Introduction to Machine Learning

Fall 2022

University of Science and Technology of China

Lecturer: Jie Wang — Homework 5

Posted: Nov. 24, 2022 — Due: Dec. 8, 2022

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

Exercise 1: Proximal Operator

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we define its proximal operator at $\mathbf{x}$ by

$$\text{prox}_f(\mathbf{x}) = \underset{\mathbf{u}\in\text{dom}\, f}{\arg\min} \left\{ f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2 \right\}.$$

1. Recall the convex optimization problem in Lecture 08.

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

   Please rewrite $p(\mathbf{x}_c)$ using proximal operator.

2. The proximal operator has the following properties.

   (a) If f is proper and close (which means $\mathbf{epi}f$ is close), then for any $\mathbf{x} \in \mathbb{R}^n$, $\text{prox}_f(\mathbf{x})$ exists and is unique. You can use the properties we have proved in Homework 4 directly.

   (b) If f is proper and close, then $\mathbf{u} = \text{prox}_f(\mathbf{x})$ if and only if $\mathbf{x} - \mathbf{u} \in \partial f(\mathbf{u})$.

   (c) Please show that if $\mathbf{u} = \text{prox}_f(\mathbf{x}), \mathbf{v} = \text{prox}_f(\mathbf{y})$, then

$$\langle \mathbf{u}-\mathbf{v}, \mathbf{x}-\mathbf{y}\rangle \geq \|\mathbf{u}-\mathbf{v}\|_2^2,$$

   which means prox_f is firmly nonexpansive. Then show that this implies nonexpansive

$$\left\|\text{prox}_f(\mathbf{x}) - \text{prox}_f(\mathbf{y})\right\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2.$$

3. The proximal operator satisfies the following equations.

   (a) For $\lambda \neq 0$ and $a \in \mathbb{R}^n$, we let $h(\mathbf{x}) = f(\lambda\mathbf{x}+\mathbf{a})$, then $\text{prox}_h(\mathbf{x}) = \frac{1}{\lambda}\left(\text{prox}_{\lambda^2 f}(\lambda\mathbf{x}+\mathbf{a}) - \mathbf{a}\right)$.

   (b) For $\lambda > 0$, we let $h(\mathbf{x}) = \lambda f\left(\frac{\mathbf{x}}{\lambda}\right)$, then $\text{prox}_h(\mathbf{x}) = \lambda\,\text{prox}_{\lambda^{-1}f}\left(\frac{\mathbf{x}}{\lambda}\right)$.

   (c) For $\mathbf{a} \in \mathbb{R}^n$, we let $h(\mathbf{x}) = f(\mathbf{x}) + \mathbf{a}^\top\mathbf{x}$, then $\text{prox}_h(\mathbf{x}) = \text{prox}_f(\mathbf{x}-\mathbf{a})$.

4. Please find the proximal operator of the following functions.

   (a) $f(\mathbf{x}) = 0$.

   (b) $f(\mathbf{x}) = \|\mathbf{x}\|_2$

   (c) $f(\mathbf{x}) = I_C(\mathbf{x})$, where C is a convex set.

   (d) $f(\mathbf{x}) = \|\mathbf{x}\|_1$.

5. (Optional) Consider the convex optimization problem

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

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 (1).

   (b) Find the relationship between (1) and the constrained optimization problem

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

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

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

Solution: ■

Exercise 2: 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 $D \subseteq \mathbf{dom}\ F$. Suppose that the problem (2) is solvable, and **we do not require the differentiability of** F.

1. If $\mathbf{x} \in \mathbf{int}\,(\mathbf{dom}\ F) \cap D$ 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 $\mathbf{x} \in \mathbf{int}\,(\mathbf{dom}\ F)$ 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

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

   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 $\min\limits_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x})$.

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

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

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

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

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

Solution: ■

Exercise 3: Projected Gradient Descent (Optional)

Consider the following problem

$$\min_{x \in D} f(x), \tag{3}$$

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 (3) 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 (3) and the sequence generated by the *projected gradient descent* algorithm. Suppose that $\mathbf{x}^*$ is the solution to the problem (3).

   (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 4: [1] ISTA with Backtracking

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

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}), \tag{4}$$

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 (4) 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

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

Algorithm 1 ISTA with Backtracking

1: **Input:** An initial point $\mathbf{x}_0$, an initial constant $L_0 > 0$, a threshold $\eta > 1$, and $k = 1$.
2: **while** the *termination condition* does not hold **do**
3: Find the smallest non-negative integer i_k such that with $\tilde{L} = \eta^{i_k} L_{k-1}$

$$F(p_{\tilde{L}}(\mathbf{x}_{k-1})) \le Q_{\tilde{L}}(p_{\tilde{L}}(\mathbf{x}_{k-1});\mathbf{x}_{k-1}). \tag{5}$$

4: $L_k \leftarrow \eta^{i_k} L_{k-1}$, $\mathbf{x}_k \leftarrow p_{L_k}(\mathbf{x}_{k-1})$,
5: $k \leftarrow k+1$,
6: **end while**

1. Show that the sequence $\{F(\mathbf{x}_k)\}$ produced by Algorithm 1 is non-increasing.

2. Show that Inequality (5) 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 5: Programming Exercise: Naive Bayes Classifier

We provide you with a data set that contains spam and non-spam emails ("hw5_nb.zip"). Please use the Naive Bayes Classifier to detect the spam emails. Finish the following exercises by programming. You can use your favorite programming language.

1. Remove all the tokens that contain non-alphabetic characters.
2. Train the Naive Bayes Classifier on the training set according to Algorithm 2.
3. Test the Naive Bayes Classifier on the test set according to Algorithm 3. You may encounter a problem that the likelihood probabilities you calculate approach 0. How do you deal with this problem?
4. Compute the confusion matrix, accuracy, precision, recall, and F-score.
5. Without the Laplace smoothing technique, complete the steps again.

Algorithm 2 Training Naive Bayes Classifier

Input: The training set with the labels $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$.

$\mathcal{V} \leftarrow$ the set of distinct words and other tokens found in $\mathcal{D}$
for each target value c in the labels set $\mathcal{C}$ **do**
  $\mathcal{D}_c \leftarrow$ the training samples whose labels are c
  $P(c) \leftarrow \frac{|\mathcal{D}_c|}{|\mathcal{D}|}$
  $T_c \leftarrow$ a single document by concatenating all training samples in $\mathcal{D}_c$
  $n_c \leftarrow |T_c|$
  for each word w_k in the vocabulary $\mathcal{V}$ **do**
    $n_{c,k} \leftarrow$ the number of times the word w_k occurs in T_c
    $P(w_k|c) = \frac{n_{c,k}+1}{n_c+|\mathcal{V}|}$
  end for
end for

Algorithm 3 Testing Naive Bayes Classifier

Input: An email $\mathbf{x}$. Let x_i be the i^{th} token in $\mathbf{x}$. $\mathcal{I} = \emptyset$.

for $i = 1, \ldots, |\mathbf{x}|$ **do**
  if $\exists\, w_{k_i} \in \mathcal{V}$ such that $w_{k_i} = x_i$ **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup i$
  end if
end for
predict the label of $\mathbf{x}$ by

$$\hat{y} = \arg\max_{c \in \mathcal{C}} P(c) \prod_{i \in \mathcal{I}} P(w_{k_i}|c)$$

Solution: ■

Exercise 6: Logistic Regression and Newton's Method

Given the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Let

$$\mathcal{I}^+ = \{i : i \in [n], y_i = 1\},$$
$$\mathcal{I}^- = \{i : i \in [n], y_i = 0\},$$

where $[n] = \{1, 2, \ldots, n\}$. We assume that $\mathcal{I}^+$ and $\mathcal{I}^-$ are not empty.
Then, we can formulate the logistic regression of the form.

$$\min_{\mathbf{w}} \; L(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log\left(\frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}{1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}\right) + (1-y_i)\log\left(\frac{1}{1+\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle)}\right)\right), \quad (6)$$

where $\mathbf{w} \in \mathbb{R}^{d+1}$ is the model parameter to be estimated and $\overline{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$.

1. (a) Suppose that the training data is strictly linearly separable, that is, there exists $\hat{\mathbf{w}} \in \mathbb{R}^{d+1}$ such that

$$\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle > 0, \, \forall \, i \in \mathcal{I}^+,$$
$$\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i\rangle < 0, \, \forall \, i \in \mathcal{I}^-.$$

   Show that problem (6) has no solution.

   (b) Suppose that the training data is NOT linearly separable, that is, for all $\mathbf{w} \in \mathbb{R}^{d+1}$, there exists $i \in [n]$ such that

$$\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle < 0, \text{ if } i \in \mathcal{I}^+,$$

   or

$$\langle \mathbf{w}, \overline{\mathbf{x}}_i\rangle > 0, \text{ if } i \in \mathcal{I}^-.$$

   Show that problem (6) always admits a solution.

2. Suppose that $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_n)^\top \in \mathbb{R}^{n\times(d+1)}$ and $\mathbf{rank}(\overline{\mathbf{X}}) = d + 1$. Show that $L(\mathbf{w})$ is strictly convex, i.e., for all $\mathbf{w}_1 \neq \mathbf{w}_2$,

$$L(t\mathbf{w}_1 + (1-t)\mathbf{w}_2) < tL(\mathbf{w}_1) + (1-t)L(\mathbf{w}_2), \forall \, t \in (0, 1).$$

3. In real applications, a widely-used method to learn the parameters' values of logistic regression is to solve the optimization problem in (6) with a regularization term, e.g.,

$$\min_{\mathbf{w}} F(\mathbf{w}) = L(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \; \lambda > 0.$$

   The Newton's method is an iterative method for optimization problems. We use the Newton's method to fit the regularized logistic regression by the following algorithm.

Algorithm 4 Newton's Method for Logistic Regression

1: **Input:** The twice-differentiable objective function $F(\mathbf{w})$, the initial point $\mathbf{w_0}$, the degree of precision ϵ.
2: Calculate the gradient $\mathbf{g}(\mathbf{w}) = \nabla F(\mathbf{w})$ and the Hessian matrix $\mathbb{H}(\mathbf{w})$ of the input $F(\mathbf{w})$.
3: **while** $\|\mathbf{g}_k(\mathbf{w}_k)\|_2 \geq \epsilon$

(a) Let $\mathbb{H}(\mathbf{w}_k) = \mathbb{H}_k$ and $\mathbf{g}(\mathbf{w}_k) = \mathbf{g}_k$ to simplify notations. Calculate the Hessian matrix $\mathbb{H}_k$, and the let $\mathbf{w}_{k+1} = \mathbf{w}_k - \mathbb{H}_k^{-1}\mathbf{g}_k$.

(b) $k = k + 1$.

(c) Calculate the gradient $\mathbf{g}_{k+1}$.

4: **Output:** $\hat{\mathbf{w}}$,the first point satisfying $\|\hat{\mathbf{g}}\|_2 < \epsilon$.

(a) Please calculate the gradient $\mathbf{g}(\mathbf{w})$ and the Hessian matrix $\mathbb{H}(\mathbf{w})$ of the regularized Logistic regression.

(b) Please show that the Hessian matrix $\mathbb{H}(\mathbf{w})$ is invertible.

(c) (Bonus) Please show the local convergence of Newton's method in logistic regression, i.e.,

$$\frac{\|\mathbf{w}_{k+1} - \mathbf{w}^*\|}{\|\mathbf{w}_k - \mathbf{w}^*\|^2} < B,$$

for some $B \in \mathbb{R}$, if the initial point is close enough to $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$.

Exercise 7: Convergence of Stochastic Gradient Descent for Convex Function

Consider an optimization problem

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}), \tag{7}$$

where the objective function F is continuously differentiable and strongly convex with convexity parameter $\mu > 0$. Suppose that the gradient of F, i.e., ∇F, is Lipschitz continuous with Lipschitz constant L, and F can attain it minimum F^* at $\mathbf{w}^*$. We use the stochastic gradient descent(SGD) algorithm introduced in Lecture 12 to solve the problem (7). Let the solution sequence generated by SGD be $(\mathbf{w}_k)$.

1. Please show that $\forall \mathbf{w} \in \mathbf{dom}\ F$, the following inequality

$$F(\mathbf{w}) - F^* \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2 \tag{8}$$

   holds, and interpret the role of strong convexity based on this.

2. Recall that with a fixed stepsize $\alpha \in [0, \frac{1}{LM_G}]$ where M_G (as well as the following M) is a parameter regarding the upper bound of the variance of stochastic gradient in SGD, the sequence $(\mathbb{E}[F(\mathbf{w}_k)])$ generated by SGD converges to a neighborhood of F^* with a linear rate, i.e,

$$\mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^*] \leq \frac{LM}{2\mu}\alpha + (1-\mu\alpha)^k(F(\mathbf{w}_0) - F^* - \frac{LM}{2\mu}\alpha) \xrightarrow{\text{linear}} \frac{LM}{2\mu}\alpha.$$

   This means that the expected optimality gap, i.e., $\mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^*]$, fails to converge to zero. In order to alleviate this problem, we consider a strategy of diminishing stepsize α_k. Suppose that the SGD method is run with a stepsize sequence (α_k) such that, for all $k \in \mathbb{N}$,$\alpha_k = \frac{\beta}{\gamma+k}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ satisfying $\alpha_0 \leq \frac{1}{LM_G}$. Please show that $\forall k \in \mathbb{N}$, we have

$$\mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^*] \leq \frac{\tau}{\gamma + k},$$

   where $\tau = \max\{\frac{\beta^2 LM}{2(\beta\mu - 1)}, \gamma(F(\mathbf{w}_0) - F^*)\}$.

3. In practice, for the same problem, SGD enjoys less time cost but more iteration steps than gradient descent methods and may suffer from non-convergence. As a trade-off between SGD and gradent descent approaches, consider using mini-batch samples to estimate the full gradient. Taking k^{th} iteration as an example, instead of picking a single sample, we randomly select a subset $\mathcal{S}_k$ of the sample indices to compute the update direction

$$\mathbf{g}_k(\xi_k) = \frac{1}{|\mathcal{S}_k|}\sum_{i \in \mathcal{S}_k} \nabla f_i(\mathbf{w}_k)$$

where ξ_k is the selected samples. For simplicity, suppose that the mini-batches in all iterations are of constant size, i.e., $|\mathcal{S}_k| = n_m$, and the stepsize α is fixed. Please show that for mini-batch SGD, there holds

$$\mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^*] \leq \frac{LM}{2\mu n_m}\alpha + (1-\mu\alpha)^k(F(\mathbf{w}_0) - F^* - \frac{LM}{2\mu n_m}\alpha) \xrightarrow{\text{linear}} \frac{LM}{2\mu n_m}\alpha.$$

Moreover, point out the advantage of mini-batch SGD compared to SGD in terms of the number of the iteration step.

4. Notice that in real applications, F is not always strongly convex. Let F be convex and continuously differentiable, and the second moment of stochastic gradient $\mathbf{g}$ be bounded, i.e.,

$$\mathbb{E}_\xi[\|\mathbf{g}(\xi)\|_2^2] \leq G^2.$$

We denote $(\mathbf{w}_k)$ as a sequence generated by SGD algorithm with a fixed stepsize α. Besides, define $\tilde{\mathbf{w}}_K = \frac{1}{K+1}\sum_{k=0}^{K}\mathbf{w}_k$ and $F^* = F(\mathbf{w}^*)$.

(a) If X is a random variable and $h : \mathbb{R} \to \mathbb{R}$ is a convex function, please show that

$$h(\mathbb{E}[X]) \leq \mathbb{E}[h(X)].$$

(b) Suppose that the stochastic gradient at k^{th} iteration is $\mathbf{g}_k$. Please show that

$$\mathbb{E}_{\xi_0:\xi_k}[F(\mathbf{w}_k) - F^*] \leq \mathbb{E}_{\xi_0:\xi_k}[\langle \mathbf{g}_k, \mathbf{w}_k - \mathbf{w}^*\rangle].$$

(c) Please show that

$$\mathbb{E}_{\xi_0:\xi_k}[F(\mathbf{w}_k) - F^*] \leq \frac{1}{2\alpha}\mathbb{E}_{\xi_0:\xi_k}[\|\mathbf{w}_k - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{k+1} - \mathbf{w}^*\|_2^2 + \alpha^2\|\mathbf{g}_k\|_2^2].$$

(d) Please show that

$$\mathbb{E}_{\xi_0:\xi_K}[F(\tilde{\mathbf{w}}_K) - F^*] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 + \alpha^2 G^2(K+1)}{2\alpha(K+1)} \to \frac{\alpha G^2}{2}.$$

Exercise 8: Programming Exercise: Logistic Regression (Optional)

We provide you with a dataset of handwritten digits[1] that contains a training set of 60000 examples and a test set of 2022 examples ("hw5_lr.mat"). Each image in this dataset has 28×28 pixels and the associated label is the handwritten digit—that is, an integer from the set $\{0, 1, \cdots, 9\}$—in the image. In this exercise, you need to build a logistic regression classifier to *predict if a given image has the handwritten digit* 6 *in it or not*. You can use your favorite programming language to finish this exercise.

1. Normalize the data matrix and please find a Lipschitz constant of $\nabla L(\mathbf{w})$, where $L(\mathbf{w})$ is the objective function of the logistic regression after normalizing and $\mathbf{w}$ is the model parameter to be estimated.

2. (a) Use the gradient descent algorithm (GD), which is a special case of ISTA introduced in Lecture 09, and SGD to train the logistic regression classifier on the training set, respectively. Evaluate the classification accuracy on the training set after each iteration. Stop the iteration when Accuracy $\geq 90\%$ or total steps are more than 5000. Please plot the accuracy of these two classifiers (the one trained by GD and the other trained by SGD) versus the iteration step on one graph.

   (b) Compare the total iteration counts and the total time cost of the two methods (GD and SGD), respectively. Please report your result.

   (c) Compare the confusion matrix, precision, recall and F1 score of the two classifiers (the one trained by GD and the other trained by SGD). Please report your result.

   (d) Use GD and SGD to train the logistic regression classifier with a 2-norm regularization term. Note that other experimental setup details is in line with 2.(a). Please plot the accuracy of these two classifiers (the one trained by GD and the other trained by SGD) versus the iteration step on one graph and compare the confusion matrix, precision, recall and F1 score of the two classifiers.

3. (a) The training set is imbalanced as the majority class has roughly nine times more images than the minority class. Imbalanced data can hurt the performance of the classifiers badly. Thus, please undersample the majority class such that the numbers of images in the two classes are roughly the same.

   (b) Use GD and SGD to train the logistic regression classifier on the new training set after undersampling. Stop the iteration when Accuracy $\geq 90\%$ or total steps are more than 5000.

   (c) Evaluate the two classifiers (the one trained with GD on the original training set and the other trained on the new training set after undersampling) on the test set. Compare the confusion matrix, precision, recall and F1 score of the two classifiers. Please report your result.

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.