Lecture 11. Logistic Regression

Lecturer: Jie Wang Date: Nov 16, 2023

Last Update: November 27, 2023

The major references of this lecture are this note by Tom Mitchell and [1].

1 Introduction

Suppose that we are given a set of data $\{(\mathbf{x}_i, y_i)\}_i^n$, where $y_i \in \{0, 1\}$. Clearly, this is a classification problem. As a commonly-used approach for classification, logistic regression aims to learn a mapping $f : X \to Y$, where $X = (X_1, \ldots, X_d)$ and $Y \in \{0, 1\}$.

2 The Probabilistic Approach

Similarly to the naïve Bayes classifier, we start again from the Bayes rule, which leads to

$$\begin{aligned} P(Y = 0|X) &= \frac{P(X|Y = 0)P(Y = 0)}{P(X|Y = 0)P(Y = 0) + P(X|Y = 1)P(Y = 1)} \\ &= \frac{1}{1 + \dfrac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 0)P(Y = 0)}}. \end{aligned} \tag{1}$$

Again, we see the class priors $P(Y = 1)$ and $P(Y = 0)$, and the conditional joint probabilities $P(X|Y = 1)$ and $P(X|Y = 0)$. We model these probabilities by a suite of assumptions as follows. We first make assumptions on the class priors.

Assumption 1.

We assume that $Y \sim \text{Bern}(p)$, that is, Y has the Bernoulli distribution with $P(Y = 1) = p$ [clearly, we have $P(Y = 0) = 1 - p$].

Next, we make assumptions on the conditional joint probabilities.

Assumption 2. *For $X = (X_1, \ldots, X_d)$,*

1. *X_i and X_j are conditionally independent given Y for $i \neq j$;*
2. *X_j is a continuous random variable, and the class-conditional distribution is Gaussian, i.e., $P(X_j|Y = 0) \sim N(\mu_{j,0}, \sigma_j^2)$ and $P(X_j|Y = 1) \sim N(\mu_{j,1}, \sigma_j^2)$.*

Notice that, for different values of Y, the conditional distributions of the random variable X_j, $j = 1, \ldots, d$, only differ in the means, while they have the same variance.

We can now continue our derivation from where we left in Eq. (1).

$$\begin{aligned} P(Y = 0|X) &= \frac{1}{1 + \exp\left(\ln \dfrac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 0)P(Y = 0)}\right)} \\ &= \frac{1}{1 + \exp\left(\sum_j \ln \dfrac{P(X_j|Y = 1)}{P(X_j|Y = 0)} + \ln \dfrac{p}{1 - p}\right)}. \end{aligned} \tag{2}$$

According to the second part of Assumption 2, we have

$$\begin{aligned}\sum_j \ln \frac{P(X_j|Y=1)}{P(X_j|Y=0)} &= \sum_j \ln \frac{\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(X_j-\mu_{j,1})^2}{2\sigma_j^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(X_j-\mu_{j,0})^2}{2\sigma_j^2}\right)} \\ &= \sum_j \ln \exp\left(\frac{(X_j-\mu_{j,0})^2 - (X_j-\mu_{j,1})^2}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{(X_j-\mu_{j,0})^2 - (X_j-\mu_{j,1})^2}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{(X_j^2 - 2X_j\mu_{j,0} + \mu_{j,0}^2) - (X_j^2 - 2X_j\mu_{j,1} + \mu_{j,1}^2)}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2} X_j + \frac{\mu_{j,0}^2 - \mu_{j,1}^2}{2\sigma_j^2}\right).\end{aligned} \tag{3}$$

Plugging Eq. (3) into Eq. (2) leads to

$$\begin{aligned}P(Y=0|X) &= \frac{1}{1+\exp\left(\ln\frac{p}{1-p} + \sum_j \left(\frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2} X_j + \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2}\right)\right)} \\ &= \frac{1}{1+\exp\left(\ln\frac{p}{1-p} + \sum_j \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2} + \sum_j \frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2} X_j\right)}.\end{aligned}$$

To simplify notations, we let

$$\begin{aligned}w_j &= \frac{\mu_{j,1}-\mu_{j,0}}{\sigma_j^2},\ j=1,\ldots,d, \\ w_0 &= = \frac{p}{1-p} + \sum_j \frac{\mu_{j,0}^2-\mu_{j,1}^2}{2\sigma_j^2}.\end{aligned}$$

Then, we can rewrite Eq. (1) as follows

$$P(Y=0|X) = \frac{1}{1+\exp(w_0 + \sum_{j=1}^d w_j X_j)}, \tag{4}$$

which implies that

$$P(Y=1|X) = 1 - P(Y=0|X) = \frac{1}{1+\exp(-(w_0 + \sum_{j=1}^d w_j X_j))}. \tag{5}$$

Thus, given a data instance $\mathbf{x}$, we compute the conditional probability $P(Y=0|X=\mathbf{x})$ and $P(Y=1|X=\mathbf{x})$, and predict its label as the one which makes the corresponding conditional probability larger.

Remark 1. We derive the logistic regression model based on Assumptions 1 and 2. We can see that logistic regression is a **linear** model, as the decision boundary—that is,

$$\{\mathbf{x} \in \mathbb{R}^d : P(Y = 0|X = \mathbf{x}) = P(Y = 1|X = \mathbf{x})\}$$

is a hyperplane in $\mathbb{R}^d$ (why?). However, in many real applications, Assumptions 1 and 2 may not hold. For example, the input may have discrete features or include both discrete and continuous features. An alternative approach to derive the logistic regression model is to simply assume that the log likelihood ratio of the class-conditional densities is linear:

$$\ln \frac{P(Y = 1|X)}{P(Y = 0|X)} = w_0 + \sum_j w_j X_j. \tag{6}$$

3 Learning the Parameters via MLE

Based on Assumptions 1 and 2, we derive the logistic regression model in the form of Eq. (4) and Eq. (5), and we also derives the value of the parameters. However, as the values of the parameters of the involved distributions are usually unknown, and Assumptions 1 and 2 may not hold in many real applications, we can not direct apply Eq. (4) and Eq. (5) to predict the label of a new data instance.

In this section, we describe how to learn the parameters, i.e., $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ from the training data, via the maximum likelihood estimation (MLE):

$$\begin{aligned}\hat{\mathbf{w}} &= \operatorname*{\mathbf{argmax}}_{\mathbf{w}\in\mathbb{R}^{d+1}} \prod_i P(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \operatorname*{\mathbf{argmax}}_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_i \ln P(y_i|\mathbf{x}_i, \mathbf{w}).\end{aligned}$$

For notational convenience, let

$$-L(\mathbf{w}) = \sum_i \ln P(y_i|\mathbf{x}_i, \mathbf{w}).$$

We can write the right hand side of the above equation in a unified form of

$$\begin{aligned}-L(\mathbf{w}) &= \sum_i \{y_i \ln P(Y = 1|\mathbf{x}_i, \mathbf{w}) + (1 - y_i) \ln P(Y = 0|\mathbf{x}_i, \mathbf{w})\} \\ &= \sum_i \left\{ y_i \ln \frac{1}{1 + \exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle)} + (1 - y_i) \ln \frac{1}{1 + \exp(\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle)} \right\} \\ &= -\sum_i \{y_i \ln(1 + \exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle)) + (1 - y_i) \ln(1 + \exp(\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle))\},\end{aligned}$$

where $\bar{\mathbf{x}}_i = (1, \mathbf{x}_i^\top)^\top$. We can see that

$$\begin{aligned}\hat{\mathbf{w}} &= \operatorname*{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}) \\ &= \operatorname*{\mathbf{argmin}}_{\mathbf{w}} \sum_i \{y_i \ln(1 + \exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle)) + (1 - y_i) \ln(1 + \exp(\langle \mathbf{w}, \bar{\mathbf{x}}_i \rangle))\}.\end{aligned} \tag{7}$$

Question 1. To solve for $\hat{\mathbf{w}}$, we can of course apply the gradient descent algorithm we introduced before. However, we need to answer one important question: does the problem in (7) always admit a solution?

3.1 Gradient Descent for Logistic Regression

Let us assume at this point that the problem in (7) admits a solution. To apply GD to find $\hat{\mathbf{w}}$, we need to compute the gradient of L:

$$\nabla L(\mathbf{w}) = \mathbf{X}^\top \mathbf{h}(\mathbf{w}),$$

where $\mathbf{X} \in \mathbb{R}^{n\times(d+1)}$ with its i^{th} row being $\bar{\mathbf{x}}_i^\top$ and $\mathbf{h}(\mathbf{w}) = (h_1(\mathbf{w}), \ldots, h_n(\mathbf{w}))^\top$ with

$$h_i(\mathbf{w}) = \frac{1}{1+\exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle)} - y_i.$$

Indeed, the problem in (7) is a **convex** optimization problem. This can be seen from the fact that the Hessian matrix of $L(\mathbf{w})$ is positive semidefinite for all $\mathbf{w}$. Specifically, the Hessian matrix of $L(\mathbf{w})$ is

$$\nabla^2 L(\mathbf{w}) = \mathbf{X}^\top \mathbf{\Sigma}_{\mathbf{w}} \mathbf{X},$$

where $\mathbf{\Sigma}_{\mathbf{w}}$ is a diagonal matrix with its i^{th} entry on its diagonal being

$$\frac{\exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle)}{(1+\exp(-\langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle))^2}.$$

Clearly, the Hessian matrix $\nabla^2 L(\mathbf{w}) \succeq 0$.

Question 2. Suppose that the problem in (7) admits a solution. Is it unique?

3.2 Regularization

In real applications, a widely-used method to learn the parameters' values of logistic regression is to solve the optimization problem in (7) with a regularization term, e.g.,

$$\hat{\mathbf{w}} = \operatorname*{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{8}$$

Besides alleviating overfitting, the regularization term also brings desirable properties in computation. To to answer the question as follows and find out why.

Question 3.

1. Does the problem in (8) always admit a solution?
2. If the problem in (8) admits a solution, is it unique?

4 Logistic Regression for Multiple Target Values

Previous sections consider the classification problems with two classes. Is logistic regression applicable to the cases with more than two classes?

Suppose that $Y \in \mathcal{C} = \{c_1, \ldots, c_K\}$. Then

$$P(Y = c_k|X) = \frac{\exp(w_{k,0} + \sum_{j=1}^d w_{k,j}X_j)}{1 + \sum_{k=1}^{K-1}\exp(w_{k,0} + \sum_{j=1}^d w_{k,j}X_j)},\ k = 1, \ldots, K-1,$$

$$P(Y = c_K|X) = \frac{1}{1 + \sum_{k=1}^{K-1}\exp(w_{k,0} + \sum_{j=1}^d w_{k,j}X_j)},\ k = K.$$

5 Stochastic Gradient Descent

Besides the classic gradient descent, we have another popular suit of methods, called stochastic gradient descent, that are widely used to solve the problems like (7). We first motivate SGD in Section 5.1. Then, we analyze the convergence property of SGD in Section 5.2. An excellent visualization of SGD can be found by following this link http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html.

5.1 Motivation

The motivation of many machine learning methods boils down to a minimization of the so-called **empirical risk**, i.e., the average of the sample losses:

$$R_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \mathbf{w}), y_i). \tag{9}$$

where $h(\cdot; \mathbf{w})$ is the parameterized model and $\ell(\cdot, \cdot)$ measures the prediction error (loss). To save notations, let

$$f_i(\mathbf{w}) = \ell(h(\mathbf{x}_i; \mathbf{w}), y_i).$$

In this section, we consider the following optimization problem:

$$\min_{\mathbf{w}} R_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}). \tag{10}$$

Remark 2. If we assume that each single sample $(\mathbf{x}_i, y_i)$ is a realization of a random vector $\xi \in \mathbb{R}^{d+1}$ with an unknown distribution $\mathcal{D}$, the objective function we would like to minimize is indeed the **expected risk**

$$R(\mathbf{w}) = \mathbb{E}_{\xi}[f(\xi; \mathbf{w})], \tag{11}$$

where

$$f(\xi; \mathbf{w}) = \ell(h(\xi_{[1:d]}; \mathbf{w}), \xi_{d+1}),$$

with $\xi_{[1:d]} = (\xi_1, \ldots, \xi_d)$.

To solve the problem in (10), we can apply the gradient descent algorithm, which is a special case of ISTA introduced in previous lectures. This requires a scan of the entire data set in each iteration to compute the full gradient

$$\nabla R_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{w}),$$

which can be quite time-consuming if the training set contains a huge amount of data instances. Moreover, in many real applications, we have no access to the full gradient, as the data instance comes in only one at a time. This motives the popular **stochastic gradient descent** (SGD) method [1].

Let us consider the k^{th} iteration. Instead of computing the full gradient $\nabla R_n(\mathbf{w}_k)$, SGD aims to find an update direction $\mathbf{g}_k$ to approximate the full gradient. **The only requirement is that the expectation of this approximate update direction equals to the full gradient**, i.e.,

$$\mathbb{E}[\mathbf{g}_k] = \nabla R_n(\mathbf{w}_k).$$

A simple choice of $\mathbf{g}_k$ is to uniformly sample a data instance $\xi_k = (\mathbf{x}_{i_k}, y_{i_k})$ from $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and set

$$\mathbf{g}_k(\xi_k) = \nabla f_{i_k}(\mathbf{w}_k) = \nabla \ell(h(\mathbf{x}_{i_k}; \mathbf{w}), y_{i_k}). \tag{12}$$

It is easy to see that

$$\mathbb{E}_{\xi_k}[\mathbf{g}_k(\xi_k)] = \nabla R_n(\mathbf{w}_k).$$

We summarize the SGD algorithm as follows.

Algorithm 1 Stochastic Gradient Descent Algorithm

Input: an initial point $\mathbf{w}_0$, the number of iterations K, stepsize $\alpha > 0$, $k = 0$

Output: $\mathbf{w}_K$

1: **repeat**
2: choose update direction $\mathbf{g}_k(\xi_k)$ by Eq. (12)
3: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \alpha \mathbf{g}_k(\xi_k)$
4: $k \leftarrow k + 1$
5: **until** $k \geq K$

5.2 Convergence Analysis

To keep the notation simple, we rewrite the problem (10) as follows.

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w}). \tag{13}$$

Similar to the analysis of ISTA, we first make some assumptions on the problem (13).

Assumption 3.

1. *The objective function F is convex and continuously differentiable, which implies that*

$$F(\mathbf{w}_1) \geq F(\mathbf{w}_2) + \langle \nabla F(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle. \tag{14}$$

2. *The gradient of function F is Lipschitz continuous, i.e., $\exists L > 0$, such that*

$$\|\nabla F(\mathbf{w}_2) - \nabla F(\mathbf{w}_1)\| \leq L \|\mathbf{w}_2 - \mathbf{w}_1\|. \tag{15}$$

3. *The function F attains its minimum at $\mathbf{w}^*$, i.e.,*

$$F(\mathbf{w}^*) = F^* = \min_{\mathbf{w}} F(\mathbf{w}). \tag{16}$$

Recall that, for GD, the above assumptions imply the descent lemma as follows.

Lemma 1. *Suppose that a function F is continuously differentiable and its gradient is Lipschitz continuous with constant $L > 0$. Then, for the sequence $(\mathbf{w}_k)$ generated by*

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha \nabla F(\mathbf{w}_k),$$

we have

$$F(\mathbf{w}_{k+1}) \leq F(\mathbf{w}_k) - \alpha(1 - \frac{L}{2}\alpha)\|\nabla F(\mathbf{w}_k)\|^2. \tag{17}$$

Thus, with stepsize $0 < \alpha < \frac{2}{L}$, the sequence $(F(\mathbf{w}_k))$ decreases monotonously.

Proof. By the Lipschitz continuity of the gradient, we have

$$\begin{aligned} F(\mathbf{w}_{k+1}) &\leq F(\mathbf{w}_k) + \langle \nabla F(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle + \frac{L}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \\ &= F(\mathbf{w}_k) - \alpha\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2}\alpha^2\|\nabla F(\mathbf{w}_k)\|^2 \\ &= F(\mathbf{w}_k) - \alpha\left(1 - \frac{L}{2}\alpha\right)\|\nabla F(\mathbf{w}_k)\|^2, \end{aligned} \tag{18}$$

which completes the proof. □

Though there is a **descent** in the name "SG**D**", it is **NOT** a descent algorithm. Due to the stochastic nature of the update direction $\mathbf{g}_k(\xi_k)$, the function values may even go up in some iterations. Can we show a *descent* property for SGD in terms of the **expectation**? The answer is still no, due to the nonnegative variance of the update direction.

Lemma 2. *Suppose that F is continuously differentiable and its gradient is Lipschitz continuous with constant $L > 0$. Then, for the sequence generated by SGD in Algorithm 1, we have*

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq -\alpha(1 - \frac{L}{2}\alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2}\alpha^2 \mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)], \tag{19}$$

where $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] = \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k) - \mathbb{E}_{\xi_k}[\mathbf{g}_k(\xi_k)]\|^2]$.

Proof. Let us consider the k^{th} iteration of SGD.

Lipschitz continuity of ∇F implies that

$$F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k) \leq \langle \nabla F(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle + \frac{L}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \tag{20}$$

Noting the update rule in Algorithm 1 and taking expectation with respect to ξ_k of both sides of (20) lead to

$$\begin{aligned} \mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq& \langle \nabla F(\mathbf{w}_k), -\alpha \mathbb{E}_{\xi_k}[\mathbf{g}_k(\xi_k)] \rangle + \frac{L}{2}\alpha^2 \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2] \\ =& -\alpha\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2}\alpha^2 \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2]. \end{aligned} \tag{21}$$

Moreover, we have

$$\begin{aligned} \mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] &= \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2] - \|\mathbb{E}_{\xi_k}[\mathbf{g}_k(\xi_k)]\|^2 \\ &= \mathbb{E}_{\xi_k}[\|\mathbf{g}_k(\xi_k)\|^2] - \|\nabla F(\mathbf{w}_k)\|^2. \end{aligned} \tag{22}$$

Thus, the inequality (19) follows immediately by plugging Eq. (22) into (21). □

This lemma shows that the **expected** difference of two successive function values consists of two terms: the descent term and the variance term. We can see that $\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1})]$ **decrease with respect to** $F(\mathbf{w}_k)$ **only if the descent term dominates the variance term**.

Notice that, even if $\nabla F(\mathbf{w}_k) = \mathbf{0}$—that is, $\mathbf{w}_k$ is one of the optimum of F—the variance $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)]$ can still be positive. This suggests that the sequence $(\mathbb{E}[F(\mathbf{w}_k)])$ may not converge to F^* with a fixed stepsize $\alpha > 0$. On the other hand, if we set $\mathbf{g}_k(\xi_k) = \nabla F(\mathbf{w}_k)$, then we have $\mathbb{V}_{\xi_k}[g_k] = 0$ and we recover the descent lemma in GD immediately.

For SGD, we cannot expect that $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] = 0$, and we cannot even expect that it is bounded. However, we can make the following reasonable assumption for the objective function F for SGD.

Assumption 4. *We assume that the upper bound of* $\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)]$ *takes the form of*

$$\mathbb{V}_{\xi_k}[\mathbf{g}_k(\xi_k)] \leq M + M_V \|\nabla F(\mathbf{w}_k)\|^2, \tag{23}$$

where M *and* M_V *are positive constants.*

Question 4. Why do we need a constant M on the RHS of Eq. (23)?

Lemma 3. *Let* $M_G = M_V + 1$. *Assumptions 3 and 4 imply that*

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k)] \leq -\alpha(1 - \frac{L}{2} M_G \alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{L}{2} M \alpha^2. \tag{24}$$

We are now ready to analyze the convergence property of SGD. To light the burden, we consider the strongly convex objective functions, i.e.,

$$F(\mathbf{w}') \geq F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\mu}{2}\|\mathbf{w}' - \mathbf{w}\|^2, \forall\, \mathbf{w}, \mathbf{w}' \in \mathbf{dom}\ F, \tag{25}$$

where $\mu > 0$. Recall that, we have following result for strongly convex functions.

Lemma 4. *Suppose that* F *is strongly convex with convexity parameter* $\mu > 0$ *and continuously differentiable. Then,*

$$F(\mathbf{w}) - F^* \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2, \forall\, \mathbf{w} \in \mathbf{dom}\ F. \tag{26}$$

Proof. If we fix $\mathbf{w}$, the RHS of (25) is clearly a quadratic function of $\mathbf{w}'$. Let

$$Q(\mathbf{w}'; \mathbf{w}) = F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\mu}{2}\|\mathbf{w}' - \mathbf{w}\|^2.$$

It is easy to see that

$$Q^*(\mathbf{w}) := \min_{\mathbf{w}'} Q(\mathbf{w}'; \mathbf{w}) = F(\mathbf{w}) - \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2.$$

By noting that $F^* \geq Q^*(\mathbf{w})$, the claim follows immediately. □

The following result shows a linear convergence rate of SGD for strongly convex objective functions.

Theorem 1. ***(Strongly Convex Objective, Fixed Stepsize)*** *Suppose that Assumptions 3 and 4 hold and* $0 < \alpha < \frac{1}{LM_G}$. *Then, the sequence* $(\mathbb{E}[F(\mathbf{w}_k)])$ *generated by SGD converges to a neighborhood of* F^* *with a linear rate. Specifically, we have*

$$\begin{aligned} \mathbb{E}_{\xi_0:\xi_{k-1}}[F(\mathbf{w}_k) - F^*] &\leq \frac{LM}{2\mu}\alpha + (1 - \mu\alpha)^k (F(x_0) - F^* - \frac{LM}{2\mu}\alpha) \\ &\xrightarrow{\text{linear}} \frac{LM}{2\mu}\alpha. \end{aligned} \tag{27}$$

Proof. Subtracting F^* from both sides of (24) and rearranging the terms yield

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] \leq F(\mathbf{w}_k) - F^* - \alpha(1 - \frac{L}{2}M_G\alpha)\|\nabla F(\mathbf{w}_k)\|^2 + \frac{LM}{2}\alpha^2. \tag{28}$$

As $0 < \alpha < \frac{1}{LM_G}$, we have

$$\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] \leq F(\mathbf{w}_k) - F^* - \frac{\alpha}{2}\|\nabla F(\mathbf{w}_k)\|^2 + \frac{LM}{2}\alpha^2. \tag{29}$$

Combining (26) and (29) leads to

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(\mathbf{w}_{k+1}) - F^*] &\leq F(\mathbf{w}_k) - F^* - \mu\alpha(F(\mathbf{w}_k) - F^*) + \frac{LM}{2}\alpha^2 \\ &= (1 - \mu\alpha)(F(\mathbf{w}_k) - F^*) + \frac{LM}{2}\alpha^2,\end{aligned} \tag{30}$$

which is equivalent to

$$\mathbb{E}_{\xi_k}\left[F(\mathbf{w}_{k+1}) - F^* - \frac{LM}{2\mu}\alpha\right] \leq (1 - \mu\alpha)\left(F(\mathbf{w}_k) - F^* - \frac{LM}{2\mu}\alpha\right).$$

Now take the expectation with respect to $\xi_{k-1}, \ldots, \xi_0$ to both sides of the above inequality, we have

$$\mathbb{E}_{\xi_0:\xi_k}\left[F(\mathbf{w}_{k+1}) - F^* - \frac{LM}{2\mu}\alpha\right] \leq (1 - \mu\alpha)^{k+1}\left[F(\mathbf{w}_0) - F^* - \frac{LM}{2\mu}\alpha\right]. \tag{31}$$

In view of the fact that $0 < \mu < L$ (**why?**), we have

$$0 < \mu\alpha < \frac{1}{M_G} = \frac{1}{M_V + 1} < 1.$$

The claim follows immediately. □

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 2018.