Mixture of Online and Offline Experts for Non-stationary Time Series

Zhilin Zhao\textsuperscript{1,2,3}, Longbing Cao\textsuperscript{2}, Yuanyu Wan\textsuperscript{4}

\textsuperscript{1} School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
\textsuperscript{2} School of Computing, Macquarie University, Sydney, Australia
\textsuperscript{3} Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
\textsuperscript{4} School of Software Technology, Zhejiang University, Ningbo, China
zhaozhl7@hotmail.com, longbing.cao@mq.edu.au, wanyy@zju.edu.cn

Abstract

We consider a general and realistic scenario involving non-stationary time series, consisting of several offline intervals with different distributions within a fixed offline time horizon, and an online interval that continuously receives new samples. For non-stationary time series, the data distribution in the current online interval may have appeared in previous offline intervals. We theoretically explore the feasibility of applying knowledge from offline intervals to the current online interval. To this end, we propose the Mixture of Online and Offline Experts (MOOE). MOOE learns static offline experts from offline intervals and maintains a dynamic online expert for the current online interval. It then adaptively combines the offline and online experts using a meta expert to make predictions for the samples received in the online interval. Specifically, we focus on theoretical analysis, deriving parameter convergence, regret bounds, and generalization error bounds to prove the effectiveness of the algorithm.

Introduction

For a non-stationary time series, the data distribution in the current time window may have appeared in the past. Therefore, can we apply the knowledge from historical data to the current time window? In this paper, we theoretically prove that this is feasible.

A common assumption in statistical learning theories for time series is that observed samples are i.i.d., or stationary in stochastic processes (Hamilton 1994). To leverage sample dependence in non-i.i.d. processes, it is often assumed that observations come from a stationary $\phi$-mixing or $\beta$-mixing sequence (Mohri and Rostamizadeh 2010). However, these assumptions may not hold as the distribution of real-life time series usually changes over time, making the hypothesis class not (agnostically) PAC learnable (Hanneke 2016). Fortunately, distribution changes in real life are often gradual, and samples in a short interval are nearly identically distributed (Kuznetsov and Mohri 2015). Therefore, we consider a realistic scenario involving non-stationary time series with several offline intervals of different distributions within a fixed offline time horizon and an online interval that continuously receives new samples. Once the number of received labeled samples reaches a predefined size, the online interval is converted to the last offline interval, and a new online interval begins.

Some existing methods (Shalev-Shwartz 2012; Zhang, Lu, and Zhou 2018) train an expert on the entire time series using off-the-shelf online optimization techniques, without considering the non-stationary nature of the data. However, the dynamic data with varying distributions can mislead the expert. Other methods (Yu 1994; Mohri and Rostamizadeh 2008, 2010) train an expert from scratch for each new online interval, which is safer but unreliable due to the scarcity of labeled samples at the early stage. Thus, it is fundamentally highly challenging to design a learning method with tight sample complexity that outputs a hypothesis with desirable generalization. For non-stationary time series, the data distribution in the current online interval may have appeared in historical offline intervals. Therefore, a natural solution is to combine the offline experts from the offline intervals with the online expert from the current interval to address the shortcomings of the aforementioned methods.

Inspired by the mixture of experts (Puigcerd et al. 2024), we propose Mixture of Online and Offline Experts (MOOE) to transfer knowledge from offline intervals to the online interval, addressing the non-stationarity issue. Following the paradigm of prediction with expert advice (Cesa-Bianchi and Lugosi 2006; van Erven and Koolen 2016), MOOE employs a meta expert to combine the online and offline experts by adaptively weighting them according to their effectiveness. Specifically, the online expert is continuously updated in the current online interval using an existing online optimization method. Additionally, when an online interval collects enough samples to become an offline interval, all samples from this interval, along with previously obtained offline experts, are used to train the offline expert for this interval.

Theoretically, we prove that the regret of MOOE is determined by the regret of the off-the-shelf online optimization method used for the online expert. However, this can be improved if the number of maintained experts is within a bound controlled by the size of intervals and the empirical errors of the offline experts. By connecting optimization with learning theory (Hazan 2016), we derive the generalization error bound by jointly exploiting the regret, the properties of the loss function, the hypothesis class, and the data distribution, thereby verifying the effectiveness of our approach. Exper-
mentally, MOOE outperforms state-of-the-art methods for handling non-stationary time series.

Related Work

Learning Theory for non-stationary Time Series

For non-i.i.d. processes, under the stationary and $\beta$-mixing assumptions, the early work (Yu 1994) establishes the convergence rate over VC-dimension, and the work in (Mohri and Rostamizadeh 2008) presents data-dependent bounds in terms of the Rademacher complexity. By exploiting the stability properties of a specific learning algorithm, generalization bounds for $\phi$-mixing and $\beta$-mixing sequences are provided in (Mohri and Rostamizadeh 2010). However, the mixing assumption is hard to be verified in practice. There are some attempts to relax the stationary and mixing assumptions. The uniform convergence under ergodicity sampling is shown in the work of (Adams and Nobel 2010). For an asymptotically stationary (mixing) process, although a generalization error is derived in (Agarwal et al. 2016) through the regret of an online algorithm, and their analysis depends on the assumption that the output from an online learning algorithm tends to be stable, which is invalid in a dynamic environment. In (Kuznetsov and Mohri 2014), the guarantee of the learning rate for nonstationary mixing processes is given by a sub-sample selection technique with the Rademacher complexity. Further, in (Kuznetsov and Mohri 2015), a more general scenario of nonstationary and nonmixing processes is considered, which proves the learning guarantees with the conditions of the discrepancies between distributions.

Regret Analysis in dynamic environments

The regret theory (Buchbinder et al. 2016) for measuring the performance has been extensively studied. The dynamic regret (Anagnostides, Farina, and Sandholm 2023) and its restricted form (Besbes, Gur, and Zeevi 2015) have been introduced to manage changing environments. A basic idea behind such regrets is to compare the cumulative loss of the learned expert with several experts rather than the best one. Along this line of study, adaptive learning for dynamic regret (Ader) (Zhang, Lu, and Zhou 2018) considers multiple experts with various learning rates updated by online gradient descent (OGD) (Zinkevich 2003), and the established regret guarantees with the conditions of the discrepancies between distributions.

Problem Statement

In non-stationary time series, an online platform containing experts will receive an input $x$ at each time step and predict its label $y$, indicating the class the input belongs to. For inputs with feedback, i.e., when the ground truth labels are revealed after predicting, the online platform will update by learning from the feedback. For inputs without feedback, the online platform merely predicts the labels. We aim to continuously update and utilize the experts in this online platform to more accurately predict class labels for samples from non-stationary time series in the current online interval.

Specifically, the considered non-stationary time series contains $G-1$ offline intervals and one online interval. Each offline interval contains $B$ samples, and the online interval contains $T$ samples ($T \in [B]$). We assume the distribution changes gradually, and the samples in each interval can be approximately drawn from a distribution. Accordingly, we set the maximal sample size $B$ as a hyperparameter, even if the time between distribution changes is not constant and usually unknown. The online interval will become offline if $T = B$, increasing the number of offline intervals. Accordingly, we have the following assumptions.

Assumption 1 Let $D_g$ be the data distribution in the $g^{th}$ interval. $D_{out} = \bigcup_{g=1}^{G} D_g$ is non-stationary since $D_g \neq D_{g'}$, $\forall g, g' \in [G], g \neq g'$.

Assumption 2 The norm of every input sample $x$ with label $y$ in the Hilbert space $i.d.$ drawn from the distribution $D_C$ of the online interval is upper bounded by a constant $D$: $\|x\| \leq D, \forall (x, y) \sim D_C$.

The eigendecomposition of the Hilbert-Schmidt operator is $E_{(x, y) \sim D_C}[xx^T] = \sum_{i=1}^{\infty} \lambda_i u_i u_i^T$, where $(u_i)_{i=1}^{\infty}$ forms an orthonormal basis of Hilbert Space and $(\lambda_i)_{i=1}^{\infty}$ corresponds to the eigenvalues in a non-increasing order.

Assumption 3 For any sample $(x, y) \sim D_{out}$, the hypothesis class is $H \triangleq \{h : x \mapsto \langle w, x \rangle | w \in W, \|w\| \leq R\}$, where the domain $W$ bounded by $R$ is a convex subspace of a Hilbert space.
**Assumption 4** For any sample \((x, y) \sim D_u\), the loss function \(L\) with the hypothesis class \(\mathcal{H}\) is bounded in the interval \([0, 1]\):
\[
\mathcal{L} = \{(x, y) \mapsto l(h(x), y) \mid h \in \mathcal{H}, l(h(x), y) \in [0, 1]\}.
\]

**Assumption 5** For any \((x, y) \sim D_u\) and all \(w, w' \in W\), \(l((\cdot, x), y)\) is convex and \(\beta\)-smooth over the domain \(W\):
\[
\|\nabla l((w, x), y) - \nabla l((w', x), y)\| \leq \beta \|w - w'\|.
\]

In the \(G\)th interval, we would like to learn an expert \(w \in W\) with a small popular risk with respect to the nonnegative loss function \(l\):
\[
L_D(w) = \mathbb{E}_{(x, y) \sim D}[l((w, x), y)],
\]
by minimizing the corresponding empirical risk using the proposed method:
\[
L_S(w) = \frac{1}{T} \sum_{t=1}^{T} l((w, x_t), y_t),
\]
where \(S = \{(x_1, y_1), \ldots, (x_T, y_T)\}\) is the data set consisting of \(T(T \in [B])\) samples in the online interval, and we use \(L_S(w)\) to denote the specific case when \(T = B\). Let \(w^* \in \arg \min_{w \in W} L_D(w)\) be an optimal solution and \(\bar{w} \in \arg \min_{w \in W} L_S(w)\) be an empirical minimizer.

Although the whole data stream is dynamic, we make an i.i.d. data assumption in Assumption 2 for the online interval that data distributions do not change drastically in practice. Assumption 4 is mild since we can ensure the loss function \(l\) is nonnegative by adding a large constant. We assume the interval of the loss function is \([0, 1]\) for convenience without loss of generality.

Because the loss function \(l\) is nonnegative as well as \(\beta\)-smooth, according to the self-bounding property (Srebro, Sridharan, and Tewari 2010) of smooth functions and Assumption 4, we obtain the following upper bound on the norm of the gradients of \(l((\cdot, x), y)\) for any \((x, y) \sim D_u\) and all \(w \in W\):
\[
\|\nabla l((w, x), y)\| \leq \sqrt{4\beta \cdot l((w, x), y)} \leq 2\sqrt{\beta}.
\]

Note that this paper aims to address the non-stationary issue rather than the widely-explored non-convex problem. We thus assume the loss function is convex for convenience and focus on providing the theoretical guarantees for the proposed learning mechanism.

**Mixture of Online and Offline Experts**

Fig. 1 introduces the working process of Mixture of Online and Offline Experts (MOOE) for the non-stationary time series. MOOE maintains several offline experts for the corresponding offline intervals and an online expert for the current online interval. It then integrates all of these experts using a meta expert with adaptive weights.

The number of maintained experts is \(K\), which is defined as,
\[
K = \begin{cases} 
K_{\max}, & \text{if } G \geq K_{\max} \\
G, & \text{if } G < K_{\max}
\end{cases}
\]

where \(K_{\max}\) is a hyperparameter denoting the maximal number of maintained experts. Therefore, MOOE contains \(K - 1\) offline experts and one online expert. In the interest of brevity, an expert and its corresponding advice are denoted as its parameters \(w\). Accordingly, we assume the \(k\)th \((k \in [K-1])\) offline expert is \(w_k^i\) and the online expert is \(w^K\). For the \(t\)th sample with feedback in the online interval, MOOE firstly selects \(K\) experts
\[
\begin{aligned}
\{w^t_1, \ldots, w^t_{K-1}, w^K\},
\end{aligned}
\]

and integrates them into a meta expert \(w_t\) for making a prediction. When \(T = B\), the online interval becomes offline, and a new online interval appears. We generate the new offline expert \(w^K\) for the just-passed complete online interval and refresh \(K - 1\) offline experts if \(G \geq K_{\max}\).

**Meta Expert**

The meta expert adjusts its strategy of integrating the \(K\) experts \((K - 1\) offline experts and one online expert) according to their losses received on labeled samples. For the online interval, we track the best expert (Herbster and Warmuth 1995) based on the exponentially weighted average forecaster (Cesa-Bianchi and Lugosi 2006) by assigning a considerable weight to the expert with a small cumulative loss, and vice versa. Accordingly, at iteration \(t\) in the online interval, the meta expert outputs a weighted average solution
\[
w_t = \sum_{k=1}^{K-1} \alpha^t_k w^t_k + \alpha^K_t w^K = \sum_{k=1}^{K} \alpha^K_t w^K,
\]

where \(\alpha^t_k\) is the weight of the \(k\)th expert \(w^K\). To lead to a compact regret bound, ensure that \(\sum_{k=1}^{K} \alpha^t_k = 1\), and provide different weights for experts according to their priorities, \(\alpha^t_k\) is initialized as
\[
\alpha^K_t = \frac{K + 1}{(K + 1 - k)(K + 2 - k)K}.
\]

Note that it is unnecessary to project \(w_t\) into the domain \(W\). Because each expert satisfies \(w^K_t \in W(k \in [K])\) and the weighting function Eq. (6) is linear, the weighted average \(w_t\) is still in the domain \(W\) according to convex properties.

After obtaining the loss at iteration \(t\), the \(K\) weights are updated according to the exponential weighting scheme
\[
\alpha^t_k = \frac{\alpha^K_t e^{-\nu f_t(w^K_t)}}{\sum_{k=1}^{K} \alpha^K_t e^{-\nu f_t(w^K_t)}},
\]
where \(\nu = 4 \sqrt{\frac{\ln K}{T}}\) is the step size. MOOE is summarized in Algorithm 1.

**Offline Expert**

We extract knowledge from offline intervals by learning an offline expert for each online interval when all of its samples are available. Each interval is coupled with its previous offline experts and online expert when its online expert has passed this interval once, and its previous offline experts
the search space of \( w \) or \( w \).

We do not delve into it here. Instead, we provide two simplifications for all \( B \) as

\begin{align*}
\alpha_{k}^{t} &= \frac{K + 1}{(K + 1 - k)(K + 2 - k)K}, \quad \forall k \in [K]
\end{align*}

for \( t = 1, \ldots, T \).

4: Receive online expert \( w_{t}^{K} \)
5: Assign offline expert \( w_{k}^{t} = w_{k}^{t}, \forall k \in [K - 1] \)
6: Output weighted average: \( w_{t} = \sum_{k=1}^{K} \alpha_{k}^{t} w_{k}^{t} \)
7: Receive the loss function \( f_{t}(\cdot) \)
8: Update expert weights:
\begin{align*}
\alpha_{k}^{t} &= \frac{\alpha_{k}^{t} e^{-\frac{1}{\alpha} f_{t}(w_{k}^{t})}}{\sum_{k=1}^{K} \alpha_{k}^{t} e^{-\frac{1}{\alpha} f_{t}(w_{k}^{t})}}, \quad \forall k \in [K]
\end{align*}

9: Send gradient \( \nabla f_{t}(w_{t}^{K}) \) to the online expert
10: end for

Algorithm 1: MOOE

1: Input: step size \( \nu \)

2: online expert \( w_{1}^{K} \)

3: Initialize \( \alpha_{1}^{t} < \alpha_{2}^{t} < \cdots < \alpha_{K}^{t} \) according to:
\begin{align*}
\alpha_{k}^{t} &= \frac{K + 1}{(K + 1 - k)(K + 2 - k)K}, \quad \forall k \in [K]
\end{align*}

4: for \( t = 1, \ldots, T \)
5: Receive online expert \( w_{t}^{K} \)
6: Assign offline expert \( w_{k}^{t} = w_{k}^{t}, \forall k \in [K - 1] \)
7: Output weighted average: \( w_{t} = \sum_{k=1}^{K} \alpha_{k}^{t} w_{k}^{t} \)
8: Receive the loss function \( f_{t}(\cdot) \)
9: Update expert weights:
\begin{align*}
\alpha_{k}^{t} &= \frac{\alpha_{k}^{t} e^{-\frac{1}{\alpha} f_{t}(w_{k}^{t})}}{\sum_{k=1}^{K} \alpha_{k}^{t} e^{-\frac{1}{\alpha} f_{t}(w_{k}^{t})}}, \quad \forall k \in [K]
\end{align*}

10: end for

Online Expert

To train an online expert for a new online interval, we can reinitialize its parameters randomly or by inheriting its solution from the just-passed complete online interval as a warm start. Recall that we can train the online expert by any off-the-shelf online optimization methods on the fly. In this paper, we use the standard Online Gradient Descent (OGD) (Zinkevich 2003) method as an instance because it is the most common and famous online optimization method. On the online interval, the online expert submits its advice \( w_{t}^{K} \) to the meta expert and receives the gradient \( \nabla f_{t}(w_{t}^{K}) \) to update its parameters by
\begin{align*}
w_{t+1}^{K} = \Pi_{W}[w_{t}^{K} - \eta_{t} \nabla f_{t}(w_{t}^{K})]
\end{align*}

\( \eta_{t} = \frac{D}{\sqrt{2T}} \) is the step size, and \( \Pi_{W} \) is the proximal operator onto space \( W \).

Theoretical Guarantees

In this section, we provide theoretical guarantees for MOOE, which match our expectations. Specifically, we analyze the properties of the regularization term \( \Omega(w) \) and provide the regret and the generalization error of the output hypothesis. To exploit the convexity, smoothness, and nonnegativity conditions of the loss function, the hypothesis class, the data distribution, and the regret, we involve the data-independent excess risk of \( \tilde{w} \), the Rademacher complexity of hypothesis class \( \mathcal{H} \) w.r.t. \( D \) and the regret for implying the generalization.

Parameter Convergence

The hyperparameter \( \gamma \) for \( \Omega(w) \) should be assigned with considerable value to ensure the validity of the regularization. To process, we derive the upper bound of this regularization.

Lemma 1 \( L_{\gamma}(w) \) is strongly-convex w.r.t. \( w \in W \), and
\begin{align*}
\Omega(w^{K}) &\leq \sum_{k=1}^{K} \alpha_{k}^{K} L_{\gamma}(w_{k}^{K}) / \gamma.
\end{align*}

We set \( \gamma \geq \left( \sum_{k=1}^{K} \alpha_{k}^{K} L_{\gamma}(w_{k}^{K}) \right) / 4R^{2} \) to ensure the validity of the regularization term.

Accordingly, the following theorem shows the benefit of this regularization, it can narrow the gap between the minimizer \( w_{\gamma}^{K} \) and the optimal solution \( w^{*} \) by applying the maintained \( K \) experts adaptively and setting \( \gamma \) carefully.

Theorem 1 By setting \( \gamma \geq \left( \sum_{k=1}^{K} \alpha_{k}^{K} L_{\gamma}(w_{k}^{K}) \right) / 4R^{2} \) and using \( w_{1}^{K}, w_{2}^{K}, \ldots, w_{K}^{K} \) as prior knowledge to obtain \( w_{\gamma}^{K} \) from \( L_{\gamma}(w) \), we have
\begin{align*}
\|w^{K} - w^{*}\| &\leq \sqrt{2\Omega(w^{*}) + \frac{32\beta}{\gamma^{2}} + \frac{6\sum_{k=1}^{K} \alpha_{k}^{K} L_{\gamma}(w_{k}^{K})}{\gamma}}.
\end{align*}

Although it is impossible for us to obtain \( w^{*} \) since the distribution for this interval is unknown, we can obtain...
an approximate solution by using the regularization term \( \Omega(w) \). If the optimal solution is close to the weight average \( \sum_{k=1}^{K} \alpha_k^{w} w_k^k \), the value of \( \Omega(w^*) \) and the upper bound of the difference \( \|w_k^k - w^*\| \) are small. Although it is also impossible for us to measure \( \Omega(w^*) \), we can measure the weighted term \( \sum_{k=1}^{K} \alpha_k^{w} L_\ell(w_k^k) \) in the above upper bound where \( L_\ell(w_k^k) \) is the empirical error of the \( k \)th expert in the latest interval. As a result, we know that the empirical minimizer \( w^K \) of \( L_\ell(w) \) approaches the optimal solution \( w^* \) of the original problem \( L_D(w) \) if these experts considered in the regularization term \( \Omega(w) \) are effective in the latest interval. To sharpen this bound, the weights for experts with small empirical errors should be larger and the design of the meta expert can meet the need. Therefore, we can draw a conclusion that \( w^K \) should be close to these experts with small empirical errors in the domain \( W \). This conclusion leads to the design of the regularization term \( \Omega(w) \).

**Regret Bound**

The following regret measures the performance of MOOE

\[
\text{Regret}_{\text{MOOE}} = \sum_{t=1}^{T} f_t(w_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^{T} f_t(w). \tag{11}
\]

However, it is hard to minimize the regret directly because the output \( w_t \) is related to a meta expert, an online expert, and \( K - 1 \) offline experts. Therefore, we decompose the regret into two regrets: \( \text{Regret}_{\text{ME}} \) w.r.t. the meta expert and \( \text{Regret}_{\text{KE}} \) w.r.t. online and offline experts. Further, we can bound \( \text{Regret}_{\text{KE}} \) by \( \text{Regret}_{\text{OE}} \) which corresponds to the online expert. Therefore, we can obtain the regret bound of \( \text{Regret}_{\text{MOOE}} \) by bounding \( \text{Regret}_{\text{ME}} \) and \( \text{Regret}_{\text{OE}} \) separately.

\[
\text{Regret}_{\text{MOOE}} = \text{Regret}_{\text{ME}} + \text{Regret}_{\text{KE}} + \text{Regret}_{\text{OE}}, \tag{12}
\]

where

\[
\text{Regret}_{\text{ME}} = \sum_{t=1}^{T} f_t(w_t) - \min_{k \in [K]} \sum_{t=1}^{T} f_t(w^k_t),
\]

\[
\text{Regret}_{\text{KE}} = \min_{k \in [K]} \sum_{t=1}^{T} f_t(w^k_t) - \sum_{t=1}^{T} f_t(\hat{w}),
\]

\[
\text{Regret}_{\text{OE}} = \sum_{t=1}^{T} f_t(w^k_t) - \sum_{t=1}^{T} f_t(\hat{w}).
\]

The online expert \( w^k_t \) never surpasses that of the best expert among all the \( K \) experts because it is also one of them. Besides, it is impossible to obtain the regret for the offline experts since they are pre-given and their parameters do not change after receiving the loss \( f_t(\cdot) \). Specifically, we have the following theorem.

**Theorem 2** The MOOE method with step sizes \( \{\nu = 4\sqrt{\frac{\ln K}{T}}, \eta_t = \frac{D}{\sqrt{\beta t}}, t \in [T]\} \) guarantees the following regret for all \( 1 \leq T \leq B \),

\[\text{Regret}_{\text{MOOE}} \leq \sqrt{T\ln K} + 6D\sqrt{T\beta},\]

and the number of experts \( K \) and samples \( T \) should satisfy

\[K \leq 2\exp\left(6D\sqrt{\beta} - \frac{\text{Regret}_{\text{KE}}}{\sqrt{T}}\right),\]

to ensure that the advice from MOOE gives an equivalent or better result than that from its online expert.

Accordingly, the regret of MOOE for the online interval is \( O(\sqrt{T}) \), which is consistent with that of the chosen online expert. However, MOOE works better, i.e., \( \text{Regret}_{\text{MOOE}} \leq \text{Regret}_{\text{OE}} \), if \( K \) and \( T \) satisfy the condition in Theorem 2. In theory, we have \( \text{Regret}_{\text{KE}} \leq \text{Regret}_{\text{OE}} \leq 6D\sqrt{T\beta} \). These offline experts are better than the online expert when their corresponding data distributions are approximately matched, or the number of observed labeled samples in the current interval is limited. The first inequality is strict, and \( K \) is bounded by a positive value. On the other hand, the number of samples in an interval \( T \leq B \) should not be too large. Although the bound of \( K \) depends on \( \text{Regret}_{\text{KE}} \), it is impossible to bound this term without any further assumptions because the \( K \) experts are trained from different data sets. Fortunately, it is unnecessary to set \( K \) strictly according to its conditions. We can apply MOOE if we believe that the regret of the best offline expert can surpass that of the online expert at least \( 6D\sqrt{T\beta} - \text{Regret}_{\text{KE}} = \sqrt{T\ln K} \). The assumption is mild since we can set a small \( K \) (like 2 or 3) even without prior knowledge. An intuitive understanding is that: if \( K \) is too large, it is difficult for the meta expert to derive effective advice because of the dilution effect from those weak experts; if \( B \) is too large, the samples in an interval may come from various distributions, and the assumption about the setting may not hold.

**Generalization Error Bound**

The MOOE performance is measured by the excess risk \( L_D(w) - L_D(w^*) \) where \( w = \frac{1}{T} \sum_{t=1}^{T} w_t \) is the average of the online interval. To derive an algorithmic bound, we introduce the intermediate term \( L_\ell(\hat{w}) \) because \( \hat{w} \) as an empirical minimizer of \( L_\ell(\cdot) \) is necessary for analyzing the regret. Taking the divide-and-conquer approach, we have

\[
L_D(w) - L_D(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} L_D(w_t) - L_D(w^*) = \frac{1}{T} \sum_{t=1}^{T} L_D(w_t) - L_D(\hat{w}) + L_D(\hat{w}) - L_D(w^*). \tag{14}
\]

The inequality is owing to the convexity of \( L_D(\cdot) \), which implies \( L_D(\frac{1}{T} \sum_{t=1}^{T} w_t) \leq \frac{1}{T} \sum_{t=1}^{T} L_D(w_t) \).

The regret of MOOE is applied to imply the upper bound of \( B_1 \) by the following lemma.

**Lemma 2** Following Theorem 2, with probability at least \( 1 - \delta \), we have

\[
\frac{1}{T} \sum_{t=1}^{T} L_D(w_t) - L_D(\hat{w}) \leq \frac{\sqrt{\ln K} + 6D\sqrt{\beta} + 4\log(4/\delta)}{\sqrt{T}}.
\]
Following the advanced study for any norm-regularized hypothesis class (Yousefi et al. 2018) and the self-bound property of smooth functions (Srebro, Sridharan, and Tewari 2010), we can derive the following data-dependent generalization bound for \( \mathcal{B}_2 \) by the following lemmas.

**Lemma 3** Exploiting the convexity, smoothness, and non-negativity conditions of the loss function family \( \mathcal{L} \), with probability at least \( 1 - \delta \), \( L_D(\bar{w}) - L_D(w^*) \) is bounded by

\[
\frac{(12\beta R^2 + 4R\sqrt{\beta}) \log(4/\delta)}{T} + 4R\sqrt{2 \beta \log(4/\delta)} \cdot \frac{\log(4/\delta)}{T}.
\]

**Lemma 4** Exploiting the hypothesis class \( \mathcal{H} \) and the distribution \( D \) of the observed data at the online interval, with probability at least \( 1 - \delta \), \( L_D(\bar{w}) - L_D(w^*) \) is bounded by

\[
42\sqrt{6\beta \log \frac{1}{\delta}} \cdot \frac{\log(4/\delta)}{T} + 3\sqrt{\frac{\log(4/\delta)}{T}}.
\]

where \( \mathcal{R}_D(\mathcal{H}) \) is the Rademacher complexity of hypothesis space \( \mathcal{H} \).

Using the excess risk bound framework in Eq. (14), we obtain the following generalization error bound by considering Lemma 2, Lemma 3 and Lemma 4.

**Theorem 3** Exploiting the loss function properties (convexity, smoothness, and nonnegativity) of \( \mathcal{L} \), the hypothesis class \( \mathcal{H} \), the data distribution \( D \) and the regret of MOOE, with probability at least \( 1 - \delta \), we have

\[
L_D(\bar{w}) - L_D(w^*) \leq \frac{(12\beta R^2 + 4R\sqrt{\beta}) \log(16/\delta)}{T} + 28R\sqrt{\beta} \log^{\frac{3}{2}}(64T) \left( \sum_{i=1}^{\infty} (TD^2 \land e\lambda_i) + D\sqrt{\epsilon} \right) + \frac{6(R + D)\sqrt{\beta} + 2}{\sqrt{T}} \sqrt{\log(16/\delta) + 4\log(8/\delta) + \ln K}.
\]

The convergence rate for the generalization error is \( O(1/\sqrt{T}) \), which is consistent with that in stationary and non-algorithmic cases (Kakade, Sridharan, and Tewari 2008). The bound reflects the best result achieved so far without any other assumptions. Furthermore, it is directly related to the sample complexity, and the result is algorithmic. This result triggers an immediate problem: Can we use fewer samples to achieve a desirable generalization error if the used off-the-shelf online optimization method can achieve a better regret? Unfortunately, the answer is not affirmative. The intuition behind the problem is that the bottleneck is not on the optimization method.

**Experimental Results**

In this section, we present empirical analysis to support our proposed theory and model.

**Regret on Synthetic and Real-world Datasets**

We address binary classification on non-stationary time series and compare MOOE with OGD using both synthetic and real-world datasets (ijcnn and cod-rna) from the LIBSVM repository (Chang and Lin 2011). On the synthetic dataset, each interval features samples from two-dimensional Gaussian distributions with dynamically changing means. For the synthetic datasets, we divide the data into 15 intervals, applying Gaussian noise to simulate dynamic changes. The samples from the first 10 intervals are presented in Fig. 3. We maintain a maximum of five experts in MOOE to ensure fairness in comparison. Theoretical analyses show that MOOE outperforms OGD in dynamic environments, maintaining a convergence rate of \( O(1/\sqrt{T}) \). As shown in Fig. 2, Empirical results indicate that MOOE achieves significantly lower loss than OGD, particularly at the early stages with few samples, due to the integration of offline expert knowledge. Additionally, MOOE exhibits smaller regret over time, adapting effectively to samples by adjusting the strategy of integrating offline and online experts.

**Predictive Accuracy**

**Real-world Non-stationary Time Series** To verify the effect of the proposed MOOE method, we perform comparison experiments following the setup (Zhao, Cai, and Zhou 2018). Specifically, we use eight real-world non-stationary time series datasets, including Usenet (Katakis, Tsoumakas, and Vlahavas 2008), Weather (Elwell and Polikar 2011), GasSensor (Vergara et al. 2012), Powersupply (Dau et al. 2019), Electricity (Harries 1999), Covertype (Sun et al. 2018), WESAD (Schmidt et al. 2018), and Kitsune (Mirsky et al. 2018). We compare MOOE with three state-of-the-

| Dataset       | NSE  | DTEL | Condor | MOOE |
|---------------|------|------|--------|------|
| Usenet        | 63.8 | 68.0 | 73.1   | **78.5** |
| Weather       | 76.0 | 68.9 | 79.4   | **82.4** |
| GasSensor     | 42.4 | 63.8 | 81.6   | **83.1** |
| Powersupply   | 74.0 | 69.9 | 72.8   | **77.9** |
| Electricity   | 79.0 | 81.0 | 84.7   | **85.6** |
| Covertype     | 79.0 | 69.4 | 89.6   | **90.0** |
| WESAD         | 70.4 | 73.8 | 86.3   | **89.9** |
| Kitsune       | 73.9 | 71.6 | 87.3   | **93.4** |
art methods, including NSE (Elwell and Polikar 2011), DTEL (Sun et al. 2018), and Condor (Zhao, Cai, and Zhou 2018). In the experiments, we adopt the maximum sample size of an interval $B = 50$ and the maximal number of maintained experts $K_{\text{max}} = 25$. The overall mean of predictive accuracy is reported, which indicates the average performance of the algorithm over the whole time series. The comparison results are reported in Table 1. The results show that the proposed MOOE method outperforms other contenders. Specifically, MOOE achieves 13.7% improvement over the other state-of-the-art methods. This is because MOOE can utilize the knowledge of the offline experts to adopt each new online interval. These experimental results show the superiority of the proposed MOOE method.

### Table 2: Comparisons on non-stationary time series with recurring concept drift.

| Algorithm | Email list | Spam filtering |
|-----------|------------|----------------|
|           | Accuracy   | Precision | Recall |
| NSE       | 70.0       | 76.5      | 76.5  |
| DTEL      | 86.2       | 88.2      | 88.2  |
| Condor    | 95.6       | 93.2      | 99.8  |
| MOOE      | 97.1       | 94.0      | 99.8  |

### Non-stationary Time Series with Recurring Concept Drift

To verify the versatility of the proposed MOOE method, we conduct the comparisons on a special case of non-stationary time series, i.e., recurring concept drift, in which previous distributions may disappear and then re-appear in the future. We consider two real-world non-stationary time series with recurring concept drift, namely Email list and Spam filtering (Katakis, Tsoumakas, and Vlahavas 2010). The concepts are decided by the personal interests of users that change in a recurring manner. The results are summarized in Table 2, which show that MOOE exhibits an encouraging performance on the two datasets regarding all measures. Specifically, MOOE achieves 11.8% improvement in terms of accuracy. This is because the offline experts can learn the knowledge on the previous distributions, and the meta expert can reuse the knowledge when the distribution re-appear.

### Non-stationary Time Series with Increasing Levels of Noise

To verify the robustness of the proposed MOOE method, we perform experiments on non-stationary time series with increasing levels of noise. Specifically, we adopt Covertype and gradually add Gaussian noise until the time series becomes completely random. The experiment results presented in Table 3 indicate that MOOE achieves the best prediction result. This is because MOOE can utilize the knowledge learned by the offline experts when the online expert is hard to learn knowledge from the noisy data.

### Effect of Maximal Number of Maintained Experts

To verify the effect of the hyper-parameter $K_{\text{max}}$, we select its value from $\{5, 10, 15, 20, 25, 30, 35\}$ and perform experiments on Covertype. The experimental results presented in Fig. 4 show that increasing $K_{\text{max}}$ can improve the classification performance and the performance stabilizes when $K_{\text{max}}$ is sufficiently large, e.g., $K_{\text{max}} = 25$. This is because a large number of maintained experts $K_{\text{max}}$ causes more knowledge can be stored and applied for the data in the online interval. Furthermore, if $K_{\text{max}}$ is sufficiently large, increasing its number cannot make MOOE obtain the new knowledge that the offline experts have not explored.

### Conclusion

In this paper, we address a general and realistic scenario involving non-stationary time series, where several offline intervals with various distributions exist alongside an online interval. We propose MOOE, which employs a meta expert to integrate static offline experts, learned from previous offline intervals, with the dynamic online expert, updated in the online interval. We provide theoretical guarantees regarding parameter convergence, regret bounds, and generalization error bounds. Our theoretical results demonstrate that MOOE achieves the same generalization error bounds in both stationary and non-stationary cases, proving that leveraging knowledge from historical intervals is effective. Future work will explore other assumptions and techniques to overcome bottlenecks in the generalization bound.
Acknowledgments

This work was supported by the Australian Research Council through the Linkage Grant (LP230201022), the Discovery Grant (DP240102050), and the Linkage Infrastructure, Equipment, and Facilities Grant (LE240100131).

References

Adams, T. M.; and Nobel, A. B. 2010. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. The Annals of Probability, 38(4): 1345–1367.

Agarwal, A.; and Duchi, J. C. 2013. The Generalization Ability of Online Algorithms for Dependent Data. IEEE Trans. Inf. Theory, 59(1): 573–587.

Anagnostides, I.; Farina, G.; and Sandholm, T. 2023. Near-Optimal $\Phi$-Regret Learning in Extensive-Form Games. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., ICML, 814–839.

Bartlett, P. L.; and Mendelson, S. 2002. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. J. Mach. Learn. Res., 3: 463–482.

Besbes, O.; Gur, Y.; and Zeevi, A. J. 2015. Non-stationary Stochastic Optimization. Operations Research, 63(5): 1227–1244.

Boucheron, S.; Lugosi, G.; and Massart, P. 2013. Concentration Inequalities - A Nonasymptotic Theory of Independence. Oxford university press.

Buchbinder, N.; Chen, S.; Naor, J.; and Shamir, O. 2016. Unified Algorithms for Online Learning and Competitive Analysis. Math. Oper. Res., 41(2): 612–625.

Cesa-Bianchi, N.; and Lugosi, G. 2006. Prediction, Learning, and Games. Cambridge University Press.

Chang, C.; and Lin, C. 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., 2(3): 1–27.

Dau, H. A.; Bagnall, A.; Kamgar, K.; Yeh, C.-C. M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C. A.; and Keogh, E. 2019. The UCR Time Series Archive.

Elwell, R.; and Polikar, R. 2011. Incremental Learning of Concept Drift in Nonstationary Environments. IEEE Trans. Neural Networks, 22(10): 1517–1531.

Hamilton, J. D. 1994. Time Series Analysis. Princeton.

Hanneke, S. 2016. The Optimal Sample Complexity of PAC Learning. J. Mach. Learn. Res., 17(38): 1–15.

Harries, M. B. 1999. SPLICE-2 Comparative Evaluation: Electricity Pricing. In Technical Report of South Wales University.

Hazan, E. 2016. Introduction to Online Convex Optimization. Foundations and Trends in Optimization, 2(3-4): 157–325.

Herbster, M.; and Warmuth, M. K. 1995. Tracking the Best Expert. In ICML, 286–294.

Hoffmann-Jørgensen, J.; Kuelbs, J.; and Marcus, M. B. 2012. On the Rademacher Series, Probability in Banach spaces, 9, volume 35. Springer Science & Business Media.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2008. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In NeurIPS, 793–800.

Katakis, I.; Tsoumakas, G.; and Vlahavas, I. P. 2008. An Ensemble of Classifiers for coping with Recurring Contexts in Data Streams. In ECAI, 763–764.

Katakis, I.; Tsoumakas, G.; and Vlahavas, I. P. 2010. Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowl. Inf. Syst., 22(3): 371–391.

Kloft, M.; and Blanchard, G. 2012. On the convergence rate of lp-norm multiple kernel learning. J. Mach. Learn. Res., 13: 2465–2502.

Kuznetsov, V.; and Mohri, M. 2014. Generalization Bounds for Time Series Prediction with Non-stationary Processes. In ALT, 260–274.

Kuznetsov, V.; and Mohri, M. 2015. Learning theory and algorithms for forecasting non-stationary time series. In NeurIPS, 541–549.

Mirsy, Y.; Doitschin, T.; Elovici, Y.; and Shabtai, A. 2018. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. In NDSS, 1–15.

Mohri, M.; and Rostamizadeh, A. 2008. Rademacher Complexity Bounds for Non-I.I.D. Processes. In NeurIPS, 1097–1104.

Mohri, M.; and Rostamizadeh, A. 2010. Stability Bounds for Stationary phi-mixing and beta-mixing Processes. J. Mach. Learn. Res., 11: 789–814.

Nesterov, Y. 2004. Introductory lectures on convex optimization: a basic course, volume 87 of Applied optimization. Kluwer Academic Publishers.

Puigcerver, J.; Ruiz, C. R.; Mustafa, B.; and Houlsby, N. 2024. From Sparse to Soft Mixtures of Experts. In ICLR.

Schmidt, P.; Reiss, A.; Dürichen, R.; Marberger, C.; and Laerhoven, K. V. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In NeurIPS.

Shalev-Shwartz, S. 2012. Online Learning and Online Convex Optimization. Foundations and Trends in Machine Learning, 4(2): 107–194.

Shalev-Shwartz, S.; and Ben-David, S. 2014. Understanding Machine Learning From Theory to Algorithms. Cambridge University Press.

Smale, S.; and Zhou, D.-X. 2009. Geometry on Probability Spaces. Constr. Approx., 30: 311–323.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Optimistic Rates for Learning with a Smooth Loss. In ArXiv, arXiv:1009.3896.

Steve, S.; and Zhou, D.-X. 2007. Learning Theory Estimates via Integral Operators and Their Approximations. Constr. Approx., 26(2): 153–172.

Sun, Y.; Tang, K.; Zhu, Z.; and Yao, X. 2018. Concept Drift Adaptation by Exploiting Historical Knowledge. IEEE Trans. Neural Networks Learn. Syst., 29(10): 4822–4832.

van Erven, T.; and Koelen, W. M. 2016. MetaGrad: Multiple Learning Rates in Online Learning. In NeurIPS, 3666–3674.
of stationary mixing sequences.
Yu, B. 1994. Rates of convergence for empirical processes. *J. Mach. Learn. Res.*

Complexity-based Learning Guarantees for Multi-Task Learning. Anagnostopoulos, G. C. 2018. Local Rademacher complexity.

Yousefi, N.; Lei, Y.; Kloft, M.; Mollahasemi, M.; and Anagnostopoulos, G. C. 2018. Local Rademacher complexity-based Learning Guarantees for Multi-Task Learning. *J. Mach. Learn. Res.*, 19: 1–47.

Yu, B. 1994. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1): 94–116.

Zhang, L.; Lu, S.; and Zhou, Z. 2018. Adaptive Online Learning in Dynamic Environments. In *NeurIPS*, 1330–1340.

Zhang, L.; Yang, T.; and Jin, R. 2017. Empirical Risk Minimization for Stochastic Convex Optimization: O(1/n2)-type of Risk Bounds. In *COLT*, 1954–1979.

Zhao, P.; Cai, L.; and Zhou, Z. 2018. Handling concept drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:167–320.

M. L.; and Huerta, R. 2012. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:167–320.

We, C.; Hong, Y.; and Lu, C. 2019. Tracking the Best Expert in Non-stationary Stochastic Environments. In *NeurIPS*, 3972–3980.

Proofs

In this section, we present the proofs of all the theorems and lemmas. Our analysis follows some advanced techniques, including the self-bound property of smooth functions (Srebro, Sridharan, and Tewari 2010), the analysis of adaptive online optimization method with multiple learning rates (van Erven and Koolen 2016), the connection between agnostic PAC learning and online convex optimization (Hazan 2016), empirical risk minimization for stochastic convex optimization (Zhang, Yang, and Jin 2017), and the bound of Rademacher complexity for any norm-regularized hypothesis class (Yousefi et al. 2018).

Proof of Lemma 1

According to the property of strong convexity (Shalev-Shwartz and Ben-David 2014, Lemma 13.5.2), we know $L_S(w)$ is convex and $\frac{\gamma}{2} \Omega(w)$ is strongly convex. Accordingly, we have

$$
\frac{\gamma}{2} \Omega(w^K) \leq L_S \left( \sum_{k=1}^{K} \alpha_{B_k} w_{B_k}^k \right) + \frac{\gamma}{2} \Omega \left( \sum_{k=1}^{K} \alpha_{B_k} w_{B_k}^k \right)
$$

$$
- L_S(w^K) - \frac{\gamma}{2} \Omega(w^K)
$$

$$
\leq \sum_{k=1}^{K} \alpha_{B_k} L_S (w_{B_k}^k) + 0 - \frac{\gamma}{2} \Omega(w^K).
$$

(15)

Above, the first inequality is due to the strongly-convex property that $\frac{\gamma}{2} \|x - x^*\|_2^2 \leq f(x) - f(x^*)$ (Shalev-Shwartz and Ben-David 2014, Lemma 13.5.3) and $w^K$ is an empirical minimizer of $L$ and $\gamma$ is the second inequality uses the condition that $L_S(w^K) \geq 0$ as assumed.

Proof of Theorem 1

By using the fact that $w^*$ minimizes $L_D(w)$ over the domain $\mathcal{W}$, we have

$$
L_D(w^*) - L_D(w^K) \leq 0.
$$

(16)

According to the property of strong convexity (Shalev-Shwartz and Ben-David 2014, Lemma 13.5.2), $L_D(w) + \frac{\gamma}{2} \Omega(w)$ is $\gamma$-strongly convex because the former term is convex and the last term is $\gamma$-strongly convex. Following the definition of strongly convex function, we have

$$
\frac{\gamma}{2} \|w^K - w^*\|_2^2
$$

$$
\leq L_D(w^*) + \frac{\gamma}{2} \Omega(w^*) - L_D(w^K) - \frac{\gamma}{2} \Omega(w^K)
$$

$$
+ \left(\nabla L_D(w^K) + \frac{\gamma}{2} \Omega(w^K) - \frac{\gamma}{2} \Omega(w^*)\right), w^K - w^*.
$$

(17)

To upper bound the last term above, we have

$$
\left\langle \nabla L_D(w^K) + \frac{\gamma}{2} \Omega(w^K) - \frac{\gamma}{2} \Omega(w^*)\right\rangle, w^K - w^*
$$

$$
\leq \left( \left\| \nabla L_D(w^K) \right\|_2 + \frac{\gamma}{2} \left\| w^K - \sum_{k=1}^{K} \alpha_{B_k} w_{B_k}^k \right\|_2 \right) \left\| w^K - w^* \right\|_2
$$

$$
\leq \frac{4 \left\| \nabla L_D(w^K) \right\|_2^2}{\gamma} + \frac{\gamma}{2} \left\| w^K - w^* \right\|_2^2
$$

$$
+ \frac{4 \left\| w^K - \sum_{k=1}^{K} \alpha_{B_k} w_{B_k}^k \right\|_2^2}{\gamma} + \frac{\left\| w^K - w^* \right\|_2^2}{\gamma}.
$$

(18)

where the first inequality uses the Cauchy-Schwarz inequality and the second inequality uses the Young’s inequality $\langle a, b \rangle \leq \frac{1}{2} \|a\|_2^2 + \frac{1}{2} \|b\|_2^2$.

Substituting Eqs. (16) and (18) into Eq. (17), we have

$$
\left\| w^K - w^* \right\|_2^2
$$

$$
\leq 2 \Omega (w^*) + \frac{8}{\gamma^2} \left\| \nabla L_D(w^K) \right\|_2^2 + 6 \Omega (w^K)
$$

$$
\leq 2 \Omega (w^*) + \frac{32L}{\gamma^2} + \frac{6}{\gamma^2} \sum_{k=1}^{K} \alpha_{B_k} L_S (w_{B_k}^k)
$$

(19)

where the second inequality is owing to Eq. (3) and Lemma 1.

Proof of Theorem 2

As shown in Eq. (12), the analysis is divided into two parts. First, we show the Regret of the meta-expert: the difference between the total cost it has incurred and that of the best existing expert of $K$ experts. Then, we
demonstrate the Regret_{oE} of the online expert: the difference between the total cost it has incurred and that of the empirical minimizer. Based on the previous study (Cesa-Bianchi and Lugosi 2006, Theorem 2.2), we define \( W_j = \sum_{k=1}^{K} \alpha_k^j e^{-\nu} \sum_{t=1}^{T} f_t(w_t^k) \) and the lower bound of the related quantities is

\[
\ln \frac{W_T}{W_0} \geq \ln \left( \sum_{k=1}^{K} \alpha_k^1 e^{-\nu} \sum_{t=1}^{T} f_t(w_t^k) \right) - \ln \left( \sum_{k=1}^{K} \alpha_k^1 \right) - \nu \min_{k \in [K]} \left( \sum_{t=1}^{T} f_t(w_t^k) + \frac{1}{\nu} \ln \frac{1}{\alpha_k^1} \right) - \ln \left( \sum_{k=1}^{K} \alpha_k^1 \right).
\]

We obtain Regret_{oE} by combining Eq. (20) with Eq. (23), we have

\[
\sum_{t=1}^{T} f_t(w_t) - \min_{k \in [K]} \left( \sum_{t=1}^{T} f_t(w_t^k) + \frac{1}{\nu} \ln \frac{1}{\alpha_k^1} \right) \leq \frac{T \nu}{8} + \ln \left( \sum_{k=1}^{K} \alpha_k^1 \right)
\]

which implies

\[
\sum_{t=1}^{T} f_t(w_t) - \min_{k \in [K]} \left( \sum_{t=1}^{T} f_t(w_t^k) \right) \leq \frac{T \nu}{8} + \ln \left( \sum_{k=1}^{K} \alpha_k^1 \right) + \max_{k \in [K]} \frac{1}{\nu} \ln \frac{1}{\alpha_k^1}.
\]

With Eq. (7), we have

\[
\sum_{k=1}^{K} \alpha_k^1 = K^2 + 1 = 1,
\]

and for any \( 2 \leq k \leq K \), we have

\[
\frac{K}{K+1} \left( \alpha_k^{k-1} - \alpha_k^1 \right) \geq \frac{1}{K-k} - \frac{1}{K+1-k} - \frac{1}{k+2-k} \geq 0,
\]

so \( \{ \alpha_k^1 \} \) is in the ascending order and

\[
\max_{k \in [K]} \ln \frac{1}{\alpha_k^1} = \ln \frac{1}{\alpha_1^1} = 2 \ln K.
\]

Substitute Eqs. (26) and (28) into Eq. (25), we have

\[
\frac{T \nu}{8} + 2 \ln K,
\]

and minimize this function toward \( \nu \) to obtain

\[
\nu = 4 \sqrt{\ln K}.
\]

Substitute Eqs. (7) and (30) into Eq. (29), we have

\[
\text{Regret}_{\text{ME}} \leq \sqrt{T \ln K}.
\]

We apply the standard OSG to optimize the online expert \( w_t^k \). Following the previous result (Hazan 2016, Theorem 3.1), setting the step size \( \eta_t = \frac{D}{\sqrt{T}} \), we have

\[
\text{Regret}_{\text{oE}} \leq 6D \sqrt{T \beta}.
\]

We obtain \( \text{Regret}_{\text{CO2}} \) by combining Eq. (31) with Eq. (32). According to Eq. (12) and the requirement that \( \text{CO2} \) should surpass its online expert, we obtain the upper bound of \( K \) by solving the following inequality

\[
\sqrt{T \ln K} + \text{Regret}_{\text{KE}} \leq 6D \sqrt{T \beta}.
\]
**Proof of Lemma 2**

To proceed, we introduce the following norm concentration inequality.

**Lemma 6** (Smale and Zhou 2009, Proposition 1) Let \( \xi \) be a random variable on \((Z, \rho)\) with values in a Hilbert space and be randomly drawn according to \( \rho \) satisfying \( \|\xi\| \leq M \leq \infty \). Then, for any \( 0 < \delta < 1 \), with a probability at least \( 1 - \delta \),

\[
\left\| \frac{1}{m} \sum_{i=1}^{m} (\xi_i - \mathbb{E}[\xi_i]) \right\| \leq \frac{4M}{\sqrt{m}} \log \frac{2}{\delta}.
\]

Using Lemma 6 with the results \( \|\frac{1}{T} \sum_{t=1}^{T} f_t(w_t)\| \leq 1 \) and \( \|\frac{1}{T} \sum_{t=1}^{T} f_t(w^*)\| \leq 1 \) implied from Assumption 4, with probability at least \( 1 - \delta \), we have

\[
\frac{1}{T} \sum_{t=1}^{T} L_D(w_t) \leq \frac{1}{T} \sum_{t=1}^{T} f_t(w_t) + \frac{4 \log(2/\delta)}{\sqrt{T}} + \frac{1}{T} \sum_{t=1}^{T} f_t(\tilde{w}) \leq \frac{1}{T} \sum_{t=1}^{T} L_D(\tilde{w}) + \frac{4 \log(2/\delta)}{\sqrt{T}}.
\]  

(33)

Putting the two results in Eq. (33) together, with probability at least \( 1 - \delta \), we have

\[
\frac{1}{T} \sum_{t=1}^{T} L_D(w_t) - L_D(\tilde{w}) \leq \frac{1}{T} \left( \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(\tilde{w}) \right) + \frac{4 \log(4/\delta)}{\sqrt{T}} \leq \frac{\sqrt{\ln K}}{\sqrt{T}} + 6D \sqrt{\beta} + \frac{4 \log(4/\delta)}{\sqrt{T}},
\]

(34)

where the first inequality is owing to the Hoeffding’s inequality (Boucheron, Lugosi, and Massart 2013, Theorem 2.8) and the second inequality follows Theorem 2.

**Proof of Lemma 3**

Our analysis is based on the techniques used in (Zhang, Yang, and Jin 2017). For simplicity, we assume \( \hat{b}_t = \nabla f_t(w^*) \) and \( \hat{b}_t = \nabla f_t(\tilde{w}) - \nabla f_t(w^*) \), so we know

\[
\nabla L_S(w^*) = \frac{1}{T} \sum_{t=1}^{T} \hat{b}_t,
\]

\[
\nabla L_D(w^*) = \mathbb{E}[\hat{b}],
\]

\[
\nabla L_S(\tilde{w}) - \nabla L_S(w^*) = \frac{1}{T} \sum_{t=1}^{T} \hat{b}_t,
\]

\[
\nabla L_D(\tilde{w}) - \nabla L_D(w^*) = \mathbb{E}[\hat{b}].
\]

By the Karush-Kuhn-Tucker (KKT) (Hazar 2016, Theorem 2.2) condition for convex function \( L_S(\cdot) \) and \( L_D(\cdot) \), we have

\[
\langle \nabla L_S(\tilde{w}), \tilde{w} - \tilde{w} \rangle \geq 0, \langle \nabla L_D(w^*), w - w^* \rangle \geq 0, \forall w, \tilde{w} \in \mathcal{W}.
\]

(36)

We first upper bound the excess risk \( L_D(\tilde{w}) - L_D(w^*) \) by two terms with the same form and then further derive the upper bounds of the two terms by similar methods.

\[
L_D(\tilde{w}) - L_D(w^*) \leq \langle \nabla L_D(\tilde{w}), \tilde{w} - w^* \rangle = \mathbb{E}[\hat{b}] + \frac{1}{T} \sum_{t=1}^{T} \hat{b}_t \| \tilde{w} - w^* \| \leq \mathbb{E}[\hat{b}] + \frac{1}{T} \sum_{t=1}^{T} \hat{b}_t \| \tilde{w} - w^* \| + \mathbb{E}[\hat{b}] + \frac{1}{T} \sum_{t=1}^{T} \hat{b}_t \| w - w^* \|.
\]

(37)

In the above, the first inequality is owing to the convexity of \( L_D(\cdot) \) over the domain \( \mathcal{W} \), the second inequality applies Eq. (37) for convex function \( L_S(\cdot) \) with respect to \( w^* \); \( \| \nabla L_S(\tilde{w}), \tilde{w} - w^* \| \leq 0 \) and uses \( \mathbb{E}[\hat{b}] + \frac{1}{T} \sum_{t=1}^{T} \hat{b}_t \| \tilde{w} - w^* \| \leq \frac{\sqrt{\ln K}}{\sqrt{T}} + 6D \sqrt{\beta} + \frac{4 \log(4/\delta)}{\sqrt{T}} \); the third inequality uses the Cauchy-Schwarz inequality.

Note that \( B_1 \) and \( B_2 \) have the same structure. To bound the variance terms in \( B_1 \) and \( B_2 \), we introduce the following norm concentration inequality in a Hilbert space.

**Lemma 7** (Steve and Zhou 2007, Lemma 2) Let \( \xi \) be a random variable on \((Z, \rho)\) with values in a Hilbert space, assume \( \|\xi\| \leq M \leq \infty \) almost surely holds, denote \( \sigma^2(\xi) = \mathbb{E}(|\xi|^2) \), and let \( \xi_i \) be independent random draws of \( \rho \), for any \( 0 \leq \delta \leq 1 \), with confidence \( 1 - \delta \),

\[
\left\| \frac{1}{m} \sum_{i=1}^{m} (\xi_i - \mathbb{E}[\xi_i]) \right\| \leq \frac{2M \log(2/\delta)}{m} + \sqrt{\frac{2\sigma^2(\xi) \log(2/\delta)}{m}}.
\]

To use this lemma to upper bound \( B_1 \) and \( B_2 \), we need the bounds for \( \|\hat{b}_t\|, \mathbb{E}[\|\hat{b}_t\|^2], \|\hat{b}_t\|, \mathbb{E}[\|\hat{b}_t\|^2] \). From Eq. (3), we have

\[
\|\hat{b}_t\| \leq 2\sqrt{\beta}, \mathbb{E}[\|\hat{b}_t\|^2] \leq 4\beta.
\]

(38)

With Assumption 5, we have

\[
\|\hat{b}_t\| \leq \beta \|\tilde{w} - w^*\|.
\]

Based on that, by using the properties of smooth function (Nesterov 2004, Theorem 2.1.5), we have

\[
\|\hat{b}_t\|^2 \leq 2\beta (f_t(\tilde{w}) - f_t(w^*) - \langle \nabla f_t(\tilde{w}), \tilde{w} - w^* \rangle).
\]

(39)

Taking the expectation on both sides, we have

\[
\mathbb{E}[\|\hat{b}_t\|^2] \leq 2\beta (L_D(\tilde{w}) - L_D(w^*) - \langle \nabla L_D(\tilde{w}), \tilde{w} - w^* \rangle) \leq 2\beta (L_D(\tilde{w}) - L_D(w^*)).
\]

(40)
where the last inequality applies Eq. (37) to the convex function $L_D(\cdot)$ with respect to $\hat{w}$.

Based on Lemma 7, we establish the uniform convergence of $\frac{1}{T}\sum_{t=1}^{T} \hat{b}_t$ to $\mathbb{E}[\hat{b}]$ and $\frac{1}{T}\sum_{t=1}^{T} \hat{b}_t$ to $\mathbb{E}[\hat{b}]$. By using Eqs. (39) and (40) with Lemma 7, with probability at least $1 - \delta$, we have

$$B_1 \leq \frac{2\beta \| \hat{w} - w^* \|^2 \log(2/\delta)}{T} + 2\| \hat{w} - w^* \| \sqrt{\frac{2L_D(\hat{w}) - L_D(w^*) \log(2/\delta)}{T}} + \frac{3\beta \| \hat{w} - w^* \|^2 \log(2/\delta)}{2T} \leq \frac{12\beta R^2 \log(2/\delta)}{T} + \frac{L_D(\hat{w}) - L_D(w^*)}{2},$$

(41)

where the second inequality uses the Young’s inequality and the third inequality is owing to that Assumption 3 implies $\| \hat{w} - w^* \| \leq 2R$. By using the same method as above with Eq. (38), with probability at least $1 - \delta$, we have

$$B_2 \leq \frac{4\sqrt{3} \log(\beta/\delta)}{T} + 4\sqrt{\frac{2\beta \log(2/\delta)}{T}}.$$

(42)

We complete the proof by substituting Eqs. (41) and (42) into Eq. (37).

**Proof of Lemma 4**

Based on the divide-and-conquer idea, we split our target into three more comfortable summands and derive their bounds respectively,

$$L_D(\hat{w}) - L_D(w^*) = (L_D(\hat{w}) - L_S(\hat{w})) + (L_S(\hat{w}) - L_S(w^*)) + (L_S(w^*) - L_D(w^*)).$$

(43)

Recall that $l$ is a nonnegative function. To bound the generation error $L_D(\hat{w}) - L_S(\hat{w})$ by the Rademacher complexity for this nonnegative function, we need the following two lemmas.

**Lemma 8** (Shalev-Shwartz and Ben-David 2014, Theorem 26.5) Assume that the loss function $l(h,z)$ is bounded by $b$ for all $z$ and $h \in \mathcal{H}$. $S$ is a $m$-samples data set. With probability of at least $1 - \delta$, for all $h \in \mathcal{H}$,

$$L_D(h) - L_S(h) \leq 2\mathcal{R}(l \circ \mathcal{H}) + b \sqrt{\frac{2\log(2/\delta)}{m}}.$$

**Lemma 9** (Srebro, Sridharan, and Tewari 2010, Lemma 2.2) For a nonnegative $\beta$-smooth loss $l$ bounded by $b$ and any function class $\mathcal{H}$, the Rademacher Complexity on a $m$-samples data set is

$$\mathcal{R}(l \circ \mathcal{H}) \leq 2\sqrt{6\beta b \log^3(64m)} \mathcal{R}(\mathcal{H}).$$

With Assumption 4, we know $|l((\hat{w}, x), y)| \leq 1$. By combining Lemmas 8 and 9 under the condition that $\hat{w} \in \mathcal{W}$, with probability at least $1 - \delta$, we have

$$L_D(\hat{w}) - L_S(\hat{w}) \leq 2\mathcal{R}_D(l \circ \mathcal{W}) + \sqrt{\frac{2\log(2/\delta)}{T}} \leq 42\sqrt{6\beta \log^3(64T)} \mathcal{R}_D(\mathcal{W}) + \sqrt{\frac{2\log(2/\delta)}{T}}.$$

(44)

Note that the above result ignores the specificity of $\hat{w}$. We utilize the specificity of $\hat{w}$ as well as $w^*$ to bound the next two summands.

Recall that $\hat{w}$ is an empirical minimizer of $L_S(w)$ over the domain $\mathcal{W}$, we have

$$L_S(\hat{w}) - L_S(w^*) \leq 0.$$

(45)

Because $w^*$ is independent of the data set $S$ and $f_1(w^*), \ldots, f_T(w^*)$ is a sequence of i.i.d. random variables, we have $E_S[L_S(w^*)] = L_D(w^*)$. With Assumption 4, we know that $P[0 \leq f_t(w^*) \leq 1]$ is true for every $t \in [T]$. The Hoeffding’s inequality (Boucheron, Lugosi, and Massard 2013, Theorem 2.8) implies that with probability at least $1 - \delta$, we have

$$L_S(w^*) - L_D(w^*) \leq \sqrt{\frac{\log(2/\delta)}{2T}}.$$

(46)

We complete the proof by substituting Eqs. (45), (46) and (44) into Eq. (43).

**Proof of Theorem 3**

To proceed, we introduce the following lemma.

**Lemma 10** Let $S$ be a set of i.i.d. samples from the distribution $D$, the Rademacher complexity $\mathcal{R}$ of hypothesis class $\mathcal{H}$ w.r.t. distribution $D$ at the online interval is bounded as

$$\mathcal{R}_D(\mathcal{H}) \leq R \sqrt{\frac{1}{T} \sum_{i=1}^{\infty} \left( D^2 / \lambda_i / T \right) + DR \sqrt{\epsilon} / T}.$$

We complete the proof by substituting the results of Lemmas 2, 3, 10 and 4 into the excess risk bound framework in Eq. (14).

**Proof of Lemma 10**

By using the definition of the Rademacher complexity (Bartlett and Mendelson 2002), we have

$$\mathcal{R}_D(\mathcal{W}) = E_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t \langle \hat{w}, x_t \rangle \right].$$

Let $\Gamma_i = \sum_{t=1}^{T} \sigma_t x_t / T$ for any $i \geq 1$ and $\theta \geq i$, by using advanced techniques of the Rademacher complexities (Yousefi et al. 2018), we have

$$E_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t \langle \hat{w}, x_t \rangle \right] = \mathcal{E}_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \left( \sum_{t=1}^{T} \sigma_t \langle \hat{w}, x_t \rangle \right) \right] \leq \mathcal{E}_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \left( \sum_{t=1}^{T} \sigma_t \langle \hat{w}, x_t \rangle \right) \right] + \mathcal{E}_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \left( \sum_{t=1}^{T} \Gamma_i u_i \right) \right] \leq \mathcal{E}_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \left( \sum_{t=1}^{T} \Gamma_i u_i \right) \right] + \mathcal{E}_{S, \sigma} \left[ \sup_{\hat{w} \in \mathcal{W}} \left( \sum_{t=1}^{T} \Gamma_i u_i \right) \right].$$
The above inequality is owing to the Jensen’s inequality. Based on the Cauchy-Schwarz inequality, we have the following upper bounds for the last two terms

\[
\mathbb{E}_{s, \sigma} \left[ \sup_{w \in \mathcal{W}} \left\langle w, \sum_{i=1}^{\theta} \Gamma_i u_i \right\rangle \right] \\
= \mathbb{E}_{s, \sigma} \left[ \sup_{w \in \mathcal{W}} \left\langle \sum_{i=1}^{\theta} \sqrt{\lambda_i} \left\langle w, u_i \right\rangle u_i, \sum_{i=1}^{\theta} \frac{1}{\sqrt{\lambda_i}} \Gamma_i u_i \right\rangle \right] \\
\leq \mathbb{E}_{s, \sigma} \left[ \sup_{w \in \mathcal{W}} \left\| \sum_{i=1}^{\theta} \sqrt{\lambda_i} \left\langle w, u_i \right\rangle u_i \right\|_2 \left\| \sum_{i=1}^{\theta} \frac{1}{\sqrt{\lambda_i}} \Gamma_i u_i \right\|_2 \right] \\
= \sup_{w \in \mathcal{W}} \left\{ \sum_{i=1}^{\theta} \lambda_i \left\langle w, u_i \right\rangle^2 \mathbb{E}_{s, \sigma} \left[ \sum_{i=1}^{\theta} \frac{1}{\lambda_i} \Gamma_i^2 \right] \right\} \\
:= U_1 \\
=: U_2
\]

and

\[
\mathbb{E}_{s, \sigma} \left[ \sup_{w \in \mathcal{W}} \left\langle w, \sum_{i=\theta+1}^{\infty} \Gamma_i u_i \right\rangle \right] \\
\leq \sup_{w \in \mathcal{W}} \left\| w \right\| \mathbb{E}_{s, \sigma} \left[ \sum_{i=\theta+1}^{\infty} \Gamma_i u_i \right] \right] \\
=: U_3
\]

**Bounding U₁:** Enlarging U₁ by replacing θ with ∞, we have the following upper bound

\[
U_1 \leq \sup_{w \in \mathcal{W}} \left\{ \sum_{i=1}^{\infty} \lambda_i \left\langle w, u_i \right\rangle \left\langle w, u_i \right\rangle \right\} \\
= \sup_{w \in \mathcal{W}} \sqrt{\mathbb{E} \left( {w}_T \left| {x}_T \right| \right)} \leq DR,
\]

where the last inequality uses Assumptions 3 and 2.

**Bounding U₂:** By using the Jensen’s inequality, we have

\[
U_2 \leq \sum_{i=1}^{\theta} \mathbb{E}_{s, \sigma} \left[ \frac{1}{\lambda_i} \left\langle \frac{1}{T} \sum_{t=1}^{T} \sigma_t x_t, u_i \right\rangle^2 \right] \\
= \sum_{i=1}^{\theta} \mathbb{E}_{s, \sigma} \left[ \frac{1}{\lambda_i T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} \sigma_t \sigma_{t'}' \left\langle x_t, u_i \right\rangle \left\langle x_{t'}, u_i \right\rangle \right] \\
= \sum_{i=1}^{\theta} \frac{1}{\lambda_i T} \left\langle \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_s \left[ x_t x_t' \right], u_i u_i' \right\rangle \\
= \sum_{i=1}^{\theta} \frac{1}{\lambda_i T} \left\langle \sum_{j=1}^{\infty} \lambda_j u_j u_j', u_i u_i' \right\rangle \\
= \sum_{i=1}^{\theta} \frac{1}{\lambda_i T} \lambda_i = \sqrt{\frac{\theta}{T}}.
\]

**Bounding U₃:** Recall the definition of \( \Gamma_i \), we have

\[
U_3 \leq \mathbb{E}_{s, \sigma} \left[ \left\| \sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle u_i \right\|_2^2 \right],
\]

where the inequality is owing to our assumption \( \sup_{w \in \mathcal{W}} \|w\| \leq R \). To derive its bound further, we introduce the Khintchine-Kahane inequality.

**Lemma 11** (Hoffmann-Jorgensen, Kuelbs, and Marcus 2012) Let \( \mathcal{H} \) be an inner-product space with the induced norm \( \| \cdot \|_{\mathcal{H}}, v_1, \ldots, v_n \in \mathcal{H} \) and \( \sigma_1, \ldots, \sigma_n \) i.i.d. Rademacher random variables. Then, for any \( p \geq 1 \), we have

\[
\mathbb{E} \left[ \left\| \sum_{i=1}^{n} \sigma_i v_i \right\|_p^p \right] \leq \left( c \sum_{i=1}^{n} \|v_i\|_{\mathcal{H}}^2 \right)^{\frac{p}{2}}
\]

where \( c := \max(1, p - 1) \). The inequality also holds for \( p \) in place of \( c \).

With Lemma 11, we have

\[
\mathbb{E}_{s, \sigma} \left[ \left\| \sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle u_i \right\|_2^2 \right] \\
\leq \frac{1}{\sqrt{T}} \mathbb{E}_{s, \sigma} \left[ \frac{1}{T} \sum_{t=1}^{T} \sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle \left\langle x_t, u_i \right\rangle \right]^2 \\
= \frac{1}{\sqrt{T}} \mathbb{E}_{s, \sigma} \left[ \frac{1}{T} \sum_{t=1}^{T} \sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle \right]^2.
\]

Although we can bound Eq. (48) further according to the following result obtained from Assumption 2

\[
\sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle \left\langle x_t, u_i \right\rangle = \left\langle \sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle u_i, u_i \right\rangle \leq \left\langle x_t, x_t \right\rangle = D^2,
\]

we pursue a tighter bound by introducing the Rosenthal-Young inequality.

**Lemma 12** (Kloft and Blanchard 2012, Lemma 3) Let the independent nonnegative random variables \( X_1, \ldots, X_n \) satisfy \( X_i \leq B \leq +\infty \) almost surely for all \( i = 1, \ldots, n \), if \( q \geq \frac{1}{2} \), \( c_q := 2(2e)^q \), then the following holds

\[
\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)^q \leq c_q \left( \left( \frac{B}{n} \right)^q + \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} X_i \right)^q \right).
\]

By combining Eq. (49) with Lemma 12, we have

\[
\mathbb{E}_{s, \sigma} \left[ \frac{1}{T} \sum_{t=1}^{T} \sum_{i=\theta+1}^{\infty} \left\langle x_t, u_i \right\rangle \right]^2 \\
\leq \sqrt{e} \left( \frac{D}{\sqrt{T}} + \frac{1}{\sqrt{T}} \sum_{i=\theta+1}^{\infty} \mathbb{E}_{s, \sigma} \left\langle x_t, u_i \right\rangle \right)^2 \\
= D \left( \sqrt{\frac{e}{T}} + \frac{1}{\sqrt{T}} \sum_{i=\theta+1}^{\infty} \lambda_i \right).
\]
To obtain the upper bound of $U_3$, we combine Eqs. (47), (48) and (50).

To sum up, we can complete the proof by

$$
R_D(W) = U_1 \cdot U_2 + U_3 
\leq DR \sqrt{\frac{\theta}{T}} + R \left( e^{\sum_{i=\theta+1}^{\infty} \lambda_i} + \frac{DR \sqrt{e}}{T} \right) 
\leq R \left( \frac{1}{T} \left( D^2 \theta + \sum_{i=\theta+1}^{\infty} \frac{e \lambda_i}{T} \right) + \frac{DR \sqrt{e}}{T} \right) 
= R \left( \frac{1}{T} \left( \sum_{i=1}^{\theta} D^2 + \sum_{i=\theta+1}^{\infty} \frac{e \lambda_i}{T} \right) + \frac{DR \sqrt{e}}{T} \right).
$$

In the above, the first inequality uses the bounds of $U_1$, $U_2$ and $U_3$ we obtained, the second inequality is owing to the inequality of arithmetic and geometric means for nonnegative numbers: $2 \sqrt{xy} \leq x + y \Leftrightarrow \sqrt{x} + \sqrt{y} \leq \sqrt{2x + 2y}$. Note that the above bound for $R_D(W)$ holds for any positive integer $\theta \geq 1$, we obtain the tightest result by minimizing the upper bound, that is

$$
R_D(W) \leq R \left( \frac{1}{T} \min_{\theta \in \mathbb{N}} \left( \sum_{i=1}^{\theta} D^2 + \sum_{i=\theta+1}^{\infty} \frac{e \lambda_i}{T} \right) + \frac{DR \sqrt{e}}{T} \right) 
= R \left( \frac{1}{T} \sum_{i=1}^{\infty} \left( D^2 \land \frac{e \lambda_i}{T} \right) + \frac{DR \sqrt{e}}{T} \right),
$$

where the equality is owing to that the sequence of eigenvalues $\{\lambda_i\}$ is in the ascending order.