



Lecture 5. Gradient Descent Method II

Advanced Optimization (Fall 2022)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

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

Part 1. GD for Smooth Optimization

Smooth and Convex

Smooth and Strongly Convex

Extension to Constrained Case

Overview

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

Function Family		Step Size	Output Sequence	Convergence Rate	
<i>G</i> -Lipschitz	convex σ -strongly convex	$\eta = \frac{D}{G\sqrt{T}}$ $\eta_t = \frac{2}{\sigma(t+1)}$	$ar{\mathbf{x}}_T = rac{1}{T} \sum_{t=1}^T \mathbf{x}_t$ $ar{\mathbf{x}}_T = \sum_{t=1}^T rac{2t}{T(T+1)} \mathbf{x}_t$	$\mathcal{O}(1/\sqrt{T})$ $\mathcal{O}(1/T)$	last lecture
<i>L</i> -smooth	convex σ -strongly convex	$\eta = \frac{1}{L}$ $\eta = \frac{2}{\sigma + L}$	$ar{\mathbf{x}}_T = \mathbf{x}_T \ ar{\mathbf{x}}_T = \mathbf{x}_T$	$\mathcal{O}(1/T)$ $\mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right)$	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 $f : \mathbb{R}^d \to \mathbb{R}$ 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 $t \geq 0$,

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

Proof:
$$\|\mathbf{x}_{t+1} - \mathbf{x}^{\star}\|^{2} = \|\Pi_{\mathcal{X}}[\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t})] - \mathbf{x}^{\star}\|^{2}$$
 (GD)
$$\leq \|\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t}) - \mathbf{x}^{\star}\|^{2}$$
 (Pythagoras Theorem)
$$= \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} - 2\eta_{t}\langle\nabla f(\mathbf{x}_{t}), \mathbf{x}_{t} - \mathbf{x}^{\star}\rangle + \eta_{t}^{2}\|\nabla f(\mathbf{x}_{t})\|^{2}$$

$$\leq \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} - 2\eta_{t}(f(\mathbf{x}_{t}) - f^{\star}) + \eta_{t}^{2}\|\nabla f(\mathbf{x}_{t})\|^{2}$$
(convexity: $f(\mathbf{x}_{t}) - f^{\star} = f(\mathbf{x}_{t}) - f(\mathbf{x}^{\star}) \leq \langle\nabla f(\mathbf{x}_{t}), \mathbf{x}_{t} - \mathbf{x}^{\star}\rangle$)

Refined Result for Smooth Optimization

Proof:
$$\|\mathbf{x}_{t+1} - \mathbf{x}^{\star}\|^{2} = \|\Pi_{\mathcal{X}}[\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t})] - \mathbf{x}^{\star}\|^{2}$$
 (GD)
$$\leq \|\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t}) - \mathbf{x}^{\star}\|^{2} \text{ (Pythagoras Theorem)}$$

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

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

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

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

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

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

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*.

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

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

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

exploiting coercivity of smoothness and unconstrained first-order optimality

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

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

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

$$\leq \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} \leq \ldots \leq \|\mathbf{x}_{1} - \mathbf{x}^{\star}\|^{2}$$
which already implies the convergence

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

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

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

$$= f(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)) - f(\mathbf{x}_t)$$

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

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

$$= -\frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 \text{ (recall that we have picked } \eta_t = \eta = \frac{1}{L})$$

one-step improvement

Proof:

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

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

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

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

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

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

Key Lemma for Smooth GD

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

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

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

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

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

average-iterated convergence vs last-iterated convergence

Key Lemma for Smooth GD

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

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

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

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

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

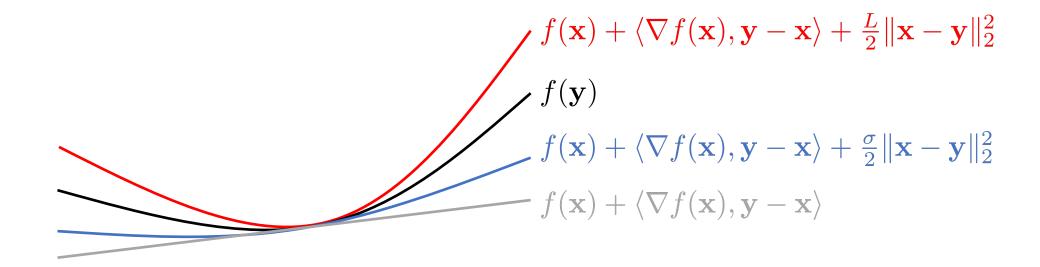
• Recall the definition of strongly convex functions (*first-order* version).

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

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

$$f \text{ is } \sigma\text{-strongly convex} \qquad \qquad f \text{ is } L\text{-smooth}$$

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



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

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

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.

Proof:
$$\|\mathbf{x}_{t+1} - \mathbf{x}^{\star}\|^{2} = \|\Pi_{\mathcal{X}}[\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t})] - \mathbf{x}^{\star}\|^{2}$$
 (GD)
$$\leq \|\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t}) - \mathbf{x}^{\star}\|^{2}$$
 (Pythagoras Theorem)
$$= \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} - 2\eta_{t} \langle \nabla f(\mathbf{x}_{t}), \mathbf{x}_{t} - \mathbf{x}^{\star} \rangle + \eta_{t}^{2} \|\nabla f(\mathbf{x}_{t})\|^{2}$$

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

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

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

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

Then, rearranging the terms finishes the proof.

Proof:
$$\|\mathbf{x}_{t+1} - \mathbf{x}^{\star}\|^{2} = \|\Pi_{\mathcal{X}}[\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t})] - \mathbf{x}^{\star}\|^{2}$$
 (GD)
$$\leq \|\mathbf{x}_{t} - \eta_{t}\nabla f(\mathbf{x}_{t}) - \mathbf{x}^{\star}\|^{2} \text{ (Pythagoras Theorem)}$$

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

$$\leq \left(1 - \frac{2\eta_{t}\sigma L}{L + \sigma}\right) \|\mathbf{x}_{t} - \mathbf{x}^{\star}\|^{2} + \left(\eta_{t}^{2} - \frac{2\eta_{t}}{L + \sigma}\right) \|\nabla f(\mathbf{x}_{t})\|^{2}$$
exploiting co-coercivity of smoothness and strong convexity

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

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

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

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 (simple) setting: $\eta_t = \eta = \frac{2}{L + \sigma}$

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

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

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

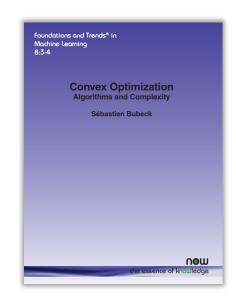
- In unconstrained case, $g(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)$.
- In unconstrained case, setting $\mathbf{u} = \mathbf{x}_t$ recovers the one-step improvement: $f(\mathbf{x}_{t+1}) f(\mathbf{x}_t) \le -\frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2$.

Constrained Optimization

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

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

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



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

Lower Bound

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

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

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

Function Family		Convergence Rate	Lower Bound	Optimal?
G-Lipschitz	convex	$\mathcal{O}(1/\sqrt{T})$	$\Omega(1/\sqrt{T})$	\checkmark
	σ -strongly convex	$\mathcal{O}(1/T)$	$\Omega(1/T)$	√
<i>L</i> -smooth	convex	$\mathcal{O}(1/T)$	$\Omega(1/T^2)$	X
	σ -strongly convex	$\mathcal{O}\left(\exp\left(-\frac{T}{\kappa}\right)\right)$	$\Omega\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right)\right)$	×

 $\qquad \qquad \Box >$

GD is suboptimal in *smooth* convex optimization!

Part 2. Nesterov's Accelerated GD

Algorithm

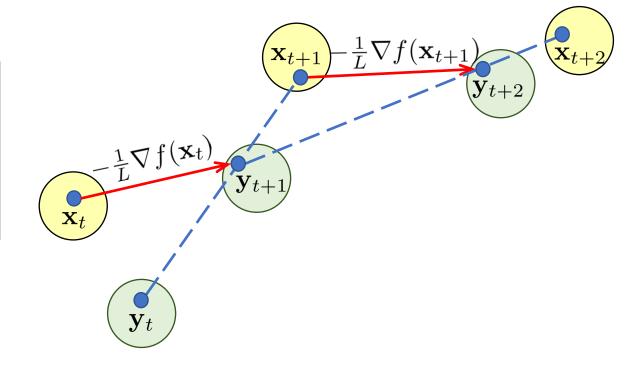
Smooth and Convex

Smooth and Strongly Convex

Nesterov's Accelerated GD

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

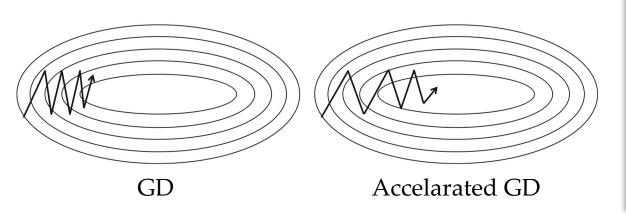
$$\mathbf{x}_{t+1} = (1 - \alpha_t) \mathbf{y}_{t+1} + \alpha_t \mathbf{y}_t$$

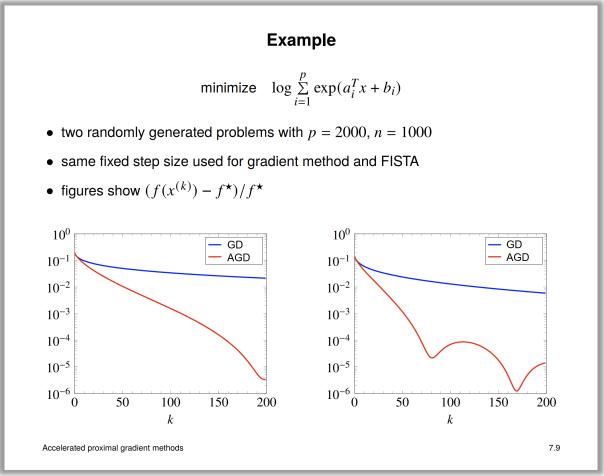


- Define $\mathbf{x}_1 = \mathbf{y}_1$.
- $\alpha_t < 0$ is a *time-varying* mixing rate of \mathbf{y}_t and \mathbf{y}_{t+1} .
- $\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \alpha_t(\mathbf{y}_t \mathbf{y}_{t+1})$ is an *extrapolated* point, i.e., with *momentum*.

Nesterov's Accelerated GD

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





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

Convergence of Nesterov's Accelerated GD

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

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

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

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

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:

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

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

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

Generalized One-Step Improvement

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

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

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

Proof:

Froof:
$$f(\mathbf{x}_{t+1}) - f(\mathbf{u}) = f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) + f(\mathbf{x}_t) - f(\mathbf{u})$$

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

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

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

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

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

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

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

That is

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

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

Proof: (continued proving Theorem 3)

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

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

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

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

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

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

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

Proof: (continued proving Theorem 3)

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

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t$$

$$\leq \frac{1}{2L} (L^2 \| \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^* \|^2 - \| L(\lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^*) - \lambda_t \nabla f(\mathbf{x}_t) \|^2)$$

$$= \frac{L}{2} \left(\| \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^* \|^2 - \| \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^* - \lambda_t \frac{\nabla f(\mathbf{x}_t)}{L} \|^2 \right)$$

$$= \frac{L}{2} (\| \lambda_t \mathbf{x}_t - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^* \|^2 - \| \lambda_t \mathbf{y}_{t+1} - (\lambda_t - 1)\mathbf{y}_t - \mathbf{x}^* \|^2)$$

Goal: design a telescoping series

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

telescope

Proof: (continued proving Theorem 3)

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

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

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

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

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

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

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

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): $\mathbf{x}_1 = \mathbf{y}_1$

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

Proof

Proof: (continued proving Theorem 3)

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

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

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

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

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

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

Requirement (3): $\lambda_0 = 0$

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

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

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

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 proven that the above construction satisfies the two properties.

Smooth and Strongly Convex

Proof sketch

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

$$\Phi_{1}(\mathbf{x}) \triangleq f(\mathbf{x}_{1}) + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}_{1}\|^{2}$$

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

$$f(\mathbf{y}_{t}) - f(\mathbf{x}^{\star}) \stackrel{(ii)}{\leq} \min_{\mathbf{x} \in \mathbb{R}^{d}} \Phi_{t}(\mathbf{x}) - f(\mathbf{x}^{\star}) \leq \Phi_{t}(\mathbf{x}^{\star}) - f(\mathbf{x}^{\star}) \qquad \text{(by property (ii))}$$

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

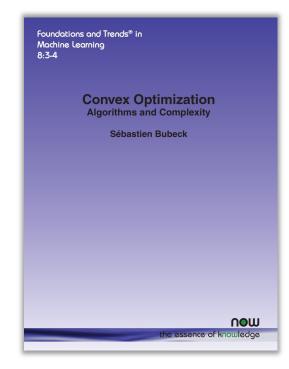
$$= (1 - \theta)^{t} \left(f(\mathbf{x}_{1}) + \frac{\sigma}{2} \|\mathbf{x}^{\star} - \mathbf{x}_{1}\|^{2} - f(\mathbf{x}^{\star}) \right) \qquad \text{(definition of } \Phi_{1})$$

$$\lesssim (\sigma + L) \|\mathbf{x}^{\star} - \mathbf{x}_{1}\|^{2} \exp(-\theta t) \qquad \text{(smoothness)}$$

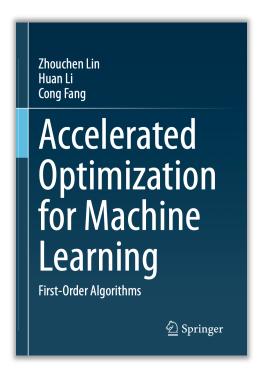
Estimate Sequence

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

References:



Chapter 3.7



Chapter 2.1



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

References for Nesterov's Accelerated GD

Nesterov's four ideas (three acceleration methods):

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



Yurii Nesterov 1956 – UCLouvain, Belgium

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

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

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

UDC 51

YU. E. NESTEROV

1. In this note we propose a method of solving a convertible the majority of convex programming this method constructs a minimizing sequence of points (a This property allows us to reduce the amount of computation At the same time, it is possible to obtain an estimate of computation 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 $x, y \in E$

(1)
$$||f'(x) - f'(y)|| \le L||x - y||.$$
 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, $a_0 = 1$, $x_{-1} = y_0$, $\alpha_{-1} = |y_0 - z|/|$ 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 > 0 for

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

b) Put

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

$$y_{k+1} = x_k + (a_k - 1)(x_k - x_{k-1})/2$$

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 $\{x_k\}_0^\infty$ 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 guarathe 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 $k \ge 0$;

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

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

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

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

Using inequality (4) and the convexity of f(x), we o

$$\langle f'(y_{k+1}), y_{k+1} - x^* \rangle \ge f(x_{k+1}) - f^*$$

 $0.5\alpha_{k+1} || f'(y_{k+1}) ||^2 \le f(y_{k+1}) - f(x_k + a_{k+1}^{-1}) || f'(y_{k+1})$

We substitute these two inequalities into the preceding

$$\begin{split} \|p_{k+1} - x_{k+1} + x^*\|^2 - \|p_k - x_k + x^*\|^2 &\leq 2(a_k) \\ -2a_{k+1}\alpha_{k+1}(f(x_{k+1} - f^*) + (a_{k+1}^2 - a_{k+1} \\ &\leq -2a_{k+1}a_{k+1}(f(x_{k+1}) - f^*) + 2(a_{k+1}^2 - 2a_{k+1}a_{k+1}^2) \\ &= 2\alpha_{k+1}a_k^2(f(x_k) - f^*) - 2\alpha_{k+1}a_{k+1}^2(f(f_k) \\ &\leq 2\alpha_k a_k^2(f(x_k) - f^*) - 2\alpha_{k+1}a_{k+1}^2(f(f_k) - f^*) - 2\alpha_{k+1}a_{k+1}^2(f(f_$$

Thus

$$\begin{split} & 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \leq 2\alpha_{k+1}a_{k+1}^2(f(x_{k+1}) - f^*) \\ & \leq 2\alpha_ka_k(f(x_k) - f^*) + \|p_k - x_k + x^*\|^2 \\ & \leq 2\alpha_0a_0^2(f(x_0) - f^*) + \|p_0 - x_0 + x^*\|^2 \leq \|y_0 - f^*\|^2 \end{split}$$

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 ratmethod (3)–(5) needs to achieve accuracy ε will be neach iteration, one gradient and at least two values of be calculated. Let us remark, however, that to each addit function corresponds a halving of α_k . Therefore the total not exceed $\lfloor \log_2(2L\alpha_{-1})\rfloor + 1$. 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 $\|\|y_0 - x^n\|\sqrt{2L/\varepsilon}\|$ –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

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

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

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

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

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 n cannot be improved (up to a dimensionless constant) ame class of strictly convex functions in $C^{1,1}(E)$ (see [1]).

3. Consider the following extremal problem:

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

where Q is a convex closed set in E, F(u), with $u \in R^m$, positive homogeneous of degree one, and $\hat{f}(x) = (f_{\dagger}(x))$ continuously differentiable functions on E. The set X assumed to be nonempty. In addition to this, we will a functions $\{F(\cdot), \hat{f}(\cdot)\}$ 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. The the convexity of the function $F(\bar{f}(x))$ on all of E.

Problem (8) can be written in minimax form:

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

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

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

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 $L^{(k)} \ge 0$. Let $L=(L^{(1)},\ldots,L^{(m)})$. Consider the function

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

where

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

and A is a positive constant. Let

$$\Phi^*(y, A) = \min\{\Phi(y, A, z) \mid z \in Q\}, \quad T(y, A) = \text{ar}$$

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

(10)
$$\Phi^*(y, A) + A\langle y - T(y, A), x - y \rangle + 0.5A||y - 5$$

for all
$$x \in Q$$
, $y \in E$ and $A \ge 0$, and if $A \ge F(L)$, then
$$\Phi^*(y, A) \ge F(\tilde{f}(T(v, A))).$$

To solve problem (8) we propose the following method. 0) Select a point $y_0 \in E$. Put

(11)
$$k = 0, \quad a_0 = 1, \quad x_{-1} = y_0, \quad A_{-1} = I$$
 where $\bar{L}_0 = (L_0^{(1)}, \dots, L_0^{(m)}), \quad L_0^{(k)} = \|f_k'(y_0) - f_k'(z)\|/\|y_0 - z\|$ in $F_0 \neq 0$.

1) kth iteration. a) Calculate the smallest index $i \ge 0$ for w

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

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

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

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 $\{x_k\}_0^{\infty}$ 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 $C_1 = 4F(\overline{L})||y_0 - x^*||^2, x^* \in X^*.$

2) To obtain accuracy ε 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

times

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

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

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

time

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_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 $F(\bar{I}(x)) - m$ (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 $\min\{\Phi(v_t, A, x) | x \in Q\}$ 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

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

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

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

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

lere

$$T(y, A) = y - A^{-1} \sum_{k=0}^{m} \lambda^{(k)}(y) f'_{k}(y),$$

where the $\lambda^{(k)}(y)$, k = 1, ..., m, remark that the set $\partial F(0)$ is usus such cases problem (14) is the sta. The author expresses his since

Received 19/JULY/82

stimulated his interest in the questions considere Central Economico-Mathematical Institute

Academy of Sciences of the USSR

Received 19/JULY/82

BIBLIOGRAPHY

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

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 ПРОІ СО СКОРОСТЬЮ СХОДИМОСТИ O

(Представлено академиком Л.В. Канторовичем

- 1. В статье предлагается метод решения задачи вания в гильберговом пространстве E. В отличее от бол лого программирования, предлагавшихся ранее, этот ме шую 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 миним $f(\mathbf{x})$. Мы будем предполагать, что функция $f(\mathbf{x})$ 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 репнения задачи $\min\{f(x) \mid x \in E\}$ с непусты X^* предлагается следующий метод.

Выбираем точку y₀ ∈ 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 Итерания.
 - а) Вычисляем наименьший номер $i \ge 0$, шля которого
- (4) $f(y_k) f(y_k 2^{-i}\alpha_{k-1}f'(y_k)) \ge 2^{-i-1}\alpha_{k-1} \| f'(y_k) \|$ 6) 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 поиска (4) ак женному в [2]. Разница лишь в том, что в (4) дроблени 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. Перес вляется с помощью "овражного" шага. Отметим также, ч печивает монотонное убывание функции f(x) на посл $\|y_k\|_{k=0}^k$.

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

- 1) для любого k≥0
- (6) $f(x_k) f^* \le C/(k+2)^2$,
- $e\partial e\ C = 4L\,\|\,y_0 x^*\,\|^2, \ f^* = f(x^*), \ x^* \in X^*;$
 - для достижения точности є по функционалу необх
 вычислить градиент целевой функции не более NG
- 5) вычислить значение целевой функции не $+ \log_2(2L\alpha_{-1})[+1$ раз.

Здесь и далее] (·) [— целая часть числа (·). $\mathbb H$ о к а з а т е л ь с т в о. Пусть $y_k(\alpha) = y_k - \alpha f'(y_k)$ 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 $p_k + a_{k+1}\alpha_{k+1}f'(y_{k+1})$. Спедовательно, $\|p_{k+1} - x_{k+1} + x_$

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

Подставим эти два неравенства в предыдущее равенс $\|p_{k+1}-x_{k+1}+x^*\|^2-\|p_k-x_k+x^*\|^2\leqslant 2(a_{k+1}-1)^2-a_{k+1}\alpha_{k+1}(f(x_{k+1}-f^*)+(a_{k+1}^2-a_{k+1})\alpha_{k+1}^2\|f'(y_k)\|^2-2a_{k+1}\alpha_{k+1}(f(x_{k+1})-f^*)+2(a_{k+1}^2-a_{k+1})\alpha_{k+1}(g_k)\|^2-2\alpha_{k+1}a_{k}^2(f(x_k)-f^*)-2\alpha_{k+1}a_{k+1}^2(f(x_{k+1})-f^*)\leqslant -2\alpha_{k+1}a_{k+1}^2(f(x_{k+1})-f^*).$

Таким образом,

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

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

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

мое методу (3)—(5) для достижения точности є, не будет При этом на каждой итерации будет вычисляться один гр. два значения целевой функции. Заметим, однако, что ка вычислению значения целевой функции соответствует у 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 утвер нутся верными пои $C = 2L \|v_n - x^*\|^2$, $NC = \|v_n - x^*\|\sqrt{2}$

3. 1

- В заключение этого раздела покажем, как мож (3)-(5) для решения задачи минимизации сильно выпукл Предположим, что для функции f(x) при всех $x \in I$
- $f(x) f^* \geqslant 0.5m \|x x^*\|^2$, где m > 0, и пусть константа n Введем 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 $\geqslant 0,5L^{-1}$, то $N\leqslant]4\sqrt{L/m}[-1.$ В то же время

$$f(x_N) - f^* \le \frac{2\|y_0 - x^*\|^2}{\alpha_N(N+2)^2} \le 0.25m\|y_0 - x^*\|^2 \le 0$$

После того как получена точка x_N , необходимо о чать счет методом (3) – (5), (7) из точки x_N как из началы В результате получаем, что за каждые $14\sqrt{L/m}$

- функции убывает выбов. Таким образом, метод (3)-(5) ется неулучшаемым (с точностью до безразмерной конс вого порядка на классе сильно выпуклых функций из $\mathbb{C}^{I,j}$. З. Рассмотрим следующую экстремальную задачу:
- (8) $\min\{F(\bar{f}(x))|x\in O\},\$

где Q — выпуклое замкнутое множество из E, F(u), u є R^m положительно-однородная степени единица функция \ldots , $f_m(x)$) — вектор выпуклых непрерывно диффере 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-одфункций справедливо тождество $F(u) \equiv \max\{\langle \lambda, u \rangle |$ предположения (*) следует выпуклость функции $F(\overline{f}(x))$ Задачу (8) можно записать в минимаксной форме;

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

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

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

мы будем называть задачей, двойственной к присть 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 $\Phi(y,A,z)=F(\bar{f}(y,z))+1$ ($f^{(1)}(y,z),f^{(2)}(y,z),\ldots,f^{(m)}(y,z),f^{(k)}(y,z)=f_k(y,z)$), $f^{(k)}(y,z)=f_k(y,z)$, $f^{(k)}(y,z)=f_k(y,z)$

 $\Phi^*(v, A) = \min \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T(v, A) = \arg \{\Phi(v, A, z) | z \in Q\}, T($

3. 174

Отметим, что отображение $y \to T(y,A)$ является естес задачи (8) "градиентного" отображения, введенного в [1 методов минимизации функций вида $\max_{k} f_k(x)$. Для

(как и для "градиентного" отображения" из [1]) при всполняется 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 мо 0) Выбираем точку $y_0 \in E$. Полагаем

- (11) k=0, $a_0=1$, $x_{-1}=y_0$, $A_{-1}=F(\bar{L}_0)$, ref $\bar{L}_0=(L_0^{(1)},L_0^{(2)},\ldots,L_0^{(m)})$, $L_0^{(k)}=\|f_k'(y_0)-f_k'(z)\|/1$ toka as $\bar{L},z\neq y_0$.
- к-я Итерация.
 Вычисляем наименьший номер i ≥ 0, для
- (12) $\Phi^*(y_k, 2^i A_{k-1}) \ge F(\bar{f}(T(y_k, 2^i A_{k-1}))).$
 - 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 $m=1,\ F(y)=y,\ Q=E)$.

 $(x_k)^{\infty} = 0$ Теорема 2. Если последовательность $\{x_k\}_{k=0}^{\infty}$

(13), to: 1) dar andoeo $k \ge 0$ $F(\bar{f}(x_k)) - F(\bar{f}(x^*))$

 $=4F(\bar{L})\|y_0-x^*\|^2,\ x^*\in X^*.$ 2) для достижения точности є по функционалу необ

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

6) вычислить набор градиентов $f_1'(y), f_2'(y) \sqrt{C_1/\epsilon}[pa3,$

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 $\alpha_k f'(\nu_k)$ будет вектор $\nu_k - T(\nu_k, A_k)$, а апа. Точно так же, как и в методе (3)—(5), в методе

информацию о константе $F(\bar{L})$ и параметре сильной выпул— 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 гламунлой 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 $T(y, A) = \operatorname{argmin} \{ || y - A^{-1} f'(y) - z || | z \in Q \}.$

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

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

$$\in \partial F(0)\Big\}.$$
 При этом $T(y,A) = y - A^{-1}\sum_{k=1}^{m}\lambda^{(k)}(y)f_k'(y)$, где $\lambda^{(k)}(y)$, $k=1,2,\ldots,m$, — 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 стимулировали его интерес к рассмотренным вопросам.

Центральный экономико-математический институт Академии наук СССР, Москва

Поступило

ЛИТЕРАТУРА

 Немировский А.С., Юдин Д.Б. Спожность задач и эффективность методов оптимизации. М.: Наука, 1979.
 Пшеничный Б.Н., Данилин Ю.М. Численные методы в экстремальных залачах 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 $\widetilde{f} \colon \ \ \mathbf{C} \ \to \ \mathbf{CP}^n.$

и пусть голоморфное отображение

$$f: \mathbb{C} \to \mathbb{C}^{n+1}, \quad f = (f_1, f_2, \dots, f_{n+1}),$$

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

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

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

547

More Explanations for Nesterov's AGD

Ordinary Differentiable Equations

- Su, W., Boyd, S., & Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. Advances in neural information processing systems 27 (NIPS).
- Berthier, R., Bach, F., Flammarion, N., Gaillard, P., & Taylor, A. (2021). A continuized view on Nesterov acceleration. arXiv preprint arXiv:2102.06035.

Variational Analysis

• Wibisono, A., Wilson, A. C., & Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences (PNAS), 113(47), E7351-E7358.

Linear Coupling of GD and MD

• Allen-Zhu, Z., & Orecchia, L. (2017). Linear coupling: An ultimate unification of gradient and mirror descent. The 8th Innovations in Theoretical Computer Science Conference (ITCS).

Part 3. Extension to Composite Optimization

Composite Optimization

• Proximal Gradient Method (PG)

Accelerated Proximal Gradient Method (APG)

Application to LASSO

Composite Optimization

Problem setup

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

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

• The composite optimization problem is common in practice.

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

Recall Non-composite Optimization

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

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

$$\triangleq Q(\mathbf{x}; \mathbf{y}) \text{ surrogate objective}$$

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

Claim. GD for smooth functions can be equivalently represented by

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

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

Another View of GD Method

Claim. GD for smooth functions can be equivalently represented by

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

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

Proof:

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

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

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

Another View of GD Method

Claim. GD for smooth functions can be equivalently represented by

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

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

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

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

Composite Optimization

Problem setup

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

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

An idea:

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

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

Formulation

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

$$\triangleq q(\mathbf{x}; \mathbf{y})$$

surrogate objective

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

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

Composite Optimization

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

surrogate objective

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

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

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

Composite Optimization

• Iteratively solve the surrogate optimization problem.

Deploying the following update rule:

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

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

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

Proximal Gradient

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

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

Proximal Gradient Method.

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

Proximal Gradient

Proximal Gradient Method.

$$\mathbf{x}_{t+1} = \mathcal{P}_{L}^{h}(\mathbf{x}_{t}) \triangleq \mathbf{prox}_{\frac{1}{L}h} \left(\mathbf{x}_{t} - \frac{1}{L} \nabla f(\mathbf{x}_{t}) \right)$$

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

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

Convergence of Proximal Gradient

Smooth Optimization

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

assumption: f is L-smooth

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

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

Smooth Composite Optimization

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

assumption: f is L-smooth, h not

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

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

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 $\mathcal{O}(1/T)$ 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 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 $\mathcal{O}(1/T)$ 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 $\mathbf{u} \in \mathcal{X}$,

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

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

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

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

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

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

Proof of PG Convergence

Proof:

What we want: $F(\mathbf{x}_T) - F(\mathbf{x}^*) \Rightarrow \text{Next step:}$ 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:

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

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

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

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

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

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

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

$$= \frac{1}{2L} \|g(\mathbf{x}_t)\|^2 \left(g(\mathbf{x}_t) \triangleq L(\mathbf{x}_t - \mathbf{x}_{t+1})\right)$$

Proof of One-Step Improvement Lemma

Proof:

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

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\operatorname{arg\,min}} \left\{ \underline{h(\mathbf{x}) + \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t) \right) \right\|^2} \right\} \\ \stackrel{\triangle}{=} H(\mathbf{x}) \\ \boxed{ by \, \textit{Fermat's} }$$

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

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

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

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

Proof of One-Step Improvement Lemma

Proof:

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

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

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

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

One-Step Improvement Lemma

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

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

specialized

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

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

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

general

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

Accelerated Proximal Gradient Method

A natural idea

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



This induces the Accelerated Proximal Gradient (APG) method.

Nesterov's Accelerated GD

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

Accelerated Proximal Gradient

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

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

Accelerated Proximal Gradient Method

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

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

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

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

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

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

commonly encountered in *signal/image processing*.

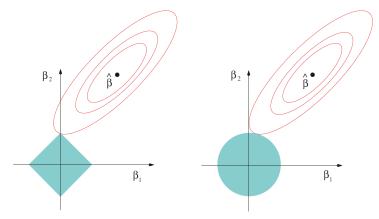
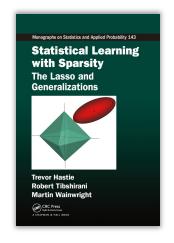
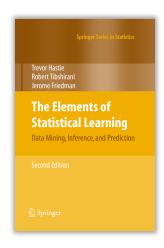
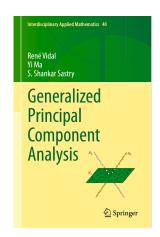


FIGURE 3.11. Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the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







Regression shrinkage and selection via the lasso

R Tibshirani

Journal of the Royal Statistical Society. Series B (Methodological), 267-288

Application to LASSO

• LASSO: ℓ_1 -regularized least squares

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

commonly encountered in *signal/image processing*.

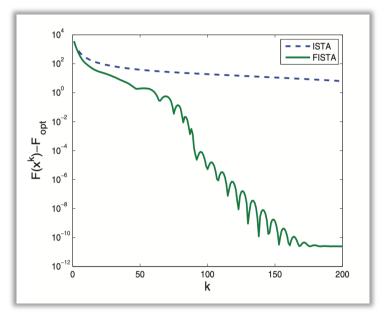
- composite optimization: first part is *smooth*, the other one is *non-smooth*
- ISTA (Iterative Shrinkage-Thresholding Algorithm): PG for LASSO
- FISTA (Fast ISTA): APG for LASSO

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

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

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 manitude.

Key words. iterative shrinkage-thresholding algorithm, deconvolution, linear inverse problem, least squares and l₁ 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

$$Ax = b + w$$
,

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.

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.).

103

A fast iterative shrinkage-thresholding algorithm for linear inverse problems

A Beck, M Teboulle

SIAM journal on imaging sciences 2 (1), 183-202

11526

2009

^{*}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.

Summary

