f)$ and $\mathbf{g} \in \partial f(\mathbf{x})$,

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

f is σ -strongly convex

f is *L*-smooth

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



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

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

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.

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

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 σ -strongly convex on \mathbb{R}^d . Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, one has

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

Advanced Optimization (Fall 2023)

Coercivity of Smooth and Strongly Convex Function

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

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

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 σ -strong convexity (see previous lecture).

-
$$h$$
 is $(L - \sigma)$ -smooth. $\nabla^2 h(\mathbf{x}) = \nabla^2 f(\mathbf{x}) - \sigma I \preceq (L - \sigma)I$.

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

by co-coercivity of smooth and convex functions

Then, rearranging the terms finishes the proof.

Advanced Optimization (Fall 2023)

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

exploiting co-coercivity of smooth and strongly convex function

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

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

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

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 $(\eta_t^2 - \frac{2\eta_t}{L + \sigma}) \le 0$, or it becomes 0 is enough. \implies a feasible (and simple) setting: $\eta_t = \eta = \frac{2}{L + \sigma}$

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

Advanced Optimization (Fall 2023)

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 $f(\mathbf{x}_T) - f(\mathbf{x}^{\star})$.

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

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

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

Advanced Optimization (Fall 2023)

Constrained Optimization

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

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

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

comparator u is introduced because now GD is not necessary "descent" due to the projection

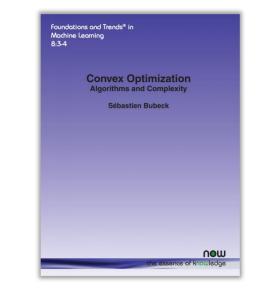
- In unconstrained case, $g(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)$.
- In unconstrained case, setting $\mathbf{u} = \mathbf{x}_t$ recovers the one-step improvement: $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2$.

Constrained Optimization

Same convergence rates as unconstrained case can be obtained in the constrained setting for smooth 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	$\mathcal{O}(1/\sqrt{T})$	$\Omega(1/\sqrt{T})$	\checkmark
	σ -strongly convex	$\mathcal{O}(1/T)$	$\Omega(1/T)$	\checkmark
<i>L-</i> smooth	convex	$\mathcal{O}(1/T)$	$\Omega(1/T^2)$	×
	σ -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. Nesterov's Accelerated GD

AGD Algorithm

• Smooth and Convex

• Smooth and Strongly Convex

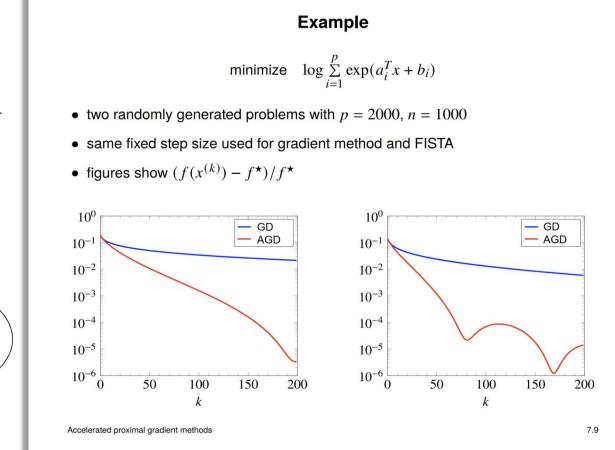
Nesterov's Accelerated GD

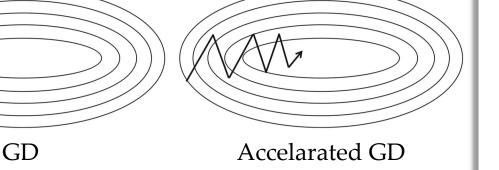
$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = (1 - \alpha_t) \mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$
$$\mathbf{y}_{t+1} = \mathbf{y}_1.$$

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

Nesterov's Accelerated GD

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





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

Advanced Optimization (Fall 2023)

Convergence of Nesterov's Accelerated GD

Theorem 3. Let
$$f$$
 be convex and L -smooth. Nesterov's accelerated GD is configured
as
 $\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1 - \alpha_t) \mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t,$
where $\lambda_0 = 0, \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2}$, and $\alpha_t = \frac{1 - \lambda_t}{\lambda_{t+1}}$. Then, we have
 $f(\mathbf{y}_T) - f(\mathbf{x}^*) \leq \frac{2L \|\mathbf{x}_1 - \mathbf{x}^*\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right).$

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

Advanced Optimization (Fall 2023)

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:

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

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

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

Advanced Optimization (Fall 2023)

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

Proof: (continued proving Theorem 3)

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

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

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

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

LHS of $(\lambda_t - 1)(\mathbf{i}) + (\mathbf{ii})$ equals: $(\lambda_t - 1)(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + f(\mathbf{y}_{t+1}) - f(\mathbf{x}^*) = \lambda_t (f(\mathbf{y}_{t+1}) - f(\mathbf{x}^*)) - (\lambda_t - 1)(f(\mathbf{y}_t) - f(\mathbf{x}^*))$ Define $\delta_t \triangleq f(\mathbf{y}_t) - f(\mathbf{x}^*)$, LHS = $\lambda_t \delta_{t+1} - (\lambda_t - 1)\delta_t$ Goal: design a telescoping series

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

Proof: (continued proving Theorem 3)

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

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

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

$$\begin{split} &(\lambda_t - 1) \left(\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{1}{2L} \| \nabla f(\mathbf{x}_t) \|^2 \right) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{1}{2L} \| \nabla f(\mathbf{x}_t) \|^2 \\ &= \langle \nabla f(\mathbf{x}_t), \lambda_t \mathbf{x}_t - (\lambda_t - 1) \mathbf{y}_t - \mathbf{x}^* \rangle - \frac{\lambda_t}{2L} \| \nabla f(\mathbf{x}_t) \|^2 \\ \end{split}$$
That is

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

Advanced Optimization (Fall 2023)

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

Proof: (continued proving Theorem 3)

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

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

$$\begin{array}{l} \textbf{Requirement (1): } \lambda_t(\lambda_t - 1) = \lambda_{t-1}^2 \\ \Rightarrow \lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{1}{2L} \left(2 \langle \lambda_t \nabla f(\mathbf{x}_t), L(\lambda_t \mathbf{x}_t - (\lambda_t - 1) \mathbf{y}_t - \mathbf{x}^*) \rangle - \|\lambda_t \nabla f(\mathbf{x}_t)\|^2 \right) \end{array}$$

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

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

Advanced Optimization (Fall 2023)

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

Proof: (continued proving Theorem 3)

Denote by $\boldsymbol{a} \triangleq \lambda_t \nabla f(\mathbf{x}_t), \boldsymbol{b} \triangleq L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*).$ $\begin{array}{l} \lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \\ \leq \frac{1}{2L} (L^2 \|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*\|^2 - \|L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*) - \lambda_t \nabla f(\mathbf{x}_t)\|^2) \\ \end{array}$ $= \frac{L}{2} \left(\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*\|^2 - \|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^* - \lambda_t \frac{\nabla f(\mathbf{x}_t)}{L}\|^2 \right) \\ = \frac{L}{2} (\|\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*\|^2 - \|\lambda_t \mathbf{y}_{t+1} - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*\|^2) \end{array}$

Goal: design a telescoping series

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

Proof: (continued proving Theorem 3)

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

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

$$\begin{split} \lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t &\leq \frac{L}{2} (\|\lambda_t \mathbf{x}_t - (\lambda_t - 1) \mathbf{y}_t - \mathbf{x}^\star \|^2 - \|\lambda_{t+1} \mathbf{x}_{t+1} - (\lambda_{t+1} - 1) \mathbf{y}_{t+1} - \mathbf{x}^\star \|^2) \\ \text{Define } \mathbf{z}_t &\triangleq \lambda_t \mathbf{x}_t - (\lambda_t - 1) \mathbf{y}_t - \mathbf{x}^\star \end{split}$$

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

Advanced Optimization (Fall 2023)

Proof of AGD Convergence

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

Proof: (continued proving Theorem 3)

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

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

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

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

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

Proof

Proof: (continued proving Theorem 3)

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

Theorem 3. Let
$$f$$
 be convex and L -smooth. Nesterov's accelerated GD is configured
as
 $\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1 - \alpha_t) \mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t,$
where $\lambda_0 = 0, \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2}$, and $\alpha_t = \frac{1 - \lambda_t}{\lambda_{t+1}}$. Then, we have
 $f(\mathbf{y}_T) - f(\mathbf{x}^*) \leq \frac{2L \|\mathbf{x}_1 - \mathbf{x}^*\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right).$

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

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

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

Requirement (3): $\lambda_0 = 0$

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

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

Smooth and Strongly Convex

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

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

satisfies

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

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 satisfy some nice properties:

- (i) $\Phi_{t+1}(\mathbf{x}) f(\mathbf{x}) \le (1 \theta)^t (\Phi_1(\mathbf{x}) f(\mathbf{x})) \Rightarrow \text{approximate } f \text{ well.}$
- (*ii*) $f(\mathbf{y}_t) \leq \min_{\mathbf{x} \in \mathbb{R}^d} \Phi_t(\mathbf{x}) \Rightarrow$ useful when giving the convergence rate.

It can be 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*)

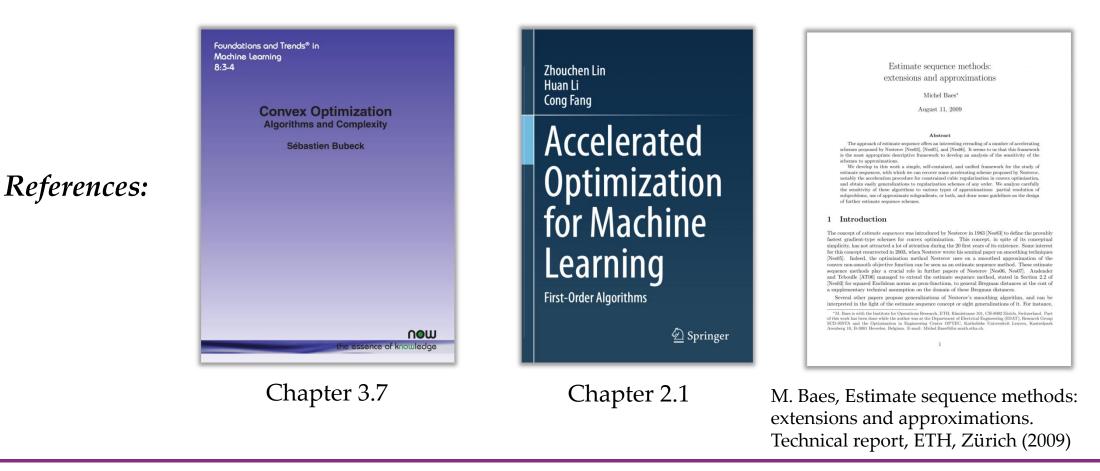
$$\Phi_1(\mathbf{x}) \triangleq f(\mathbf{x}_1) + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}_1\|^2$$
$$\Phi_{t+1}(\mathbf{x}) \triangleq (1 - \theta) \Phi_t(\mathbf{x}) + \theta \left(f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}_t\|^2 \right)$$

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

Advanced Optimization (Fall 2023)

Estimate Sequence

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



Advanced Optimization (Fall 2023)

References for Nesterov's Accelerated GD

Nesterov's four ideas (three acceleration methods):

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



Yurii Nesterov 1956 – UCLouvain, Belgium

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

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

UDC 51

A METHOD OF SOLVING A CONVEX PROGRAMMING PR WITH CONVERGENCE RATE O

YU. E. NESTEROV

1. In this note we propose a method of solving a conve Hilbert space E. Unlike the majority of convex programmi 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 co improved for the class of problems under consideration (see

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

 $||f'(x) - f'(y)|| \le L||x - y||.$ (1) From (1) it follows that for all $x, y \in E$

(2) $f(y) \le f(x) + \langle f'(x), y - x \rangle + 0.5L$ To solve the problem $\min\{f(x) | x \in E\}$ with a nonempty the following method. 0) Select a point $y_0 \in E$. Put

(3) $k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad \alpha_{-1} = ||y_0 - z|| \neq ||$ where z is an arbitrary point in E, $z \neq v_0$ and $f'(z) \neq f'(v_0)$. 1) kth iteration. a) Calculate the smallest index $i \ge 0$ for y

(4) $f(y_k) - f(y_k - 2^{-i}\alpha_{k-1}f'(y_k)) \ge 2^{-i-1}\alpha_k.$ b) Put

5)
$$a_{k} = 2 \ a_{k-1}, \ x_{k} = y_{k} - a_{k}f$$

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

$$y_{k+1} = x_{k} + (a_{k} - 1)(x_{k} - x_{k})$$

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

1980 Mathematics Subject Classification, Primary 90C25.

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 sequence $\{x_k\}_{0}^{\infty}$ is constructed by method (3)–(5), then 1) For any $k \ge 0$;

 $f(x_k) - f^* \leq C / (k + C)$ where $C = 4L ||v_0 - x^*||^2$ and $f^* = f(x^*), x^* \in X^*$. 2) In order to achieve accuracy ε with respect to the [a) to compute the gradient of the objective function n b) to evaluate the objective function no more than N

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

 $f(y_k) - f(y_k(\alpha)) \ge 0.5\alpha(2 - 1)$

Consequently, as soon as $2^{-i}\alpha_{k-1}$ becomes less than and α_k will not be further decreased. Thus $\alpha_k \ge 0.5L^2$ Let $p_k = (a_k - 1)(x_{k-1} - x_k)$. Then $p_{k+1} -$ Consequently,

 $||p_{t+1} - x_{t+1} + x^*||^2 = ||p_t - x_t + x^*||^2 + 2(a_{t+1})^2$ $+2a_{k+1}\alpha_{k+1}\langle f'(y_{k+1}), x \rangle$ Using inequality (4) and the convexity of f(x), we o $\langle f'(y_{k+1}), y_{k+1} - x^* \rangle \ge f(x_{k+1}) - f^*$ $0.5\alpha_{k+1} \|f'(y_{k+1})\|^2 \le f(y_{k+1}) - f(x_{k+1})$ $-a_{k+1}^{-1} \langle f'(y_{k+1}) \rangle$ We substitute these two inequalities into the preceding $\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \le 2(a_k)$ $-2a_{k+1}\alpha_{k+1}(f(x_{k+1}-f^*)+(a_{k+1}^2-a_{k+1}))$ $\leq -2a_{k+1}\alpha_{k+1}(f(x_{k+1}) - f^*) + 2(a_{k+1}^2 - f^*) + 2(a_{k+1}^$ $= 2\alpha_{k+1}a_k^2(f(x_k) - f^*) - 2\alpha_{k+1}a_{k+1}^2(f(x_k) - f^*) - 2\alpha_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{k+1}a_{$ $\leq 2\alpha_k a_k^2 (f(x_k) - f^*) - 2\alpha_{k+1} a_{k+1}^2 (f(x_{k+1}) - f^*) - 2\alpha_{k+1} a_{k+1} (f(x_{k+1}) - f^*) - 2\alpha_{k+1} ($ Thus

```
2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \le 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}))
     \leq 2\alpha_k a_k (f(x_k) - f^*) + ||p_k - x_k + x^*||^2
     \leq 2\alpha_0 a_0^2 (f(x_0) - f^*) + ||p_0 - x_0 + x^*||^2 \leq ||y_0 - x_0|^2
It remains to observe that a_{k+1} \ge a_k + 0.5 \ge 1 + 0.5
 It follows from the estimate of the convergence ra
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 addi function corresponds a halving of α_{i} . Therefore the total not exceed $\left|\log_2(2L\alpha_{-1})\right| + 1$. This completes the proof of If the Lipschitz constant L is known for the gradient of can take $\alpha_k \equiv L^{-1}$ in the method (3)–(5) for any $k \ge 0$. In to hold, and therefore Theorem 1 remains valid $\|y_0 - x^*\| \sqrt{2L/\epsilon} [-1]$ and NF = 0.

To conclude this section we will show how one may me the problem of minimizing a strictly convex function. Assume that $f(x) - f^* \ge 0.5m||x - x^*||^2$ for all $x \in A$ constant m is known.

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

(7) $k \ge 2\sqrt{2}/(m\alpha_{1}) - 2.$

Suppose that the halting has occurred in the Nth step. (3)–(5), one has $N \leq \left|4\sqrt{L/m}\right| - 1$. At the same time,

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

After the point x_N has been obtained, it is necessary begin calculating, by the method (3)-(5), (7), from the po As a result we obtain that after each $\left|4\sqrt{L/m}\right| - 1$ ite to the function decreases by a factor of 2. Thus the i 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:

(8)

$$\min\left\{F(\tilde{f}(x))\right\}$$

 $x \in O$

where O is a convex closed set in E, F(u), with $u \in \mathbb{R}^m$, positive homogeneous of degree one, and $f(x) = (f_1(x))$ continuously differentiable functions on E. The set . assumed to be nonempty. In addition to this, we will a functions $\{F(\cdot), f(\cdot)\}$ 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 $F(u) \equiv \max\{\langle \lambda, u \rangle | \lambda \}$ tions that are positive homogeneous of degree one. Then the convexity of the function F(f(x)) on all of E. Problem (8) can be written in minimax form:

 $\min\{\max\{\langle \lambda, \tilde{f}(x)\rangle | \lambda \in \partial F(0)\}$ (9)

One can show that the fact that the set X^* is nonemp the existence of a saddle point (λ^*, x^*) for problem (9). of problem (9) can be written as $\Omega^* = \Lambda^* \times X^*$, where $\Lambda^* = \operatorname{Arg\,max} \{ \Psi(\lambda) | \lambda \in \partial F(0) \}, \quad \Psi(\lambda) =$

The problem

where

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

```
will be called the problem dual to (8).
  Suppose the functions f_k(x), k = 1, ..., m, in problem (8)
with constants L^{(k)} \ge 0. Let \overline{L} = (L^{(1)}, \dots, L^{(m)}).
  Consider the function
```

 $\Phi(y, A, z) = F(\hat{f}(y, z)) + 0.5A \|y\|$

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

```
and A is a positive constant. Let
```

 $\Phi^*(y, A) = \min\{\Phi(y, A, z) | z \in Q\}, \quad T(y, A) = a$ 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 $\max_{1 \le k \le m} f_k(x)$. For the r as for the "gradient" mapping of [1]) we have (10) $\Phi^*(y, A) + A \langle y - T(y, A), x - y \rangle + 0.5A \|y -$

```
for all x \in O, y \in E and A \ge 0, and if A \ge F(L), then
```

```
\Phi^*(v, A) \ge 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, a_0 = 1, x_{-1} = y_0, A_{-1} =
(11)
where \overline{L}_0 = (L_0^{(1)}, \dots, L_0^{(m)}), \ L_0^{(k)} = ||f_k'(y_0) - f_k'(z)|| / ||y_0 - dx_0^{(k)}|| + ||f_0|| +
in E, z \neq y_0.
```

```
1) kth iteration. a) Calculate the smallest index i \ge 0 for v
(12)
                                 \Phi^{*}(y_{k}, 2^{i}A_{k-1}) \geq F(\tilde{f}(T(y_{k}, 2^{i}A_{k-1})))
```

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

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

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 $\{x_k\}_{k=1}^{\infty}$ is constructed by method assertions are true:

```
1) For any k \ge 0
```

```
F(\tilde{f}(x_k)) - F(\tilde{f}(x^*)) \le C_1 / (k + 1)
where C_1 = 4F(\overline{L})||y_0 - x^*||^2, x^* \in X^*.
```

975

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

 $\sqrt{C_{1}/\epsilon} [+] \max \{ \log_{2}(F(L)/A_{-1}), 0 \} [$

```
times.
```

a) to

b) to evaluate the collection of gradients $f'_1(y), \ldots, f'_m(y)$ no more than $\sqrt[3]{C_1/\varepsilon}$ [times, and c) to evaluate the vector-valued function f(x) at most

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

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 α_{i} the values of A_{i}^{-1} .

Just as in the method (3)-(5), in the method (11)-(13) one can take into account information about the constant $F(\overline{L})$ 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 $\{\Phi(y_1, A, x) | x \in Q\}$ 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

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

in the method (11)-(13), where

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

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

(14)
$$\max\left\{-0.5A^{-1}\left\|\sum_{k=1}^{m}\lambda^{(k)}f_{k}^{\prime}(y)\right\|^{2}+\sum_{k=1}^{m}\lambda^{(k)}f_{k}^{\prime}(y)\mid \left(\lambda^{(1)},\lambda^{(2)},\ldots,m^{(m)}\right)\in\partial F(0)\right\}.$$

where the $\lambda^{(k)}(v)$, $k = 1, \dots, n$

remark that the set $\partial F(0)$ is usu

such cases problem (14) is the sta

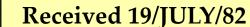
stimulated his interest in the que

French transl., "Mir", Moscow, 1977.

The author expresses his since

Here

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



```
Central Economico-Mathematical Institute
                                                      Received 19/JULY/82
  Academy of Sciences of the USSR
```

BIBLIOGRAPHY

1. A. S. Nemirovskii and D. B. Yudin, Complexity of problems and efficiency of optimization methods, "Nauka" Moscow, 1979. (Russian) 2. B. N. Pshenichnyi and Yu. M. Danilin, Numerical methods in extremal problems, "Nauka", Moscow, 1975;

```
Translated by A. ROSA
```

372

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

УДК 51

ю.е. нестеров метод решения задачи выпуклого про со скоростью сходимости о

(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]²_k=0, 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 миними

L. Гасмогран сначыя задачу скупновно видат f (x). Мы будем предполагать, что функция f (x) принад что существует константа L > 0, для которой при вс неравенство

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

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

(2) $f(y) \le f(x) + \langle f'(x), y - x \rangle + 0.5L ||y - x||^2$.

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, $a_0 = 1$, $x_{-1} = y_0$, $\alpha_{-1} = ||y_0 z|| / ||f'(y_0)||$
- где z любая точка из $E, z \neq y_0 f'(z) \neq f'(y_0)$. 1) k-я Итерация.
- к-я итерация.
 вычисляем наименьший номер i≥0, для которого
- (4) $f(y_k) f(y_k 2^{-i}\alpha_{k-1}f'(y_k)) \ge 2^{-i-1}\alpha_{k-1} \|f'(y_k)\| \le 0$ (4) Полагаем

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

(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 понска (4) ан женному в [2]. Разница лицъ в том, что в (4) дробленин изводится, начиная с α_{k-1} (а не с единицы, как в [2]) тельство теоремы 1) при построении методом (3)-(5) п будет сделано не более $O(\log_2 L)$ таких дроблений. Перест вляется с помощъю "овражного" шага. Отметим также, ч печивает монотонное убывание функции f(x) на пост $|y_k|_{k=0}^k$.

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

1) для любого $k \ge 0$ (6) $f(x_k) - f^* \le C/(k+2)^2$.

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

 для достижения точности є по функционалу необх а) вычислить градиент целевой функции не более NG б) вычислить значение целевой функции не +]log₂(2Lα₋₁)[+1 paз.

Здесь и далее] (•) [– целая часть числа (•). Д о к а з а т е ль с т в о. Пусть $y_k(\alpha) = y_k - \alpha f'(y)$ получаем $f(y_k) - f(y_k(\alpha)) \ge 0.5\alpha(2 - \alpha L) \| f'(y_k) \|^2$. С. $2^{-l}\alpha_{k-1}$ ставет меньше, чем L^{-1} , неравенство (4) выпол уменьшаться не будут. Таким образом, $\alpha_k \ge 0.5L^{-1}$ 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, $||p_{k+1} - x_{k+1} + x_{k+1}\alpha_{k+1}(f'(y_{k+1}), p_k) + 2a_{k+1}\alpha_{k+1}(f'(y_{k+1}), x_{k+1}(f'(y_{k+1})))$.

Пользуясь неравенством (4) и выпуклостью функция $\langle f'(y_{k+1}), y_{k+1} - x^* \rangle \ge f(x_{k+1}) - f^* + 0.5\alpha_{k+1} \| f'(y_{k+1}) \|^2 \le f(y_{k+1}) - f(x_{k+1}) \le f(x_k) - a_{k+1}^{-1} \langle f'(y_{k+1}), p_k \rangle.$

Подставим эти два неравенства в предыдущее равенс $\|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 \le 2(a_{k+1} - 1)$

 $\begin{aligned} & = 2a_{k+1}a_{k+1}(f(x_{k+1}) - f^*) + (a_{k+1}^2 - a_{k+1})a_{k+1}^2 \mathbf{1}^f(y) \\ & \leq -2a_{k+1}a_{k+1}(f(x_{k+1}) - f^*) + 2(a_{k+1}^2 - a_{k+1})a_{k+1}(f(x_{k+1}) - f^*) \\ & = 2a_{k+1}a_{k}^2(f(x_k) - f^*) - 2a_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \leq \\ & - 2a_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*). \end{aligned}$

Таким образом,

544

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

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

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

Осталось заметить, что $a_{k+1} \ge a_k + 0.5 \ge 1 + 0.5(k+1)$. Из оценки скорости сходимости (6) следует, что

по оцелки скорасни скоданости (о) следуст, что ч мое методу (3) – (5) для достижения точности є, не будет При этом на каждой игерации будет вычисляться один гра два значения целевой функции. Заметим, однако, что ка вычислению значения целевой функции соответствует у вдвое. Поэтому общее число таких вычислений не превзо Теорема доказана.

Если для градиента целевой функции известна кометоде (3) – (5) можно положить $\alpha_k\equiv L^{-1}$ при любом kвенство (4) будет заведомо выполнено и позтому утвер нутся верными при $C=2L\,\|\,y_0\,-\,x^\star\,\|^2, NG=\|\,y_0\,-\,x^\star\,\|\sqrt{2}$

В заключение этого раздела покажем, как мож (3)-(5) для решения задачи минимизации сильно выпукл Предположим, что для функции f(x) при всех x ∈.

 $f(x) - f^* \ge 0.5m \|x - x^*\|^2$, где $m \ge 0$, и пусть константа nВведем в метод (3) – (5) следующее правило преры в) Останавливаемся, если

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

Пусть прерывание произошло на *N*-м шаге. Так к $\ge 0.5L^{-1}$, то $N \le]4\sqrt{L/m}[-1$. В то же время

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

После того как получена точка x_N , необходимо с чать счет методом (3) – (5), (7) из точки x_N как из началь

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

```
где Q – выпуклое замкнутое множество из E, F(u), u
```

 K^m положительно-однородная степени единица функция ..., $f_m(x)$ — вектор выпуклых непрерывно дифферен Миожество X^{*} решений задачи (8) всегда предполагает мы всегда будем предполагать, что система функций | дующим свойством:

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

```
Через \partial F(0) в (*) обозначен сублифференциал фун
Как известно, для выпуклых положительно-одр
функций справедливо тождество F(u) \equiv \max\{\langle \lambda, u \rangle \}
предположения (*) следует выпуклость функции F(\bar{f}(x))
Задачу (8) можно записать в минимаксной форме:
```

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

```
Можно показать, что из непустоты множества X<sup>*</sup> и прел
ществование у задачи (9) седловой точки (\lambda^*, x^*). По:
точек задачи (9) представимо в виле \Omega^* = \Lambda^* \times X^*, гл
\in \partial F(0), \Psi(\lambda) = \min\{\lambda, f(x)\}| 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, ..., C^{l,1}(E) сконстантами L^{(k)} \ge 0.06означим \bar{L} = (L^{(1)}, L)
Рассмотрим функцию 0(y, A, z) = F(\bar{C}(y, z)) + = (f^{(1)}(y, z), f^{(2)}(y, z), ..., f^{(m)}(y, z)), f^{(k)}(y, z) = f_k(z)
```

..., т, А – положительная константа. Обозначим

```
\Phi^*(y, A) = \min \{\Phi(y, A, z) \mid z \in Q\}, \quad T(y, A) = \arg \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) = \max \{\Phi(y, A) \mid z \in Q\}, \quad T(y, A) \in A\}
```

```
3.174
```

Отметим, что отображение $y \to T(y, A)$ является естес залачи (8) "градиентного" отображения, введенного в [1 методов минимизации функций вида max $f_k(x)$. Для 1 < k < m

(как и для "градиентного" отображения" из [1]) при все полняется неравенство

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

```
причем если A \ge F(L), то

\Phi^{\bullet}(y, A) \ge F(\overline{f}(T(y, A))).
```

Для решения задачи (8) предлагается следующий м 0) Выбираем точку $y_0 \in E$. Полагаем

```
(11) k = 0, a_0 = 1, x_{-1} = y_0, A_{-1} = F(\overline{L}_0),

rme \overline{L}_0 = (L_0^{(1)}, L_0^{(2)}, \dots, L_0^{(m)}), L_0^{(k)} = \|f'_k(y_0) - f'_k(z)\|

точка на E, z \neq y_0.
```

 k-я Итерация.
 а) Вычисляем наяменьший номер i ≥ 0, для равенство

```
(12) \Phi^{\bullet}(y_k, 2^i A_{k-1}) \ge F(\bar{f}(T(y_k, 2^i A_{k-1}))).
```

```
6) Полагаем A_k = 2^i A_{k-1}, x_k = T(y_k, A_k),
a_{k+1} = (1 + \sqrt{4a_k^2 + 1})/2,
```

```
(13) \frac{a_{k+1} = (1 + \sqrt{4a_k} + 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) $\partial_{AR} a \mu \delta \delta \sigma \sigma k \ge 0 \quad F(\bar{f}(x_k)) - F(\bar{f}(x^*))$ = $4F(\bar{L}) \| y_0 - x^* \|^2, \ x^* \in X^*.$

```
2) для достижения точности є по функционалу нео 
а) решить вспомогательную задачу \min \{ \Phi(y_k,
```

 $\sqrt{C_{1}} (\epsilon +] \max \log_{2}(F(\bar{L})/A_{-1}), 0 \} [pa3, 6]$ 6) вычислить набор градиентов $f'_{1}(y), f'_{2}(y)$

 $\sqrt{C_1/\epsilon}$ [pas,

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

Теорема 2 доказывается практически так же, как только вместо неравенства (2) использовать неравенства 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 метода

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 { $\Phi(y_k, A, x)$ | $x \in Q$ } оказы а) Минимизация гладкой выпуклой функции на

простым множеством мы понимаем такое множество, д ектирования записывается в явном виде. В этом случае в

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

лировали его интерес к рассмотренным вопросам.

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

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

Iри этом
$$T(y, A) = y - A^{-1} \sum_{k=1}^{\infty} \lambda^{(k)}(y) f'_k(y)$$
, где $\lambda^{(k)}(y)$, $k = 1, 2, ..., m, -$ ре-

шения задачи (14) при фиксированном $y \in E$. Отметим, что множество $\partial F(0)$ обычно задается простыми ограничениями – линейными либо квадратичными. В таких случаях задача (14) – стандартная задача квадратичного программирования. Автор искренне признателен А.С. Немировскому за беседы, которые стиму-

```
Центральный экономико-математический институт Поступило
Академии наук СССР, Москва 19 VII 1982
```

ЛИТЕРАТУРА

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

Е.И. НОЧКА

к теории мероморфных кривых

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

1. Пусть задана мероморфная кривая, т.е. мероморфное отображение $\tilde{f}: \mathbf{C} \to \mathbf{CP}^n$.

```
f : C \rightarrow CP'
```

УЛК 515.1

и пусть голоморфное отображение

```
f: C \to C^{n+1}, f = (f_1, f_2, \dots, f_{n+1}),
является редуцированным представлением кривой \tilde{f}. Характеристическую функ-
цию \tilde{f} определим, следуя А. Картану [1]:
```

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

Пусть A – гиперплоскость в **СР**^{*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
```

МАТЕМАТИКА

Lecture 4. Gradient Descent Method II

546

More Explanations for Nesterov's AGD

- Ordinary Differentiable Equations
 - Su, W., Boyd, S., & Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In NIPS 27.
 - Berthier, R., Bach, F., Flammarion, N., Gaillard, P., & Taylor, A. (2021). A continuized view on Nesterov acceleration. ArXiv preprint, arXiv:2102.06035.
- Variational Analysis
 - Wibisono, A., Wilson, A. C., & Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences (PNAS), 113(47), E7351-E7358.
- Linear Coupling of GD and MD
 - Allen-Zhu, Z., & Orecchia, L. (2017). Linear coupling: An ultimate unification of gradient and mirror descent. The 8th Innovations in Theoretical Computer Science Conference (ITCS).
 - Cutkosky A. (2022). Chapter 14 Momentum & Chapter 15 Acceleration. In Lecture Notes for EC525: Optimization for Machine Learning.

Part 3. Extension to Composite Optimization

Composite Optimization

• Proximal Gradient Method (PG)

• Accelerated Proximal Gradient Method (APG)

• Application to LASSO

• 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*: $F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}^{\top}X - \mathbf{y}\|_{2}^{2} + \lambda \|\mathbf{w}\|_{1}$, where $X = [\mathbf{x}_{1}, \dots, \mathbf{x}_{n}], \mathbf{y} = [y_{1}, \dots, y_{n}]^{\top}$.

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

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, \dots, 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.

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

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

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

$$\triangleq U_t(\mathbf{x}) \quad surrogate \ objective$$

 \Box 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

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

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 .

Claim. GD for smooth functions can be equivalently represented by

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

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:

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

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

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

Advanced Optimization (Fall 2023)

Claim. GD for smooth functions can be equivalently represented by

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

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 .

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

linear approximation of f at \mathbf{x}_t prevent \mathbf{x}_t from getting too far

Advanced Optimization (Fall 2023)

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

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

surrogate objective

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

-

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

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

surrogate objective

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

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

$$\underset{\mathbf{x}}{\operatorname{arg\,min}} U_{t}(\mathbf{x}) = \underset{\mathbf{x}}{\operatorname{arg\,min}} \left\{ \frac{L}{2} \left(-2 \left\langle \mathbf{x}_{t} - \frac{\nabla f(\mathbf{x}_{t})}{L}, \mathbf{x} \right\rangle + \|\mathbf{x}\|^{2} \right) + h(\mathbf{x}) \right\}$$
$$= \left[\underset{\mathbf{x}}{\operatorname{arg\,min}} \left\{ \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{x}_{t} - \frac{\nabla f(\mathbf{x}_{t})}{L} \right) \right\|^{2} + h(\mathbf{x}) \right\} \right]$$

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

• Iteratively solve the surrogate optimization problem.

Deploying the following update rule:

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

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

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

Proximal Gradient

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

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

Proximal Gradient Method

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

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

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

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

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

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*

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

Proximal gradient can also achieve an 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 σ -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 $\mathbf{u} \in \mathcal{X}$,

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

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

Proof of PG Convergence

Proof:

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

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

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

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

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

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

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

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

Proof of PG Convergence

Proof:

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

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: $F(\mathbf{x}_T) - F(\mathbf{x}^*)$

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

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

Advanced Optimization (Fall 2023)

Proof of PG Convergence

Proof:

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

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

What we have:

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

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

Advanced Optimization (Fall 2023)

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

$$\int_{U_{t}(\mathbf{x}_{t+1}) = f(\mathbf{x}_{t}) + \langle \nabla f(\mathbf{x}_{t}), \mathbf{x}_{t+1} - \mathbf{x}_{t} \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|_{2}^{2} + h(\mathbf{x}_{t+1}) \\
F(\mathbf{u}) = f(\mathbf{u}) + h(\mathbf{u}) \ge f(\mathbf{x}_{t}) + \langle \nabla f(\mathbf{x}_{t}), \mathbf{u} - \mathbf{x}_{t} \rangle + h(\mathbf{x}_{t+1}) + \langle \nabla h(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1} \rangle \quad \text{(convexity)}$$

$$\Box > U_{t}(\mathbf{x}_{t+1}) - F(\mathbf{u}) \le \langle \nabla f(\mathbf{x}_{t}) + \nabla h(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{u} \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|_{2}^{2} = \frac{1}{2L} \|g(\mathbf{x}_{t})\|^{2} \quad (g(\mathbf{x}_{t}) \triangleq L(\mathbf{x}_{t} - \mathbf{x}_{t+1}))$$

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: $F(\mathbf{x}) \leq U_t(\mathbf{x})$ 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*

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

$$\mathbf{x}_{t+1} = \operatorname*{arg\,min}_{\mathbf{x}} \left\{ \frac{h(\mathbf{x}) + \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t) \right) \right\|^2}{\triangleq H(\mathbf{x})} \right\}$$
$$\stackrel{\text{Definition}}{=} H(\mathbf{x})$$

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*

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

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

Advanced Optimization (Fall 2023)

One-Step Improvement Lemma

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

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

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

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

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

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

Accelerated Proximal Gradient Method

• A natural idea

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

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

Nesterov's Accelerated GD $\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t), \quad \mathbf{x}_{t+1} = (1 - \alpha_t) \mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$

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

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

Advanced Optimization (Fall 2023)

Accelerated Proximal Gradient Method

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

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

Suppose that h is convex and f is σ -strongly convex and L-smooth. Setting the parameters properly, APG enjoys

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

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: ℓ_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*.

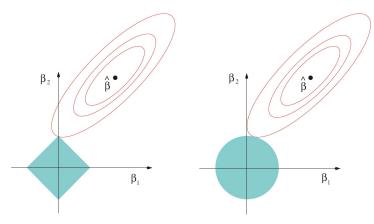
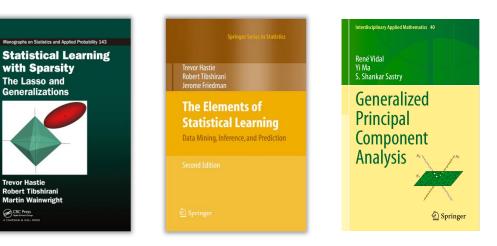


FIGURE 3.11. Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the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

Advanced Optimization (Fall 2023)





Application to LASSO

• **LASSO:** ℓ_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*.

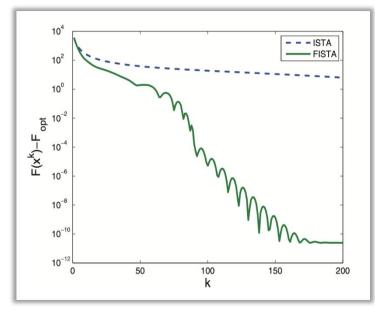
composite optimization: first part is *smooth*, the other one is *non-smooth*

- ISTA (Iterative Shrinkage-Thresholding Algorithm): PG for LASSO
- FISTA (Fast ISTA): APG for LASSO

Closed-form solution: $[\mathcal{P}_{L}^{h}(\mathbf{w}_{t})]_{i} = \operatorname{sign}\left(\left[\mathbf{w}_{t} - \frac{1}{L}\nabla f(\mathbf{w}_{t})\right]_{i}\right)\left(\left|\left[\mathbf{w}_{t} - \frac{1}{L}\nabla f(\mathbf{w}_{t})\right]_{i}\right| - \frac{\lambda}{L}\right)_{+}$

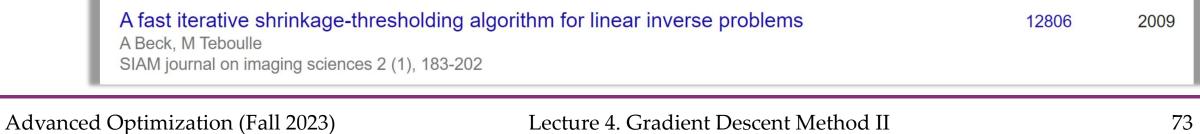
Application to LASSO

• Comparison of **ISTA** and **FISTA**



Comparison of ISTA and FISTA.

SIAM J. IM/ Vol. 2, No. 1	AGING SCIENCES 1, pp. 183–202	© 2009 Society for Industrial and Applied Mathem
		tive Shrinkage-Thresholding Algorithm or Linear Inverse Problems*
)	Amir Beck [†] and Marc Teboulle [‡]
Abstract.	problems arising in signal/ tension of the classical grad- solving large-scale problem converge quite slowly. In the (FISTA) which preserves the which is proven to be signi- merical results for wavelet-	rative shrinkage-thresholding algorithms (ISTA) for solving linear inve- image processing. This class of methods, which can be viewed as an lent algorithm, is attractive due to its simplicity and thus is adequate seven with dense matrix data. However, such methods are also known is paper we present a new fast iterative shrinkage-thresholding algorith e computational simplicity of ISTA but with a global rate of convergen ficantly better, both theoretically and practically. Initial promising based image deblurring demonstrate the capabilities of FISTA which TA by several orders of magnitude.
Key word		olding algorithm, deconvolution, linear inverse problem, least squares a optimal gradient method, global rate of convergence, two-step iterati ng
AMS sub	ject classifications. 90C25,	90C06, 65F22
DOI 10		
	ntroduction. Linear inv	verse problems arise in a wide range of applications such a
1. I astrophy few. Th which in the mon	ntroduction. Linear inv ysics, signal and image he interdisciplinary nati neludes a large body of r nograph [13] and the refe	verse problems arise in a wide range of applications such a processing, statistical inference, and optics, to name just are of inverse problems is evident through a vast literatuu nathematical and algorithmic developments; see, for instance rences therein. m leads us to study a discrete linear system of the form
1. I astrophy few. Th which in the mon A ba	ntroduction. Linear inv ysics, signal and image he interdisciplinary nati neludes a large body of r nograph [13] and the refe	processing, statistical inference, and optics, to name just are of inverse problems is evident through a vast literatur nathematical and algorithmic developments; see, for instance rences therein.
1. If astrophy few. Th which in the mon A ba (1.1) where A and x is for exan whose s formed applicat	ntroduction. Linear inv ysics, signal and image he interdisciplinary nat- net the interdisciplinary nat- nograph [13] and the refe asic linear inverse proble $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$ is the "true" and unknow nple, $\mathbf{b} \in \mathbb{R}^m$ represents ize is assumed to be th by stacking the column ions, the matrix \mathbf{A} desci- presents a two-dimension	processing, statistical inference, and optics, to name just are of inverse problems is evident through a vast literatur nathematical and algorithmic developments; see, for instance rences therein. en leads us to study a discrete linear system of the form



Summary

