Optimization Theory and Algorithm **Lecture 11 - 06/01/2021**

Lecture 11

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Proximal Gradient Descent for Nonsmooth and Convex Function

1.1 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.

$$\min_{\mathbf{x}} h(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \tag{1}$$

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 (1).

Example 1 *(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 $\mathbf{x} = (x_1, \dots, x_n)^\top$ *is denoted as regression coefficient.*
- *Matrix Form: denote that* $\mathbf{b} = (b_1, \dots, b_m)^\top \in \mathbb{R}^m$, $\mathbf{A} = (a_{ij}) = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m)^\top \in \mathbb{R}^{m\times n}$, *and*
$$\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon},$$
where $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_m)^\top$.
- *Optimization Formulation:*
$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{2}$$
- *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:*
$$\mathbf{x}^*(\lambda) = (A^\top A + \lambda I_n)^{-1}A^\top \mathbf{b}, \tag{3}$$
and let $\lambda \to 0$.
- *This is the solution of the optimization prolbem:*
$$\min_{\mathbf{x}} \left\{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2\right\}. \tag{4}$$

Example 2 *(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
$$\min_{\mathbf{x}} -\log \mathbb{P}(A,\mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x}) = \min_{\mathbf{x}} \left\{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2\right\}.$$
- *Numerical Solution:*
$$\mathbf{x}^*(\lambda) = (A^\top A + \lambda I_n)^{-1} A^\top \mathbf{b}. \tag{5}$$

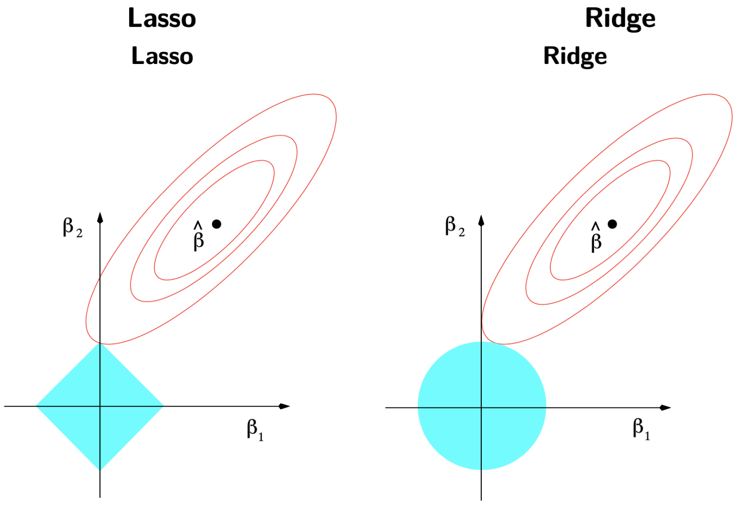

Figure 1: LASSO vs. Ridge

Example 3 *(Least Absolute Shrinkage Selection Operator (LASSO) [Tibshirani, 1996]) 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 $\mathbf{x} = (x_1, \dots, x_n)^\top$ *is denoted as regression coefficient and* $\epsilon_i \sim \mathcal{N}(0,1)$.
- *From optimization perspective:*
$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 \tag{6}$$
$$s.t.\ \|\mathbf{x}\|_1 \le t. \tag{7}$$
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
$$\min_{\mathbf{x}} -\log \mathbb{P}(A, \mathbf{b}|\mathbf{x})\mathbb{P}(\mathbf{x}) = \min_{\mathbf{x}} \left\{\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1\right\}.$$

1.2 Proximal Gradient Algorithm

We consider the

$$\min_{\mathbf{x}} h(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \tag{8}$$

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

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

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

$$\min_{\mathbf{x}} \frac{\beta}{2}\|\mathbf{x} - \mathbf{z}^t\|^2 + g(\mathbf{x}). \tag{9}$$

Definition 1 *Assume that g is convex, the proximal operator of g is*

$$prox_{\gamma g}(\mathbf{z}) = \arg\min_{\mathbf{x} \in dom(g)} \left\{ g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{z}\|^2 \right\}. \tag{10}$$

Based on the definition, actually

$$prox_{1/\beta g}(\mathbf{z}^t) = \arg\min_{\mathbf{x} \in dom(g)} \left\{ g(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{z}^t\|^2 \right\} = \arg\min\{m_t(\mathbf{x}) + g(\mathbf{x})\}. \tag{11}$$

Proximal Gradient Descent Algorithm:

$$\mathbf{z}^t = \mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t), \tag{12}$$

$$\mathbf{x}^{t+1} = prox_{1/\beta g}(\mathbf{z}^t). \tag{13}$$

Definition 2 *Suppose that $\Omega \subseteq \mathbb{R}^n$, the indicator function of Ω is*

$$\delta_\Omega(\mathbf{x}) = \begin{cases} +\infty, & \mathbf{x} \notin \Omega \\ 0, & \mathbf{x} \in \Omega. \end{cases} \tag{14}$$

Definition 3 *The projection of a point $\mathbf{z}$ onto a set Ω is defined as*

$$\pi_\Omega(\mathbf{z}) = \arg\min_{\mathbf{x} \in \Omega} \|\mathbf{x} - \mathbf{z}\|_2. \tag{15}$$

Example 4 *Projection examples:*

- $\Omega = \{\mathbf{x} | \mathbf{x} \geq 0\}$, *then* $\pi_\Omega(\mathbf{z}) = \max\{\mathbf{z}, 0\}$.
- $\Omega = \{\mathbf{x} | l \leq \mathbf{x} \leq u\}$, *then* $\pi_\Omega(\mathbf{z}) = \max(\min\{\mathbf{z}, u\}, l)$.
- $\Omega = B_2 = \{\mathbf{x} | \ \|\mathbf{x}\|_2 \leq 1\}$, *then*
$$\pi_\Omega(\mathbf{z}) = \begin{cases} \mathbf{z}, & \|\mathbf{z}\|_2 \leq 1, \\ \frac{\mathbf{z}}{\|\mathbf{z}\|_2} & \|\mathbf{z}\|_2 > 1. \end{cases}$$
- $\Omega = \{\mathbf{b} | \mathbf{b} = \sum_{i=1}^m x_i \mathbf{a}_i, \mathbf{a}_i \in \mathbb{R}^n, x_i \in \mathbb{R}\} = \{\mathbf{b} | \mathbf{b} = A\mathbf{x}\} = Col(A)$. **Q:** *What is the $\pi_\Omega(\mathbf{z})$??*

Example 5 *(Projected Gradient Descent)*

Let us consider a general optimization problem

$$\min_{x} \; f(\mathbf{x}),$$
$$s.t.\; \mathbf{x} \in \Omega.$$

This is equivalent to

$$\min_{\mathbf{x}} \{f(\mathbf{x}) + \delta_\Omega(\mathbf{x})\}. \tag{16}$$

Obviously, δ_Ω is convex and non-smooth. So, it has the form of Eq.(1). Let us compute the proximal operator of δ_Ω as follows.

$$prox_{1/\beta\delta_\Omega}(\mathbf{z}^t) = \arg\min_{\mathbf{x}\in dom(\delta_\Omega)} \left\{\delta_\Omega(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{z}^t\|^2\right\} = \arg\min_{\mathbf{x}\in\Omega} \|\mathbf{x} - \mathbf{z}^t\|^2 := \pi_\Omega(\mathbf{z}^t). \tag{17}$$

Obviously, $\pi_\Omega(\mathbf{z}^t)$ is the projection of $\mathbf{z}_t$ onto Ω.

- $\Omega = \{\mathbf{x}|\mathbf{x} \geq 0\}$*, then* $\mathbf{x}^{t+1} = prox_{1/\beta\delta_\Omega}(\mathbf{z}^t) = \pi_\Omega(\mathbf{z}^t) = \max\{\mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t), 0\}$.
- $\Omega = \{\mathbf{x}|l \leq \mathbf{x} \leq u\}$*, then* $\mathbf{x}^{t+1} = prox_{1/\beta\delta_\Omega}(\mathbf{z}^t) = \pi_\Omega(\mathbf{z}^t) = \max(\min\{\mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t), u\}, l)$.
- *The same with B_2 or $Col(A)$.*

These algorithms are called projected gradient descent.

References

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