Introduction to Machine Learning

Spring 2021

University of Science and Technology of China

Lecturer: Jie Wang	Homework 4
Posted: Nov. 16, 2021	Due: Nov. 29, 2021

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Convex Functions

1. (Optional) For each of the following functions, determine whether it is convex.
   (a) $f(x) = x^2 \log x$ on $\mathbb{R}_{++}$, where $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$.
   (b) $f(x_1, x_2) = x_1 x_2$ on $\mathbb{R}^2$.
   (c) $f(x_1, x_2) = \frac{x_1}{x_2}$ on $\mathbb{R}^2_{++}$, where $\mathbb{R}^2_{++} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 > 0\}$.
   (d) $f(x_1, x_2) = \frac{x_1^2}{x_2}$ on $\mathbb{R} \times \mathbb{R}_{++}$.
   (e) $f(x_1, x_2) = x_1^{\alpha} x_2^{1-\alpha}$ on $\mathbb{R}^2_{++}$, where $0 \leq \alpha \leq 1$.

2. Please show that the following functions are convex.
   (a) $f(\mathbf{x}) = \log \sum_{i=1}^{n} e^{x_i}$ on $\mathbf{dom}\ f = \mathbb{R}^n$, where x_i denotes the i^{th} component of $\mathbf{x}$.
   (b) $f(\mathbf{x}) = \sum_{i=1}^{k} x_{[i]}$ on $\mathbf{dom}\ f = \mathbb{R}^n$, where $1 \leq k \leq n$ and $x_{[i]}$ denotes the i^{th} largest component of $\mathbf{x}$.
   (c) The extended-value extension of the indicator function of a convex set $C \subseteq \mathbb{R}^n$, i.e.,

$$\tilde{I}_c(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in C, \\ \infty, & \mathbf{x} \notin C. \end{cases}$$

   (d) The negative entropy, i.e.,

$$f(\mathbf{p}) = \sum_{i=1}^{n} p_i \log p_i$$

   on $\mathbf{dom}\ f = \{\mathbf{p} \in \mathbb{R}^n : 0 < p_i \leq 1, \sum_{i=1}^{n} p_i = 1\}$, where p_i denotes the i^{th} component of $\mathbf{p}$.
   (e) The spectral norm, i.e.,

$$f(\mathbf{X}) = \|\mathbf{X}\|_2 = \sigma_{\max}(\mathbf{X})$$

   on $\mathbf{dom}\ f = \mathbb{R}^{m \times n}$, where $\sigma_{\max}$ denotes the largest singular value of $\mathbf{X}$.
   (f) $f(\mathbf{X}) = \mathrm{tr}(\mathbf{X}^{-1})$ on $\mathbf{dom}\ f = \mathbb{S}^n_{++}$, where $\mathbb{S}^n_{++}$ is the space of all $n \times n$ real positive definite matrices.

3. Please show that a continuously differentiable function f is strongly convex with parameter $\mu > 0$ if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

4. Suppose that f is twice continuously differentiable and strongly convex with parameter $\mu > 0$. Please show that $\mu \leq \lambda_{\min}(\nabla^2 f(\mathbf{x}))$ for any $\mathbf{x} \in \mathbb{R}^n$, where $\lambda_{\min}(\nabla^2 f(\mathbf{x}))$ is the smallest eigenvalue of $\nabla^2 f(\mathbf{x})$.

5. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and the gradient of f is Lipschitz continuous, i.e.,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

where $L > 0$ is the Lipschitz constant. Please show that $\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$, where $\lambda_{\max}(\nabla^2 f(\mathbf{x}))$ is the largest eigenvalue of $\nabla^2 f(\mathbf{x})$.

6. Consider the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and convex, and **dom** f is closed.

(a) Please show that the α-sublevel set of f, i.e., $C_\alpha = \{\mathbf{x} \in \textbf{dom}\ f : f(\mathbf{x}) \leq \alpha\}$ is closed.

(b) Please give an example to show that Problem (1) may be unsolvable even if f is strictly convex.

(c) Suppose that f can attain its minimum. Please show that the optimal set $\mathcal{C} = \{\mathbf{y} : f(\mathbf{y}) = \min_{\mathbf{x}} f(\mathbf{x})\}$ is closed and convex. Does this property still hold if **dom** f is not closed?

(d) Suppose that f is strongly convex with parameter $\mu > 0$. Please show that Problem (1) admits a unique solution.

Solution: ■

Exercise 2: Operations that Preserve Convexity

1. (a) Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be a given convex function, $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Please show that

$$F(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b}), \quad \mathbf{x} \in \mathbb{R}^n.$$

is convex.

(b) Let $f_i : \mathbb{R}^n \to (-\infty, +\infty]\,, i = 1, \dots, m$, be given convex functions. Please show that

$$F(\mathbf{x}) = \sum_{i=1}^{m} w_i f_i(\mathbf{x})$$

is convex, where $w_i \geq 0$, $i = 1, \dots, m$.

(c) Let $f_i : \mathbb{R}^n \to (-\infty, +\infty]$ be given convex functions for $i \in I$, where I is an arbitrary index set. Please show that the supremum

$$F(\mathbf{x}) = \sup_{i\in I} f_i(\mathbf{x})$$

is convex.

2. (Optional) Let $\mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{x}_0 \in \mathbb{R}^n$. The restriction of $f : \mathbb{R}^n \to \mathbb{R}$ to the affine set $\{\mathbf{A}\mathbf{z} + \mathbf{x}_0 | \mathbf{z} \in \mathbb{R}^m\}$ is defined as the function $F : \mathbb{R}^m \to \mathbb{R}$ with

$$F(\mathbf{z}) = f(\mathbf{A}\mathbf{z} + \mathbf{x}_0)$$

on $\mathbf{dom}\ F = \{\mathbf{z} | \mathbf{A}\mathbf{z} + \mathbf{x}_0 \in \mathbf{dom}\ f\}$. Suppose f is twice differentiable with a convex domain.

(a) Show that F is convex if and only if for all $\mathbf{z} \in \mathbf{dom}\ F$, we have

$$\mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{z} + \mathbf{x}_0)\mathbf{A} \succeq 0.$$

(b) Suppose $\mathbf{B} \in \mathbb{R}^{p\times n}$ is a matrix whose nullspace is equal to the range of $\mathbf{A}$, i.e., $\mathbf{A}\mathbf{B} = \mathbf{0}$ and $\text{rank}(\mathbf{B}) = n - \text{rank}(\mathbf{A})$. Show that F is convex if for all $\mathbf{z} \in \mathbf{dom}\ F$, there exists a $\lambda \in \mathbb{R}$ such that

$$\nabla^2 f(\mathbf{A}\mathbf{z} + \mathbf{x}_0) + \lambda \mathbf{B}^\top \mathbf{B} \succeq 0.$$

(**Hint:** you can use the result as follows. If $\mathbf{C} \in \mathbb{S}^n$ and $\mathbf{D} \in \mathbb{R}^{p\times n}$, then $\mathbf{x}^\top \mathbf{C}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathcal{N}(\mathbf{D})$ if there exists a λ such that $\mathbf{C} + \lambda \mathbf{D}^\top \mathbf{D} \succeq 0$.)

3. (Optional)

(a) Consider the function $f(\mathbf{X}) = \lambda_{\max}(\mathbf{X})$, with $\mathbf{dom}\ f = \mathbb{S}^n$, where $\lambda_{\max}(\mathbf{X})$ is the largest eigenvalue of $\mathbf{X}$ and $\mathbb{S}^n$ is the set of $n \times n$ real symmetric matrices. Show that f is a convex function.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, with **dom** $f = \mathbb{R}^n$. Show that it can be represented as the pointwise supremum of a family of affine functions, i.e.,

$$f(\mathbf{x}) = \sup\{g(\mathbf{x}) : g \text{ is affine}, g(\mathbf{z}) \leq f(\mathbf{z}) \text{ for all } \mathbf{z} \in \mathbb{R}^n\}.$$

4. Suppose that the training set is $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i^{th} data instance and $y_i \in \mathbb{R}$ is the corresponding label. Recall that Lasso is the regression problem:

$$\min_{\mathbf{w}} \frac{1}{2n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1,$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ with its i^{th} row being $\mathbf{x}_i^\top$, $\mathbf{w} \in \mathbb{R}^d$, and $\lambda > 0$ is the regularization parameter. Show that the objective function in the above problem is convex.

Solution: ■

Exercise 3: Subdifferentials

1. Calculation of subdifferentials.

   (a) Let $H \subset \mathbb{R}^n$ be a hyperplane. The extended-value extension of its indicator function I_H is

   $$\tilde{I}_H(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in H, \\ \infty, & \mathbf{x} \notin H. \end{cases}$$

   Find $\partial \tilde{I}_H(\mathbf{x})$.

   (b) Let $f(\mathbf{x}) = \exp \|\mathbf{x}\|_1$, $\mathbf{x} \in \mathbb{R}^n$. Find $\partial f(\mathbf{x})$.

   (c) Let $f(x) = \max\{0, x\}$, $x \in \mathbb{R}$. Find $\partial f(x)$.

   (d) For $\mathbf{x} \in \mathbb{R}^n$, let $x_{[i]}$ be the i^{th} largest component of $\mathbf{x}$. Find the subdifferential of

   $$f(\mathbf{x}) = \sum_{i=1}^{k} x_{[i]}.$$

   (e) Let $f(\mathbf{x}) = \max\limits_{1 \le i \le n} x_i$, $\mathbf{x} \in \mathbb{R}^n$. Find $\partial f(\mathbf{x})$.

   (f) (Optional) Let

   $$f(\mathbf{x}) = \left(\sum_{i=1}^{k} x_i^2 \right)^{\frac{1}{2}} + \left(\sum_{i=k+1}^{n} x_i^2 \right)^{\frac{1}{2}}, \quad \mathbf{x} \in \mathbb{R}^n,$$

   where $1 \le k \le n-1$. Find $\partial f(\mathbf{x})$.

   (g) (Optional) Let $f(\mathbf{X}) = \|\mathbf{X}\|_*$ be the trace norm of $\mathbf{X} \in \mathbb{R}^{m \times n}$. Find $\partial f(\mathbf{X})$.

2. In Example 5 of Lecture 08, we got two forms of $\partial f(\mathbf{x})$ by two approaches. Please show that they are the same, i.e., the following two sets are the same for $\forall \mathbf{x} \in \mathbb{R}^n$.

   $$A_{\mathbf{x}} = \{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|_\infty, \|\mathbf{y}\|_1 \le 1\}$$
   $$B_{\mathbf{x}} = \begin{cases} \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\| \le 1\}, & \mathbf{x} = \mathbf{0}, \\ \{\mathbf{y} \in \mathbb{R}^n : \sum_{i=1}^n \varepsilon_i y_i = 1, \varepsilon_i y_i \ge 0, y_i = 0 \,\text{if}\, \varepsilon_i = 0\}, & \mathbf{x} \ne \mathbf{0}, \end{cases}$$

   where ε_i is defined as

   $$\varepsilon_i = \begin{cases} 1, & x_i = \|\mathbf{x}\|_\infty, \\ 0, & |x_i| < \|\mathbf{x}\|_\infty, \\ -1, & x_i = -\|\mathbf{x}\|_\infty. \end{cases}$$

Solution: ■

Exercise 4: Proximal Gradient

Consider the following convex optimization problem

$$\begin{aligned} &\min_{\mathbf{x}} F(\mathbf{x}) \\ &\text{s.t.} \mathbf{x} \in D \end{aligned} \tag{2}$$

where $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper convex function and $D \subseteq \mathbb{R}^n$ is a nonempty convex set with $D \subseteq \textbf{dom}\ F$. Suppose that the problem (2) is solvable, and **we do not require the differentiability of** F.

1. If $\mathbf{x} \in \textbf{int}\,(\textbf{dom}\ F) \cap D$ and there exists a $\mathbf{g} \in \partial F(\mathbf{x})$ such that

   $$\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \geq 0, \forall\, \mathbf{y} \in D,$$

   show that $\mathbf{x}$ is optimal.

2. (Optional) If $\mathbf{x} \in \textbf{int}\,(\textbf{dom}\ F)$ and $\mathbf{x}$ is optimal, show that $\mathbf{x} \in D$ and there exists a $\mathbf{g} \in \partial F(\mathbf{x})$ such that

   $$\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \geq 0, \forall\, \mathbf{y} \in D.$$

3. Please give an example to show that $\partial F(\mathbf{x})$ can be empty.

4. If $\mathbf{x}^*$ is an interior point of D, show that

   $$\mathbf{x}^* \in \underset{\mathbf{x} \in D}{\textbf{argmin}}\ F(\mathbf{x}) \Leftrightarrow 0 \in \partial F(\mathbf{x}^*).$$

   You can use the conclusion of Problems 1 and 2.

In many cases, the function F can be decomposed into $F = f + g$, where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a continuous convex function, and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and continuously differentiable function, whose gradient is Lipschitz continuous with the constant L. We can use ISTA, which has been introduced in Lecture 08, to find $\min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x})$.

5. For a **given** point $\mathbf{x}_c$, we consider the following quadratic approximation of F:

   $$Q(\mathbf{x}; \mathbf{x}_c) = f(\mathbf{x}_c) + \langle \nabla f(\mathbf{x}_c), \mathbf{x} - \mathbf{x}_c \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_c\|^2 + g(\mathbf{x}).$$

   Please show that it always admits a unique minimizer

   $$p(\mathbf{x}_c) = \underset{\mathbf{x} \in \mathbb{R}^n}{\textbf{argmin}}\, Q(\mathbf{x}; \mathbf{x}_c).$$

6. (Optional) We can think of the update step of ISTA, i.e., $\mathbf{x}^+ = p(\mathbf{x})$, as two steps:

   (a) Take a step in the opposite direction of the gradient of f at $\mathbf{x}$, i.e.,

   $$\mathbf{z} = \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x}).$$

(b) Project $\mathbf{z}$ on some set C, i.e.,

$$\mathbf{x}^+ = p(\mathbf{x}) = \Pi_C(\mathbf{x}).$$

Find the set C. Is it closed, open or neither? Is it convex or not?

7. Consider the Lasso problem

$$\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1.$$

Suppose that $\hat{\mathbf{w}}$ solves the problem. Write down the optimality condition at $\hat{\mathbf{w}}$.

8. If we use ISTA to solve the Lasso problem, show that

$$w_i^+ = \begin{cases} z_i + \frac{\lambda}{L}, & \text{if } z_i < -\frac{\lambda}{L}, \\ 0, & \text{if } |z_i| \leq \frac{\lambda}{L}, \\ z_i - \frac{\lambda}{L}, & \text{if } z_i > \frac{\lambda}{L}, \end{cases}$$

where $\mathbf{z} = \mathbf{w}_k - \frac{2}{Ln}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_k - \mathbf{y})$.

9. (Optional) Consider the convex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \tilde{I}_D(\mathbf{x}), \tag{3}$$

where $D \subseteq \mathbb{R}^n$ is a closed convex set and $\tilde{I}_D(\mathbf{x})$ is the extended-value extension of its indicator function $I_D(\mathbf{x})$.

(a) Write down the optimality condition and the proximal operator of Problem (3).

(b) Find the relationship between (3) and the constrained optimization problem

$$\min_{\mathbf{x}\in D} f(\mathbf{x}).$$

10. (Optional) Write down the proximal operator of the following convex optimization problems.

(a)

$$\min_{w\in\mathbb{R}^n} \frac{1}{n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2 I_{\mathbb{R}^n_+}(\mathbf{w}),$$

(b)

$$\min_{w\in\mathbb{R}^n} \frac{1}{n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\left[\left(\sum_{i=1}^{k} x_i^2\right)^{\frac{1}{2}} + \left(\sum_{i=k+1}^{n} x_i^2\right)^{\frac{1}{2}}\right].$$

Solution: ■

Exercise 5: Projected Gradient Descent (Optional)

Consider the following problem

$$\min_{x \in D} f(x), \tag{4}$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, continuously differentiable and strongly convex with convexity parameter $\mu > 0$. We assume that the gradient of f is Lipschitz with a constant $L > 0$.

A commonly used approach to solve the constrained optimization problem (4) is the so-called *projected gradient descent*, in which each iteration improves the current estimation $\mathbf{x}_k$ of the optimum by

$$\mathbf{x}_{k+1} = \Pi_D(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)),$$

where $\alpha > 0$ is the step size.

1. Show that
$$f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2, \ \forall \, \mathbf{x}, \mathbf{y} \in D.$$

2. Consider the problem (4) and the sequence generated by the *projected gradient descent* algorithm. Suppose that $\mathbf{x}^*$ is the solution to the problem (4).
   (a) Find the range of α such that the function values $f(\mathbf{x}_k)$ converge linearly to $f(\mathbf{x}^*)$.
   (b) When does the (projected) gradient descent always achieve the optimal solution in one iteration wherever the intial point $\mathbf{x}_0$ is?

Solution: ■

Exercise 6: [1] ISTA with Backtracking

Suppose that we would like to apply ISTA to solve the convex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}), \tag{5}$$

where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a continuous convex function, and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and continuously differentiable function, whose gradient is Lipschitz continuous with the constant L. We assume that Problem (5) is solvable, i.e., there exists $\mathbf{x}^*$ such that

$$F(\mathbf{x}^*) = F^* = \min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}).$$

In practice, however, a possible drawback of ISTA is that the Lipschitz constant L is not always known or computable. For instance, if $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, the Lipschitz constant for ∇f depends on $\lambda_{\max}(\mathbf{A}^\top\mathbf{A})$, which is not always easily computable for large-scale problems. To tackle this problem, we always equip ISTA with the backtracking stepsize rule as shown in Algorithm 1.

Note that in Algorithm 1, Q_L and p_L are defined as

$$\begin{aligned} &Q_L(\mathbf{x};\mathbf{x}_c) = f(\mathbf{x}_c) + \langle \nabla f(\mathbf{x}_c), \mathbf{x} - \mathbf{x}_c\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_c\|_2^2 + g(\mathbf{x}) \\ &p_L(\mathbf{x}_c) = \operatorname*{\mathbf{argmin}}_{\mathbf{x}\in\mathbb{R}^n} Q_L(\mathbf{x};\mathbf{x}_c). \end{aligned}$$

Algorithm 1 ISTA with Backtracking

```
Input: An initial point $\mathbf{x}_0$, an initial constant $L_0 > 0$, a threshold $\eta > 1$, and $k = 1$.
while the termination condition does not hold do
    Find the smallest non-negative integer $i_k$ such that with $\tilde{L} = \eta^{i_k} L_{k-1}$
        $F(p_{\tilde{L}}(\mathbf{x}_{k-1})) \le Q_{\tilde{L}}(p_{\tilde{L}}(\mathbf{x}_{k-1}); \mathbf{x}_{k-1}).$    (6)
    $L_k \leftarrow \eta^{i_k} L_{k-1}$, $\mathbf{x}_k \leftarrow p_{L_k}(\mathbf{x}_{k-1})$,
    $k \leftarrow k + 1$,
end while
```

1. Show that the sequence $\{F(\mathbf{x}_k)\}$ produced by Algorithm 1 is non-increasing.
2. Show that Inequality (6) is satisfied for any $\tilde{L} \ge L$, where L is the Lipschitz constant of ∇f, thus showing that for Algorithm 1 one has $L_k \le \eta L$ for every $k \ge 1$.
3. Let $\{\mathbf{x}_k\}$ be the sequence generated by Algorithm 1. Show that for any $k \ge 1$ we have

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \le \frac{\eta L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2k}, \ \forall \mathbf{x}^* \in \operatorname*{\mathbf{argmin}}_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}).$$

The above result means that the number of iterations of Algorithm 1 required to obtain an ε-optimal solution, i.e., an $\hat{\mathbf{x}}$ such that $F(\hat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \varepsilon$, is at most

$$\left\lceil \frac{\eta L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\varepsilon} \right\rceil .$$

Solution: ■

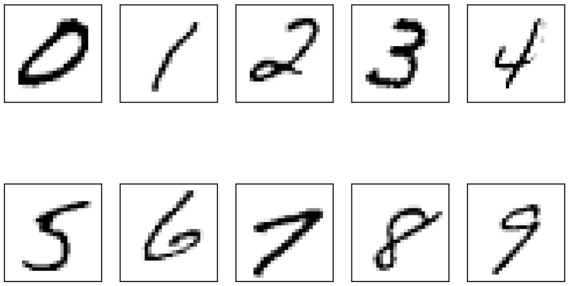
Exercise 7: Programming Exercise: Handwritten Digits Recognition

MNIST is a widely used data set—which consists of grey images of handwritten digits—in machine learning and pattern recognition. The training set and the testing set have 60000 and 10000 images, respectively. We show ten images in MNIST in Figure (a).

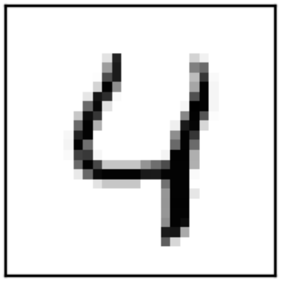
In this exercise, we would like to predict the label of a given handwritten image. Notice that, the labels of the images in the training set come for granted. For example, given an image from the training set with a handwritten digit “7” in it, we know the number (the label) in the image is 7. Our task is to predict the number in a handwritten image outside of the training set.

To do so, we first transform the 28 × 28 gray images in the training set into a set of 784-dimensional column vectors, and then we put them together to construct the input feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $n = 784$ and $d = 60000$. We also provide you labels of all images in the training set. Then, we randomly choose an image in the testing set and transform it into a column vector in the same way as the corresponding response vector $\mathbf{y} \in \mathbb{R}^n$. We show this image in Figure (b).

To predict the label of the response vector, we use Lasso to fit this data. Please follow the instructions step by step. You can use your favorite programming language.



(a) Ten images in the training set.



(b) The image we choose from the testing set.

1. We first normalize the data matrix $\mathbf{X}$ and the response vector $\mathbf{y}$, such that each entries are within $[0, 1]$. Specifically, let

$$\mathbf{Z} = \mathbf{X}/255, \mathbf{h} = \mathbf{y}/255.$$

   Then, Lasso takes the form of

$$\min_{\mathbf{u}} F(\mathbf{u}) = \frac{1}{n}\|\mathbf{h} - \mathbf{Z}\mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1, \tag{7}$$

   where $\mathbf{u} \in \mathbb{R}^d$. Let $f(\mathbf{u}) = \frac{1}{n}\| \ \mathbf{h} - \mathbf{Z}\mathbf{u}\|_2^2$ and $g(\mathbf{u}) = \lambda\|\mathbf{u}\|_1$.

2. Please find the Lipschitz constants of $\nabla f(\mathbf{u})$ and write down the corresponding quadratic approximation function of $F(\mathbf{u})$.

3. From 0.002 to 0.2, uniformly pick out 100 different λs and then implement the ISTA algorithm in **Exercise** 6 to solve Problem (7), with different λs. Terminate the

iteration after 2000 steps and denote the $\mathbf{u}_{2000}$ by $\mathbf{u}'$. Please plot scatter diagrams of $f(\mathbf{u}')$ versus λ, $g(\mathbf{u}')$ versus λ and $F(\mathbf{u}')$ versus λ, respectively. Besides, please plot a scatter diagram of the ration of nonzero elements in $\mathbf{u}'$ versus λ.

4. Please explain the practical meaning of the zero elements and nonzero elements in $\mathbf{u}'$ and give your prediction about the digit in $\mathbf{y}$ using our provided labels.

5. (Optional) There exists a threshold value, which satisfies that if λ exceeds this value, $\mathbf{u}' = \mathbf{0}$. Please find this threshold in theory.

Solution: ■

References

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