最適化問題に対する一次法とその反復計算量の理論

日本大学　伊藤勝

第 14 回 三部会連携「応用数理セミナー」
2022/12/23

Outline

最適化問題

最適化問題

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \mathbb{R}^n$$

- $f : \mathbb{R}^n \to \mathbb{R}$ の関数値 $f(x)$ を最小にする $x \in \mathbb{R}^n$ を (近似的に) 計算したい
- 最適値 を $f^* = \min_x f(x)$ とおく
- 最適解の集合 $X^* = \{x \mid f(x) = f^*\}$ は非空とする

- 効率的なアルゴリズムを作りたい:
  最適解に収束する $\{x_k\}$ を効率よく計算したい
- 実際には停留点 (勾配の零点) の計算を目指す:
  $\nabla f(x_k) \to \mathbf{0}$ なる $\{x_k\}$ を効率よく計算したい

一次法の位置付け

- 0 次法: $f(x)$ だけが使える
  $\rightarrow$ Nelder-Mead 法, 勾配推定法 など
- 1 次法: $f(x)$, $\nabla f(x)$ だけが使える $(f \in C^1)$
  $\rightarrow$ 最急降下法, 共役勾配法, 準 Newton 法 など
- 2 次法: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$ だけが使える $(f \in C^2)$
  $\rightarrow$ Newton 法 など
- アルゴリズムの性能 $=$ 反復回数 $\times$ 一反復のコスト

一次法の特徴

● 一反復のコストが抑えられるかも
▲ 収束が速くない (多くの反復回数を要する)

応用例: 高次元の最適化で高精度を要求しない場合

- 機械学習, 画像・信号処理, 圧縮センシング

仮定: f の平滑性

仮定: $f : \mathbb{R}^n \to \mathbb{R}$ は L-平滑, すなわち, C^1 級で ∇f は L-リプシッツ連続

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y$$

同値な性質

- $|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2, \quad \forall x, y$
- $|\nabla^2 f(x)$ の固有値 $| \leq L, \quad \forall x$
- $f(x) + \frac{L}{2}\|x\|^2$ は凸関数

基本的な一次法：最急降下法

最急降下法:

$$x_0 \in \mathbb{R}^n, \quad x_{k+1} = x_k - \lambda_k \nabla f(x_k), \quad k = 1, 2, \ldots,$$

ステップ幅 $\lambda_k > 0$ をうまく選んで反復を繰り返す. たとえば

- 定数ステップ幅: $\lambda_k = 1/L$
- Armijo 直線探索: 以下を満たす十分小さい $\lambda_k > 0$ を探す.

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1/\lambda_k}{2} \|x_{k+1} - x_k\|^2$$

  リプシッツ定数 L がわからないときに有効.

最急降下法の収束率

最急降下法の収束率 [Levitin & Polyak 1966], [Nesterov 2004]

(1) f が L-平滑のとき

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\| \leq O\left(\sqrt{\frac{L(f(x_0) - f^*)}{k}}\right), \quad \forall k = 0, 1, 2, \ldots$$

最急降下法の収束率

最急降下法の収束率 [Levitin & Polyak 1966], [Nesterov 2004]

(1) f が L-平滑のとき

$$\min_{0\le i\le k} \|\nabla f(x_i)\| \le O\left(\sqrt{\frac{L(f(x_0)-f^*)}{k}}\right), \quad \forall k = 0,1,2,\dots$$

(2) f がさらに凸関数のとき

$$f(x_k) - f^* \le O\left(\frac{L\,\mathrm{dist}(x_0, X^*)^2}{k}\right), \quad \min_{0\le i\le k} \|\nabla f(x_i)\| \le O\left(\frac{L\,\mathrm{dist}(x_0, X^*)}{k}\right)$$

- (1) は L-平滑関数に対する “一次法” の中で最適 (定数倍の違いで) [Carmon et al. 2020].
- (2) を改善する一次法が存在: 加速勾配法 [Nesterov 1983]

Nesterov の一次法

Nesterov の一次法 [Nesterov 1983, 2004; Beck & Teboulle 2009]:

$x_0 = y_0 \in \mathbb{R}^n$

$y_k^+ = y_k - \lambda_k \nabla f(y_k)$　(y_k からの最急降下ステップ)

$x_{k+1} = x_k$ と y_k^+ のうち目的関数値の小さい方.

$$y_{k+1} = x_{k+1} + \underbrace{\frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})}_{\text{慣性項}}$$

ステップ幅 $\lambda_k > 0$ は最急降下法と同じ.　$t_1 = 1$, $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$

Nesterov の一次法の収束率

Nesterov の一次法の収束率

f が L-平滑な凸関数のとき,

$$(*) \quad f(x_k) - f^* \leq O\left(\frac{L\,\mathrm{dist}(x_0, X^*)^2}{k^2}\right), \quad \forall k = 0, 1, 2, \ldots$$

$$(**) \quad \min_{0 \leq i \leq k} \|\nabla f(x_i)\| \leq O\left(\frac{L\,\mathrm{dist}(x_0, X^*)}{k^{1.5}}\right), \quad \forall k = 0, 1, 2, \ldots$$

- $(*)$ は最適 [Nemirovsky & Yudin 1979]
- $(**)$ は $k^{1.5}$ から k^2 (最適) に改良する一次法が存在 [Kim & Fessler 2020]

反復回数の最適な上界の比較

- f: L-平滑のとき, 最急降下法は

$$\|\nabla f(x_k)\| \leq \varepsilon \text{ の達成に } O\left(\frac{L(f(x_0)-f^*)}{\varepsilon^2}\right) \text{ 反復で十分}$$

- f: L-平滑な凸関数のとき,

  Nesterov の一次法は

$$f(x_k)-f^* \leq \varepsilon \text{ の達成に } O\left(\sqrt{\frac{L\,\mathrm{dist}(x_0,X^*)^2}{\varepsilon}}\right) \text{ 反復で十分}$$

  [Kim-Fessler 2020] の一次法は

$$\|\nabla f(x_k)\| \leq \varepsilon \text{ の達成に } O\left(\sqrt{\frac{L\,\mathrm{dist}(x_0,X^*)}{\varepsilon}}\right) \text{ 反復で十分}$$

f に強凸関数という仮定を追加すると, これは更に改善される.

強凸関数の最小化

f が μ-強凸関数であるとは,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y$$

同値な条件

- $\nabla^2 f(x)$ の固有値 $\geq \mu, \quad \forall x$
- $f(x) - \frac{\mu}{2}\|x\|^2$ が凸関数
- 凸共役 $f^*(x) = \sup_y \{\langle x, y \rangle - f(y)\}$ が $\frac{1}{\mu}$-平滑 (Baillon-Haddad theorem)

強凸関数の最小点は一意的: $X^* = \{x^*\}$

強凸関数に対する一次法

強凸関数に対する一次法の収束率

f は L-平滑かつ μ-強凸とする (言い換え: $\nabla^2 f(x)$ の固有値 $\in [\mu, L]$)

(1) 最急降下法は, 一次収束する:

$$f(x_k) - f^* \leq O\left(L\|x_0 - x^*\|^2 \exp\left(-\frac{\mu}{L} \cdot k\right)\right)$$

利点: ステップ幅 $\lambda_k = 1/L$ や直線探索は, 強凸定数 μ を必要としない.

強凸関数に対する一次法

強凸関数に対する一次法の収束率

f は L-平滑かつ μ-強凸とする (言い換え: $\nabla^2 f(x)$ の固有値 $\in [\mu, L]$)

(1) 最急降下法は, 一次収束する:

$$f(x_k) - f^* \leq O\left(L\|x_0 - x^*\|^2 \exp\left(-\frac{\mu}{L} \cdot k\right)\right)$$

利点: ステップ幅 $\lambda_k = 1/L$ や直線探索は, 強凸定数 μ を必要としない.

(2) μ を知っていれば, [Nesterov 2018] の一次法は, 一次収束する:

$$f(x_k) - f^* \leq O\left(\mu\|x_0 - x^*\|^2 \exp\left(-\sqrt{\frac{\mu}{L}} \cdot k\right)\right) \quad \text{最適}$$

- μ に依存せず, 最良の収束率を実現する一次法は知られていない.
- $\frac{\mu}{2}\|x_k - x^*\|^2 \leq f(x_k) - f^*$ を使うと $\|x_k - x_0\|$ の収束率も示せる (最適)
- $\frac{1}{2L}\|\nabla f(x_k)\|^2 \leq f(x_k) - f^*$ を使うと $\|\nabla f(x_k)\|$ の収束率も示せる (ほぼ最適)

課題

- f: L-平滑な凸関数のとき, [Kim-Fessler 2020] の一次法は

  $\|\nabla f(x_k)\| \leq \varepsilon$ の達成に $O\left(\sqrt{\dfrac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\right)$ 反復で十分 (最適)

  → ただし, 反復回数が固定

- f: L-平滑かつ μ-強凸のとき, Nesterov の一次法は

  $\|\nabla f(x_k)\| \leq \varepsilon$ の達成に $O\left(\sqrt{\dfrac{L}{\mu}}\log\dfrac{L\|x_0 - x^*\|}{\varepsilon}\right)$ 反復で十分 (ほぼ最適)

  → ただし, μ を知っている必要がある

アプローチ：適応的正則化 [I. & Fukuda 2021]

$f(x) + \dfrac{\sigma}{2}\|x - x_0\|^2$ の最小化を解く　&　σ を適応的に決める

正則化

正則化した目的関数の最小化

$$\text{minimize} \quad f_{\sigma,x_0}(x) := f(x) + \frac{\sigma}{2}\|x - x_0\|^2$$

- 最適値 $\min_x f_{\sigma,x_0}(x) = e_\sigma f(x_0)$: Moreau envelope
- 最適解 $x^*_{\sigma,x_0} = \text{prox}_{f/\sigma}(x_0)$: 近接写像

$$\|\nabla f(x^*_{\sigma,x_0})\| \leq \sigma \,\text{dist}(x_0, X^*)$$

正則化 [Nesterov 2012]

正則化した目的関数の最小化

$$\text{minimize} \quad f_{\sigma,x_0}(x) := f(x) + \frac{\sigma}{2}\|x - x_0\|^2$$

- f が L-平滑な凸関数 $\Longrightarrow$ f_{σ,x_0} は $(L+\sigma)$-平滑かつ σ-強凸
- そこで, f_{σ,x_0} に Nesterov の一次法を使う $\quad\rightarrow\quad$ 一次収束

$$k = O\left(\sqrt{\frac{L}{\sigma}}\log\frac{L+\sigma}{\sigma}\right) \text{ 反復すると } \|\nabla f(x_k)\| \le 2\sigma\mathrm{dist}(x_0, X^*).$$

[Nesterov 2012] $\quad \sigma = \dfrac{\varepsilon}{2\mathrm{dist}(x_0, X^*)}$ とするとき, $\|\nabla f(x_k)\| \le \varepsilon$ を得るには

$$O\left(\sqrt{\frac{L\mathrm{dist}(x_0, X^*)}{\varepsilon}}\log\frac{L\mathrm{dist}(x_0, X^*)}{\varepsilon}\right) \quad \text{反復で十分 (準最適)}$$

適応的 正則化 [I. & Fukuda 2021]

$$\text{minimize}_x \quad f_{\sigma,x_0}(x) := f(x) + \frac{\sigma}{2}\|x - x_0\|^2, \qquad \sigma \leftarrow L \text{ と初期化}$$

Algorithm I: Adaptive regularization scheme

(a) $\bar{x} \leftarrow$ Nesterov 一次法で $\min_x f_{\sigma,x_0}(x)$ を解く:

初期点 x_0 として $O(\sqrt{L/\sigma}\log(L+\sigma)/\sigma)$ 反復する.

$(\Longrightarrow \qquad \|\nabla f(\bar{x})\| \le 2\sigma\,\mathrm{dist}(x_0, X^*)$ が成り立っている)

(b) IF $\|\nabla f(\bar{x})\| > \varepsilon \Rightarrow \sigma > \dfrac{\varepsilon}{2\,\mathrm{dist}(x_0, X^*)}$ so restart (a) letting $\sigma \leftarrow \sigma/2$.

ELSE: $\|\nabla f(\bar{x})\| \le \varepsilon$ なので終了

Main result I: $\|\nabla f(\bar{x})\| \le \varepsilon$ となるには

$$\text{合計で} \quad O\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\log\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}\right) \quad \text{反復で十分} \quad (\text{準最適})$$

課題 (再掲)

- f: L-平滑な凸関数のとき, [Kim-Fessler 2020] の一次法は

  $\|\nabla f(x_k)\| \leq \varepsilon$ の達成に $O\left(\sqrt{\dfrac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\right)$ 反復で十分 (最適)

  → ただし, 反復回数が固定

- f: L-平滑かつ μ-強凸のとき, Nesterov の一次法は

  $\|\nabla f(x_k)\| \leq \varepsilon$ の達成に $O\left(\sqrt{\dfrac{L}{\mu}}\log\dfrac{L\|x_0 - x^*\|}{\varepsilon}\right)$ 反復で十分 (ほぼ最適)

  → ただし, μ を知っている必要がある

アプローチ：適応的正則化 [I. & Fukuda 2021]

$f(x) + \dfrac{\sigma}{2}\|x - x_0\|^2$ の最小化を解く　&　σ を適応的に決める

Hölderian error bound condition

仮定: Hölderian Error Bound (HEB)

初期点 $x_0 \in \mathbb{R}^n$ に対して, $\exists \kappa > 0$, $\exists \rho \geq 1$ such that

$$f(x) - f^* \geq \kappa \, \mathrm{dist}(x, X^*)^\rho, \quad \forall x \text{ with } f(x) \leq f(x_0).$$

ただし X^* は最適解集合

Hölderian error bound condition

仮定: Hölderian Error Bound (HEB)

初期点 $x_0 \in \mathbb{R}^n$ に対して, $\exists \kappa > 0$, $\exists \rho \geq 1$ such that

$$f(x) - f^* \geq \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \text{ with } f(x) \leq f(x_0).$$

ただし X^* は最適解集合

- 強凸性の一般化になっている:

  f: μ-強凸 $\iff$ $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \dfrac{\mu}{2}\|x - y\|^2$,

  $\implies f(x) - f^* \geq \dfrac{\mu}{2} \operatorname{dist}(x, X^*)^2, \ \forall x$

  $\implies$ HEB with $\kappa = \dfrac{\mu}{2}$, $\rho = 2$

Hölderian error bound condition

仮定: Hölderian Error Bound (HEB)

初期点 $x_0 \in \mathbb{R}^n$ に対して, $\exists \kappa > 0$, $\exists \rho \geq 1$ such that

$$f(x) - f^* \geq \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \text{ with } f(x) \leq f(x_0).$$

ただし X^* は最適解集合

- 強凸性の一般化になっている:

  f: μ-強凸 $\iff$ $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,

  $\implies f(x) - f^* \geq \frac{\mu}{2} \operatorname{dist}(x, X^*)^2, \ \forall x$

  $\implies$ HEB with $\kappa = \frac{\mu}{2}$, $\rho = 2$

f が連続, 凸, 強圧的, 半代数的 $\Rightarrow \forall x_0 \in \mathbb{R}^n$, $\exists \kappa, \rho$ such that HEB holds.

f : 半代数的 $\iff$ $\operatorname{graph}(f)$: 半代数的 $\iff$
$\operatorname{graph}(f) = \bigcup_i^{\text{finite}} \bigcap_j^{\text{finite}} \{x : \mathrm{p}_{ij}(x) \leq 0\}$, p_{ij}: 多項式

Hölderian error bound condition

仮定: Hölderian Error Bound (HEB)

初期点 $x_0 \in \mathbb{R}^n$ に対して, $\exists \kappa > 0$, $\exists \rho \geq 1$ such that

$$f(x) - f^* \geq \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \text{ with } f(x) \leq f(x_0).$$

ただし X^* は最適解集合

Łojasiewicz 不等式との関係 (Bolte et al. 2017)

C^1 凸関数 $f : \mathbb{R}^n \to \mathbb{R}$, $x_0 \in \mathbb{R}^n$, $\rho \geq 1$ に対して,

HEB が成り立つ $\kappa > 0$ が存在 $\iff$ $\exists c > 0$ such that

$$\|\nabla f(x)\| \geq c(f(x) - f^*)^\alpha, \quad \forall x \text{ with } f(x) \leq f(x_0), \quad \alpha = 1 - \frac{1}{\rho} \in [0, 1)$$

これらを組み合わせると,

$$\operatorname{dist}(x, X^*) \leq \kappa^{-\frac{1}{\rho-1}} \|\nabla f(x)\|^{\frac{1}{\rho-1}}$$

適応的な一次法

- HEB は多くの応用で成り立つ
- HEB の定数 κ と ρ を予め知ることは一般に難しい.

	仮定	適応する定数	近似尺度
Nesterov '07 Lin & Xiao '15	μ-強凸	μ	$\|\nabla f(x)\|$
Fercoq & Qu '17	HEB with $\rho = 2$	係数 κ	$\|\nabla f(x)\|$
Liu & Yang '17	HEB (ρ が既知)	係数 κ	$\|\nabla f(x)\|$
This work	HEB	係数 κ と指数 ρ	$\|\nabla f(x)\|$
Roulet & d'Aspremont '17 Renegar & Grimmer '18	HEB	係数 κ と指数 ρ	$f(x) - f^*$

アルゴリズムの導出

HEB は以下を満たす:

$$\mathrm{dist}(x, X^*) \leq \kappa^{-\frac{1}{\rho-1}} \|\nabla f(x)\|^{\frac{1}{\rho-1}}$$

ここで Algorithm I で $\varepsilon = \|\nabla f(x_0)\|/2$ としたときの反復回数の上界は

$$O\left(\sqrt{\frac{L\mathrm{dist}(x_0, X^*)}{\varepsilon}} \log \frac{L\mathrm{dist}(x_0, X^*)}{\varepsilon}\right) \leq O\left(\sqrt{\frac{L\|\nabla f(x_0)\|^{\frac{2-\rho}{\rho-1}}}{\kappa^{\frac{1}{\rho-1}}}} \log \frac{L\|\nabla f(x_0)\|^{\frac{2-\rho}{\rho-1}}}{\kappa^{\frac{1}{\rho-1}}}\right)$$

HEB に対する適応的な一次法

仮定: f: L-平滑な凸関数で HEB を満たす κ, ρ が存在 (L は既知, κ, ρ は未知)

Algorithm II

$x_0 \in \mathbb{R}^n$, $\sigma := L$. Set $x_0^+ := x_0 - \nabla f(x_0)/L$

外部反復 $t = 0, 1, 2, \ldots$: x_t から x_{t+1} への更新

(a) $x_t^{(0)}, x_t^{(1)}, \ldots \leftarrow$ Nesterov 一次法で f_{σ, x_t^+} を最小化:

初期点 x_t^+, 反復回数 $K_t := O\left(\sqrt{L/\sigma} \log \dfrac{L+\sigma}{\sigma}\right)$,

($*$) IF $\|\nabla f(x_t^{(k)})\| \le \|\nabla f(x_t)\|/2$ がある反復 k で成り立てば,

$x_{t+1} := x_t^{(k)}$, $x_{t+1}^+ := x_{t+1} - \nabla f(x_{t+1})/L$ として $(t+1)$-外部反復へ

(b) IF ($*$) does not hold until K_t iteration, set $\sigma \leftarrow \sigma/2$ and retry t-th stage.

反復回数の上界

簡略版:

x_{t+1}, $\sigma \leftarrow$ Algorithm 1: 目的関数 f_{σ, x_t^+}, 初期点 x_t^+, 誤差 $\|\nabla f(x_t)\|/2$

$x_{t+1}^+ := x_{t+1} - \nabla f(x_{t+1})/L$

反復回数の上界

簡略版:

x_{t+1}, $\sigma \leftarrow$ Algorithm 1: 目的関数 f_{σ,x_t^+}, 初期点 x_t^+, 誤差 $\|\nabla f(x_t)\|/2$

$x_{t+1}^+ := x_{t+1} - \nabla f(x_{t+1})/L$

Main result II

$\|\nabla f(x)\| \leq \varepsilon$ を得るまでの, パフォーマンス:

	$\rho = 1$	$1 < \rho < 2$	$\rho = 2$	$\rho > 2$
$\|\nabla f(x_t)\|$ の収束	finite	superlinear	linear	sublinear
反復回数 (w.r.t. ε)	const	$O(\log\log\frac{1}{\varepsilon})$	$O(\log\frac{1}{\varepsilon})$*1	$O(\varepsilon^{-\frac{\rho-2}{2(\rho-1)}}\log\frac{1}{\varepsilon})$*2

$$(*1) = O\left(\sqrt{\frac{L}{\kappa}}\log\frac{L}{\kappa}\log\frac{1}{\varepsilon}\right), \quad (*2) = O\left(\sqrt{\frac{L}{\kappa^{\frac{1}{\rho-1}}\varepsilon^{\frac{\rho-2}{\rho-1}}}}\log\frac{1}{\varepsilon}\right) : \quad \text{準最適}$$

最急降下法との比較

最急降下法: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k),\ k = 0, 1, 2, \ldots.$

$\|\nabla f(x_k)\| \leq \varepsilon$ を得るまでのパフォーマンス [Liu-Yang 2017]:

	$\rho = 1$	$1 < \rho < 2$	$\rho = 2$	$\rho > 2$
$\|\nabla f(x_k)\|$ の収束	finite	superlinear	linear	sublinear
反復回数 (w.r.t. ε)	const	$O(\log\log\frac{1}{\varepsilon})$	$O(\frac{L}{\kappa}\log\frac{1}{\varepsilon})$	$O(L\kappa^{-\frac{1}{\rho-1}}\varepsilon^{-\frac{\rho-2}{\rho-1}})$

- 最急降下法は適応的
- 提案手法の反復回数 $\approx \sqrt{\text{最急降下法の反復回数}}$

まとめと課題

- 課題: L-平滑な凸関数 f に対して $O\left(\sqrt{\frac{L\|x_0 - x^*\|}{\varepsilon}}\right)$ 反復で
  $\|\nabla f(x)\| \leq \varepsilon$ を達成できる, 反復回数を固定しない一次法はあるか?
  → 提案手法 $O\left(\sqrt{\frac{L\|x_0 - x^*\|}{\varepsilon}} \log \frac{L\|x_0 - x^*\|}{\varepsilon}\right)$
- 課題: L-平滑な μ-強凸関数 f に対して $O\left(\sqrt{\frac{L}{\mu}} \log \frac{\mu\|x_0 - x^*\|}{\varepsilon}\right)$ 反復で
  $\|\nabla f(x)\| \leq \varepsilon$ を達成できる一次法はあるか (L, μ を知らなくても)
  → 提案手法 $O\left(\sqrt{\frac{L}{\mu}} \log \frac{L}{\mu} \log \frac{L\|x_0 - x^*\|}{\varepsilon}\right)$

他の興味

- 非平滑の場合の解析
- 他のエラーバウンド
- 他のアルゴリズムへの応用: Frank-Wolfe 法 [Carderera et al. 2021]

References I

- Amir Beck, Mark Teboulle, A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *SIAM J. Imaging Sciences*, **2**(1):183–202, 2009.
- J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, From error bounds to the complexity of first-order descent methods for convex functions, *Math. Program.*, **165**, pp. 471–507, 2017.
- Y. Carmon, J. C. Duchi, O. Hinder, A. Sidford, Lower bounds for finding stationary points I, Mathematical Programming, **184**:71–120, 2020.
- A. Carderera, J. Diakonikolas, C. Y. Lin, S. Pokutta, Parameter-free Locally Accelerated Conditional Gradients, arXiv:2102.06806, 2021.
- O. Fercoq and Z. Qu, Adaptive restart of accelerated gradient methods under local quadratic growth condition, arXiv:1709.02300, 2017.
- M. Ito and M. Fukuda, Nearly Optimal First-Order Methods for Convex Optimization under Gradient Norm Measure: an Adaptive Regularization Approach, *Journal of Optimization Theory and Applications* **188**:770–804 (2021).

References II

- D. Kim and J. A. Fessler, Optimizing the Efficiency of First-Order Methods for Decreasing the Gradient of Smooth Convex Functions, Journal of Optimization Theory and Applications **188**:192–219 (2021).
- E.S. Levitin, B.T. Polyak, Constrained minimization methods, *Zh. vychisl. Mat. mat. Fiz.*, **6**(5):787–823, 1966.
- Q. Lin and L. Xiao, An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization, *Comput. Optim. Appl.*, **60**, pp. 633–674, 2015.
- M. Liu and T. Yang, Adaptive accelerated gradient converging methods under Hölderian error bound condition, arXiv:1611.07609, 2017.
- A. S. Nemirovsky, On optimality of Krylov's information when solving linear operator equations, *Journal of Complexity*, **7**, pp. 121–130, 1991.
- A. Nemirovski and Y. Nesterov, Optimal methods of smooth convex optimization, *U.S.S.R. Comput. Maths. Math. Phys.*, **25**(2), pp. 21–30, 1985.

References III

- Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, *Soviet Math. Dokl.*, **27**(2):372–376, 1983.
- Y. Nesterov, *Introductry Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publisher, 2004.
- Y. Nesterov, How to make the gradients small *Optima* **88**, 2012
- Y. Nesterov, Gradient methods for minimizing composite functions, *Mathematical Programming*, **140**, pp. 125–161, 2013.
- Y. Nesterov, *Lectures on Convex Optimization*, Springer, 2018.
- J. Renegar and B. Grimmer, A Simple Nearly-Optimal Restart Scheme For Speeding-Up First Order Methods, arXiv:1803.00151, 2018.
- V. Roulet and A. d'Aspremont, Sharpness, Restart and acceleration, in *Advances in Neural Information Processing Systems*, pp. 1119–1129, 2017.