Optimization Theory and Algorithm II **September 16, 2022**

Lecture 3

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Subgradient and Subdifferential

In the last subsection, we have shown that how to use gradient descent algorithms to solve smooth and convex objective function.

Q: How about non-smooth objective function?

Example 1 *Least Absolute Deviation Regression (LAD Regression), it is similar to the Least Squares problems with the optimization formulation as:*

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_1. \tag{1}$$

We need a way to measure stationarity in the non-smooth case. For convex functions, a natural notion is that of the subgradient/subdifferential.

Definition 1 *A subgradient of a convex possible non-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^n$ is a vector $\mathbf{g} \in \mathbb{R}^n$ if*

$$f(\mathbf{y}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + f(\mathbf{x})$$

for all $\mathbf{y}$.

Definition 2 *The subdifferential of f at $\mathbf{x}$ is the set of all subgradients, denoted $\partial f(\mathbf{x})$. Equivalently*

$$\partial f(\mathbf{x}) := \{\mathbf{g} \in \mathbb{R}^n : f(\mathbf{y}) \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + f(\mathbf{x}) \text{ for all } \mathbf{y}\}.$$

Theorem 1 *$\mathbf{x}^*$ is a global minimal point of the convex possible non-smooth function f if $0 \in \partial f(\mathbf{x}^*)$.*

1.1 Subgradient Descent

Subgradient descent algorithm should be

$$\mathbf{x}^{t+1} = \mathbf{x}^t - s_t \mathbf{g}^t \tag{2}$$

where $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$.

Compared with the standard gradient descent algorithm, we need to consider the following problems:

- How to select $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$?
- How to choice the step size s_t?
- How to stop the algorithm?

We will answer these questions for the specific non-smooth objective function which is a Lipschitz continuous function.

Definition 3 *Function $f:\mathbb{R}^n \to$ is Lipschitz function with respect to a constant $G>0$ if for any $\mathbf{x},\mathbf{y}\in(f)$*

$$|f(\mathbf{x})-f(\mathbf{y})|\le G\|\mathbf{x}-\mathbf{y}\|_2, \tag{3}$$

where G is referred as to Lipschitz constant of f.

Theorem 2 *Assume that f is a convex and G-Lip function, $\mathbf{x}^*=\arg\min f(\mathbf{x}), f^*=f(\mathbf{x}^*)>-\infty$, then $\{\mathbf{x}^t\}_{t=0}^{\infty}$ is generated form the subgradient descent algorithm, then for any $T>0$, it has*

$$f(\mathbf{x}^{t^*})-f^*\le\frac{\|\mathbf{x}^0-\mathbf{x}^*\|^2+G^2\sum_{t=0}^{T}s_t^2}{2\sum_{t=0}^{T}s_t}, \tag{4}$$

where $t^*=\arg\min\limits_{0\le t\le T} f(\mathbf{x}^t)$.

Remark 1 *Let us discuss the above theorem.*

- *See that $f(\mathbf{x}^t)-f(\mathbf{x}^*)$ may be not decreasing!*
- *Let $\|\mathbf{x}^0-\mathbf{x}^*\|^2=R^2, s_t=s$, then*

$$f(\mathbf{x}^{t^*})-f^*\le\frac{R^2}{2Ts}+\frac{sTG^2}{2}:=\Phi(s). \tag{5}$$

  Obvisouly, if $s=\frac{R}{G\sqrt{T}}$, then $\min\Phi(s)=\frac{GR}{\sqrt{T}}$. Thus,

$$f(\mathbf{x}^{t^*})-f^*\le\inf_s\Phi(s)=\frac{GR}{\sqrt{T}}.$$

  This indicates that the convergence speed is the same with the only β-smooth objective function.

- *To $f(\mathbf{x}^{t^*})-f^*\to 0$, it should be $\sum_{t=1}^{\infty}s_t=+\infty$ and $\sum_{t=1}^{\infty}s_t^2\le M$, where M is a constant.*
  Q: *Could you please give us an example of $\{s_t\}_{t=0}^{\infty}$.*

2 Proximal Gradient Descent for Nonsmooth and Convex Function

2.1 Motivation

Convergence speed is $O(\frac{1}{\sqrt{T}})$ of subgradient descent for convex, non-smooth, and Lip objective function. Comparing with the speed $O(\frac{1}{T})$ of GD for smooth and convex objective functions, it is relatively slow.

Q: Can we improve the convergence speed?

Let us consider a specific type of non-smooth optimization problems.

$$\min_{\mathbf{x}} h(\mathbf{x}):=f(\mathbf{x})+g(\mathbf{x}) \tag{6}$$

where $f(\mathbf{x})$ is *convex and β-smooth*, and g is *convex and possibly non-smooth.*

Next we will show some examples for demonstrating the importance of the optimization formulation (6).

Example 2 *(Ridge Regression) Let us consider the linear regression example again.*

- *Data:* $\{\mathbf{a}_i, b_i\}_{i=1}^m$, *where* $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.
- *Suppose that*
$$b_i = \mathbf{a}_i^\top \mathbf{x} + \epsilon_i,$$
where $\mathbf{x} = (x_1, \ldots, x_n)^\top$ *is denoted as regression coefficient.*
- *Matrix Form: denote that* $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$, $\mathbf{A} = (a_{ij}) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m)^\top \in \mathbb{R}^{m\times n}$, *and*
$$\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon},$$
where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^\top$.
- *Optimization Formulation:*
$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{7}$$
- *Solution:* $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{b}$. *However, if* $rank(A) < n$, *then it is not invertable. This is called col-linearity.*
- *Numerical Solution:*
$$\mathbf{x}^*(\lambda) = (A^\top A + \lambda I_n)^{-1}A^\top \mathbf{b}, \tag{8}$$
and let $\lambda \to 0$.
- *This is the solution of the optimization prolbem:*
$$\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \right\}. \tag{9}$$

Example 3 *(Statistical Perspective for Ridge Regression) From the statistical modeling framework: we suppose that:*

- *Data:* $\{\mathbf{a}_i, b_i\}_{i=1}^m$, *where* $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.
- *Suppose that*
$$b_i = \mathbf{a}_i^\top \mathbf{x} + \epsilon_i,$$
where $\mathbf{x} = (x_1, \ldots, x_n)^\top$ *is denoted as regression coefficient and* $\epsilon_i \sim \mathcal{N}(0, 1)$.
- *Prior distribution:* $\mathbf{x} \sim \mathcal{N}(0, \frac{1}{\lambda} I_n)$.
- *Posterior distribution:*
$$\mathbb{P}(\mathbf{x}|A, \mathbf{b}) = \frac{\mathbb{P}(A, \mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x})}{\mathbb{P}(A, \mathbf{b})},$$
where
$$\mathbb{P}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{\lambda\|\mathbf{x}\|^2}{2} \right\},$$
and
$$\mathbb{P}(A, \mathbf{b}|\mathbf{x}) = \prod_{i=1}^n \mathbb{P}(\mathbf{a}_i, \mathbf{b}_i|\mathbf{x}) = \prod_{i=1}^n \frac{1}{2\pi} \exp\left\{ -\frac{(b_i - \mathbf{a}_i^\top \mathbf{x})^2}{2} \right\}.$$
- *Maximal Posterior (MAP) Estimation:*
$$\max_{\mathbf{x}} \mathbb{P}(\mathbf{x}|A, \mathbf{b}) \propto \mathbb{P}(A, \mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x}).$$
So, it is equivalent to
$$\min_{\mathbf{x}} -\log \mathbb{P}(A, \mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x}) = \min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \right\}.$$

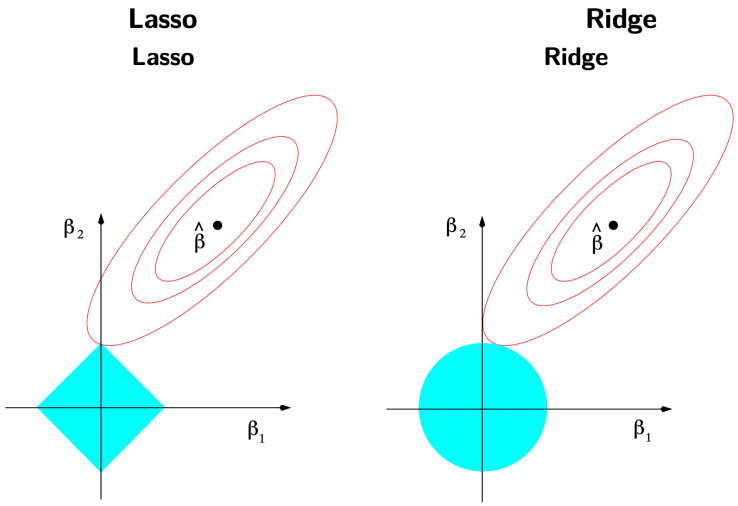

Figure 1: LASSO vs. Ridge

- *Numerical Solution:*

$$\mathbf{x}^*(\lambda) = (A^\top A + \lambda I_n)^{-1} A^\top \mathbf{b}. \tag{10}$$

Example 4 *(Least Absolute Shrinkage Selection Operator (LASSO) [1]) Let us consider a high-dimensional case study in a business setting. Assume that we have collected many customer's data for constructing the user portrait in a big company. This means that we will use* $\mathbf{x} \in \mathbb{R}^n$ *to represent one consumer and* n *is really big. Consider a common research question: which features (variables) will effect the consumer's purchase behavior for one product. How to do? If we use the linear regression model to handle the problem, it is called the variable selection problem for linear regression. Which is the best model? Actually, we have* $2^n - 1$ *candidate models that can be selected. How to handle such a huge problem? We suppose that:*

- *Data:* $\{\mathbf{a}_i, b_i\}_{i=1}^m$, *where* $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.
- *Suppose that*

$$b_i = \mathbf{a}_i^\top \mathbf{x} + \epsilon_i,$$

  where $\mathbf{x} = (x_1, \ldots, x_n)^\top$ *is denoted as regression coefficient and* $\epsilon_i \sim \mathcal{N}(0, 1)$.

- *From optimization perspective:*

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 \tag{11}$$

$$s.t.\ \|\mathbf{x}\|_1 \le t. \tag{12}$$

  See Figure 1 for the geometric interpretation.

- *Prior distribution:* $\mathbf{x} \sim \mathcal{L}(0, \frac{1}{\lambda} I_n)$, *where*

$$\mathbb{P}(\mathbf{x}) = \frac{1}{g(\lambda)} \exp\left\{-\frac{\lambda\|\mathbf{x}\|_1}{2}\right\}.$$

- *Posterior distribution:*

$$\mathbb{P}(\mathbf{x}|A, \mathbf{b}) = \frac{\mathbb{P}(A, \mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x})}{\mathbb{P}(A, \mathbf{b})}.$$

- *Maximal Posterior (MAP) Estimation:*

$$\max_{\mathbf{x}} \mathbb{P}(\mathbf{x}|A,\mathbf{b}) \propto \mathbb{P}(A,\mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x}).$$

So, it is equivalent to

$$\min_{\mathbf{x}} -\log \mathbb{P}(A,\mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x}) = \min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \right\}.$$

2.2 Proximal Gradient Algorithm

We consider the

$$\min_{\mathbf{x}} h(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \tag{13}$$

where $f(\mathbf{x})$ is *convex and β-smooth*, and g is *convex and possibly non-smooth.*

Let us go back to review the GD algorithm in advance. Because of the convexity of f, it has that

$$\begin{aligned} f(\mathbf{x}) \le m_t(\mathbf{x}) &= f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x}-\mathbf{x}^t\rangle + \frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^t\|^2 \\ &= f(\mathbf{x}^t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}^t)\|^2 + \frac{\beta}{2}\|\mathbf{x} - (\mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t))\|^2. \end{aligned}$$

So, $\mathbf{x}^* = \mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t)$ is the GD.

Let us go back to consider $h(\mathbf{x})$, it has

$$\begin{aligned} h(\mathbf{x}) \le m_t(\mathbf{x}) + g(\mathbf{x}) &= f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x}-\mathbf{x}^t\rangle + \frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^t\|^2 + g(\mathbf{x}) \\ &= f(\mathbf{x}^t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}^t)\|^2 + \frac{\beta}{2}\left\{ \|\mathbf{x} - (\mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t))\|^2 + g(\mathbf{x}) \right\}. \end{aligned}$$

If we set $\mathbf{z}^t = \mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t)$, then target optimization problem is:

$$\min_{\mathbf{x}} \frac{\beta}{2}\|\mathbf{x}-\mathbf{z}^t\|^2 + g(\mathbf{x}). \tag{14}$$

Definition 4 *Assume that g is convex, the proximal operator of g is*

$$prox_{\gamma g}(\mathbf{z}) = \arg\min_{\mathbf{x}\in(g)} \left\{ g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{z}\|^2 \right\}. \tag{15}$$

Based on the definition, actually

$$prox_{1/\beta g}(\mathbf{z}^t) = \arg\min_{\mathbf{x}\in(g)} \left\{ g(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x}-\mathbf{z}^t\|^2 \right\} = \arg\min\{m_t(\mathbf{x}) + g(\mathbf{x})\}. \tag{16}$$

Proximal Gradient Descent Algorithm:

$$\mathbf{z}^t = \mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t), \tag{17}$$

$$\mathbf{x}^{t+1} = prox_{1/\beta g}(\mathbf{z}^t). \tag{18}$$

Theorem 3 *Consider problem (6), if f is β-smooth and g is convex, the sequence generated by the proximal gradient descent algorithm satisfies,*

$$h(\mathbf{x}^T) - h^* \leq \frac{\beta}{2T}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

If further we assume f to be α-strongly convex, we have,

$$\|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \exp\left(-\frac{\alpha T}{\beta}\right)\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Where we h^* denote the optimal function value, and $\mathbf{x}^*$ optimal solution.

2.3 Accelerate Gradient Descent

What is the fastest convergence speed of an optimization algorithm? We should know the lower bound

$$O(T^s) \leq f(\mathbf{x}^T) - f^* \leq O(T^s).$$

Then the optimal convergence speed is $O(T^s)$.

Theorem 4 *[2] Let $T \leq \frac{n-1}{2}$, $\beta > 0$. Then there exists a β-smooth convex quadratic f such that any black-box method stasifies*

$$\min_{1\leq t\leq T} f(\mathbf{x}^t) - f^* \geq \frac{3\beta\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{32(1+T)^2}. \tag{19}$$

This means we have a chance to make an algorithm to achieve the convergence rate $O(T^{-2})$.

This is called the *accelerate (proximal) gradient descent* algorithm:

- Initial: $\mathbf{y}^1 = \mathbf{x}^0, a_1 = 1$ and $t = 1$.
- Step 1:
$$\mathbf{x}^t = \mathbf{y}^t - \frac{1}{\beta}\nabla f(\mathbf{y}^t) \text{ or } \mathbf{x}^t = prox_{g/\beta}(\mathbf{y}^t - \frac{1}{\beta}\nabla f(\mathbf{y}^t)). \tag{20}$$
- Step 2:
$$a_{t+1} = \frac{1+\sqrt{1+4a_t^2}}{2}. \tag{21}$$
- Step 3:
$$\mathbf{y}^{t+1} = \mathbf{x}^t + \frac{a_t - 1}{a_{t+1}}\underbrace{(\mathbf{x}^t - \mathbf{x}^{t-1})}_{momentum}. \tag{22}$$

Theorem 5 *[3] Let $\{\mathbf{x}^t, \mathbf{y}^t\}$ be generated by AGD or FISTA. Then for any $T \geq 1$,*

$$h(\mathbf{x}^T) - h^* \leq \frac{2\beta\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(1+T)^2}. \tag{23}$$

3 Optimization with Linear Equality Constrains

Let us consider a special case which is called "quadratic programing".

Example 5

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top P\mathbf{x} + q^\top \mathbf{x} + r, \tag{24}$$

$$s.t.\ A\mathbf{x} = \mathbf{b}, \tag{25}$$

where $P \succ 0$. If we disregard the quality constrain, the optimality condition of unconstrained optimization says: $\nabla f(\mathbf{x}^*) = P\mathbf{x}^* + q = 0$, that is $\mathbf{x}^* = -P^{-1}q$. Thus, a natural question should be asked that what optimality conditions of Eq.(24).

To this end, the optimality conditions of general convex optimization formulation are provided via the following theorem.

Theorem 6 *$\mathbf{x}^*$ is optimal of the convex optimization problem if and only if*

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0, \ \textit{for all } \mathbf{y} \in \mathcal{X}. \tag{26}$$

Proof 1 *(i) If $\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0$ for all $\mathbf{y} \in \mathcal{X}$, then we have $f(\mathbf{y}) \geq f(\mathbf{x}^*)$ due to the convexity of f, namely*

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle. \tag{27}$$

(ii) Suppose that $\mathbf{x}^*$ is optimal, but the condition (26) does not hold, i.e., there exists $\mathbf{y} \in \mathcal{X}$ such that

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle < 0.$$

Let $\mathbf{z} = \lambda\mathbf{y} + (1-\lambda)\mathbf{x}^*$, then

$$\begin{aligned} \frac{\partial f(\mathbf{z})}{\partial \lambda}|_{\lambda=0} &= \langle \nabla f(\lambda\mathbf{y} + (1-\lambda)\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle|_{\lambda=0} \\ &= \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle < 0. \end{aligned}$$

This implies that $f(\mathbf{z}) < f(\mathbf{x}^*)$. Contradiction!

Remark 2

- *Theorem 6 shows that $-\nabla f(\mathbf{x}^*)$ defines a supporting hyperplane to the feasible set at $\mathbf{x}^*$.*

Figure 2: Geometric Interpretation of Optimality Condition

- *If $\mathcal{X} = \mathbb{R}^n$, then the condition (26) reduces to the unconstrained optimality condition, $\nabla f(\mathbf{x}^*) = 0$.*

Example 6 *Let us consider the following general convex optimization with linear equality constrains.*

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{28}$$

$$s.t.\ A\mathbf{x} = \mathbf{b}. \tag{29}$$

We will write down the optimality condition of (28) according to Theorem 6.

First, Theorem 6 shows that

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0, A\mathbf{x}^* = \mathbf{b}, A\mathbf{y} = \mathbf{b}.$$

So, $A(\mathbf{x}^* - \mathbf{y}) = 0$ *and* $\mathbf{y} - \mathbf{x}^* \in \mathcal{N}(A)$*. Let* $\mathbf{v} = \mathbf{y} - \mathbf{x}^*$*, then* $\mathbf{v}^\top \nabla f(\mathbf{x}^*) \geq 0$*. However,* $\mathcal{N}(A)$ *is a linear space, we thus have* $\mathbf{y}'$ *such that* $\mathbf{y}' - \mathbf{x}^* = -\mathbf{v}$*, then* $\mathbf{v}^\top \nabla f(\mathbf{x}^*) \leq 0$*. Finally, we have* $\mathbf{v}^\top \nabla f(\mathbf{x}^*) = 0$ *and* $\nabla f(\mathbf{x}^*) \perp \mathcal{N}(A)$*. Thus,* $\nabla f(\mathbf{x}^*) \in \mathcal{C}(A^\top)$*, there exists* $\lambda \in \mathbb{R}^n$ *such that*

$$\nabla f(\mathbf{x}^*) + A^\top \lambda = 0 \text{ (}\textit{Optimality Condition}\text{)}.$$

To obtain the optimal point, we have to solve the following equations.

$$(*) = \begin{cases} Ax^* = b, \\ \nabla f(\mathbf{x}^*) + A^\top \lambda = 0. \end{cases}$$

For Example 5, it becomes a linear equation system:

$$\begin{cases} Ax^* = b, \\ P\mathbf{x}^* + q + A^\top \lambda = 0. \end{cases}$$

Actually, variable λ *is called* ***dual variable*** *which will be denoted in the next section.*

Q: How to solve the general equation system (*)?

References

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[2] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.