CHAPTER 6. STOCHASTIC GRADIENT DESCENT

Jinghua Huang Pengfei Wu Kun Yuan

October 24, 2023

1 Problem formulation

This chapter considers the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} \quad f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[F(x;\xi)] \tag{1}$$

where $\xi \sim \mathcal{D}$ denotes the random data sample and $\mathcal{D}$ denotes the data distribution. Since $\mathcal{D}$ is typically unknown in machine learning, the closed-form of $f(x)$ is also unknown.

Notation. We introduce the following notations:

- Let $x^\star := \arg\min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal solution to problem (1).
- Let $f^\star := \min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal function value.
- Let $\mathcal{F}_k = \{x_k, \xi_{k-1}, x_{k-1}, \cdots, \xi_0\}$ be the filtration containing all historical variables at and before iteration k. Note that ξ_k does not belong to $\mathcal{F}_k$.

2 Stochastic gradient descent

Since $f(x)$ does not have a closed-form, we cannot access its gradient. However, since $F(x;\xi)$ is known, we can use $\nabla_x F(x;\xi)$ to approximate the true gradient $\nabla f(x)$. Throughout this lecture, we let $\nabla F(x;\xi) = \nabla_x F(x;\xi)$ for notation simplicity. Given any arbitrary initialization variable x_0, stochastic gradient descent (SGD) iterates as follows

$$x_{k+1} = x_k - \gamma \nabla F(x_k;\xi_k), \quad \forall\, k = 0, 1, 2, \cdots \tag{2}$$

where γ is the learning rate, and $\xi_k \sim \mathcal{D}$ is a random data sampled at iteration k. Since ξ_k is a random variable for any $k = 0, 1, \cdots$, each variable x_k is also a random variable for $k = 1, 2, \cdots$.

3 Convergence analysis

To facilitate convergence analysis, we introduce the following assumption:

Assumption 3.1. Given the filtration $\mathcal{F}_k$, we assume

$$\mathbb{E}[\nabla F(x_k; \xi_k)|\mathcal{F}_k] = \nabla f(x_k) \tag{3}$$

$$\mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k)\|^2|\mathcal{F}_k] \leq \sigma^2 \tag{4}$$

The above assumption indicates that, conditioned on the filtration $\mathcal{F}_k$, the stochastic gradient $\nabla F(x_k; \xi_k)$ is an unbiased estimate on $\nabla f(x_k)$, and the variance is bounded by σ^2. Under the above assumption, it is easy to verify that

$$\begin{aligned}
\mathbb{E}[\|\nabla F(x_k; \xi_k)\|^2|\mathcal{F}_k] &= \mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k) + \nabla f(x_k)\|^2|\mathcal{F}_k] \\
&= \|\nabla f(x_k)\|^2 + \mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k)\|^2|\mathcal{F}_k] \\
&\leq \|\nabla f(x_k)\|^2 + \sigma^2
\end{aligned} \tag{5}$$

where the second equality holds due to (3) and the last inequality holds due to (4).

3.1 Smooth and non-convex problem

Theorem 3.2. Suppose $f(x)$ is L-smooth and Assumption 3.1 holds. If $\gamma \leq 1/L$, SGD will converge at the following rate

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2\Delta_0}{\gamma(K+1)} + \gamma L\sigma^2, \tag{6}$$

where $\Delta_0 = f(x_0) - f^\star$. If we further choose $\gamma = \left[\left(\frac{2\Delta_0}{(K+1)L\sigma^2}\right)^{-\frac{1}{2}} + L\right]^{-1}$, SGD converges as

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \sqrt{\frac{8L\Delta_0\sigma^2}{K+1}} + \frac{2L\Delta_0}{K+1}. \tag{7}$$

Remark. If $\sigma^2 = 0$, the stochastic gradient reduces to the true gradient, and hence, SGD reduces to GD. Substituting $\sigma^2 = 0$ to SGD rate (7), we recover the rate $O(L/K)$ for GD. In other words, our convergence rate for SGD is consistent with GD.

Proof. Since $f(x)$ is L-smooth, we have

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})|\mathcal{F}_k] &\leq f(x_k) + \mathbb{E}[\langle \nabla f(x_k), x_{k+1} - x_k \rangle | \mathcal{F}_k] + \frac{L}{2}\mathbb{E}[\|x_{k+1} - x_k\|^2 | \mathcal{F}_k] \\
&= f(x_k) - \gamma \mathbb{E}[\langle \nabla f(x_k), \nabla F(x_k; \xi_k) \rangle | \mathcal{F}_k] + \frac{L\gamma^2}{2}\mathbb{E}[\|\nabla F(x_k; \xi_k)\|^2 | \mathcal{F}_k] \\
&\overset{(a)}{\leq} f(x_k) - \gamma(1 - \frac{L\gamma}{2})\|\nabla f(x_k)\|^2 + \frac{L\gamma^2\sigma^2}{2} \\
&\overset{(b)}{\leq} f(x_k) - \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{L\gamma^2\sigma^2}{2} \qquad (8)
\end{aligned}$$

where inequality (a) holds due to Assumption 3.1, and inequality (b) holds if $\gamma \leq 1/L$. By taking expectations over the filtration $\mathcal{F}_k$, we have

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \frac{\gamma}{2}\mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L\gamma^2\sigma^2}{2} \qquad (9)$$

which is equivalent to

$$\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2}{\gamma}\left(\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]\right) + \gamma L\sigma^2 \qquad (10)$$

Taking averaging over $k = 0, 1, \cdots, K$, we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2(f(x_0) - f^\star)}{\gamma(K+1)} + \gamma L\sigma^2. \qquad (11)$$

Defining $\Delta_0 := f(x_0) - f^\star$, if we set

$$\gamma = \left[\left(\frac{2\Delta_0}{(K+1)L\sigma^2}\right)^{-\frac{1}{2}} + L\right]^{-1}, \qquad (12)$$

it then holds that

$$\gamma \leq \min\left\{\frac{1}{L}, \gamma_1\right\}, \quad \text{where} \quad \gamma_1 = \left(\frac{2\Delta_0}{(K+1)L\sigma^2}\right)^{\frac{1}{2}}. \qquad (13)$$

Substituting (12) and (13) into (11), we have

$$\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] &\leq \frac{2(f(x_0) - f^\star)}{\gamma(K+1)} + \gamma_1 L\sigma^2 \\
&= 2\sqrt{\frac{2L\Delta_0\sigma^2}{K+1}} + \frac{2L\Delta_0}{K+1}. \qquad (14)
\end{aligned}$$

□

3.2 Smooth and convex problem

Theorem 3.3. Suppose $f(x)$ is convex and L-smooth. Under Assumption 3.1, if $\gamma \leq 1/(2L)$, SGD will converge at the following rate

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f(x^\star)] \leq \frac{\Delta_0}{\gamma(K+1)} + \gamma\sigma^2 \tag{15}$$

where $\Delta_0 = \|x_0 - x^\star\|^2$. If we further choose $\gamma = \left[\left(\frac{\Delta_0}{(K+1)\sigma^2}\right)^{-\frac{1}{2}} + 2L\right]^{-1}$, SGD converges as follows

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k)-f(x^\star)] \leq 2\sqrt{\frac{\sigma^2\Delta_0}{K+1}} + \frac{2L\Delta_0}{K+1}. \tag{16}$$

Proof. According to Lemma 3.4 in Chapter 1, if $f(x)$ is L-smooth, we have

$$\|\nabla f(x_k)\|^2 \leq 2L(f(x_k)-f^\star), \quad \forall k = 0, 1, \cdots. \tag{17}$$

Also, since $f(x)$ is convex, we have

$$f^\star - f(x_k) \geq \langle \nabla f(x_k), x^\star - x_k\rangle, \quad \forall k = 0, 1, \cdots. \tag{18}$$

With the recursion of SGD, we have

$$\begin{aligned}
\mathbb{E}[\|x_{k+1}-x^\star\|^2|\mathcal{F}_k] &= \mathbb{E}[\|x_k - x^\star - \gamma\nabla F(x_k;\xi_k)\|^2|\mathcal{F}_k] \\
&\leq \|x_k - x^\star\|^2 - 2\gamma\mathbb{E}[\langle x_k - x^\star, \nabla F(x_k;\xi_k)\rangle|\mathcal{F}_k] + \gamma^2\mathbb{E}[\|\nabla F(x_k;\xi_k)\|^2|\mathcal{F}_k] \\
&\overset{(a)}{=} \|x_k - x^\star\|^2 - 2\gamma\langle x_k - x^\star, \nabla f(x_k)\rangle + \gamma^2\|\nabla f(x_k)\|^2 + \gamma^2\sigma^2 \\
&\overset{(b)}{\leq} \|x_k - x^\star\|^2 - 2\gamma(f(x_k)-f^\star) + 2\gamma^2 L(f(x_k)-f^\star) + \gamma^2\sigma^2 \\
&= \|x_k - x^\star\|^2 - 2\gamma(1-\gamma L)(f(x_k)-f^\star) + \gamma^2\sigma^2 \\
&\overset{(c)}{\leq} \|x_k - x^\star\|^2 - \gamma(f(x_k)-f^\star) + \gamma^2\sigma^2
\end{aligned} \tag{19}$$

where (a) holds due to Assumption 3.1, (b) holds due to (17) and (19), and (c) holds if we choose $\gamma \leq 1/(2L)$. Taking expectation over the filtration $\mathcal{F}_k$, we have

$$\mathbb{E}[\|x_{k+1}-x^\star\|^2] \leq \mathbb{E}[\|x_k - x^\star\|^2] - \gamma\mathbb{E}[f(x_k)-f(x^\star)] + \gamma^2\sigma^2 \tag{20}$$

which is equivalent to

$$\mathbb{E}[f(x_k)-f(x^\star)] \leq \frac{\mathbb{E}[\|x_k - x^\star\|^2] - \mathbb{E}[\|x_{k+1}-x^\star\|^2]}{\gamma} + \gamma\sigma^2 \tag{21}$$

Taking averaging over $k = 0, 1, \cdots, K$, we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k) - f(x^\star)] \le \frac{\|x_0 - x^\star\|^2}{\gamma(K+1)} + \gamma\sigma^2 \tag{22}$$

Similar to the arguments in (12)–(14), if we choose

$$\gamma = \left[\left(\frac{\Delta_0}{(K+1)\sigma^2}\right)^{-\frac{1}{2}} + 2L\right]^{-1}, \tag{23}$$

SGD will converge as follows

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k) - f(x^\star)] \le 2\sqrt{\frac{\sigma^2\Delta_0}{K+1}} + \frac{2L\Delta_0}{K+1}. \tag{24}$$

□

Remark. If $\sigma^2 = 0$, we recover the rate $O(L/K)$ of GD in convex scenarios.

3.3 Smooth and strongly convex problem

Theorem 3.4. Suppose $f(x)$ is μ-strongly convex and L-smooth. Under Assumption 3.1, if $\gamma \le 1/L$, SGD will converge at the following rate

$$\mathbb{E}[f(x_k)] - f^\star \le (1-\gamma\mu)^k\Delta_0 + \frac{\gamma L\sigma^2}{\mu}. \tag{25}$$

where $\Delta_0 = f(x_0) - f^\star$. If we further choose $\gamma = \min\{\frac{1}{L}, \frac{1}{\mu K}\ln\big(\frac{\mu^2\Delta_0 K}{L\sigma^2}\big)\}$, SGD will converge at the following rate

$$\mathbb{E}[f(x_K)] - f^\star = \tilde{O}\left(\frac{L\sigma^2}{\mu^2 K} + \Delta_0\exp(-\frac{\mu}{L}K)\right) \tag{26}$$

where the $\tilde{O}(\cdot)$ notation hides all logarithm terms.

Proof. Since $f(x)$ is μ-strongly convex, it holds from Theorem 3.8 of our notes "Ch0" that

$$f(y) \le f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2. \tag{27}$$

Let $y = x_k$ and $x = x^\star$, we have

$$\|\nabla f(x_k)\|^2 \ge 2\mu\big(f(x^k) - f(x^\star)\big). \tag{28}$$

Substituting the above inequality to (9), we have

$$\mathbb{E}[f(x_{k+1})] - f^\star \le (1-\gamma\mu)\big(\mathbb{E}[f(x_k)] - f^\star\big) + \frac{L\gamma^2\sigma^2}{2}. \tag{29}$$

Keep iterating the above inequality, we have

$$\mathbb{E}[f(x_K)] - f^\star \leq (1-\gamma\mu)^K (f(x_0) - f^\star) + \frac{\gamma L \sigma^2}{\mu}. \tag{30}$$

We let $\Delta_0 = f(x_0) - f^\star$. With the fact that $(1-x) \leq \exp(-x)$ when $x \in (0,1)$, the above inequality becomes

$$\mathbb{E}[f(x_K)] - f^\star \leq \Delta_0 \exp(-\gamma\mu K) + \frac{\gamma L \sigma^2}{\mu}. \tag{31}$$

Now we let

$$\gamma = \min\{\gamma_1, \frac{1}{L}\} \leq \gamma_1 \quad \text{where} \quad \gamma_1 = \frac{1}{\mu K} \ln\big(\frac{\mu^2 \Delta_0 K}{L\sigma^2}\big). \tag{32}$$

Since $\exp(-\min\{a,b\}) \leq \exp(-a) + \exp(-b)$ for any $a \in \mathbb{R}$ and $b \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}[f(x_K)] - f^\star &\leq \Delta_0 \exp(-\frac{\mu}{L}K) + \Delta_0 \exp(-\gamma_1 \mu K) + \frac{\gamma_1 L \sigma^2}{\mu} \\
&\leq \frac{L\sigma^2}{\mu^2 K}[1 + \ln\big(\frac{\mu^2 \Delta_0 K}{L\sigma^2}\big)] + \Delta_0 \exp(-\frac{\mu}{L}K) \\
&= \tilde{O}\left(\frac{L\sigma^2}{\mu^2 K} + \Delta_0 \exp(-\frac{\mu}{L}K)\right)
\end{aligned} \tag{33}$$

where we hide the logarithm term in the $\tilde{O}(\cdot)$ notation.

□

4 Mini-batch stochastic gradient descent

When training deep neural network, it is common to sample a batch of data to estimate the true gradient. SGD with mini-batch will iterate as follows:

$$g_k = \frac{1}{B} \sum_{b=1}^{B} \nabla F(x_k; \xi_k^{(b)}), \tag{34a}$$

$$x_{k+1} = x_k - \gamma g_k. \tag{34b}$$

where B is the batch size.

4.1 Mini-batch SGD suffers smaller variance

The following lemma establishes that mini-batch samples enable SGD to have a more accurate gradient estimate. To state the lemma, we first introduce

$$\mathcal{F}_k^B = \{x_k, \{\xi_{k-1}^{(b)}\}_{b=1}^B, x_{k-1}, \{\xi_{k-2}^{(b)}\}_{b=1}^B, \cdots, x_0\} \tag{35}$$

Assumption 4.1. Given the filtration $\mathcal{F}_k^B$, we assume

$$\mathbb{E}[\nabla F(x_k;\xi_k^{(b)})|\mathcal{F}_k^B] = \nabla f(x_k), \tag{36}$$

$$\mathbb{E}[\|\nabla F(x_k;\xi_k^{(b)}) - \nabla f(x_k)\|^2|\mathcal{F}_k^B] \le \sigma^2. \tag{37}$$

Moreover, we assume $\{\xi_k^{(b)}\}_{b=1}^B$ are independent of each other for any $k = 0, 1, \cdots$.

With the above assumption, it is easy to verify that

$$\mathbb{E}[g_k|\mathcal{F}_k^B] = \frac{1}{B}\sum_{b=1}^B \mathbb{E}[\nabla F(x_k;\xi_k^{(b)})|\mathcal{F}_k] = \nabla f(x_k), \tag{38}$$

and the variance is derived as

$$\mathbb{E}[\|g_k - \nabla f(x_k)\|^2|\mathcal{F}_k^B] \overset{(a)}{=} \frac{1}{B^2}\sum_{b=1}^B \mathbb{E}[\|\nabla F(x_k;\xi_k^{(b)}) - \nabla f(x_k)\|^2|\mathcal{F}_k^B] \overset{(b)}{\le} \frac{\sigma^2}{B} \tag{39}$$

where the first equality (a) holds due to the independence between $\{\xi_k^{(b)}\}_{b=1}^B$ and Eq.(36), and (b) holds due to Eq.(37). According to (39), the variance of the mini-batch stochatic gradient g_k is B-times smaller than single-sample stochastic gradient $\nabla F(x;\xi)$.

4.2 Convergence of mini-batch SGD

The convergence analysis of mini-batch SGD is almost the same as vanilla SGD, except that the variance of stochastic gradient g_k becomes σ^2/B.

Theorem 4.2. Under Assumption 4.1, mini-batch SGD will converge as follows

- If $f(x)$ is L-smooth, mini-batch SGD converge as follows

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(x_k)\|^2] = O\left(\sqrt{\frac{L\Delta_0^f\sigma^2}{B(K+1)}} + \frac{L\Delta_0^f}{K+1}\right). \tag{40}$$

  where $\Delta_0^f = f(x_0) - f^\star$.

- If $f(x)$ is L-smooth and convex, mini-batch SGD converge as follows

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[f(x_k) - f^\star] = O\left(\sqrt{\frac{L\Delta_0^x\sigma^2}{B(K+1)}} + \frac{L\Delta_0^x}{K+1}\right). \tag{41}$$

  where $\Delta_0^x = \|x_0 - x^\star\|^2$.

- If $f(x)$ is L-smooth and μ-strongly convex, mini-batch SGD converges as follows

$$\mathbb{E}[f(x_K)] - f^\star = \tilde{O}\left(\frac{L\sigma^2}{\mu^2 BK} + \Delta_0^f \exp(-\frac{\mu}{L}K)\right) \tag{42}$$

  where $\Delta_0^f = f(x_0) - f^\star$ and $\tilde{O}(\cdot)$ hides all logarithm terms.

5 Experiments

5.1 Linear regression

- We use different learning rates to implement stochastic gradient descent for linear regression task, where the loss function is $\ell(X\theta; y) = \frac{1}{2}(X\theta - y)^2$. We randomly generated a dataset $X(N = 100, d = 5)$ and a label set $y(N = 100, d = 1)$, and set the initial parameter θ. In the experiment, we set the number of epochs to 10 and the batch-size to 1. We used two different learning rates, varying from low to high, and the results are shown in Figure 1.

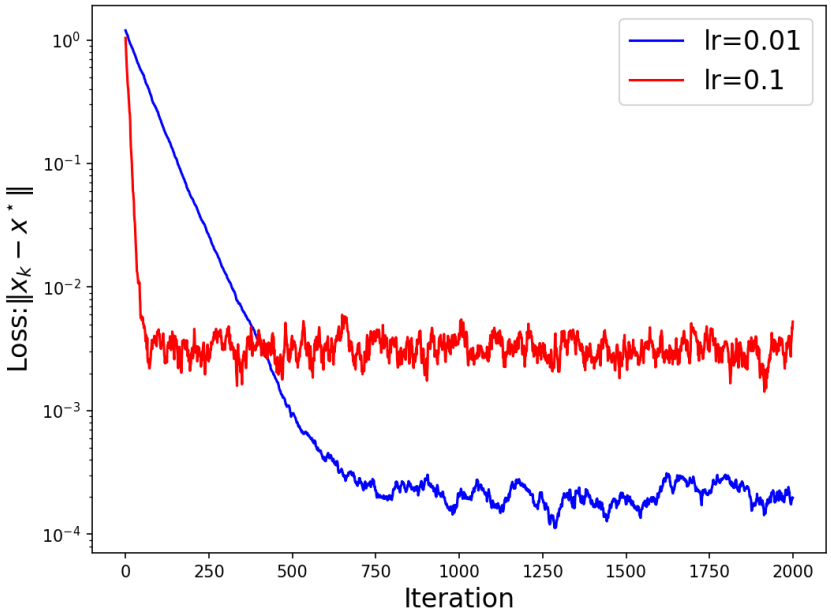

Figure 1: The loss descent plots for stochastic gradient descent with different learning rates.

We can observe that a smaller learning rate can lead to more cautious and stable parameter updates, and the final convergence results are better, but the training speed is slower. Properly increasing the learning rate can accelerate training speed, but it will sacrifice a certain convergence effect. However, excessive learning rates may lead to unstable training, causing oscillations during training, and may cause the model to fall into local minima or result in divergence. Therefore, we usually use the decaying learning rate in our daily life, which converges quickly in the early stage and ensures a good convergence result in the later stage.

- We set decaying learning rate:$\gamma_k = \gamma_0^{0.2k+1}$ and $\gamma_k = \gamma_0/(0.2k+1)$,where k represents epochs. The results of different learning rate decaying strategies are also different. The experimental results are shown in Figure 2.

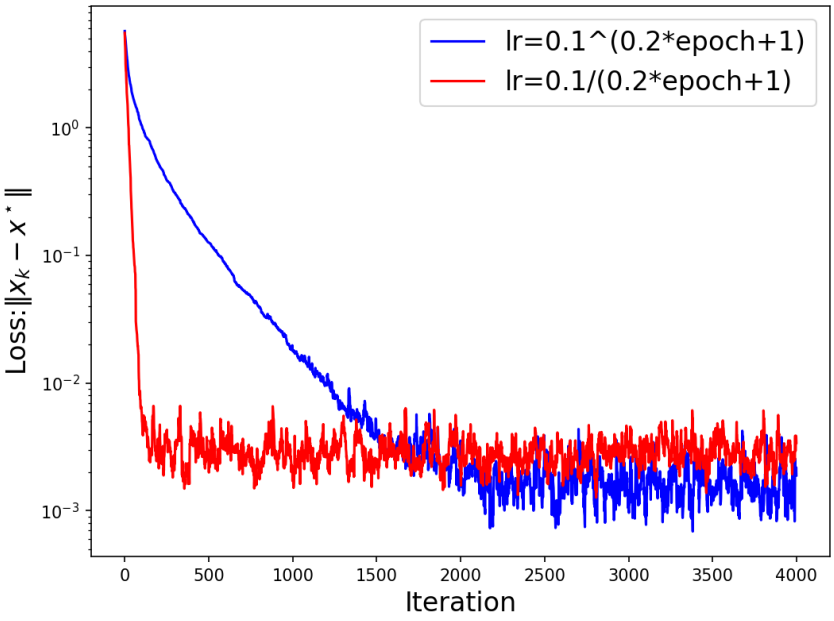

Figure 2: Loss curves trained using two learning rate reduction strategies in the SGD algorithm.

- In the previous task, we set the learning rate to 0.05, epochs to 10, and conducted stochastic gradient descent experiments with different batch-size, as shown in Figure 3.

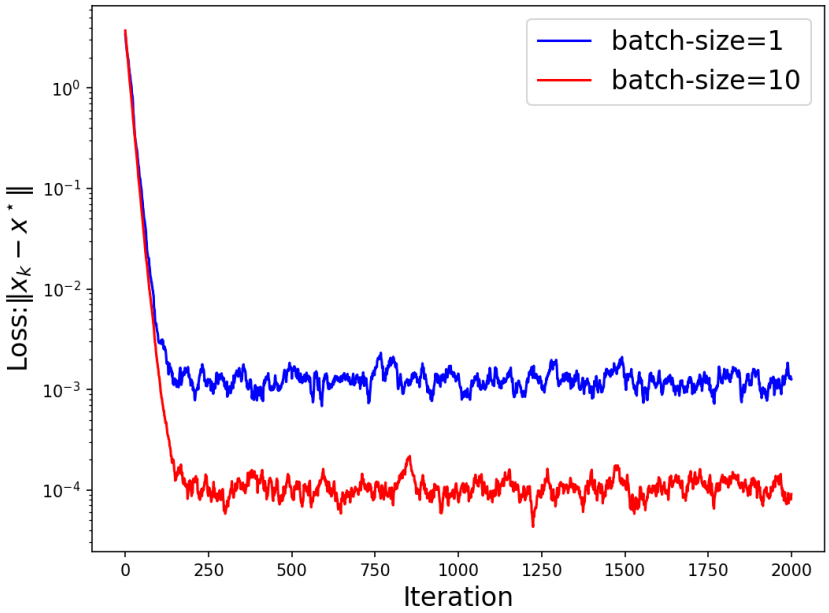

Figure 3: The loss descent plots for stochastic gradient descent with different batch sizes.

We can observe that when the batch size is larger, the convergence speed is slightly

faster and ultimately better convergence results can be obtained, because the more data used for each parameter update, the better it represents the gradient of the overall loss function, resulting in higher gradient accuracy.

5.2 Image classification

- We conducted training on the CIFAR-10 dataset using the ResNet-18 architecture. Now, we choose SGD as the optimizer and we are investigating the impact of different batch sizes on the results. We performed a comparative experiment by setting the batch size to 16 and 128, while keeping the learning rate constant at 0.005. We trained for nearly 1200 steps and recorded the loss during the training process. And the loss in the training process is shown as the Figure 4.

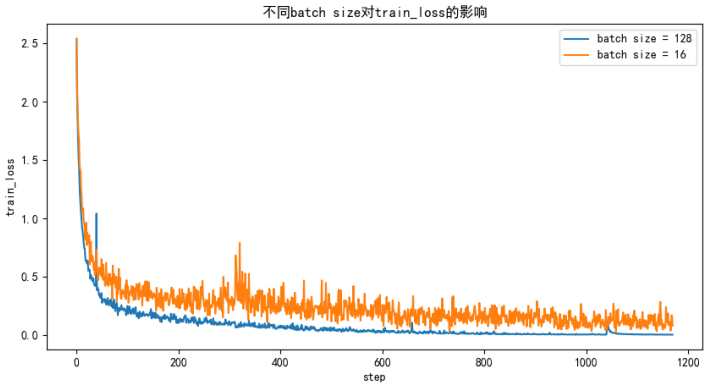

Figure 4: Loss curves trained using two batch size in the CIFAR-10 classification experiment.

We can observe that increasing the batch size can improve the convergence speed and effectiveness of the model with a fixed number of iterations. This is because with a larger batch size, the information each batch captured is closer to the true distribution, resulting in more accurate gradient calculations. As a result, the model can obtain better convergence effectiveness.