Appendix-Adaptive Curriculum Learning

In the appendix, we will introduce experiments and theoretical analysis following the structure:

- **A Experiments**
  - **A.1 Experiments settings**
    In this part, we will introduce the details of experiments settings, including the datasets, networks and hyperparameters settings.
  - **A.2 Supplementary experimental results**
    The supplementary experimental results include:
    1) The difficulty scores of examples;
    2) Comparison of different pacing functions;
    3) The validation accuracy curve of Adaptive CL and other curriculum learning methods;
    4) Comparison of the running time.
    5) Comparison of Adaptive CL and FCL on three additional datasets;
- **B Theoretical Analysis**

  In this section, we provide the proof of Theorems 1 and 2, respectively. We arrange the contents as follow:

  1) Definitions and notations;
  2) Proof of Theorem 1;
  3) Proof of Theorem 2;

A. Experiments

A.1. Experiments settings

A.1.1 Datasets and network architectures

In this subsection, we introduce the datasets and network architectures in detail. In the experiments, we use five kinds of datasets: two-class datasets, small ImageNet, the superclasses of CIFAR-100, CIFAR-10, and CIFAR-100. The two-class datasets are small datasets containing two classes, each of which has 250 training examples and 50 test examples selected from ImageNet [6]. The two-class datasets include three datasets: a) "Airplane" & "Car"; b) "Dog" & "Cat"; c) "Elephant" & "Cat". The small Imagenet is a subset from the ImageNet dataset ILSVRC 2012 [6] and contains five randomly selected classes: "Cat", "Dog", "Airplane", "Elephant", and "Car". Each class has 250 training images and 50 test images. The images are resized to 56×56 color images for faster performance. In the experiments, we use six superclasses of CIFAR-100 in total. Each superclass contains five fine classes of CIFAR-100. We show the superclasses and their corresponding five fine classes in Table 3.

Superclass	Classes
"Aquatic Mammals"	"beaver", "dolphin", "otter", "seal", "whale"
"Flowers"	"orchids", "poppies", "roses", "sunflowers", "tulips"
"Household Electrical Devices"	"clock", "computer keyboard", "lamp", "telephone', "television"
"Large Natural Outdoor Scenes"	"cloud", "forest", "mountain", "plain", "sea"
"Medium-Sized Mammals"	"fox", "porcupine", "possum", "raccoon, skunk"
"Small Mammals"	"hamster", "mouse", "rabbit", "shrew", "squirrel"

Table 3. The superclasses and their corresponding five fine classes we used in our experiments.

For a fair comparison, the networks we use in the paper cover the networks used in [11]: the hand-crafted network and the VGG-16 [37]. The hand-crafted network contains 8 convolutional layers, which have 32, 32, 64, 64, 128, 128, 256, and 256 filters, respectively. The kernel size of the first six convolutional layers is 3×3, and the kernel size of the last two layers is 2×2. There are a 2×2 maxpool layer and a 0.25 dropout layer every 2 convolutional layers . The layers after

the convolutional layers are a flatten layer, two fully connected layers, and a softmax layer. For the experiments of binary classification, we use an MLP network containing two fully connected layers. The hidden unit is 20. Following the first fully connected layer is an elu activation layer, and the second is a softmax layer. Moreover, we also use ResNet-v1-14 [12] and ResNet18 [12].

A.1.2 Settings of Hyperparameters

Unless specified otherwise, for a fair comparison, the hyperparameters such as learning rate, batch size, pacing function, dropout rate are kept unchanged. We tune the method-specific hyperparameters for each method. In the experiments, we use the optimizer of SGD and the batch size of all experiments is 100.

learning rate: For fairness, we use the same learning rate settings as [11], which contains three parameters: (i) the initial learning rate; (ii) the factor by which the learning rate is decreased; (iii) the batch number with constant learning rate. The settings of learning rate in different experiments are as follow: MLP network trained on two-class datasets:(i) 0.005; (ii) 3; (iii) 40; hand-crafted network trained on the small ImageNet or the superclasses: (i) 0.03; (ii) 1.1; (iii) 100; hand-crafted network trained on the CIFAR-10: (i) 0.12; (ii) 1.1; (iii) 700; hand-crafted network trained on the CIFAR-100: (i) 0.12; (ii) 1.1; (iii) 400; ResNet-v1-14 trained on the CIFAR-10: (i) 0.12; (ii) 1.1; (iii) 1000. For the experiments on VGG and ResNet18, the learning rate parameters are left as publicly determined.

Paramenters α, λ, and inv: We tune the parameters by grid search. The range of $\alpha : -0.03 \sim 0.0$, $\lambda : 0.0 \sim 0.02$. α and λ can be constants or functions of batch number. All of the experiments begin the update of difficulty score after the iteration of 150 while MLP is 2, Resnet_v1_14 is 200, VGG and Resnet18 are 500. After the iteration, the model begins update the difficulty score every inv iterations. Except for the experiments of two-class datasets where we set $inv = 4$, the inv of all other experiments is 50.

Initial difficulty score: In the experiments, the initial difficulty score is determined by the confidence score. We use *sklearn.svm.libsvm.predict proba* to obtain the confidence score of the SVM classifier. For the VGG-based initial difficulty score, we first obtain the output probabilities of the five classes of the superclass from the pre-trained VGG-16 model, which is pretrained on the CIFAR-100. Then we divide the sum of the output probabilities of the five classes to make the sum of the output probabilities be 1. The confidence score of each image is obtained from the output probability of its corresponding label.

Pacing function: Pacing function determines the datasize of the sample pool from which the mini-batch will be randomly sampled. In the paper, we use exponential functions as our pacing functions:

$$p(i) = n \times min(p_0 \times q^{\lfloor i/r_0 \rfloor}, 1), \tag{4}$$

where $\lfloor \cdot \rfloor$ denotes rounding down, n is the total number of data, p_0 is the fraction of the data in the initial step, q and r_0 are the exponential factors that are used to control the increasing speed of pacing functions. The illustration of the pacing functions are shown in Figure 7[4].

We set $r_0 = 100$ for all experiments except the experiments on two-class datasets where we set $r_0 = 2$. The settings of other parameters of pacing function are as follow: (a) hand-crafted network trained on the small ImageNet or the superclasses: $p_0 : 0.04$; $q : 1.9$; (b) hand-crafted network trained on CIFAR-10 and CIFAR-100: $p_0 : 0.04$; $q : 1.2$; (c) ResNet_v1_14: $p_0 : 0.04$; $q : 1.3$; (d) VGG network trained on CIFAR-100 and ResNet18 trained on CIFAR-10: $p_0 : 0.04$; $q : 1.1$; (e) MLP network on two-class datasets: $p_0 : 0.8$; $q : 1.1$.

A.2. Supplementary experimental results

A.2.1 The difficulty scores of examples

In this part, we want to show that α will affect the difficulty score during training, and the effect of α on different examples is different. Figure 8 shows different representations of the difficulty score. All of the experiments are trained on the superclass of "Small Mammals" and α is set to $0.01, 0.5$, and 0.03, respectively. We set inv=50 and train the dataset on the hand-crafted network. We use positive α just because it's more convenient to illustrate. In Figure 8(a) and 8(b) we show the consecutive 17 scores which are sampled every fifty iterations. According to the figures, when α is small, the scores change, but the changing ranges are small and vice-versa. The results indicate that the magnitude of α has effects on the change of the difficult score during training. Figure 8(a) and 8(b) also show that the score ranges of different examples with different α are different. For example, the scores of example 5 and example 7 almost keep unchanged both in the cases of $\alpha = 0.01$ and $\alpha = 0.5$, but

[4]The figure refers to [11].

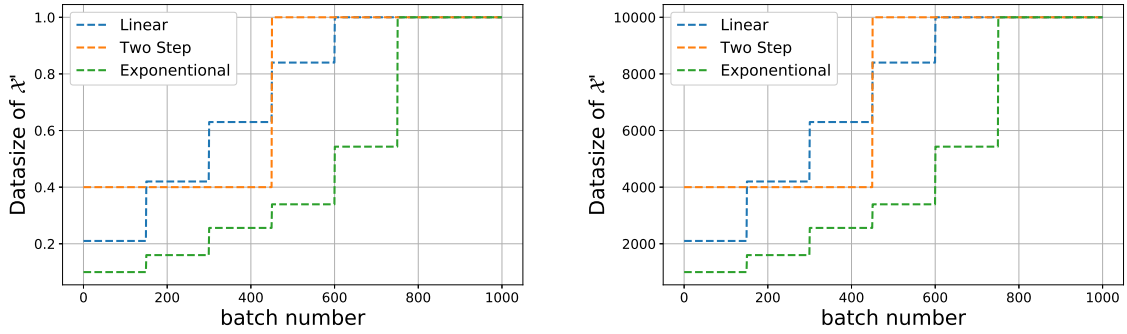
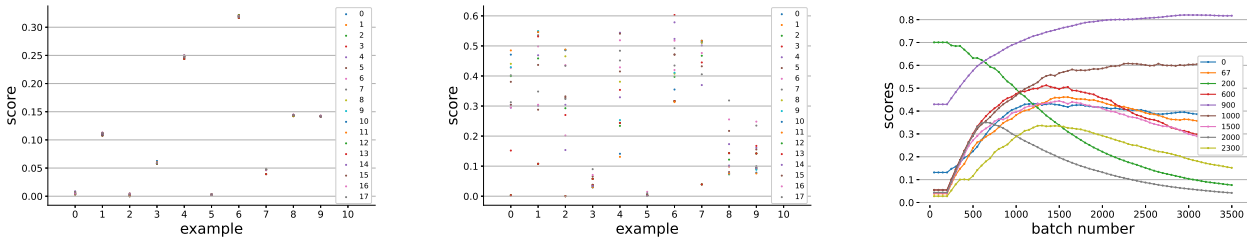


Figure 7. The illustration of three kinds of pacing functions. The exponential pacing function is the pacing function we used in the paper and the representation is given in Eq. (4). Left: The fraction of the data in the sample pooling. Right: The datasize of the sample pooling $\mathcal{X}'$ during training. In the figure, we set the full datasize $n = 10000$ and the total iterations $M = 1000$. The values were chosen arbitrarily, for illustration only.



(a) Difficulty score with α=0.01 (b) Difficulty score with α=0.5 (c) Difficulty score with α = 0.03

Figure 8. Different representations of difficulty score during training. The experiments train the hand-crafted network on the superclass of "Small Mammals". Left: Difficulty scores of the first ten examples with α=0.01. Middle: Difficulty scores of the first ten examples with α=0.5. We show the consecutive 17 scores which are sampled every fifty iterations. The x-axis is the index of examples, the y-axis is their corresponding difficulty scores, and the numbers in the legend present the index of the 17 stages. Right: The difficulty score changing with iterations when α = 0.03. The y-axis is the difficulty score, the x-axis is the iterations, and the numbers in legend are the indexes of selected examples.

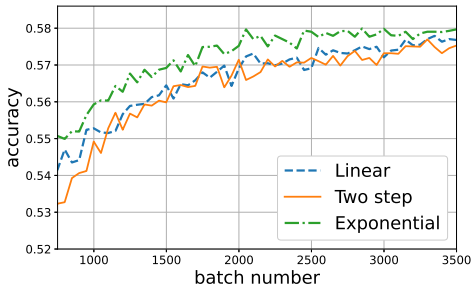
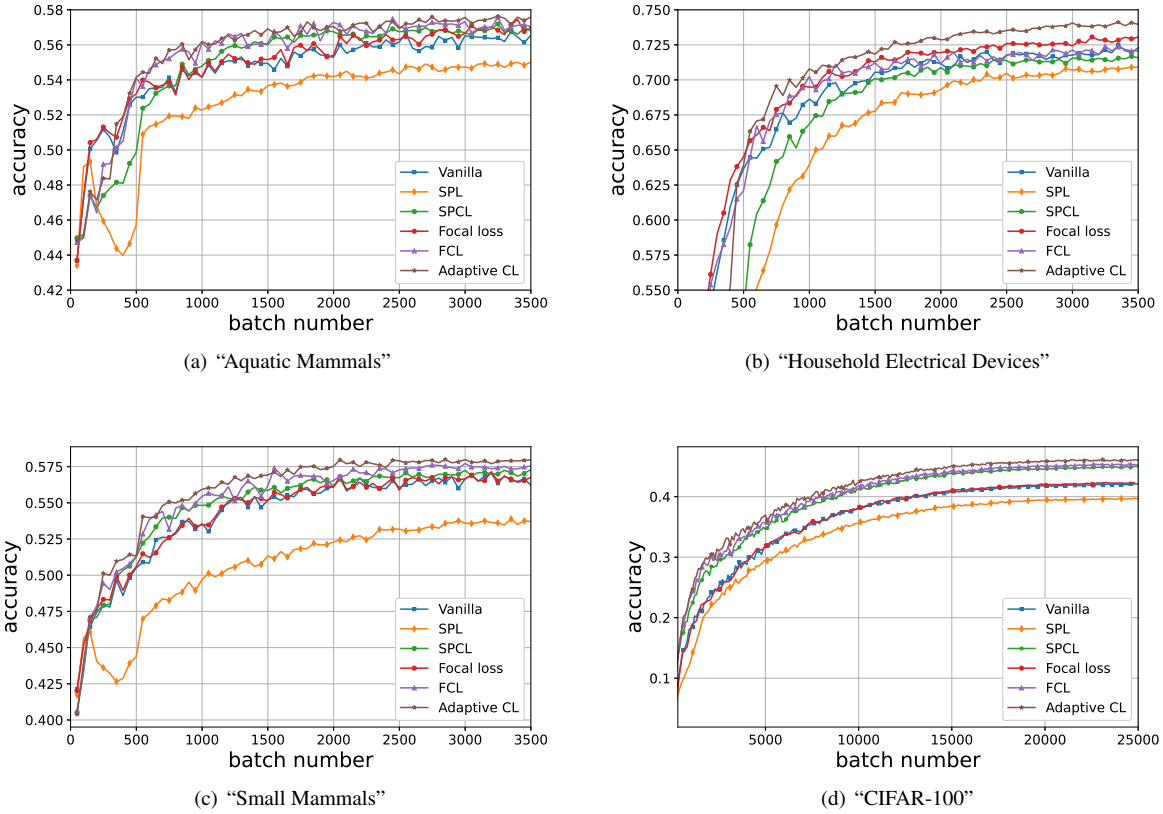


Figure 9. The average validation accuracy curve of Adaptive CL using different kinds of pacing functions. The dataset is Small Mammals, and the network is the hand-crafted network. The repetition of experiments is 25.

example 0's score changes a lot in the case of $\alpha = 0.5$, compared to the case of $\alpha = 0.01$. The results indicate that the initial scores of some examples are proper to the current task, while others are not. Thus the aim of our method is to adapt the score of each example to the current task according to the learning progress of each example. Figure 8(c) shows the scores changing curves of different examples with iterations, demonstrating that the difficulty scores are dynamic during training.



(a) "Aquatic Mammals" (b) "Household Electrical Devices"

(c) "Small Mammals" (d) "CIFAR-100"

Figure 10. The average accuracy of the proposed method and other methods. The datasets are "Aquatic Mammals", "Household Electrical Devices", "Small Mammals", and "CIFAR-100", respectively. The network is the hand-crafted network. The y-axis is the validation accuracy, and the x-axis is the batch number.

A.2.2 Comparison of pacing functions

In this part, we explore the effect three kinds of pacing functions: exponential function, linear function, and two-step function. The figure illustration of the three pacing functions are shown in Figure 7[5]. The formula of the exponential pacing function is Eq. (4) and we set $p_0 = 0.04$, $q = 1.9$ and $r_0 = 100$ in the experiment. The linear pacing function is a function linear to iterations: $p(i) = n \times min(p_0 + (1.0 - p_0) * i/c, 1)$, where p_0 is the fraction of the data in the initial step, c is the linear coefficient. We set $p_0 = 0.04$, $c = 600$, and update the datasize of the sample pool according to the function every 100 iterations. A network using two-step pacing function means that the network will first train on a sub-dataset, which contains certain numbers of easier examples. After certain iterations, the network will train on the full dataset until the network converges. The formula of the two step function is: $p(i) = n \times p_0^{\mathbf{1}_{[i<len]}}$. In the experiment, we set $p_0 = 0.6$ and $len = 600$. From Figure 9 shows that the exponential function performs better, which is coincide with the phenomenon we observe in daily life, that a student could learn harder concepts quickly if he or she has a solid foundation during the early stages of learning.

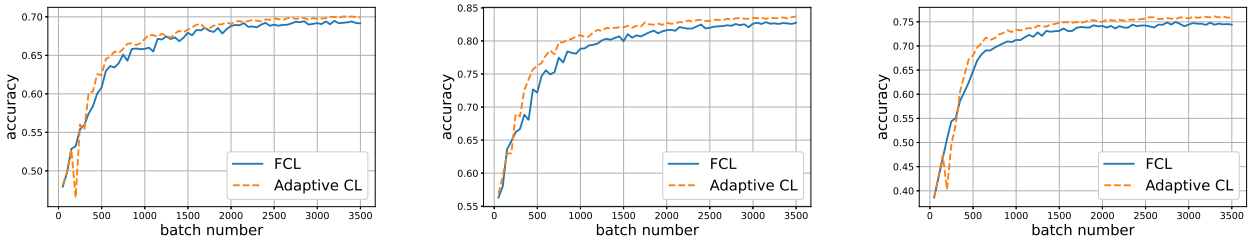
A.2.3 The validation accuracy curve of Adaptive CL and baselines

Figure 10 shows the average accuracy curve of the proposed method and other methods during training. The datasets are "Aquatic Mammals" , "Household Electrical Devices", "Small Mammals" and "CIFAR-100", respectively. The repetition of experiments is 6 in (d), and 25 in (a), (b), and (c). The network is the hand-crafted network. The y-axis is the validation

[5]The figure refers to [11].

Table 4. Comparison of running time. We train the hand-crafted network on the superclass HE devices with a single GeForce GTX 1080 Ti GPU. We train the dataset 3500 batches (140 epochs) and set $inv = 50$. The results are average time and STD over five runs.

Method	Vanilla	SPL	SPCL	FCL	Focal loss	Ours
Time (s)	53.252 ($\pm$0.906)	54.129 ($\pm$2.114)	56.147 ($\pm$1.725)	54.485 ($\pm$0.975)	53.789 ($\pm$1.511)	55.497 ($\pm$0.886)



(a) "Flowers" (b) "Large Natural Outdoor Scenes" (c) "Medium-Sized Mammals"

Figure 11. Results of the three superclasses: "Flowers", "Large Natural Outdoor scenes", and "Medium-Sized Mammals" , respectively. The network is the hand-crafted network. The y-axis is the validation accuracy, and the x-axis is the batch number. The repetition of experiments is 25.

accuracy, and the x-axis is the batch number. Results in Figure 10 show that the proposed method performs better than other competitive methods.

A.2.4 Comparison of the running time

We also compare the running time of the proposed method and baselines. The results in Table 4 show that the running time of the proposed algorithm is comparable to that of other curriculum learning approaches.

A.2.5 Comparison of Adaptive CL and FCL on three additional datasets

In addition to the datasets used in the paper, we perform the proposed method on three additional datasets: superclasses of "Flowers", "Large Natural Outdoor Scenes", and "Medium-Sized Mammals", respectively. We compare Adaptive CL to FCL, and the results are shown in Figure 11. The results further validate that Adaptive CL performs better than FCL.

B. Theoretical Analysis

B.1. Definitions and notations

Let $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^n$ be the training data, where $x_i \in \mathbb{R}^d$ is the i-th data point and y_i is its corresponding label. Let ψ be the ideal difficulty score and we define the ideal difficulty score as

$$\psi = \left| g\left(\bar{\mathbf{w}}^\top \boldsymbol{x}_i\right) - y_i \right|, \tag{5}$$

where $g(\cdot)$ is the activation function, $\boldsymbol{x_i} = \begin{bmatrix} x_i^\top & 1 \end{bmatrix}^\top$ [6], and $\bar{\mathbf{w}}$ is the global solution of the empirical loss. In an analogous way, let γ be the current difficulty score and we define the current difficulty score at time t as

$$\gamma = \left| g\left(\mathbf{w}_t^\top \boldsymbol{x}_i\right) - y_i \right|, \tag{6}$$

where $\mathbf{w}_t$ is the model parameter at time t.

B.2. Convergence rate with an ideal difficulty score

In this subsection, we analyze the relationship between the expected convergence rate and the ideal difficulty score. We prove that under mild assumptions, the expected convergence rate monotonically decreases with respect to the ideal difficulty score in the early stages of learning.

[6] $\mathbf{w}^\top \boldsymbol{x}_i = w^\top x_i + b$, where $\mathbf{w} = \begin{bmatrix} w^\top & b \end{bmatrix}^\top$. Therefore, $\boldsymbol{x}_i = \begin{bmatrix} x_i^\top & 1 \end{bmatrix}^\top$.

Curriculum Learning trains the model on a sequence of training points $X_t = \{\boldsymbol{x}_{i_t}, y_{i_t}\}, t \in [T]$, sampled from the training dataset. As we analyze a non-linear model with the least squares loss, the loss function can be represented as $L(X_t, \mathbf{w}) = (g(\mathbf{w}^\top \boldsymbol{x}_{i_t}) - y_{i_t})^2$. Following the stochastic gradient descent rule, the update step at time t:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial L(X_t, \mathbf{w})}{\partial \mathbf{w}}|_{\mathbf{w}=\mathbf{w}_t}, \tag{7}$$

where η is the learning rate.

We denote $\mathbf{r}$ as the gradient step at time t. Then, $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{r}$ and from (7), we obtain

$$\mathbf{r} = -2\eta \boldsymbol{x}_{i_t} \nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}) \left(g\left(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}\right) - y_{i_t}\right). \tag{8}$$

Let $\Delta(\psi)$ be the expected convergence rate at time t. Given difficulty score ψ, $\Delta(\psi)$ can be defined as

$$\Delta(\psi) = \mathbb{E}\left[\|\mathbf{w}_t - \bar{\mathbf{w}}\|^2 - \|\mathbf{w}_{t+1} - \bar{\mathbf{w}}\|^2 |_\psi\right]. \tag{9}$$

For convenience, we introduce the notation $\beta = \|\mathbf{w}_t - \bar{\mathbf{w}}\|$.

Lemma 1. *Let r_o be the projection of the gradient vector $\boldsymbol{r}$ on the vector $\bar{\boldsymbol{w}} - \boldsymbol{w}_t$, r be the length of $\boldsymbol{r}$, and $\beta = \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|$. Given the ideal difficulty score ψ, the expected convergence rate $\Delta(\psi)$ could be represented as*

$$\Delta(\psi) = 2\beta \mathbb{E}\left[r_o|_\psi\right] - \mathbb{E}\left[r^2|_\psi\right]. \tag{10}$$

Proof. From (8) and (9) we obtain

$$\begin{aligned}
\mathbb{E}[\Delta] &= \beta^2 - \mathbb{E}[(\beta - r_o)^2 + r_\perp^2] \\
&= \beta^2 - (\beta^2 - 2\beta \mathbb{E}[r_o] + \mathbb{E}[r_o^2] + \mathbb{E}[r_\perp^2]) \\
&= 2\beta \mathbb{E}\left[r_o\right] - \mathbb{E}\left[r^2\right].
\end{aligned}$$

□

Such a form of the expected convergence rate can facilitate the proof of the following theorem.

Theorem. ***1.*** *Let $h(\boldsymbol{w}) = g\left(\boldsymbol{w}^\top \boldsymbol{x}\right)$ be a non-linear model whose label belongs to $\{-1, 1\}$, $\{(\boldsymbol{x}_{i_t}, y_{i_t})\}$ is the sampled data point at time t, where $i_t \in [n]$ and $t \in [T]$. $g(\cdot)$ is the tanh activation function: $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. Let $\gamma = |g(\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t}) - y_{i_t}|$ be the current difficulty score and $\psi = |g(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_{i_t}) - y_{i_t}|$ be the ideal difficulty score. $\boldsymbol{w}_t$ is the model parameter at time t, and $\bar{\boldsymbol{w}}$ is the optimal model parameter. In the early stages of learning, the expected convergence rate at time t monotonically decreases with respect to the ideal difficulty score ψ of $\boldsymbol{x}_{i_t}$, i.e., $\frac{\partial \Delta(\psi)}{\partial \psi} \leq 0$, under the assumptions that*

1) the current example, i.e., the "easy" example, could be correctly labeled;

2) when $y_{i_t} = -1$, $g(\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t}) - g(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_{i_t}) \geq 0.2$;

3) when $y_{i_t} = 1$, $g(\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t}) - g(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_{i_t}) \leq -0.2$.

Proof. To prove that the expected convergence rate monotonically decreases with respect to the ideal difficulty score ψ, we need to prove $\frac{\partial \Delta(\psi)}{\partial \psi} \leq 0$. From (10), we can derive that

$$\frac{\partial \Delta(\psi)}{\partial \psi} = \frac{\partial 2\beta \mathbb{E}[r_o|_\psi]}{\partial \psi} + \frac{\partial - \mathbb{E}[r^2|_\psi]}{\partial \psi}. \tag{11}$$

We prove the theorem in two cases:

1. $-\frac{\partial \mathbb{E}[r^2|_\psi]}{\partial \psi} \leq 0$;

2. $\frac{\partial 2\beta\mathbb{E}[r_o|_\psi]}{\partial\psi} \leq 0$.

Different from [46], which assumed that given ideal difficulty score ψ the probability of $y_{i_t} = -1$ and $y_{i_t} = 1$ is equal, i.e., $\frac{1}{2}$. We will prove that the theorem holds for both $y_{i_t} = -1$ and $y_{i_t} = 1$.

Proof of step 1:

In this part, we prove step 1. We will prove $-\frac{\partial\mathbb{E}[r^2|_\psi]}{\partial\psi} \leq 0$ holds for the cases of $y_{i_t} = -1$ and $y_{i_t} = 1$, respectively. We first compute r^2. From (5) and (8), we can derive that

$$\begin{aligned}\mathbf{r} &= -2\eta\boldsymbol{x}_{i_t}\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t})(g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}) - y_{i_t}) \\ &= -2\eta\boldsymbol{x}_{i_t}\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t})\left(g\left(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}\right) - g\left(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}\right) + g\left(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}\right) - y_{i_t}\right) \\ &= -2\eta\boldsymbol{x}_{i_t}\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t})\left(m \pm \psi\right),\end{aligned}$$

where $m = g\left(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}\right) - g\left(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}\right)$.

So $r^2 = 4\eta^2\|\boldsymbol{x}_{i_t}\|^2(\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}))^2(m^2 \pm 2m\psi + \psi^2)$. From this, we can derive that

$$\begin{aligned}\frac{\partial\mathbb{E}\left[r^2|_\psi\right]}{\partial\psi} &= \frac{\partial\mathbb{E}[4\eta^2(\pm 2m\psi + \psi^2)\|\boldsymbol{x}_{i_t}\|^2(\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}))^2]}{\partial\psi} \\ &= \mathbb{E}[8\eta^2(\pm m + \psi)(\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}))^2\left\|\boldsymbol{x}_{i_t}\right\|^2].\end{aligned} \tag{12}$$

Now we prove that $\frac{-\partial\mathbb{E}[r^2|_\psi]}{\partial\psi} \leq 0$ holds for both $y_{i_t} = -1$ and $y_{i_t} = 1$.

- When $y_{i_t} = -1$, $\psi = g(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}) - y_{i_t}$ and $m + \psi = g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}) - y_{i_t} \geq 0$. From (12) we obtain

$$\frac{\partial\mathbb{E}\left[r^2|_\psi\right]}{\partial\psi}|_{y_{i_t}=-1} = \mathbb{E}[8\eta^2(m + \psi)(\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}))^2\left\|\boldsymbol{x}_{i_t}\right\|^2] \geq 0. \tag{13}$$

- When $y_{i_t} = 1$, similarly, $-\psi = g(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}) - y_{i_t}$ and $-m + \psi = y_{i_t} - g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}) \geq 0$. From (12) we have

$$\frac{\partial\mathbb{E}\left[r^2|_\psi\right]}{\partial\psi}|_{y_{i_t}=1} = \mathbb{E}[8\eta^2(-m + \psi)(\nabla g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}))^2\left\|\boldsymbol{x}_{i_t}\right\|^2] \geq 0. \tag{14}$$

From (13) and (14), we can conclude that $-\frac{\partial\mathbb{E}[r^2|_\psi]}{\partial\psi} \leq 0$ holds.

Proof of Step 2:

Now we prove that $\frac{\partial 2\beta\mathbb{E}[r_o|_\psi]}{\partial\psi} \leq 0$ holds under the assumptions:

1. the current example, i.e., the "easy" example, could be correctly labeled;
2. when $y_{i_t} = -1$, $m = g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}) - g(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}) \geq 0.2$;
3. when $y_{i_t} = 1$, $m = g(\mathbf{w}_t^\top\boldsymbol{x}_{i_t}) - g(\bar{\mathbf{w}}^\top\boldsymbol{x}_{i_t}) \leq -0.2$.

First we introduce some functions we will use later. $g(\cdot)$ is the tanh activation function. For $z \in \mathbb{R}$,

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \in (-1, +1).$$

It is easy to verify that the derivative of tanh function is positive, i.e., $\nabla g(z) > 0, z \in \mathbb{R}$. Let $f(\cdot) = g^{-1}(\cdot)$ be the inverse function of tanh, then

$$f(x) = \frac{1}{2}ln\left(\frac{1+x}{1-x}\right). \tag{15}$$

$\nabla f(x)$ is the derivative of $f(\cdot)$:

$$\nabla f(x) = \frac{1}{1-x^2}, x \in (-1, 1) \tag{16}$$

Following, we will prove $\frac{\partial 2\beta\mathbb{E}[r_o|_\psi]}{\partial\psi} \leq 0$ holds for both $y_{i_t} = -1$ and $y_{i_t} = 1$.

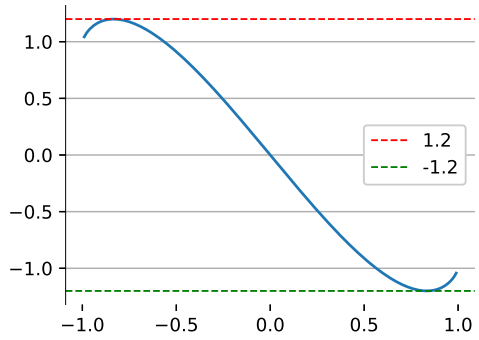


Figure 12. The curve of $\frac{1-v^2}{2} ln\left(\frac{1-v}{1+v}\right) - v, v \in (-1, 1)$.

- When $y_{i_t} = -1$, from (5) and (6) we can derive that:

$$\begin{aligned} \bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} &= f\left(\psi - 1\right); \\ \mathbf{w}_t^\top \boldsymbol{x}_{i_t} &= f\left(\gamma - 1\right). \end{aligned} \tag{17}$$

From (11), (17), and the definition of r_o, we can derive that

$$\begin{aligned} r_o &= \frac{\bar{\mathbf{w}}^\top - \mathbf{w}_t^\top}{\beta}\mathbf{r} \\ &= -\frac{2\eta}{\beta}\left(m + \psi\right)\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})\left(\bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} - \mathbf{w}_t^\top \boldsymbol{x}_{\boldsymbol{i_t}}\right) \\ &= -\frac{2\eta}{\beta}\left(m + \psi\right)\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})\left(f\left(\psi - 1\right) - f\left(\gamma - 1\right)\right). \end{aligned}$$

So

$$\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \psi}|_{y_{i_t}=-1} = -4\eta\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})((m+\psi)\nabla f(\psi-1) + f(\psi-1) - f(\gamma-1))].$$

Because $\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{\boldsymbol{i_t}}) > 0$, to prove $\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \psi}|_{y_{i_t}=-1} \leq 0$, we only need to prove $(m+\psi)\nabla f(\psi-1) + f(\psi-1) - f(\gamma-1) \geq 0$. Under the first assumption that the current example, i.e., the "easy" example, could be correctly labeled. In other words, $\gamma - 1 < 0$. From (15) we can derive that $f(\gamma-1) < 0$. So we only need to prove $(m+\psi)\nabla f(\psi-1) + f(\psi-1) \geq 0$. Let $v = \psi - 1, v \in (-1, 1)$, from (15) and (16), it is equal to prove

$$m + 1 \geq -f(v)/\nabla f(v) - v = \frac{1-v^2}{2} ln\left(\frac{1-v}{1+v}\right) - v, v \in (-1, 1). \tag{18}$$

The blue line of Figure 12 is the function of the right side of the inequality. From the plot, we can see that the max value of the right side of the inequality is than 1.2. So when $m \geq 0.2$, the inequality (18) holds, i.e.,

$$\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \psi}|_{y_{i_t}=-1} = -4\eta\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})((m+\psi)\nabla f(\psi-1) + f(\psi-1) - f(\gamma-1))] \leq 0. \tag{19}$$

Now we prove $\frac{\partial 2\beta \mathbb{E}[r_o|_\psi]}{\partial \psi} \leq 0$ also holds for the case of $y_{i_t} = 1$.

- When $y_{i_t} = 1$, from (5) and (6), we can derive that

$$\begin{aligned} \bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} &= f\left(1 - \psi\right), \\ \mathbf{w}_t^\top \boldsymbol{x}_{i_t} &= f\left(1 - \gamma\right). \end{aligned}$$

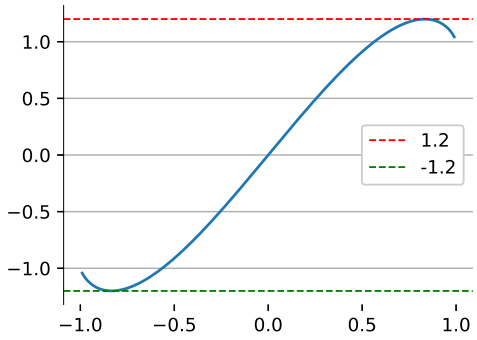


Figure 13. The curve of $\frac{1-v^2}{2} ln\left(\frac{1+v}{1-v}\right) + v, v \in (-1, 1)$.

From the definition of r_o, we can derive that

$$\begin{aligned} r_o &= \frac{\bar{\mathbf{w}}^\top - \mathbf{w}_t^\top}{\beta}\mathbf{r} \\ &= -\frac{2\eta}{\beta}(m-\psi)\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})\left(\bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} - \mathbf{w}_t^\top \boldsymbol{x}_{i_t}\right) \\ &= -\frac{2\eta}{\beta}(m-\psi)\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})(f(1-\psi) - f(1-\gamma)). \end{aligned}$$

So

$$\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \psi}|_{y_{i_t}=1} = -4\eta\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})\mathbb{E}[-m\nabla f(1-\psi) + \psi\nabla f(1-\psi)) - f(1-\psi) + f(1-\gamma)].$$

Because $\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}) > 0$, to prove $\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \psi}|_{y_{i_t}=1} \leq 0$, we only need to prove $(-m+\psi)\nabla f(1-\psi) - f(1-\psi) + f(1-\gamma) \geq 0$. Under the first assumption that the current example, i.e., the "easy" example, could be correctly labeled. In other words, $\gamma - 1 < 0$. From (15) we can derive that $f(1-\gamma) > 0$. So we only need to prove $(-m+\psi)\nabla f(1-\psi) - f(1-\psi) \geq 0$. Let $v = 1-\psi, v \in (-1, 1)$, from (15) and (16), it is equal to prove

$$-m + 1 \geq f(v)/\nabla f(v) + v = \frac{1-v^2}{2} ln\left(\frac{1+v}{1-v}\right) + v, v \in (-1, 1). \tag{20}$$

The blue line of Figure 13 is the function of the right side of the inequality. From the plot, we can see that the max value of the right side of the inequality is 1.2. So when $m \leq -0.2$, the inequality (20) holds. Therefore

$$\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \psi}|_{y_{i_t}=1} = -4\eta\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})\mathbb{E}[-m\nabla f(1-\psi) + \psi\nabla f(1-\psi)) - f(1-\psi) + f(1-\gamma)] \leq 0. \tag{21}$$

From (19) and (21), we can conclude that $\frac{\partial 2\beta \mathbb{E}[r_o|\psi]}{\partial \psi} \leq 0$ holds. Then the proof of step 2 is finished.

Together with the proof of step 1 and step 2, we can conclude that under some mild assumptions, $\frac{\partial \Delta(\psi)}{\partial \psi} \leq 0$ holds for both $y_{i_t} = -1$ and $y_{i_t} = 1$, i.e., the expected convergence rate at time t monotonically decreases with respect to the ideal difficulty score ψ of $\boldsymbol{x}_{i_t}$. □

B.3. Convergence rate with a current difficulty score

In this subsection, we analyze the relationship between the expected convergence rate and the current difficulty score. Different from the ideal difficulty score that each example is fixed for a certain optimal hypothesis, the current difficulty score of each example is changing with the state of the current model. We prove that if the gradient step η is small enough and given the ideal difficulty score, the expected convergence rate monotonically increases with respect to the current difficulty score.

Lemma 2. *Let r_o be the projection of the gradient vector $\mathbf{r}$ on the vector $\bar{\boldsymbol{w}} - \boldsymbol{w}_t$ and $\beta = \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|$. Given the ideal difficulty score ψ and the current difficulty score γ, the expected convergence rate $\Delta(\psi, \gamma)$ can be represented as*

$$\Delta(\psi, \gamma) = 2\beta \mathbb{E}\left[r_o|_{\psi,\gamma}\right] - \mathbb{E}\left[r^2|_{\psi,\gamma}\right]. \tag{22}$$

The proof of Lemma 2 is the same as the proof of Lemma 1. Such a form of convergence rate can facilitate the proof of the following theorem.

Theorem. 2. *Let $h(\boldsymbol{w}) = g\left(\boldsymbol{w}^\top \boldsymbol{x}\right)$ be a non-linear model whose label belongs to $\{-1, 1\}$, $\{(\boldsymbol{x}_{i_t}, y_{i_t})\}$ is the sampled data point at time t, where $i_t \in [n]$ and $t \in [T]$. $g(\cdot)$ is a tanh activation function: $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. Let $\gamma = |g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}) - y_{i_t}|$ be the current difficulty score and $\psi = |g(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_{i_t}) - y_{i_t}|$ be the ideal difficulty score. $\boldsymbol{w}_t$ is the model parameter at time t, and $\bar{\boldsymbol{w}}$ is the optimal model parameter. Assume that the gradient step η is small enough. Given the ideal difficulty score ψ, the expected convergence rate at time t monotonically increases with respect to the current difficulty score of $\boldsymbol{x}_{i_t}$, i.e., $\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma} \geq 0$.*

Proof. We prove theorem 2 holds for both $y_{i_t} = -1$ and $y_{i_t} = 1$. First we compute $\frac{\partial \mathbb{E}\left[r^2|_{\psi,\gamma}\right]}{\partial \gamma}$. From (8) we can derive that

$$\mathbf{r} = -2\eta \boldsymbol{x}_{i_t} \nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})(g(\mathbf{w}_t^\top \boldsymbol{x}_{\boldsymbol{i_t}}) - y_{i_t}) = -2\eta(\pm\gamma)\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})\boldsymbol{x}_{i_t}. \tag{23}$$

So

$$\begin{aligned} \frac{\partial \mathbb{E}\left[r^2|_{\psi,\gamma}\right]}{\partial \gamma} &= \frac{\partial \mathbb{E}\left[4\eta^2\gamma^2 \left(\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{\boldsymbol{i_t}})\right)^2 \|\boldsymbol{x}_{\boldsymbol{i_t}}\|^2\right]}{\partial \gamma} \\ &= 8\eta^2\gamma(\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{\boldsymbol{i_t}}))^2 \mathbb{E}\left[\|\boldsymbol{x}_{\boldsymbol{i_t}}\|^2\right]. \end{aligned} \tag{24}$$

Then we compute $\frac{\partial 2\beta \mathbb{E}[r_o|_{\psi,\gamma}]}{\partial \gamma}$ for $y_{i_t} = -1$ and $y_{i_t} = 1$ respectively.

- When $y_{i_t} = -1$, from (5) and (6) we can derive that

$$\begin{aligned} \bar{\mathbf{w}}^\top \boldsymbol{x}_{\boldsymbol{i_t}} &= f(\psi - 1), \\ \mathbf{w}_t^\top \boldsymbol{x}_{\boldsymbol{i_t}} &= f(\gamma - 1). \end{aligned} \tag{25}$$

  From the definition of r_o, (23), and (25) we can compute r_o:

$$\begin{aligned} r_o|_{y_{i_t}=-1} &= \frac{\bar{\mathbf{w}}^\top - \mathbf{w}_t^\top}{\beta}\mathbf{r} \\ &= -\frac{2\eta}{\beta}\gamma\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}) \left(\bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} - \mathbf{w}_t^\top \boldsymbol{x}_{i_t}\right) \\ &= -\frac{2\eta}{\beta}\gamma\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}) \left(f(\psi - 1) - f(\gamma - 1)\right). \end{aligned}$$

  So

$$\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \gamma}|_{y_{i_t}=-1} = -4\eta\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})(f(\psi - 1) - f(\gamma - 1) - \gamma\nabla f(\gamma - 1))]. \tag{26}$$

- When $y_{i_t} = 1$, from (5) and (6), we can derive that

$$\begin{aligned} \bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} &= f(1 - \psi), \\ \mathbf{w}_t^\top \boldsymbol{x}_{i_t} &= f(1 - \gamma). \end{aligned} \tag{27}$$

  From the definition of r_o, (23), and (27), we can compute r_o:

$$\begin{aligned} r_o|_{y_{i_t}=1} &= \frac{\bar{\mathbf{w}}^\top - \mathbf{w}_t^\top}{\beta}\mathbf{r} \\ &= \frac{2\eta}{\beta}\gamma\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t}) \left(\bar{\mathbf{w}}^\top \boldsymbol{x}_{i_t} - \mathbf{w}_t^\top \boldsymbol{x}_{i_t}\right) \\ &= \frac{2\eta}{\beta}\gamma\nabla g(\mathbf{w}_t^\top \boldsymbol{x}_{i_t})(f(1 - \psi) - f(1 - \gamma)). \end{aligned}$$

So

$$\frac{\partial \mathbb{E}[2\beta r_o]}{\partial \gamma}|_{y_{i_t}=1} = 4\eta\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})(f(1-\psi) - f(1-\gamma) + \gamma\nabla f(1-\gamma))]. \tag{28}$$

Now we prove that if the gradient step η is small enough, the expected convergence rate at time t monotonically increases with respect to the current difficulty score γ of $\boldsymbol{x}_{i_t}$, i.e., $\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma} \geq 0$.

- When $y_{i_t} = -1$, from (22), (24), and (26), we can derive that

$$\begin{aligned}\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma}|_{y_{i_t}=-1} &= \frac{\partial 2\beta \mathbb{E}\left[r_o|_{\psi,\gamma}\right]}{\partial \gamma}|_{y_{i_t}=-1} - \frac{\partial \mathbb{E}\left[r^2|_{\psi,\gamma}\right]}{\partial \gamma}|_{y_{i_t}=-1} \\ &= -4\eta\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})(f(\psi-1) - f(\gamma-1) - \gamma\nabla f(\gamma-1) + 2\eta\gamma\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})||x_{i_t}||^2)].\end{aligned}$$

  So when

$$\eta \leq \frac{\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})(\gamma\nabla f(\gamma-1) - f(\psi-1) + f(\gamma-1))]}{2\gamma\mathbb{E}[(\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}}))^2 \left\|\boldsymbol{x_{i_t}}\right\|^2]}, \tag{29}$$

  the inequality $\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma}|_{y_{i_t}=-1} \geq 0$ holds.

- When $y_{i_t} = 1$, from (22), (24), and (28), we can derive that

$$\begin{aligned}\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma}|_{y_{i_t}=1} &= \frac{\partial 2\beta \mathbb{E}\left[r_o|_{\psi,\gamma}\right]}{\partial \gamma}|_{y_{i_t}=1} - \frac{\partial \mathbb{E}\left[r^2|_{\psi,\gamma}\right]}{\partial \gamma}|_{y_{i_t}=1} \\ &= 4\eta\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})(f(1-\psi) - f(1-\gamma) + \gamma\nabla f(1-\gamma) - 2\eta\gamma\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})||x_{i_t}||^2)].\end{aligned}$$

  So when

$$\eta \leq \frac{\mathbb{E}[\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}})(f(1-\psi) - f(1-\gamma) + \gamma\nabla f(1-\gamma))]}{2\gamma\mathbb{E}[(\nabla g(\mathbf{w}_t^\top \boldsymbol{x_{i_t}}))^2 \left\|\boldsymbol{x_{i_t}}\right\|^2]}, \tag{30}$$

  the inequality $\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma}|_{y_{i_t}=1} \geq 0$ holds.

Comparing with (29) and (30), and according to the parity of f and ∇f, we can conclude that with the same ψ and γ, η needs to satisfy the same upper bound for both $y_{i_t} = -1$ and $y_{i_t} = 1$.

In conclusion, when the gradient step η satisfies the inequality (29),

$$\frac{\partial \Delta(\psi,\gamma)}{\partial \gamma} \geq 0$$

holds, i.e., the expected convergence rate at time t monotonically increases with respect to the current difficulty score γ of $\boldsymbol{x}_{i_t}$. □