THEORETICAL GENERALIZATION BOUNDS FOR IMPROVING THE EFFICIENCY OF DEEP ONLINE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

In the era of data explosion, online machine learning in which learning models are updated in real-time has become essential due to the growth of data in practice. In particular, it is more challenging to collect and annotate new massive data accurately and timely compared to traditional offline supervised training settings. Although this online training framework has been shown to be practically beneficial, there has been a lack of theoretical guarantees for the learning performance, especially for the case with noisy labels. This paper aims to investigate a learning theory for both original deep online training and online training with noisy labels. We first introduce a theoretical bound of the gaps of empirical risks and gaps of generalization risks in micro-batch online training when learning with both clean and noisy labels. Those bounds will efficiently help guide the online training scheme when receiving new data. We next analyze the impact of micro-batch size on the learning performance of models with noisy labels through our experimental results on CIFAR10, and CIFAR100 datasets using different noise, which consistently demonstrates the merit of the bounds above in the online training setting.

1 INTRODUCTION

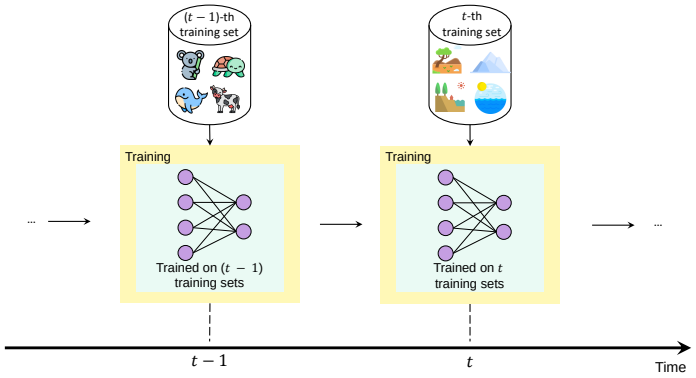
1.1 MOTIVATION

Deep learning has shown dominant performance in many domains compared to other traditional machine learning approaches. One of the key limitations of training a deep model is that it usually relies on the availability of a significant amount of data before training. Designing effective deep learning approaches, in which the model is trained over time as the number of data samples is continuously increased is critical since that will increase the use of deep learning in many real-life applications. To handle the ever-riching streams of data involved when solving real-time applications, one of the most successful approaches is online learning. Despite the abundance of these systems in deployment, the practice of online learning itself heavily lacks the rigorous theoretical studies that could serve as the foundation for both research and engineering innovations.

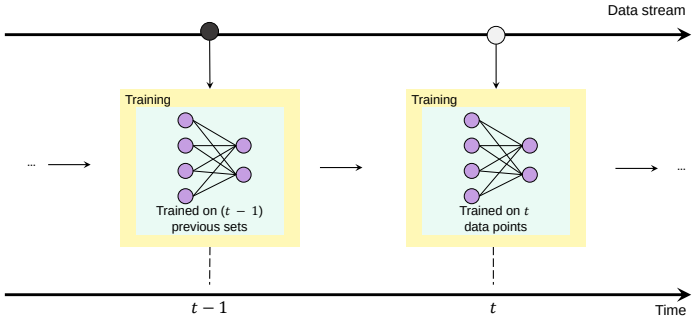
The online learning scheme considered in this paper should not be confused with continual learning (Van de Ven & Tolias, 2019), a related, but different research direction that enjoys a more active body of research. As illustrated in **Figure** 1, the goal of online learning is to progressively update a learning model using the continuous stream of data coming from the business interface (e.g., user activities, user reviews, etc.,) to solve a fixed specific task. Continual learning, on the other hand, aims to use a single deep neural network for solving a sequence of different tasks, or a task with either a sequence of different data domains or different problem requirements (e.g., different sets of classes) (Van de Ven & Tolias, 2019). Another distinguishing characteristic of these two fields is that while online learning cannot possibly access the "full training set" due to the nature of real-time problems, continual learning is free from this constraint and can be seen as repurposing the model from one training set to another.

In practice, the online learning framework can be further categorized into two main approaches, namely "streaming online learning" and "micro-batch online learning". The former considers updating the model as soon as a new data point has arrived, while the latter employs data buffers as training micro-batches and only updates the model by these micro-batches (Hoi et al., 2018). Lastly, one essential goal of streaming online learning is to perform training and inference simultaneously

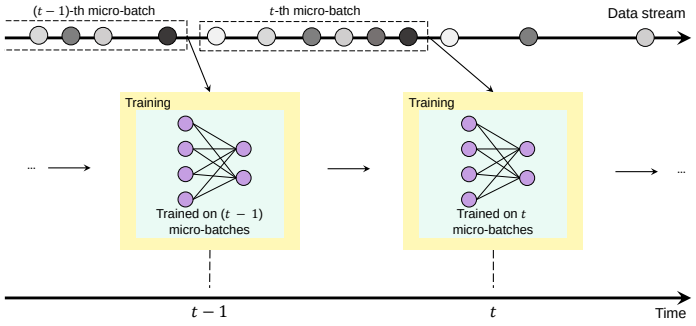
in real-time systems, while micro-batch online learning can tolerate a delay period and separate the training and inference routines for the sake of accumulated performance.



(a)



(b)



(c)

Figure 1: Differences between three learning settings: (a) Continual learning (b) Stream online learning (c) Micro-batch online learning

In this paper, we focus on the micro-batch online learning setting due to its suitability for many real-life applications. The main goal of our paper is to introduce a theoretical guarantee for the learning performance of that online training framework. Furthermore, we extend our theoretical results to the case of online learning with noisy labels. This data issue can naturally happen in many practical settings due to the fact that when collecting real-time and crowd-sourcing data from users, labels are inherently noisy with incorrect annotations. The learning performance of a model has been shown to be negatively affected by noisy labels (Ghosh et al., 2015; 2017). It is worth noting that our paper can be positioned as a theoretical extension of (Cesa-Bianchi et al., 2004) for micro-batch online

learning with deep neural networks, adding to the current short list of micro-batch online learning research.

1.2 CONTRIBUTIONS AND ORGANIZATION

Contribution: The contributions of our work are threefold:

(i) We investigate the learning theory for deep online training settings. In particular, our work provides the first lower bound for the difference between gaps of empirical risks and gaps of truth risks in micro-batch online learning, which could be used as a guideline for machine learning practitioners on when they should retrain their model with new labeled data

(ii) We extent the theoretical results above for the online training under noisy labels setting. Interestingly, we obtain similar analyses for that framework.

(iii) We provide experimental results on two benchmark datasets in machine learning to demonstrate the effects of online batch size in micro-batch online training problems with several noise levels.

Organization: The remainder of the paper is organized as follows. We provide some related work in Section 2. In Section 3, we derive detailed settings of micro-batch online learning, noisy learning, and theorems about the gap between generalization risks and empirical risks of online training problems. Section 4 benchmarks the performance of the training model with different online batch sizes by extensive experiments on some datasets including CIFAR10 (Krizhevsky et al., 2009), and CIFAR100 (Krizhevsky et al., 2009), and followed by discussions. Finally, we draw conclusions from this work in Section 5.

2 RELATED WORK

2.1 ONLINE LEARNING

Online algorithms for machine learning attracted great research interest in the past (Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz, 2007; Hoi et al., 2018) with studies ranging from theoretical bounds (Cesa-Bianchi et al., 2004; Ben-David et al., 2009), to online convex optimization theory (Bottou, 1998; Crammer et al., 2006; Zinkevich, 2003; Duchi et al., 2011), to shallow neural networks (Polikar et al., 2001; Elwell & Polikar, 2011) and kernel-based online learning (Cauwenberghs & Poggio, 2000; Kivinen et al., 2004; Read et al., 2012). However, the similar achievement for online learning with the modern deep learning approach is underdeveloped.

Most of the traditional online machine learning literature only concerns the instance-by-instance learning paradigm, which resembles modern streaming online learning, due to computational resource inadequacy. Among the representative works, Cesa-Bianchi et al. (2004) derive a probably approximately correct (PAC) generalization bound for learning by a sequence of data points that depends on the number of instances observed so far. Ben-David et al. (2009) study the learnability theory of hypotheses classes in online learning, which yields several online regret bounds. Bottou (1998) develops an online convex optimization framework for classical neural networks, while Zinkevich (2003) introduces Greedy Projection for online optimization with general convex functions. Along this line, Crammer et al. (2006) also propose Passive-Aggressive – a family of margin-based streaming online learning algorithms, before Duchi et al. (2011) present a family of subgradient methods adaptive to the evolving characteristics of newly arrived data.

A few classical research approaches, which go beyond linear classifiers include Learn++ and its variants (Polikar et al., 2001; Elwell & Polikar, 2011), which consider ensembles of multilayer perceptrons, and online kernel-based methods (Cauwenberghs & Poggio, 2000; Kivinen et al., 2004; Read et al., 2012). Some of these kernel-based methods are the first works that consider micro-batch online learning (Cauwenberghs & Poggio, 2000; Read et al., 2012), which requires the capacity of modern memory hardware but is substantially easier to adapt from offline learning in practice. Several recent works with micro-batch considerations are (Dekel et al., 2012) which analyzes the regret bounds of micro-batch online learning in the sequential setting before developing a parallel speed-up approach, and (Zhou et al., 2012) which aims to develop a micro-batch online learning algorithm for denoising autoencoders.

2.2 LEARNING WITH NOISY LABELS

To handle the learning issues with noisy labels, some considerable solutions have been proposed which can be categorized into two main approaches. The first approach relies on the robustness of loss functions or training processes, in which they are modified to improve the performance of the model in noisy label settings. For example, it has been shown that smooth losses such as hinge loss, squared loss, exponential loss, and log-loss can be more easily affected by label noise than hard losses (Patrini et al., 2016; Manwani & Sastry, 2013). The second approach is called "*filter framework*" and the algorithms consist of two separated parts. First, the label noise detection part detects the incorrect labels in the training dataset. Then, the cleaned samples will be used by the processing part to train a model.
The algorithms in each part are very diverse, and these exiting algorithm in this type is just the combination of two individual methods in each part, which also is the most beneficial advantage of this framework because it leverages the strengths of each part. Other interesting approaches are outlier detection-based algorithms (Sun et al., 2007; Zhao et al., 2019) and additional expert's knowledge-based approach (Nguyen et al., 2021). A more complex algorithm is to weigh the samples in the training set, which comes from the fact that the effect of each sample on the model can be different. Ren et al. (2018) and Shen & Sanghavi (2019) down-weigh the data samples that are likely noisy and train the model from the weighted dataset. However, there is a lack of paper working on dealing with the noisy label in online training, where the number of newly arrived samples is small.

(Kim et al., 2021) is the first paper mentioning the continual noisy learning problem. The main idea of the proposed method in that paper is based on a combination of using self-supervised learning and queue to store cleaned data. Although that approach has been shown to work well in practice, one of its key drawbacks is the lack of deep investigation of model learning behaviors. To the best of our knowledge, our paper is the first work that aims to provide theoretical and experimental results of insightful relations and boundaries for risks for online noisy learning. Our results can be used to guide both machine learning researchers and engineers through making decisions related to various settings in online training such as training frequency, sizes of online micro-batches, and many others.

3 METHODOLOGY

In this section, we first recap the empirical risk minimization framework and use it for the presentation of online training using both clean and noisy labels. We then present our theoretical development for the relation between the gap of true risks and the gap of empirical risks for both online training settings above. Finally, we provide an algorithm that depicts our proposed online training method above.

3.1 ONLINE LEARNING PROBLEM SETTING

Let us consider a supervised learning framework in which each data point is presented by a pair of random variables (x, y), where $x \in \mathcal{X} \subset \mathbb{R}^d$ is the input feature, and $y \in \mathcal{Y}$ is its corresponding label. We assume that all the data points (x, y) are i.i.d. samples from an unknown but fixed joint distribution $p(x, y)$. The ultimate goal of a supervised learning framework is to find a prediction function (or hypothesis) $f : X \to Y$, that represents the dependence of the output y on the input x. In the risk minimization scheme (Vapnik, 1998), that learning process is equivalent to minimizing the generalization risk of the model f given a loss function L, namely $R_L(f)$ or $R(f)$ for short defined by

$$R(f) = R_L(f) = \mathbb{E}_{(x,y)\sim p(x,y)}[L(y, f(x)].$$

In practice, we can only have access to a specific (finite) training dataset D, and the induced empirical risk minimization (ERM) framework will be represented by

$$f^*_{emp} = \arg\min_f R_{emp}(f),$$

where $R_{emp}(f) = \dfrac{1}{|D|}\sum_{i=1}^{|D|} L(f(x_i, y_i))$ is the empirical version of the ground truth risk $R(f)$. The performance of the learning process is then evaluated by the generalization error, which is defined as the gap between $R(f)$ and R_{emp}, i.e., $|R(f) - R_{emp}(f)|$.

Algorithm 1 micro-batch online training process

Initialize the prediction function f_0
for $t = 1, 2, ...T$ **do**
Continuously receive instance $\{(x_i^t, y_i^t)\}_{i=1}^N$ and then concatenate with available data samples D_{i-1} to form training set D_i
Train model f_{t-1} with training set D_i to minimize empirical risk $R_{emp}(f) = \frac{1}{|D|}\sum_{i=1}^{|D|} L(f(x_i, y_i))$
Use current optimal function f_t for inference when the number of arrived data samples is smaller than N
end for

In an online training procedure, we denote the original data set by D_0, and N_t is the number of new data samples at the time step t, $t = 1, \ldots, T$. In our setting, we assume that $N_t = N, \ \forall t$ for simplicity. Then the available dataset at time step t is defined as $D_t = D_{t-1} \cup \{(x_i^t, y_i^t)\}_{i=1}^N, \forall t = 1, 2, ..., T$. Then the learning idea is to estimate the model f that minimizes the empirical risk

$$R_{emp}^t(f) = \sum_{(x,y)\in D_t} \frac{1}{|D_t|} L(y, f(x)),$$

i.e., $f_t^* = \arg\min_f (R_{emp}^t)$. At the next time steps, when more N data samples have arrived, we aim to achieve a better performance model provided by the corresponding optimal model f_{t+1}^* trained with dataset D_{t+1}. In other words, $f_{t+1}^* = \arg\min_f (R_{emp}^{t+1})$. The **Algorithm** 1 presents the pseudo-code for micro-batch online training process.

3.2 Theoretical development

3.2.1 Relationship between the gap of true risks and the gap of empirical risks in micro-batch online training

Theorem 1. *Assuming that the loss function L is bounded, i.e., there exist two real number a, b such that $a \le L(y, x) \le b \ \forall (x, y)$. Then we have the relationship between gap of true risks and the gap of empirical risks of two optimal models at two time steps t and $t + B$, $(1 \le B \le T)$ as follow:*

$$P[|R(f_t^*) - R(f_{t+B}^*)| < 2\epsilon + |R_{emp}^t(f_t^*) - R_{emp}^{t+B}(f_{t+B}^*)|] > \left(1 - 2\exp\left(\frac{-2|D_t|\epsilon^2}{(b-a)^2}\right)\right)\left(1 - 2\exp\left(\frac{-2|D_{t+B}|\epsilon^2}{(b-a)^2}\right)\right) \quad (1)$$

Proof. We have $L(y_i, f(x_i)) \ \forall i$ are independent and bounded between a and b. By applying the Hoeffding's inequality (Hoeffding, 1994) with independent random variables $L(y_i, f(x_i))$ with the sum of variable is the empirical risk $R_{emp}^t(f_t^*)$, and its expectation is the generalization risk $R(f_t^*)$, we obtain

$$P[|R_{emp}^t(f_t^*) - R(f_t^*)| < \epsilon] > 1 - 2\exp\left(\frac{-2|D_t|\epsilon^2}{(b-a)^2}\right), \text{ and}$$

$$P[|R_{emp}^t(f_{t+B}^*) - R(f_{t+B}^*)| < \epsilon] > 1 - 2\exp\left(\frac{-2|D_{t+B}|\epsilon^2}{(b-a)^2}\right) \text{ for all } \epsilon > 0.$$

Hence

$$\begin{aligned} &P[(|R(f_t^*) - R_{emp}^t(f_t^*)| < \epsilon) \cap (|R(f_{t+B}^*) - R_{emp}^t(f_{t+B}^*)| < \epsilon)] \\ &= P[|R(f_t^*) - R_{emp}^t(f_t^*)| < \epsilon] \cdot P[|R(f_{t+B}^*) - R_{emp}^t(f_{t+B}^*)| < \epsilon] \\ &> \left(1 - 2\exp\left(\frac{-2|D_t|\epsilon^2}{(b-a)^2}\right)\right)\left(1 - 2\exp\left(\frac{-2|D_{t+B}|\epsilon^2}{(b-a)^2}\right)\right) \end{aligned}$$

Note that the left hand side of the equation above can be rewritten as:

$$P[|R(f_t^*) - R(f_{t+B}^*)| < 2\epsilon + |R_{emp}^t(f_t^*) - R_{emp}^{t+B}(f_{t+B}^*)|].$$

We thus get equation 1. □

In the micro-batch online training settings, it is crucial to know in advance the amount of improvement regarding the performance if we use newly arrived data to immediately train the model. The theorem 1 above provides an estimation of that improvement so that both researchers and practitioners can make decisions on the proper number of data to train the model when considering the time for training and computational cost. It is worth noting that the parameter ϵ could be used to adjust the tightness or the certainty of the difference between those two gaps.

As clearly showed in the theorem 1, the right-hand-side of equation 1 depends on the size of the set of trained data samples $|D_t|$ and the number of data samples will be used to train $|D_{t+B}| = |D_t| + B \cdot N$. With a fixed ϵ, the value of that probability increases when $|D_t|$ or B rises.

Corollary 1. *With probability at least* $(1 - \delta_t)(1 - \delta_{t+B})$*,*

$$|R(f_t^*) - R(f_{t+B}^*)| - |R_{emp}^t(f_t^*) - R_{emp}^{t+B}(f_{t+B}^*)| < \sqrt{\frac{\ln(\frac{2}{\delta_t})(b-a)^2}{2|D_t|}} + \sqrt{\frac{\ln(\frac{2}{\delta_t})(b-a)^2}{2(|D_t| + \Delta)}}, \quad (2)$$

where $\Delta = |D_{t+B}| - |D_t| = B \cdot N$*,* $\delta_t = 2\exp\left(\frac{-2|D_t|\epsilon_t^2}{(b-a)^2}\right)$*, and* $\delta_{t+B} = 2\exp\left(\frac{-2|D_{t+B}|\epsilon_{t+B}^2}{(b-a)^2}\right)$

Corollary 1 explicitly indicates the dependence of the gap with respect to the difference Δ in the number of samples of the dataset between two timesteps. As Δ tends to increase, the gap tends to be tighter with a probability of at least $(1 - \delta_t)(1 - \delta_{t+B})$.

We next derive a similar result for online training under noisy labels.

3.2.2 RELATIONSHIP BETWEEN THE GAP OF TRUTH RISKS AND THE GAP OF EMPIRICAL RISKS IN MICRO-BATCH ONLINE TRAINING

In noisy learning Ghosh et al. (2015; 2017), a noisy label $\tilde{y}$ is observed instead of the ground truth label y and the noisy data set is defined as $\tilde{D}_t = \tilde{D}_{t-1} \cup \{(x_i, \tilde{y_i})\}_{i=1}^N, \forall t = 1, 2, ..., T$. The (noisy) empirical risk of function f with $\tilde{D}_t$ then becomes $\tilde{R}_{emp}^t(f) = \frac{1}{|\tilde{D}_t|}\sum_{(x,\tilde{y})\in\tilde{D}_t} L(\tilde{y}, f(x))$. The noisy generalization risk of model f under loss function L is $\tilde{R}(f) = \mathbb{E}_{(x,\tilde{y})\sim P(x,\tilde{y})}[L(\tilde{y}, f(x)]$. We denote $\tilde{f}_t^* = \arg\min_f(\tilde{R}_{emp}^t)$, and $\tilde{f}_{t+1}^* = \arg\min_f(\tilde{R}_{emp}^{t+1})$ are the optimal functions at time step t^{th} and $(t+1)^{th}$.

The following theorem shows that Theorem 1 is also eligible for online training under noisy labels.

Theorem 2. *Assuming that the loss function L satisfies:* $a \leq L(y, x)) \leq b$*,* $\forall(x, y)$*. Then we have the relationship between gap of truth risks and the gap of empirical risks of two optimal models at two time steps t and* $t + B$ *under noisy label as follow :*

$$P[|\tilde{R}(\tilde{f}_t^*) - \tilde{R}(\tilde{f}_{t+B}^*)| < 2\epsilon + |\tilde{R}_{emp}^t(\tilde{f}_t^*) - \tilde{R}_{emp}^{t+B}(\tilde{f}_{t+B}^*)|]$$
$$> \left(1 - 2\exp\left(\frac{-2|\tilde{D}_t|\epsilon^2}{(b-a)^2|}\right)\right)\left(1 - 2\exp\left(\frac{-2|\tilde{D}_{t+B}|\epsilon^2}{(b-a)^2}\right)\right) \quad (3)$$

Proof. We have $L(\tilde{y}_i, \tilde{f}(x_i))$ are independent $\forall i$. By applying Hoeffding's inequality with independent random variables $L(\tilde{y}_i, \tilde{f}_i^*(x_i))$, where the sum of variable is the noisy empirical risk $\tilde{R}_{emp}^t(\tilde{f}_t^*)$, and its expectation is the noisy generalization risk $\tilde{R}(\tilde{f}_t^*)$ we have:

$$P[|\tilde{R}_{emp}^t(\tilde{f}_t^*) - \tilde{R}(\tilde{f}_t^*)| < \epsilon] > 1 - 2\exp\left(\frac{-2|\tilde{D}_t|\epsilon^2}{(b-a)^2}\right), \text{ and}$$

$$P[|\tilde{R}_{emp}^t(\tilde{f}_{t+B}^*) - \tilde{R}(\tilde{f}_{t+B}^*)| < \epsilon] > 1 - 2\exp\left(\frac{-2|\tilde{D}_{t+B}|\epsilon^2}{(b-a)^2}\right)$$

for all $\epsilon > 0$. Hence

$$
\begin{aligned}
&P[(|\tilde{R}(\tilde{f}_t^*) - \tilde{R}_{emp}^t(\tilde{f}_t^*)| < \epsilon) \cap (|\tilde{R}(\tilde{f}_{t+B}^*) - \tilde{R}_{emp}^t(\tilde{f}_{t+B}^*)| < \epsilon)] \\
&= P[|\tilde{R}(\tilde{f}_t^*) - \tilde{R}_{emp}^t(\tilde{f}_t^*)| < \epsilon] \cdot P[|\tilde{R}(\tilde{f}_{t+B}^*) - \tilde{R}_{emp}^t(\tilde{f}_{t+B}^*)| < \epsilon] \\
&> \left(1 - 2\exp\left(\frac{-2|\tilde{D}_t|\epsilon^2}{(b-a)^2}\right)\right)\left(1 - 2\exp\left(\frac{-2|\tilde{D}_{t+B}|\epsilon^2}{(b-a)^2}\right)\right)
\end{aligned}
$$

.

We can rewrite the left-hand side of the equation above as:

$$
P[|\tilde{R}(\tilde{f}_t^*) - \tilde{R}(\tilde{f}_{t+B}^*)| < 2\epsilon + |\tilde{R}_{emp}^t(\tilde{f}_t^*) - \tilde{R}_{emp}^{t+B}(\tilde{f}_{t+B}^*)|].
$$

We, therefore, get the results in equation 2. □

Theorem 2 showed that gaps between generalization risks of the two optimal functions f_t^* and f_{t+B}^* at time step t^{th} and $(t+B)^{th}$ will be bounded by the corresponding gaps between two minimal empirical risks of the same functions (which is essentially the optimal training loss).

Corollary 2. *With probability at least* $(1-\delta_t)(1-\delta_{t+B})$*,*

$$
|\tilde{R}(\tilde{f}_t^*) - \tilde{R}(\tilde{f}_{t+B}^*)| - |\tilde{R}_{emp}^t(\tilde{f}_t^*) - \tilde{R}_{emp}^{t+B}(\tilde{f}_{t+B}^*)| < \sqrt{\frac{\ln(\frac{2}{\delta_t})(b-a)^2}{2|D_t|}} + \sqrt{\frac{\ln(\frac{2}{\delta_t})(b-a)^2}{2(|D_t|+\Delta)}}, \quad (4)
$$

where $\Delta = |D_{t+B}| - |D_t| = B \cdot N$*,* $\delta_t = 2\exp\left(\frac{-2|D_t|\epsilon_t^2}{(b-a)^2}\right)$*, and* $\delta_{t+B} = 2\exp\left(\frac{-2|D_{t+B}|\epsilon_{t+B}^2}{(b-a)^2}\right)$

Similar to Corollary 1, Corollary 2 explicitly indicates the dependence of the gap with respect to the difference Δ in number of samples of the dataset between two time steps.

4 Experiment

4.1 Experimental Setting

In this section, we empirically evaluate the effects of online batch size by the generalization performance of models trained online with different sizes Δ of micro-batches. In particular, we consider the performance based on the top-1 classification accuracy of ResNet-34 (He et al., 2016) on two datasets: CIFAR10 (Krizhevsky et al., 2009) (containing $50,000$ training and $10,000$ testing images of 10 visual classes) and CIFAR100 (Krizhevsky et al., 2009) (consisting of $50,000$ training and $10,000$ testing images of 100 visual classes).

To simulate the microbatch-based online learning practice, we first train the model with 50% of the training set (to achieve a reasonable performance at the initial training stage), corresponding to training the based model with public datasets in practice. We then split the second half of the training set into different equal parts, corresponding to different micro-batch sizes, and incrementally train the model with one micro-batch at a time (see Figure 2). To separate evenly, three values of micro-batch size is used: 1562, 3125, and 6250 which correspond to dividing the remaining half of the training set into 4, 8 and 32 parts. The training routine is replicated identically, regardless of the micro-batch size. Based on the condition of Theorem 1, we employ the mean squared error (MSE) as the bounded training loss, with the Adam optimizer. The learning rate is fixed at 0.0001, while the number of epochs is fixed at 50. The batch size for each training epoch is 256 samples.

We further extend our experiments by replicating the above routine for different ratios of symmetric label noise to demonstrate the applicability of our theory to the pragmatic scenario of training with noisy labels. We corrupt the labels of training data by injecting uniform label noise (Ghosh et al., 2017) at 4 noise rates r_{noise}: 0%, 20%, 40%, 60% before training the model. All experiments are conducted with three random seeds $\{3, 4, 5\}$. The visual results are represented in **Figures** 3 and 4. Targeting a fair comparison, we run each experiment 3 times using different initializations, and the numerical results presented in the form of the means and standard deviations aggregated from these trials are reported in **Table**1.

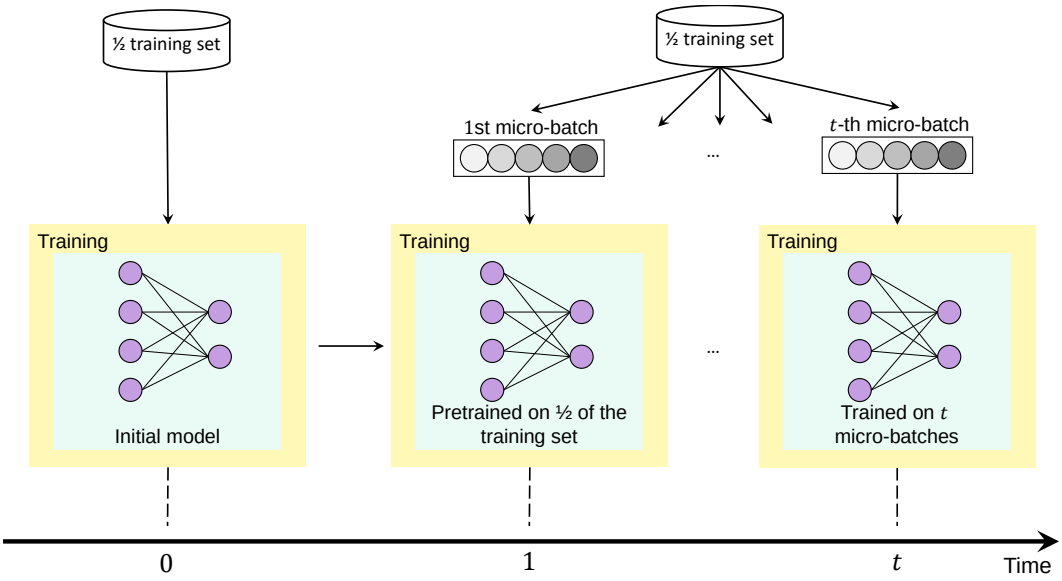


Figure 2: An illustration of experiments

4.2 Results and Discussion

First, it can be clearly seen that, in all cases, the gap between training with different micro-batches agrees with our theoretical results in that the performance gap between consecutive checkpoints should tend to be narrower as we increase the micro-batch size. As expected, the results are stable across initializations as the standard deviation is relatively insignificant across trials.

In addition, when the micro-batch size increases, the performance of the model after each time step increases. With the same training epoch, the larger the micro-batch size is, the higher performance of a model is. Furthermore, this trade-off is extendable in the case of noisy labels. As clearly showed in the **Figures** 3, 4, when the noise rate increases, the performance of models reduced, but the same patterns are observed.

Keep in mind that, unlike streaming online learning where the model is put into inference while training, our micro-batch setting only puts the model into inference after a training routine is concluded. On one hand, to have a better model within a reasonable computational cost, one should wait for more data before training. On the other hand, this might restrain its capability in catching up with real-time problems. Thus, our results should serve as a trade-off frontier in terms of micro-batch in online learning to assist machine learning engineers in deciding when to update models with new data.

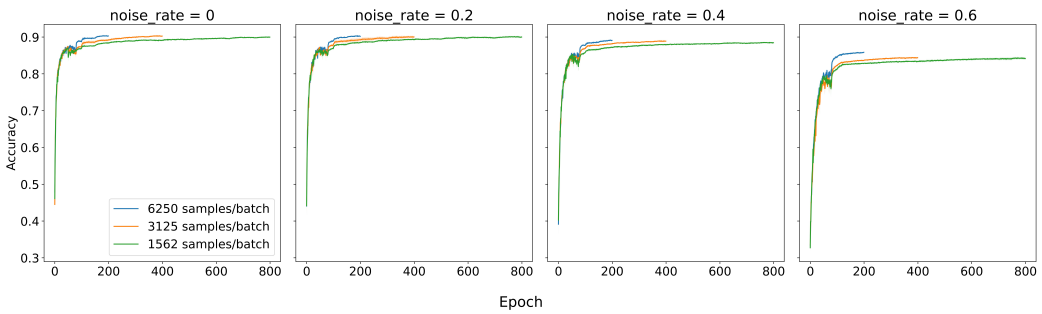


Figure 3: CIFAR10 test accuracy across micro-batch sizes and noise rates

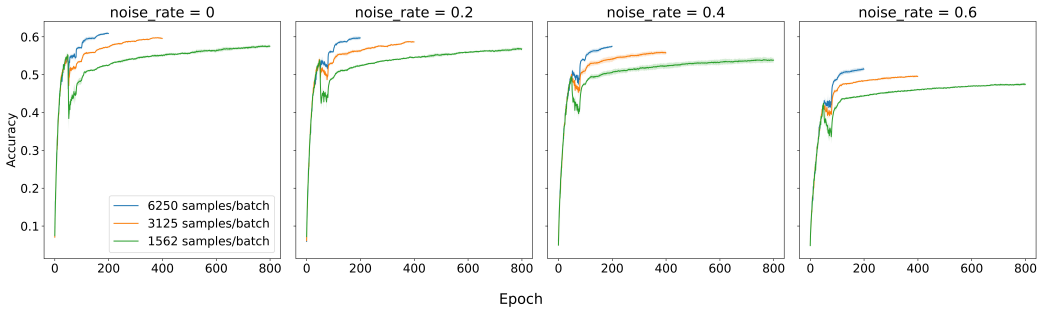


Figure 4: CIFAR100 test accuracy across micro-batch sizes and noise rates

Dataset		$\Delta = 6250$	$\Delta = 3125$	$\Delta = 1562$
Name	r_{noise}			
CIFAR10	0.0	0.903 ± 0.001	0.903 ± 0.001	0.900 ± 0.002
	0.2	0.902 ± 0.003	0.900 ± 0.003	0.900 ± 0.002
	0.4	0.891 ± 0.002	0.889 ± 0.000	0.885 ± 0.000
	0.6	0.859 ± 0.002	0.844 ± 0.001	0.842 ± 0.003
CIFAR100	0.0	0.609 ± 0.002	0.596 ± 0.001	0.575 ± 0.004
	0.2	0.597 ± 0.004	0.587 ± 0.002	0.568 ± 0.004
	0.4	0.575 ± 0.003	0.557 ± 0.006	0.538 ± 0.006
	0.6	0.514 ± 0.005	0.495 ± 0.002	0.474 ± 0.004

Table 1: Test accuracy across micro-batch sizes and noise rates on CIFAR10 and CIFAR100.

5 Conclusion and Future Works

In this paper, we deeply investigate micro-batch online training framework for learning deep models based on the empirical risk minimization learning principle. Our theoretical and empirical results demonstrate that the generalization performance gap between consecutive checkpoints critically depends on the sizes of incremental data batches. This should serve as strategic guidance for online updatable machine learning algorithms deploying in practice, e.g., whether or not to update your "real-time" models within seconds or minutes. In future works, we look foward to using these results as the foundation for investigating methods with more realistic assumptions and exploits.

References

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, pp. 1, 2009.

Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. *Advances in neural information processing systems*, 13, 2000.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of machine learning research*, 7, 2006.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Ryan Elwell and Robi Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.

Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

SC Hoi, D Sahoo, J Lu, and P Zhao. Online learning: A comprehensive survey. arxiv. *arXiv preprint arXiv:1802.02871*, 2018.

Chris Dongjoo Kim, Jinseo Jeong, Sangwoo Moon, and Gunhee Kim. Continual learning on noisy data streams via self-purified replay. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 537–547, 2021.

Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013.

Quoc Nguyen, Tomoaki Shikina, Daichi Teruya, Seiji Hotta, Huy-Dung Han, and Hironori Nakajo. Leveraging expert knowledge for label noise mitigation in machine learning. *Applied Sciences*, 11(22):11040, 2021.

Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *International conference on machine learning*, pp. 708–717. PMLR, 2016.

Robi Polikar, Lalita Upda, Satish S Upda, and Vasant Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, 31(4):497–508, 2001.

Jesse Read, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. Batch-incremental versus instance-incremental learning in dynamic and evolving data. In *International symposium on intelligent data analysis*, pp. 313–323. Springer, 2012.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University, 2007.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019.

Jiang-wen Sun, Feng-ying Zhao, Chong-jun Wang, and Shi-fu Chen. Identifying and correcting mislabeled training instances. In *Future generation communication and networking (FGCN 2007)*, volume 1, pp. 244–250. IEEE, 2007.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Vladimir Vapnik. Statistical learning theory new york. *NY: Wiley*, 1(2):3, 1998.

Zilong Zhao, Sophie Cerf, Robert Birke, Bogdan Robu, Sara Bouchenak, Sonia Ben Mokhtar, and Lydia Y Chen. Robust anomaly detection on unreliable data. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 630–637. IEEE, 2019.

Guanyu Zhou, Kihyuk Sohn, and Honglak Lee. Online incremental feature learning with denoising autoencoders. In *Artificial intelligence and statistics*, pp. 1453–1461. PMLR, 2012.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.