Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods

Feng-Yi Liao

Joint work with Lijun Ding, Yang Zheng

ECE department, UC San Diego

ISMP 2024, Montreal, Canada
July 23, 2024

Outline

1. Motivation
2. Equivalent regularity conditions for weakly convex functions
3. Proximal point method: linear convergence
4. Conclusion

Motivation

- Machine learning has shown impressive performance

- (Sub)gradient-based methods and their variants are the workhorse algorithms.
  - Gradient descent (GD), stochastic GD, coordinate descent, etc.
- Well-known: For a **L-smooth** and **strongly convex** function $f : \mathbb{R}^n \to \mathbb{R}$, the basic GD

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), k = 1, 2, \ldots$$

  enjoys linear convergence, i.e.,

$$\begin{aligned} f(x_{k+1}) - f^\star &\leq \omega_1 (f(x_k) - f^\star), \quad 0 < \omega_1 < 1, \\ \|x_{k+1} - x^\star\| &\leq \omega_2 \|x_{k+1} - x^\star\|, \quad 0 < \omega_2 < 1. \end{aligned}$$

- However, **smoothness** and **strong convexity** are often not satisfied in practice.

Motivation

Alternative regularity conditions (weaker than strong convexity)

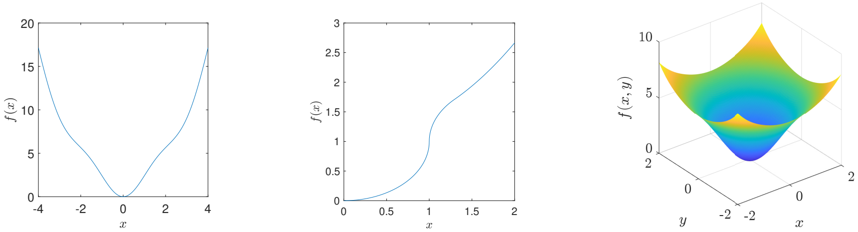
- Polyak-Łojasiewicz (PL) inequality (Polyak 1963)

$$2\beta(f(x) - f^\star) \leq \|\nabla f(x)\|^2, \ \forall x \in \mathbb{R}^n.$$

- Restricted secant inequality (RSI) (H. Zhang and Yin 2013)

$$\langle \nabla f(x), x - \hat{x} \rangle \geq \mu \cdot \mathrm{dist}^2(x, S), \ \forall x \in \mathbb{R}^n, \ \forall \hat{x} \in \Pi_S(x),$$

  where $S = \mathrm{argmin}_x \, f(x)$ and $\Pi_S(x) = \mathrm{argmin}_{y \in S} \|x - y\|$.
- PL & RSI are sufficient to ensure GD converges linearly (H. Zhang 2020).
- Functions can be nonconvex



Motivation

- Both PL and RSI ensure linear convergence. A natural question:

What is the relationship between them?

- With other regularity conditions, the relationship in the class of **smooth** functions is known

(Karimi, Nutini, and Schmidt 2016) Let f be a L-**smooth** function. Then

$$\text{(SC)} \to \text{(RSI)} \to \text{(EB)} \equiv \text{(PL)} \to \text{(QG)}.$$

Furthermore, if f is convex, then

$$\text{(RSI)} \equiv \text{(EB)} \equiv \text{(PL)} \equiv \text{(QG)}.$$

- However, it only works for L-**smooth** functions!
- Many practical functions are nonsmooth, e.g.,

$$f(\cdot) = |\cdot| \quad \text{or} \quad f(\cdot) = \langle C, \cdot \rangle + \rho \max\{0, \lambda_{\max}(-\cdot)\}.$$

This talk

- **Message 1:**

Let f be a proper closed ρ-**weakly** convex function. Then

$$\text{(SC)} \to \text{(RSI)} \to \text{(EB)} \equiv \text{(PL)} \to \text{(QG)}.$$

Furthermore, if the (QG) coefficient satisfies $\mu_{\mathrm{q}} > \rho$ (including the case where the function f is convex), then the following equivalence holds

$$\text{(RSI)} \equiv \text{(EB)} \equiv \text{(PL)} \equiv \text{(QG)}.$$

- **Message 2:**

The proximal point method (PPM) enjoys linear convergence under RSI/EB/PL/QG for (weakly) convex optimization!

Outline

Equivalent regularity conditions

- Go beyond **convexity** and **smoothness**.
- Consider the class of **weakly** convex (possibly nondifferentiable) functions.
- A function $f : \mathbb{R}^n \to \mathbb{R}$ is called ρ-weakly convex if the function

$$f + \frac{\rho}{2}\|\cdot\|^2 \quad \text{is convex.}$$

- The class of weakly convex functions is wide
  - Convex functions ($\rho = 0$), e.g., $|x|$
  - L-smooth functions, e.g., $-x^2 + \sin^2(x)$
  - Certain compositions of convex functions with smooth functions
- Fréchet subdifferential of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\hat{\partial} f(x) = \left\{ s \in \mathbb{R}^n \mid \liminf_{y \to x} \frac{f(y) - f(x) - \langle s, y - x \rangle}{\|y - x\|} \geq 0 \right\}.$$

Nonsmooth regularity conditions

- Let $S = \operatorname{argmin}_{x\in\mathbb{R}^n} f(x)$.

1. **Strong Convexity (SC)**: there exists a positive constant $\mu_{\rm s} > 0$ such that
$$f(x) + \langle g, y - x\rangle + \frac{\mu_{\rm s}}{2}\cdot\|y - x\|^2 \le f(y), \quad \forall x, y \in \mathbb{R}^n, g \in \hat{\partial} f(x). \tag{SC}$$
2. **Restricted Secant Inequality (RSI)**: there exists a positive constant $\mu_{\rm r} > 0$ s.t.
$$\mu_{\rm r}\cdot \mathrm{dist}^2(x, S) \le \langle g, x - h\rangle\,, \quad \forall x \in \mathbb{R}^n, g \in \hat{\partial} f(x), h \in \Pi_S(x). \tag{RSI}$$
3. **Error bound (EB)**: there exists a constant $\mu_{\rm e} > 0$ such that
$$\mathrm{dist}(x, S) \le \mu_{\rm e}\cdot \mathrm{dist}(0, \hat{\partial} f(x)), \quad \forall x \in \mathbb{R}^n. \tag{EB}$$
4. **Polyak-Łojasiewicz (PL) inequality**: there exists a constant $\mu_{\rm p} > 0$ such that
$$2\mu_{\rm p}\cdot (f(x) - f^\star) \le \mathrm{dist}^2(0, \hat{\partial} f(x)), \quad \forall x \in \mathbb{R}^n. \tag{PL}$$
5. **Quadratic Growth (QG)**: there exists a constant $\mu_{\rm q} > 0$ such that
$$\frac{\mu_{\rm q}}{2}\cdot \mathrm{dist}^2(x, S) \le f(x) - f^\star, \quad \forall x \in \mathbb{R}^n. \tag{QG}$$

Equivalent regularity conditions

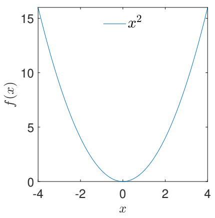
Theorem 1 (Liao, Ding, and Zheng 2023)

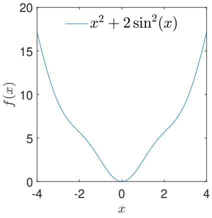
*Let f be a proper closed ρ-**weakly** convex function. Then*

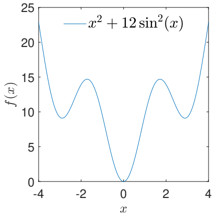
$$(\mathrm{SC}) \to (\mathrm{RSI}) \to (\mathrm{EB}) \equiv (\mathrm{PL}) \to (\mathrm{QG}).$$

Furthermore, if the (QG) *coefficient satisfies $\mu_{\mathrm{q}} > \rho$ (including the case where the function f is convex), then* $(\mathrm{RSI}) \equiv (\mathrm{EB}) \equiv (\mathrm{PL}) \equiv (\mathrm{QG})$.

- **Example 1:** $f(x) = x^2$. f is convex and all properties hold!
- **Example 2:** $f(x) = x^2 + 2\sin^2(x)$. All properties hold but f is not convex!
- **Example 3:** $f(x) = x^2 + 12\sin^2(x)$. All properties fail except (QG)!







Equivalent regularity conditions

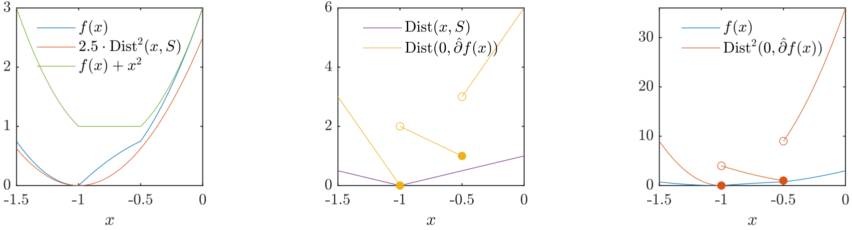
Theorem 1 (Liao, Ding, and Zheng 2023)

*Let f be a proper closed ρ-**weakly** convex function. Then*

$$(\text{SC}) \to (\text{RSI}) \to (\text{EB}) \equiv (\text{PL}) \to (\text{QG}).$$

Furthermore, if the (QG) *coefficient satisfies $\mu_{\mathrm{q}} > \rho$ (including the case where the function f is convex), then* (RSI) $\equiv$ (EB) $\equiv$ (PL) $\equiv$ (QG).

- If f is **convex** and satisfies (QG), then all the equivalence hold, as $\mu_{\mathrm{q}} > \rho = 0$.
- Example of a **nonsmooth** and **nonconvex** function satisfying $\mu_{\mathrm{q}} = 5 > \rho = 2$.



Literature and comparison

A huge body of literature (an incomplete list of references)

- **Smooth case**: A nice summary in (Karimi, Nutini, and Schmidt 2016, Theorem 2), which is a special case of our result on ρ-weakly convex functions.
- **Nonsmooth but convex case**:
  - (EB) $\equiv$ (QG) : (Drusvyatskiy and Lewis 2018; Artacho and Geoffroy 2008)
  - (PL) $\equiv$ (QG) : (Bolte et al. 2017, Theorem 5)
  - Thus, (EB) $\equiv$ (PL) $\equiv$ (QG) : (Ye et al. 2021; Zhu, Zhao, and S. Zhang 2023)
- **Nonsmooth and nonconvex**: The most closely related work is (Drusvyatskiy, Ioffe, and Lewis 2021) on nonsmooth optimization using taylor-like models.
- Our proof utilizes the notion of *slope* in (Drusvyatskiy, Ioffe, and Lewis 2021).
- **Concurrent work:** (Jin 2023) uses the proximal point method to analyze all the equivalencies.

Proof sketches

- Most of the directions are not difficult to prove.
- (RSI) $\Rightarrow$ (EB): Let $x \in \mathbb{R}^n$, g be the minimal norm element in $\hat{\partial} f(x)$, and $\hat{x} \in \Pi_S(x)$. By the definition of (RSI), we have
$$\langle g, x - \hat{x} \rangle \geq \mu_{\mathrm{r}} \cdot \mathrm{dist}^2(x, S)$$
Applying Cauchy-Schwarz on the left side yields the desired EB inequality
$$\mathrm{dist}(0, \hat{\partial} f(x)) \geq \mu_{\mathrm{r}} \cdot \mathrm{dist}(x, S).$$
- (QG) with $\mu_{\mathrm{q}} > \rho \Rightarrow$ (RSI): Let $x \in \mathbb{R}^n$, $\hat{x} \in \Pi_S(x)$, and $g \in \hat{\partial} f(x)$. From the assumption of (QG) and that f is ρ-weakly convex, we have
$$\frac{\mu_{\mathrm{q}}}{2} \cdot \mathrm{dist}^2(x, S) \leq f(x) - f^\star \leq \langle g, x - \hat{x} \rangle + \frac{\rho}{2} \mathrm{dist}^2(x, S).$$
Rearranging terms yields the desired RSI ineuality
$$\left(\frac{\mu_{\mathrm{q}} - \rho}{2} \right) \cdot \mathrm{dist}^2(x, S) \leq \langle g, x - \hat{x} \rangle .$$
- Only the direction (PL) $\Rightarrow$ (EB) requires some sophisticated arguments (slope).

Outline

Proximal point method (PPM)

Consider the optimization problem

$$f^\star = \min_{x \in \mathbb{R}^n} \; f(x),$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper closed convex function. Let $S = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$.

- PPM generates a sequence of points following

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \; f(x) + \frac{1}{2c_k} \left\| x - x_k \right\|^2,$$

  where $\{c_k\}_{k \geq 0}$ is a sequence of positive numbers.

- PPM is a conceptually simple algorithm, guiding other algorithm design
    - Proximal bundle method (Lemarechal, Strodiot, and Bihain 1981)
    - Augmented Lagrangian method (Rockafellar 1976a)
    - Proximally guided stochastic subgradient method (Davis and Grimmer 2019)

Proximal point method (PPM)

- The convergence of PPM can be traced back to (Rockafellar 1976b)
- Sublinear convergence rate (Güler 1991)

Theorem 2 (Sublinear convergence $\mathcal{O}(1/k)$ (Güler 1991, Theorem 2.1))

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function, and $S \neq \varnothing$. Then, the iterates of PPM with a positive sequence $\{c_k\}_{k \geq 0}$ satisfy

$$f(x_k) - f^\star \leq \frac{\operatorname{dist}^2(x_0, S)}{2 \sum_{t=0}^{k-1} c_t}.$$

If we further have $\lim_{k \to \infty} \sum_{t=0}^{k-1} c_t = \infty$, the iterates converge to an optimal solution $\bar{x}$, i.e., $\lim_{k \to \infty} x_k = \bar{x}$, where $\bar{x} \in S$.

- Choosing a constant step size $c_k = c > 0$ recovers the rate $\mathcal{O}(1/k)$
- Linear convergence can be shown if f satisfies some regularity conditions.

Proximal point method (PPM)

Different assumptions exist for linear convergence:

- $(\partial f)^{-1}$ is *Lipstchitz continuous*, which requires a unique solution (Rockafellar 1976b).
- $(\partial f)^{-1}$ is *upper Lipstchitz continuous*, allowing multiple solutions (Luque 1984).
- f satisfies *error bound* condition (Leventhal 2009).
- f satisfies *proximal error bound* condition (Drusvyatskiy and Lewis 2018).
- We use (PL) and (QG) to show the linear convergence

Theorem 3 (Linear convergence of PPM (Liao, Ding, and Zheng 2023))

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function, and $S \neq \varnothing$. Suppose f satisfies (PL) *(or* (EB), (RSI), (QG)*). Then, the iterates of PPM with a positive sequence $\{c_k\}_{k \geq 0}$ satisfies*

$$f(x_{k+1}) - f^\star \leq \omega_k \cdot (f(x_k) - f^\star), \quad \omega_k < 1,$$
$$\mathrm{dist}(x_{k+1}, S) \leq \theta_k \cdot \mathrm{dist}(x_k, S), \quad \theta_k < 1.$$

Proof sketches

- Optimality condition of the subproblem:

$$-(x_{k+1} - x_k)/c_k \in \partial f(x_{k+1}).$$

- Simple proof under (PL): From by definition of the subproblem, we have

$$f(x_k) - f^\star \geq f(x_{k+1}) - f^\star + \frac{1}{2c_k}\|x_{k+1} - x_k\|^2$$
$$\overset{\text{(O.C.)}}{\geq} f(x_{k+1}) - f^\star + \frac{c_k}{2}\mathrm{dist}^2(0, \partial f(x_{k+1})) \overset{\text{(PL)}}{\geq} (1 + c_k\mu_{\mathrm{p}})(f(x_{k+1}) - f^\star).$$

- Simple proof under (QG): $f + \frac{1}{2c_k}\|\cdot - x_k\|^2$ is $1/c_k$ strongly convex, so

$$f^\star + \frac{1}{2c_k}\mathrm{dist}^2(x_k, S)$$
$$\geq f(x_{k+1}) + \frac{1}{2c_k}\|x_{k+1} - x_k\|^2 + \langle 0, \Pi_S(x_k) - x_{k+1}\rangle + \frac{1}{2c_k}\|\Pi_S(x_k) - x_{k+1}\|^2$$
$$\geq f(x_{k+1}) + \frac{1}{2c_k}\mathrm{dist}^2(x_{k+1}, S) \overset{\text{(QG)}}{\Longrightarrow} \frac{1}{2c_k}\mathrm{dist}^2(x_k, S) \geq (\mu_{\mathrm{q}} + \frac{1}{2c_k})\mathrm{dist}^2(x_{k+1}, S).$$

Proximal point method (PPM)

Theorem 4 (PPM for weakly convex functions (Liao, Ding, and Zheng 2023))
Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper ρ-weakly convex, and $S \neq \varnothing$. Suppose f satisfies (QG) *with $\mu_\mathrm{q} > \rho$. The iterates of PPM with $\{c_k\}_{k\geq 0}$ with $\frac{1}{c_k} > \rho, \forall k \geq 0$, satisfy*

$$f(x_{k+1}) - f^\star \leq \omega_k \cdot (f(x_k) - f^\star), \quad \omega_k < 1,$$
$$\mathrm{dist}(x_{k+1}, S) \leq \theta_k \cdot \mathrm{dist}(x_k, S), \quad \theta_k < 1.$$

- f satisfies (QG) with $\mu_\mathrm{q} > \rho$ implies (RSI), (EB), (PL), and (QG) hold!
- The choice $\frac{1}{c_k} > \rho$ ensures the existence of the optimality condition

$$-(x_{k+1} - x_k)/c_k \in \hat{\partial} f(x_{k+1}).$$

- The proof then follows exactly as the convex case.

Numerical examples

Three machine learning instances

- Linear support vector machine (SVM) (Y. Zhang and Lin 2015)
- Lasso (ℓ_1-regularization) (Tibshirani 1996)
- Elastic-Net ($\ell_1 - \ell_2^2$-regularization) (Zou and Hastie 2005)

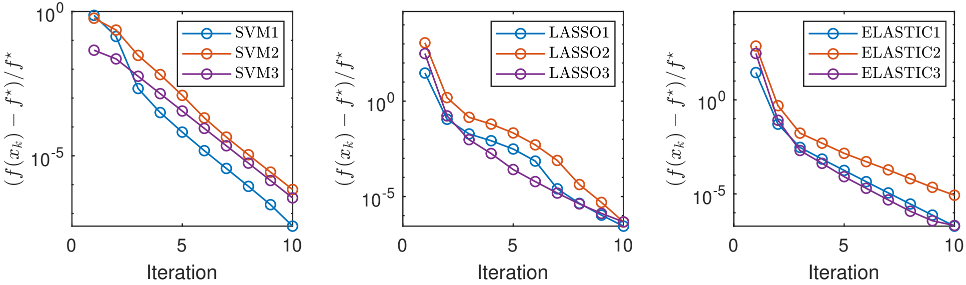


Figure: Linear convergences of cost value gaps for linear SVM (left), lasso (middle), and elastic-net (right).

Numerical examples

- Consider a 2-weakly convex function satisfying (QG) with $\mu_{\mathrm{q}} = 5 > \rho = 2$

$$f(x) = \begin{cases} -x^2 + 1 & \text{if } -1 < x < -0.5, \\ 3(x+1)^2 & \text{otherwise.} \end{cases}$$

- We run PPM with $\frac{1}{c_k} > \rho, \forall k \geq 0$.

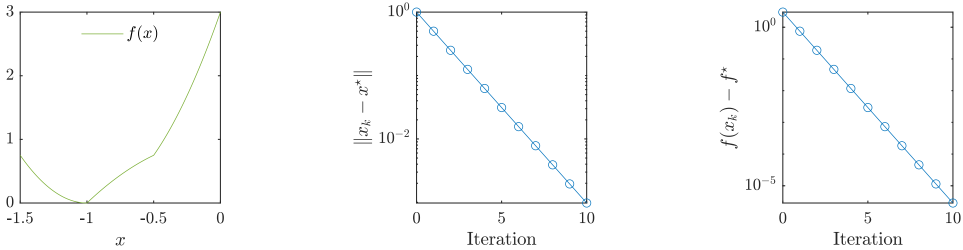


Figure: A 2-weakly convex function (left). Linear convergence of the distance to the solution set (middle). Linear convergence of the cost value gap (right).

Outline

Conclusion

- Equivalent regularity conditions in the class of weakly convex functions

> Let f be a proper closed ρ-**weakly** convex function. Then
>
> $$(\mathrm{SC}) \to (\mathrm{RSI}) \to (\mathrm{EB}) \equiv (\mathrm{PL}) \to (\mathrm{QG}).$$
>
> Furthermore, if the (QG) coefficient satisfies $\mu_{\mathrm{q}} > \rho$ (including the case where the function f is convex), then
>
> $$(\mathrm{RSI}) \equiv (\mathrm{EB}) \equiv (\mathrm{PL}) \equiv (\mathrm{QG}).$$

- PPM enjoys linear convergence under RSI/EB/PL/QG for convex optimization!
- Linear convergence extends to weakly convex functions!

Thank you for your attention!

Q & A

- Feng-Yi Liao, Lijun Ding, and Yang Zheng (2023). "Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods". In: *arXiv preprint arXiv:2312.16775*

Supported by NSF ECCS-2154650; NSF CMMI-2320697

References I

- Artacho, FJ Aragón and Michel H Geoffroy (2008). "Characterization of metric regularity of subdifferentials". In: *Journal of Convex Analysis* 15.2, p. 365.
- Bolte, Jérôme et al. (2017). "From error bounds to the complexity of first-order descent methods for convex functions". In: *Mathematical Programming* 165, pp. 471–507.
- Davis, Damek and Benjamin Grimmer (2019). "Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems". In: *SIAM Journal on Optimization* 29.3, pp. 1908–1930.
- Drusvyatskiy, Dmitriy, Alexander D Ioffe, and Adrian S Lewis (2021). "Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria". In: *Mathematical Programming* 185, pp. 357–383.
- Drusvyatskiy, Dmitriy and Adrian S Lewis (2018). "Error bounds, quadratic growth, and linear convergence of proximal methods". In: *Mathematics of Operations Research* 43.3, pp. 919–948.
- Güler, Osman (1991). "On the convergence of the proximal point algorithm for convex minimization". In: *SIAM journal on control and optimization* 29.2, pp. 403–419.
- Jin, Qinian (2023). "On growth error bound and Kurdyka-{\L} ojasiewicz condition". In: *arXiv preprint arXiv:2310.03947*.

References II

Karimi, Hamed, Julie Nutini, and Mark Schmidt (2016). "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, pp. 795–811.

Lemarechal, Claude, Jean-Jacques Strodiot, and André Bihain (1981). "On a bundle algorithm for nonsmooth optimization". In: *Nonlinear programming 4*. Elsevier, pp. 245–282.

Leventhal, D (2009). "Metric subregularity and the proximal point method". In: *Journal of Mathematical Analysis and Applications* 360.2, pp. 681–688.

Liao, Feng-Yi, Lijun Ding, and Yang Zheng (2023). "Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods". In: *arXiv preprint arXiv:2312.16775*.

Luque, Fernando Javier (1984). "Asymptotic convergence analysis of the proximal point algorithm". In: *SIAM Journal on Control and Optimization* 22.2, pp. 277–293.

Polyak, Boris T (1963). "Gradient methods for the minimisation of functionals". In: *USSR Computational Mathematics and Mathematical Physics* 3.4, pp. 864–878.

Rockafellar, R Tyrrell (1976a). "Augmented Lagrangians and applications of the proximal point algorithm in convex programming". In: *Mathematics of operations research* 1.2, pp. 97–116.

References III

Rockafellar, R Tyrrell (1976b). "Monotone operators and the proximal point algorithm". In: *SIAM journal on control and optimization* 14.5, pp. 877–898.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.

Ye, Jane J et al. (2021). "Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems". In: *Set-Valued and Variational Analysis*, pp. 1–35.

Zhang, Hui (2020). "New analysis of linear convergence of gradient-type methods via unifying error bound conditions". In: *Mathematical Programming* 180.1-2, pp. 371–416.

Zhang, Hui and Wotao Yin (2013). "Gradient methods for convex minimization: better rates under weaker conditions". In: *arXiv preprint arXiv:1303.4645*.

Zhang, Yuchen and Xiao Lin (2015). "Stochastic primal-dual coordinate method for regularized empirical risk minimization". In: *International Conference on Machine Learning*. PMLR, pp. 353–361.

Zhu, Daoli, Lei Zhao, and Shuzhong Zhang (2023). "A Unified Analysis for the Subgradient Methods Minimizing Composite Nonconvex, Nonsmooth and Non-Lipschitz Functions". In: *arXiv preprint arXiv:2308.16362*.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320.