Adapting to Smoothness: A More Universal Algorithm for Online Convex Optimization

Guanghui Wang, 1 Shiyin Lu, 1 Yao Hu, 2 Lijun Zhang, 1, *

1 National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China, 2 Youku Cognitive and Intelligent Lab, Alibaba Group, Beijing 100102, China

Abstract

We aim to design universal algorithms for online convex optimization, which can handle multiple common types of loss functions simultaneously. The previous state-of-the-art universal method has achieved the minimax optimality for general convex, exponentially concave and strongly convex loss functions. However, it remains an open problem whether smoothness can be exploited to further improve the theoretical guarantees. In this paper, we provide an affirmative answer by developing a novel algorithm, namely UFO, which achieves \( O(\sqrt{T}) \), \( O(d \log L_s) \) and \( O(\log L_s) \) regret bounds for the three types of loss functions respectively under the assumption of smoothness, where \( L_s \) is the cumulative loss of the best comparator in hindsight, and \( d \) is dimensionality. Thus, our regret bounds are much tighter when the comparator has a small loss, and ensure the minimax optimality in the worst case. In addition, it is worth pointing out that UFO is the first to achieve the \( O(\log L_s) \) regret bound for strongly convex and smooth functions, which is tighter than the existing small-loss bound by an \( O(d) \) factor.

Introduction

Online Convex Optimization (OCO) is a powerful paradigm for modeling sequential decision making (Shalev-Shwartz et al., 2012). It can be considered as a repeated game between a learner and an adversary: In each round \( t = 1, \ldots, T \), firstly the learner chooses an action \( x_t \) from a convex set \( X \subseteq \mathbb{R}^d \), at the same time the adversary reveals a loss function \( f_t : X \mapsto \mathbb{R} \), and consequently the learner suffers a loss \( f_t(x_t) \). The learner’s goal is to minimize regret, which is defined as the cumulative loss of the learner and that of the best action in hindsight (Hazan et al., 2016):

\[
R(T) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in X} \sum_{t=1}^{T} f_t(x).
\]

In the literature, various algorithms have been developed for minimizing the regret under OCO, based on different assumptions on the properties of loss functions, including general convexity (Zinkevich, 2003), exponential concavity (abbr. exp-concavity) and strongly convexity (Hazan et al., 2007). However, existing methods can only deal with one type of loss functions. Moreover, for exp-concave and strongly convex functions, they require prior knowledge of loss functions as inputs for parameter tuning. This lack of universality not only leaves a heavy burden to users, but also impedes the applications to board domains. To overcome this obstacle, recent advances in OCO have developed a series of universal algorithms, such as AOGD (Hazan et al., 2008) and MetaGrad (van Erven and Koolen, 2016), which are able to handle various types of loss functions simultaneously. Among them, the state-of-the-art method is Maler (Wang et al., 2019), which can adapt to general convex, exp-concave and strongly convex functions, and achieve \( O(\sqrt{T}) \), \( O(d \log T) \) and \( O(\log T) \) regret bounds respectively. These bounds are known to be minimax optimal, as matching lower bounds have been established (Ordentlich and Cover, 1998; Abernethy et al., 2008).

On the other hand, in a wide range of online learning tasks such as online least square and \( \ell_2 \)-regularized regressions, the loss functions enjoy the property of smoothness. While there do exist several algorithms that can exploit this property for one single type of loss functions, such as convex and smooth (Srebro et al., 2010), or exp-concave and smooth (Orabona et al., 2012), it remains unclear whether universal algorithms can make use of smoothness to achieve better performance. In this paper, we provide an affirmative answer by developing a novel algorithm, named UFO, which can automatically attain tighter bounds for smooth functions, and thus achieves broader universality.

Following previous universal methods, our proposed algorithm adopts the classic Learning with Experts Framework (LEF) (Cesa-Bianchi and Lugosi, 2006). The basic idea is to maintain multiple algorithms with different learning rates in parallel as experts, and employ a meta-algorithm to identify the best on the fly. However, directly incorporating algorithms for smooth functions into the LEF of universal methods does not provide tight results, because of the following technical challenges:

- The current state-of-the-art method for smooth and
strongly convex functions (Orabona et al., 2012) is sub-optimal in the worst case, as there exists a large $O(d)$ gap.

- In existing universal methods, the experts are performed on a series of surrogate loss functions instead of $f_t$, and thus the smoothness of $f_t$ cannot be exploited.

To address these problems, we first propose the Smooth and Strongly convex Online Gradient Descent (S$^2$OGD) algorithm, and show that it enjoys an $O(\log(\log L_{\star}))$ small-loss regret bound, where $L_{\star} = \min_{x \in X} \sum_{t=1}^T f_t(x)$ is the cumulative loss of the the best action in hindsight. Thus, this bound matches the minimax optimal bound in the worst case, and automatically becomes tighter whenever $L_{\star}$ is small, i.e., $L_{\star} = o(T)$. Then, we develop a Universal algorithm For Online smooth and convex optimization (UFO), which follows from the basic LEF, while employing carefully designed novel surrogate loss functions and expert algorithms to exploit smoothness. Theoretical analysis shows that, under the assumption of smoothness, UFO achieves $O(\sqrt{\log T})$, $O(d \log L_{\star})$ and $O(\log L_{\star})$ regret bounds for general convex, exp-concave, and strongly convex functions, respectively. Moreover, UFO can still attain minimax optimal regret bounds in the worst case, even when the smoothness assumption is violated. Thus, it is strictly better than existing universal methods.

**Notation.** Throughout the paper, we use $\| \cdot \|$ to denote the $\ell_2$-norm. The weighted $\ell_2$-norm associated with a positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$ is defined as $\| x \|^2_A = x^T A x$. Given a positive semidefinite matrix $B$, the $B$-weighted projection $\Pi_B^S[x]$ of $x$ onto $X$ is defined as $\Pi_B^S[x] = \arg\min_{y \in X} \| y - x \|^2_B$. For the sake of clarity, we denote the gradient of $f_t$ at $x_t$ as $g_t$, i.e., $g_t = \nabla f_t(x_t)$. The best action in hindsight is denoted as $x_{\star} = \arg\max_{x \in X} \sum_{t=1}^T f_t(x)$, and the $d$-dimension identity matrix is denoted as $I_d$.

**Related Work**

For general convex functions, the classic Online Gradient Descent (OGD) (Zinkevich, 2003) with step size proportional to $O(1/\sqrt{T})$ (referred to as convex OGD) attains an $O(\sqrt{T})$ regret bound. If the loss functions are strongly convex, OGD with step size on the order of $O(1/t)$ (referred to as strongly convex OGD) achieves a regret bound of $O(\log T)$. For exp-concave functions, the Online Newton Step (ONS) (Hazan et al., 2007) enjoys an $O(d \log T)$ regret bound.

While the above bounds are minimax optimal, tighter bounds are attainable if the loss functions are smooth. Specifically, for general convex and smooth functions, (Srebro et al., 2010) show that OGD with a constant step size attains an $O(\sqrt{L_{\star}})$ regret bound, where $L_{\star}$ is an upper bound of $L_{\star}$. However, this method requires the modulus of smoothness as well as $L_{\star}$ as inputs to tune the step size, which are typically unavailable in practice. To tackle this problem, (Zhang et al., 2019) propose the Scale-free Online Gradient Descent (SOMG) algorithm, which is a special case of the Scale-free Online Mirror Descent (SOMD) algorithm (Orabona and Pál, 2018). SOMG achieves the $O(\sqrt{L_{\star}})$ regret bound, and is parameter-free to the modulus of smoothness and $L_{\star}$. For exp-concave and smooth functions, (Orabona et al., 2012) prove that an $O(d \log L_{\star})$ regret bound can be achieved by employing the ONS algorithm. Although this result also implies an $O(d \log L_{\star})$ regret bound for strongly convex and smooth functions, there still exists an $O(d)$ gap from the $O(\log T)$ lower bound in the worst case. Aside from achieving small-loss bounds, there are studies work on the variation bounds (Hazan and Kale, 2010; Chiang et al., 2012). For convex and smooth functions, (Chiang et al., 2012) propose the extra-gradient descent algorithm, which attains an $O(\sqrt{D_T})$ regret bound, where $D_T = \sum_{t=2}^T \max_{x \in X} \| \nabla f_t(x) - \nabla f_{t-1}(x) \|^2$ measures the variation in gradients of loss functions. Thus, the regret bounds automatically become tighter than $O(\sqrt{T})$ when $D_T$ is small. They also develop a variant of ONS which achieves $O(d \log(D_T))$ variation bound. In this paper, we mainly focus on the small-loss bound, and it is an interesting direction to investigate whether the variation bounds can also be obtained by universal methods.

To cope with different types of loss functions simultaneously, (Hazan et al., 2008) establish the adaptive online gradient descent (AOOG), which can deal with general convex and strongly convex functions, and attains $O(\sqrt{T})$ and $O(\log T)$ regret bounds, respectively. However, AOOG requires the curvature of $f_t$ as input in each round $t$, and fails to achieve an logarithmic regret bound for exp-concave functions. Another milestone is the multiple eta gradient (MetaGrad) (van Erven and Koolen, 2016), which can automatically adapt to convex and exp-concave functions, and guarantees the corresponding minimax optimal bounds. However, Metagrad treats strongly convex functions as exp-concave functions, and thus suffers the suboptimal $O(d \log T)$ regret bound for strongly convex functions. This limitation is addressed by (Wang et al., 2019), who propose the multiple sub-algorithms and learning rates (Maler). This algorithm achieves minimal optimal regret bounds for general convex, exp-concave and strongly convex functions.

In this paper, we are devoted to designing algorithms that can adapt to the structure of the loss functions. A parallel line of research considers adapting to the structure inherent in data, such as sparsity (Duchi et al., 2011; Tieleman and Hinton, 2012; Kingma and Ba, 2015; Reddi et al., 2018). For general convex functions, these algorithms are able to achieve regret bounds which are tighter than $O(\sqrt{T})$ when the gradients are sparse.

In the definition of regret, we compare the performance of the learner with that of a fixed action. However, in some cases the best action may drift over time. To address this problem, recently some more stringent performance metrics are proposed. One of them is dynamic regret (Zinkevich, 2003; Hall and Willett, 2013), which is defined as the difference between the cumulative loss of the learner with that of any sequence of comparators. Another is adaptive regret (Hazan and Seshadhri, 2007; Daniely et al., 2015; Jun et al., 2017), which is the maximum “local” regret over any contiguous time interval.
Main Results

In this section, we first provide some assumptions and definitions, then investigate how to utilize smoothness to improve the regret bound when the loss functions are strongly convex, and finally present our universal algorithm for multiple types of smooth functions and its theoretical guarantees.

Preliminary

Following previous work (Orabona et al., 2012; Zhang et al., 2019; Wang et al., 2019), we introduce the following assumptions and definitions (Boyd and Vandenberghe, 2004).

Assumption 1. The domain $\mathcal{X}$ is convex, and its diameter is bounded by $D$, i.e.,
$$\max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\| \leq D.$$

Assumption 2. The gradients of all loss functions are bounded by $G$, i.e.,
$$\max_{x \in \mathcal{X}} \|f_t(x)\| \leq G, \forall t \in \{1, \ldots, T\}.$$

Definition 1. A function $f : \mathcal{X} \mapsto \mathbb{R}$ is convex if
$$f(x_1) \geq f(x_2) + \nabla f(x_2)\top(x_1 - x_2), \forall x_1, x_2 \in \mathcal{X}. \ (2)$$

Definition 2. A function $f : \mathcal{X} \mapsto \mathbb{R}$ is $\lambda$-strongly convex if $\forall x_1, x_2 \in \mathcal{X}$,
$$f(x_1) \geq f(x_2) + \nabla f(x_2)\top(x_1 - x_2) + \frac{\lambda}{2}\|x_1 - x_2\|^2. \ (3)$$

Definition 3. A function $f : \mathcal{X} \mapsto \mathbb{R}$ is $\alpha$-exp-concave if $\exp(-\alpha f(x))$ is concave.

Definition 4. A function $f : \mathcal{X} \mapsto \mathbb{R}$ is $h$-smooth if
$$\|
abla f(x_1) - \nabla f(x_2)\| \leq h\|x_1 - x_2\|, \forall x_1, x_2 \in \mathcal{X}.$$

Finally, we introduce some useful properties.

Lemma 1 (Hazan et al., 2007, Lemma 3). Suppose Assumptions 1 and 2 hold, and $f : \mathcal{X} \mapsto \mathbb{R}$ is $\alpha$-exp-concave. Then, the following holds: $\forall \beta \leq \frac{1}{2} \min\{\frac{\lambda}{2G^2}, \alpha\},$
$$f(x_1) \geq f(x_2) + (x_1 - x_2)\top\nabla f(x_2) + \frac{\beta}{2}(x_1 - x_2)\top\nabla f(x_2))^2, \forall x_1, x_2 \in \mathcal{X}. \ (4)$$

Lemma 2 (Srebro et al., 2010, Lemma 3.1). For an $h$-smooth and nonnegative function $f : \mathcal{X} \mapsto \mathbb{R}$, we have
$$\|
abla f(x)\| \leq \sqrt{4hf(x)}, \forall x \in \mathcal{X}. \ (5)$$

Smooth and Strongly Convex OGD

In this section, we propose a novel algorithm for smooth and strongly convex functions. The proposed algorithm is built upon the SOGD algorithm (Zhang et al., 2019), which is designed for smooth and convex functions. In SOGD, the action in each round $t$ is updated by the following projected gradient descent step
$$x_{t+1} = \Pi_{\mathcal{X}}^H[x_t - \alpha_t g_t]$$
where the step size $\alpha_t$ is configured as
$$\alpha_t = \frac{\gamma}{\sqrt{\delta + t} \sum_{i=1}^{t} \|g_i\|^2} = \frac{\gamma}{\delta + \sum_{i=1}^{t} \|g_i\|^2},$$
and $\delta, \gamma > 0$ are constant parameters. Similarly to convex OGD, the step size of SOGD decreases in general on the order of $O(1/\sqrt{T})$, but is adjusted by the average of past gradients. This enables SOGD to automatically adapt to smoothness, and achieve the $O(\sqrt{T})$ regret bound.

For strongly convex and smooth functions, mimicking the behavior of strongly convex OGD, where the step size decreases on the order of $O(1/t)$, we modify the step size of SOGD as
$$\alpha_t = \frac{\gamma}{\delta + t \sum_{i=1}^{t} \|g_i\|^2} = \frac{\gamma}{\delta + \sum_{i=1}^{t} \|g_i\|^2} \ (6)$$
so that its step size decays approximately proportional to $O(1/t)$, which is similar to that in the strongly convex OGD.

The new algorithm, named Smooth and Strongly Convex Online Gradient Descent ($S^2$OGD), is summarized in Algorithm 1. For $S^2$OGD, we prove the following theorem.

Theorem 1. Suppose Assumptions 1 and 2 hold, and all loss functions are $\lambda$-strongly convex. Let $\gamma = \frac{G^2}{2\lambda}$ and $\delta = G^2$.
Then, $S^2$OGD guarantees the following regret bound:
$$R(T) \leq \lambda D^2 + \frac{G^2}{2\lambda} \log \left(\frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1\right).$$
Moreover, if all loss functions are also nonnegative and $h$-smooth, $S^2$OGD enjoys
$$R(T) \leq \frac{G^2}{2\lambda} \log \left(\frac{8h}{G^2} \sum_{t=1}^{T} f_t(x_t) + \mu\right) + \lambda D^2$$
where
$$\mu = \frac{8\lambda h D^2}{G^2} + \frac{4h}{\lambda} \ln \left(\frac{4h + \lambda e\lambda}{e\lambda} + 2\right)$$
is a constant.

Remark 1. Theorem 1 implies that $S^2$OGD guarantees an $O(\log L_*)$ small-loss regret bound, which reduces to the minimax optimal $O(\log T)$ regret bound in the worst case,
A Parameter-Free Algorithm for Smooth and Strongly Convex Optimization

While $S^2$OGD successfully achieves the $O(\log L_*)$ regret bound, it requires the modulus of strong convexity $\lambda$ as input to tune the step size. In this section, we propose a novel algorithm which ensures the $O(\log L_*)$ regret bound while being parameter-free to $\lambda$. Then, in the next section, we extend the proposed algorithm to support more types of functions.

Our Algorithm is inspired by Maler (Wang et al., 2019), so we first briefly review some intuition behind this algorithm below.

Review of Maler  To deal with strongly convex functions, Maler introduces the following surrogate loss function:

$$s_\eta^g(x) = \eta(g_t - x)^T g_t + \eta^2 G^2 \|x_t - x\|^2$$

where $\eta \in (0, \frac{\lambda}{4G^2}]$ is a constant. It can be easily seen that $s_\eta^g$ is $2\eta^2 G^2$-strongly convex. Thus, by applying strongly convex OGD on $s_\eta^g$, we obtain

$$\sum_{t=1}^T s_\eta^g(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^T s_\eta^g(x) \leq O(\log T).$$

On the other hand, by the definition of $s_\eta^g$, we have

$$R(T) \leq \sum_{t=1}^T g_t^T (x_t - x_*)$$

$$\leq \sum_{t=1}^T s_\eta^g(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^T s_\eta^g(x)$$

$$\eta \sum_{t=1}^T s_\eta^g(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^T s_\eta^g(x) \leq O(\log T) + \eta V_\eta^g$$

where $V_\eta^g = \sum_{t=1}^T G^2 \|x_t - x_*\|^2$. By configuring $\eta$ as

$$\eta^*_\eta = \sqrt{\frac{\log T}{V_\eta^g}}$$

we have

$$\sum_{t=1}^T \|\nabla s_\eta^g(x_t^{\eta,\hat{s}})\|^2$$

$$R(T) \leq \sum_{t=1}^T g_t^T (x_t - x_*) \leq O(\sqrt{V_\eta^g \log T})$$

which, combining with Definition 2, automatically reduces to $O(\log T)$ for strongly convex functions. However, the optimal $\eta^*_\eta$ cannot be obtained, since it depends on $x_*$ and the whole learning history. To resolve this problem, Maler maintains multiple strongly convex OGD as experts, each of which runs on $s_\eta^g$ with a different $\eta$, and then utilizes a meta-algorithm to identify the expert with the best $\eta$ adaptively.

Algorithm 2 A Parameter-free algorithm for Strongly convex and Smooth functions (PASS)

1. **Input**: Learning rates $\eta_1, \eta_2, \ldots$, prior weights $\pi_{t}^{\eta_1,\hat{s}}, \pi_{t}^{\eta_2,\hat{s}}, \ldots$
2. for $t = 1, \ldots, T$
3. Get predictions $x_t^{\eta,\hat{s}}$ from Algorithm 3 for all $\eta$
4. Play
   $$x_t = \frac{\sum_{\eta} \eta \pi_{t}^{\eta,\hat{s}} x_t^{\eta,\hat{s}}}{\sum_{\eta} \eta \pi_{t}^{\eta,\hat{s}}}$$
5. Observe gradient $g_t$ and send it to all experts
6. Update weights:
   $$\pi_{t+1}^{\eta,\hat{s}} = \frac{\pi_{t}^{\eta,\hat{s}} e^{-s_t^g(x_t^{\eta,\hat{s}})}}{\sum_{\eta} \eta \pi_{t}^{\eta,\hat{s}} e^{-s_t^g(x_t^{\eta,\hat{s}})}}$$
for all $\eta$
7. **end for**

Algorithm 3 Strongly convex expert algorithm ($S^2$OGD)

1. $x_1^{\eta,\hat{s}} = 0$, $\delta = 2\eta^2 G^2$
2. for $t = 1, \ldots, T$
3. Send $x_t^{\eta,\hat{s}}$ to the meta-algorithm
4. Receive gradient $g_t$ from the meta-algorithm
5. Update
   $$x_{t+1}^{\eta,\hat{s}} = \Pi^{\eta}_\mathcal{X} \left( x_t^{\eta,\hat{s}} - \alpha_t^{\eta} \nabla s_t^g(x_t^{\eta,\hat{s}}) \right)$$
   where $\alpha_t^{\eta}$ is defined in (9).
6. **end for**

Specifically, to deal with strongly convex functions, Maler keeps $C = \lceil \frac{1}{\delta} \log T \rceil + 1$ experts. In the $t$-th round, each expert receives $s_\eta^g$ as the loss function, and runs strongly convex OGD to output an action $x_t^{\eta,\hat{s}}$. Maler then calculates the final action $x_t$ according to the Exact Exponential Weighted Algorithm (TEWA) $^1$:

$$x_t = \frac{\sum_{\eta} \pi_{t}^{\eta,\hat{s}} x_t^{\eta,\hat{s}}}{\sum_{\eta} \pi_{t}^{\eta,\hat{s}}}$$

where $\pi_{t}^{\eta,\hat{s}} \propto \exp(-\sum_{t=1}^T s_\eta^g(x_t^{\eta,\hat{s}}))$. Theoretical analysis shows that, in this way, we can successfully achieve an $O(\sqrt{\log T \log T})$ regret bound.

Our Algorithm  Following Maler, to exploit the smoothness of strongly convex loss functions, one may consider directly replacing the strongly convex OGD with our $S^2$OGD, and hope to obtain an $O(\sqrt{\log L_* \log T})$ regret bound, which can reduce to $O(\log L_*)$ for strongly convex functions. However, this procedure can not achieve our goal due to the following issues.

$^1$Here we only provide the main idea of Maler on how to deal with strongly convex functions. The exact weighting technique in Maler can be found in the Step 4 of Algorithm 1 in (7).
• Since the expert algorithm is performed on \( s_t^\eta \), in the derived regret bound, the small-loss part is related to \( s_t^\eta \) rather than \( f_t \).

• Although \( s_t^\eta \) is \( 2\eta^2 G^2 \)-smooth, it could be negative, which makes the self-bounding property in Lemma 2 inapplicable.

Thus, under this method, one can only hope to obtain a regret bound on the order of \( O(\sqrt{T \log(T)} \| \nabla s_t^\eta(x_t^\eta) \|^2) \), which does not lead to a small-loss regret bound with respect to \( f_t \).

To tackle this problem, we introduce a new surrogate loss function:

\[
\hat{s}_t^\eta(x) = -\eta(x_t - x)^T g_t + \eta^2 \| g_t \|^2 \| x_t - x \|^2 \tag{8}
\]

which is a lower bound of \( s_t^\eta \). We first observe two important properties of \( \hat{s}_t^\eta \).

**Lemma 3.** Suppose Assumptions 1 and 2 hold. Then, we have \( \forall \eta \in (0, \frac{1}{\sqrt{5GD}}), x \in X, \)

\[
\| \nabla \hat{s}_t^\eta(x) \|^2 \leq 2\eta^2 \| g_t \|^2.
\]

**Lemma 4.** Let \( \lambda_t^\eta = 2\eta \| g_t \|^2 \). We have \( \forall x_1, x_2 \in X, \)

\[
\hat{s}_t^\eta(x_1) \geq \hat{s}_t^\eta(x_2) + \nabla \hat{s}_t^\eta(x_2)^T(x_1 - x_2) + \frac{\lambda_t^\eta}{2} \| x_1 - x_2 \|^2.
\]

Lemma 3 reveals that, if we can derive the gradient-dependent \( O(\sqrt{T \log(T)} \| \nabla \hat{s}_t^\eta(x_t^\eta) \|^2) \) regret bound, we can obtain a regret bound of \( O(\sqrt{T \log(T)} \| g_t \|^2) \). This bound, together with the self-bounding property in Lemma 2, can further lead to the desired small-loss regret bound. However, the new surrogate loss brings another challenge: Since the overall “strongly convex” parameter of \( s_1^\eta, \ldots, s_T^\eta \), i.e.,

\[
\lambda_t^\eta = \min_{i \in \{1, \ldots, T\}} \lambda_i^\eta
\]

is unavailable and even might be zero, \( S^2 \text{OGD} \) cannot be applied. To handle this issue, inspired by Hazan et al., (2008), we propose to perform the following update rule on \( s_t^\eta \):

\[
x_t^{\eta, i} = \Pi_{X} \left[ x_t^{\eta, i} - \alpha_t \nabla \hat{s}_t^\eta(x_t^{\eta, i}) \right]
\]

where

\[
\alpha_t^\eta = \frac{1}{\delta + \sum_{i=1}^{T} \lambda_i^\eta} \tag{9}
\]

and \( \lambda_t^\eta = 2\eta \| g_t \|^2 \) is the curvature of \( \hat{s}_t^\eta \), which is defined in Lemma 3. We name this algorithm Surrogate loss \( S^2 \text{OGD} \) (\( S^2 \text{OGD} \)), summarized in Algorithm 3. The main advantage of \( S^2 \text{OGD} \) is that it does not require \( \lambda^\eta \) to tune the step size, and can achieve the following regret bound (Hazan et al., 2008):

\[
\sum_{t=1}^{T} \hat{s}_t(x_t^{\eta, i}) - \sum_{t=1}^{T} \hat{s}_t(x) \leq O \left( \frac{1}{\delta + \sum_{i=1}^{T} \lambda_i^\eta} \right).
\]

For general strongly convex cases, it only leads to an \( O(\log(T)) \) regret bound. Nevertheless, thanks to the nice properties of \( \hat{s}_t^\eta \), i.e., Lemma 3 and Lemma 4, for an expert running on \( \hat{s}_t^\eta \), we can derive the following tighter regret bound, which is on the order of \( O(\log(\sum_{i=1}^{T} \| g_t \|^2)) \).

**Theorem 2.** Suppose Assumptions 1 and 2 hold. Let \( \delta = 2\eta^2 G^2 \). Then, for an expert that runs \( S^2 \text{OGD} \) on \( s_t^\eta \), we have \( \forall x \in X, \)

\[
\sum_{t=1}^{T} \hat{s}_t(x_t^{\eta, i}) - \sum_{t=1}^{T} \hat{s}_t(x) \leq 1 + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \| g_t \|^2 + 1 \right). \tag{10}
\]

Based on Theorem 2, we propose our Parameter-free algorithm for Strongly convex and Smooth functions (PASS), which is summarized in Algorithm 2. In each round \( t \), we keep \( \frac{1}{2} \log(T) + 1 \) experts, each of which runs \( S^2 \text{OGD} \) on \( s_t^\eta \) with a different \( \eta \), and outputs an action, denoted as \( x_t^{\eta, s} \) (Step 3). For expert \( i \in \{0, \ldots, \lceil \frac{1}{2} \log(T) \rceil \} \), we configure

\[
\pi_1^{\eta, i} = \frac{1}{G}, \eta_i = \frac{2^{-i}}{5GD}.
\]

Then, PASS updates \( x_t \) by TWEA (Step 4):

\[
x_t = \sum_{i=1}^{\eta} \pi_i^{\eta, i} x_t^{\eta, i}.
\]

Finally, the algorithm observes \( g_t \) (Step 5) and updates the weights of experts according to their losses on \( s_t^\eta \) (Step 6).

We obtain the following regret bound for our PASS algorithm:

**Theorem 3.** Suppose Assumptions 1 and 2 hold. Then, PASS achieves the following regret bound:

\[
R(T) \leq 3 \sqrt{T \left( \ln(\log(T) + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \| g_t \|^2 + 1 \right) + 1 \right)} + 10GD \left( \ln(\log(T) + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \| g_t \|^2 + 1 \right) + 1 \right)
\]

\[
= O \left( \sqrt{T \left( \ln(\log(T) + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \| g_t \|^2 + 1 \right) \right)} \right).
\]

Further, if all loss functions are \( \lambda \)-strongly convex, we have

\[
R(T) \leq \left( 10GD + \frac{9G^2}{2\lambda} \right) \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \| g_t \|^2 + 1 \right)
\]

\[
+ \left( 10GD + \frac{9G^2}{2\lambda} \right) \left( \ln(\log(T) + 3) + 1 \right)
\]

\[
= O \left( \frac{1}{\lambda} \log \left( \sum_{t=1}^{T} \| g_t \|^2 \right) \right).
\]

Moreover, if all loss functions are also nonnegative and smooth, PASS enjoys

\[
R(T) \leq m \ln \left( \frac{8h}{G^2} \sum_{t=1}^{T} f_t(x_t) + \frac{8hn}{G^2} + 8mh \ln \left( \frac{5mh + eG^2}{eG^2} \right) + 2 \right) + n = O \left( \frac{1}{\lambda} \log L_s \right).
\]
**Algorithm 4** Universal algorithm For Online smooth optimization (UFO)

1: **Input:** Learning rates $\eta_1, \eta_2, \ldots$, prior weights $\pi^{\eta_1, \delta}_1, \pi^{\eta_2, \delta}_2, \ldots$, $\pi^{\eta_1, \delta}_1, \pi^{\eta_2, \delta}_2, \ldots$, and $\pi^{\eta_1, \delta}_{1, \delta}, \pi^{\eta_2, \delta}_{2, \delta}, \ldots$

2: for $t = 1, \ldots, T$ do

3: Get predictions $x_t^{\eta, c}, x_t^{\eta, c}$ and $x_t^{\eta, \delta}$ from Algorithms 5, 6, and 3 for all $\eta$

4: Play $x_t = \sum_{\eta} (\pi^{\eta, c}_t x_t^{\eta, c} + \pi^{\eta, c}_t x_t^{\eta, c}) / \sum_{\eta} (\pi^{\eta, c}_t x_t^{\eta, c} + \pi^{\eta, c}_t x_t^{\eta, c})$

5: Observe gradient $g_t$, and send it to all experts

6: Update weights:

\[
\pi^{\eta, c}_{t+1} = \frac{e^{-\ell_t^{\eta, c}}(x_t^{\eta, c})}{\Phi_t} \quad \forall \eta
\]

\[
\pi^{\eta, \delta}_{t+1} = \frac{e^{-\ell_t^{\eta, \delta}}(x_t^{\eta, \delta})}{\Phi_t} \quad \forall \eta
\]

\[
\pi^{\eta, \delta}_{t+1} = \frac{e^{-\ell_t^{\eta, \delta}}(x_t^{\eta, \delta})}{\Phi_t} \quad \forall \eta
\]

where

\[
\Phi_t = \sum_{\eta} \left( \pi^{\eta, c}_t e^{-\ell_t^{\eta, c}}(x_t^{\eta, c}) + \pi^{\eta, \delta}_t e^{-\ell_t^{\eta, \delta}}(x_t^{\eta, \delta}) + \pi^{\eta, \delta}_t e^{-\ell_t^{\eta, \delta}}(x_t^{\eta, \delta}) \right)
\]

7: end for

where

\[
m = \left( 10GD + \frac{9G^2}{2\lambda} \right)
\]

and

\[
n = \left( 10GD + \frac{9G^2}{2\lambda} \right) \left( \ln(\log T + 3) + 1 \right).
\]

**Remark 2.** Theorem 3 indicates that PASS achieves the $O\left( \sqrt{T \log(\sum_{t=1}^T \|g_t\|^2) + \log(\sum_{t=1}^T \|g_t\|^2)} \right)$ regret bound. For $\lambda$-strongly convex functions, it implies the $O\left( \frac{1}{\lambda} \log T \right)$ minimax optimal regret bound. When the loss functions are also smooth and nonnegative, PASS enjoys the $O\left( \frac{1}{\lambda} \log L \right)$ small-loss regret bound, which becomes much tighter whenever $L$ is small. Compared to S2OGD, PASS doesn’t require the modulus of strong convexity to tune parameters, and is thus parameter-free to $\lambda$.

**A Universal Algorithm for Online Smooth Optimization**

In this section, we extend PASS to support more types of loss functions.

To adapt to exp-concave functions, we employ the following surrogate loss function proposed by (van Erven and Koolen, 2016):

\[
\ell_t^\eta(x) = -\eta (x_t - x)^\top g_t + \eta^2 \left( (x_t - x)^\top g_t \right)^2.
\]

We introduce the following two lemmas about $\ell_t^\eta$. The first lemma implies that $\ell_t^\eta$ is also exp-concave for small $\eta$, and the second lemma reflects a direct connection between the gradients of $\ell_t^\eta$ and $g_t$.

**Algorithm 5** Convex expert algorithm (SOGD)

1: $x_1^{\eta, c} = 0, \delta = \eta^2, \gamma = \frac{D}{\sqrt{2}}$

2: for $t = 1, \ldots, T$ do

3: Send $x_t^{\eta, c}$ to Algorithm 4

4: Receive gradient $g_t$ from Algorithm 4

5: Update

\[
x_t^{\eta, c} = \Pi_{x_t^{\eta, c} - \frac{\gamma}{\delta} \nabla c_t^\eta(x_t^{\eta, c})} \left[ x_t^{\eta, c} - \frac{\gamma}{\delta} \sum_{t=1}^T \|\nabla c_t^\eta(x_t^{\eta, c})\|^2 \right]
\]

6: end for

**Algorithm 6** Exp-concave expert algorithm (ONS)

1: $x_1^{\eta, c} = 0, \beta = \frac{1}{2} \min \left\{ \frac{1}{G^\eta D^\eta}, 1 \right\} = \frac{25}{59}$, where $G^\eta = \frac{1}{25D^\eta}, \Sigma_1 = \frac{1}{25D^\eta} I_d$

2: for $t = 1, \ldots, T$ do

3: Send $x_t^{\eta, c}$ to Algorithm 4

4: Receive gradient $g_t$ from Algorithm 4

5: Update

\[
\Sigma_{t+1} = \Sigma_t + \nabla \ell_t^\eta \left( x_t^{\eta, c} \right) \left( \nabla \ell_t^\eta \left( x_t^{\eta, c} \right) \right)^\top
\]

\[
x_t^{\eta, c} = \Pi_{x_t^{\eta, c} - \frac{1}{\beta} \Sigma^{-1}_{t+1} \nabla \ell_t^\eta \left( x_t^{\eta, c} \right)}
\]

where

\[
\nabla \ell_t^\eta \left( x_t^{\eta, c} \right) = \eta g_t + 2\eta^2 g_t (x_t^{\eta, c} - x_t).
\]

6: end for

**Lemma 5** (Wang et al., 2019)). For $\eta \in (0, \frac{1}{4GD})$, $\ell_t^\eta$ is $l$-exp-concave.

**Lemma 6.** Suppose Assumptions 1 and 2 hold. Then, $\forall \eta \in (0, \frac{1}{4GD})$, $\forall x \in \mathcal{X}$, $\|\nabla \ell_t^\eta(x)\|^2 \leq 2\eta^2 \|g_t\|^2$. The two lemmas motivate us to adopt ONS as expert algorithm on $\ell_t^\eta$, and derive the following regret bound:

\[
R(T) \leq \sum_{t=1}^T g_t^\top (x_t - x_*)
\]

\[
\leq O \left( \sqrt{V_T^d \log(\sum_{t=1}^T \|g_t\|^2) + d \log(\sum_{t=1}^T \|g_t\|^2)} \right)
\]

where $V_T^d = \sum_{t=1}^T \left( g_t^\top (x_t - x_*) \right)^2$. When the loss functions are exp-concave and smooth, the above regret bound can reduce to $O(d \log(L_\lambda))$ by exploiting Lemma 1 and Lemma 2.

For convex functions, (Wang et al., 2019) propose the following convex surrogate loss function:

\[
c_t(x) = -\eta^c (x_t - x)^\top g_t + (\eta^c)^2 G^2 D^2
\]

where $\eta^c = \frac{1}{2GD\sqrt{2}}$. A naive approach is to directly apply algorithms for convex and smooth functions such as SOGD.
on \( c_t(x) \) as experts. However, due to the constant term (the second term of \( c_t \)), incorporating this approach into LEF only gives to an \( O(\sqrt{T}) \) regret bound, and fails to achieve the desired small-loss bound. To address this problem, we design a new surrogate loss:
\[
c^\eta_t(x) = -\langle x_t - x \rangle^T g_t + \eta^2 \|g_t\|^2 D^2
\]
where the second term is dependent on \( g_t \). We then employ SOGD on \( c^\eta_t \) as experts. In this way, we are able to obtain an \( O(\sqrt{\sum_{t=1}^T \|g_t\|^2}) \) regret bound, which attains \( O(\sqrt{L_\ast}) \) under the assumption of general convexity and smoothness.

Our Universal algorithm For Online smooth optimization (UFO) is summarized in Algorithm 4, which is an extension of PASS algorithm by incorporating more types of experts. Similar to PASS, our proposed UFO follows from the LEF, while utilizing different types of algorithms as experts to deal with multiple types of loss functions. Each expert algorithm is associated with a carefully designed surrogate loss. Specifically, in each round \( t \), we maintain three types of experts:

- **Strongly convex experts.** For strongly convex functions, similarly to Pass, we keep \( C = \left[ \frac{1}{\alpha} \log T \right] + 1 \) strongly convex experts, each of which runs S\(^2\)GD (Algorithm 3) on \( s^\eta_t \) to output an action \( x^\eta_t \). For each expert \( i \in \{0, \ldots, C - 1\} \), we configure its learning rate and prior weight as
  \[
  \eta_i = \frac{2^{-i}}{5GD}, \quad \pi_i^{\eta,\ell} = \frac{1}{3C}.
  \]

- **Exp-concave experts.** To handle exp-concave functions, we also maintain \( C = \left[ \frac{1}{\alpha} \log T \right] + 1 \) experts, each of which receives \( \ell^\eta_t \) as the loss function, and runs the standard Online Newton Step (ONS) (Algorithm 6) to output an action, denoted as \( x^\eta_t^{\ell,\ell} \). For each expert \( i \in \{0, \ldots, C - 1\} \), we set
  \[
  \eta_i = \frac{2^{-i}}{5GD}, \quad \pi_i^{\eta,\ell,\ell} = \frac{1}{3C}.
  \]

- **Convex experts.** To deal with general convex functions, we maintain \( C \) convex experts. Each expert receives \( c^\eta_t \) as the loss function, and runs SOGD to output an action, denoted by \( x^\eta_t^{\gamma,c} \). For each expert \( i \in \{0, \ldots, C - 1\} \), we let
  \[
  \eta_i = \frac{2^{-i}}{5GD}, \quad \pi_i^{\gamma,c} = \frac{1}{3C}.
  \]

In each round \( t \), UFO firstly receives the outputs of all experts (Step 3), then submits the following action (Step 4):
\[
x_t = \sum_{i=0}^{C-1} \eta_i x_t^{\eta,\ell,\ell} + \pi_t^{\eta,\ell,\ell} x_t^{\gamma,c} + \pi_t^{\gamma,c} x_t^{\eta,\gamma,c} + \pi_t^{\eta,c} x_t^{\eta,\gamma} + \pi_t^{\gamma} x_t^{\eta,\gamma} \frac{1}{\sum_i \eta_i x_t^{\eta,\ell,\ell} + \eta_t^{\gamma,c} x_t^{\gamma,c} + \pi_t^{\gamma,c} x_t^{\eta,\gamma,c} + \pi_t^{\gamma} x_t^{\eta,\gamma}}
\]
which is the weighted sum of the outputs of experts titled by their own \( \eta \). After the gradient \( g_t \) is observed (Step 5), UFO updates the weights based on the historical performance of the experts (Step 6) on their own surrogate loss functions.

For UFO, we can derive the following regret bound.

**Theorem 4.** Suppose Assumptions 1 and 2 hold. Then, the regret of UFO is simultaneously bounded by
\[
R(T) = \begin{cases}
O \left( \sqrt{\sum_{t=1}^T \|g_t\|^2} \right) \\
O \left( V_t^g d \log \left( \sum_{t=1}^T \|g_t\|^2 \right) + d \log \left( \sum_{t=1}^T \|g_t\|^2 \right) \right) \\
O \left( V_t^g \log \left( \sum_{t=1}^T \|g_t\|^2 \right) + \log \left( \sum_{t=1}^T \|g_t\|^2 \right) \right)
\end{cases}
\]
where \( V_t^g = \sum_{t=1}^T G^2 \|x_t - x_\ast\|^2 \), and \( V_t^g = \sum_{t=1}^T (x_t - x_\ast)^T g_t \). Moreover, for \( \lambda \)-strongly convex functions, \( R(T) = O \left( \frac{1}{\sqrt{T}} \log \left( \sum_{t=1}^T \|g_t\|^2 \right) \right) \). For \( \alpha \)-exp-concave functions, \( R(T) = O \left( \frac{d}{\alpha} \log \left( \sum_{t=1}^T \|g_t\|^2 \right) \right) \). Furthermore, assume all loss functions are nonnegative and \( h \)-smooth. Then, for general convex, \( \alpha \)-exp-concave, and \( \lambda \)-strongly convex functions, the regret bounds of UFO reduce to \( O(\sqrt{hL_\ast}) \), \( O \left( \frac{\alpha}{h} \log (hL_\ast) \right) \), and \( O \left( \frac{1}{h} \log (hL_\ast) \right) \), respectively.

**Remark 3.** Since \( \sum_{t=1}^T \|g_t\|^2 \leq TG^2 \). Theorem 4 implies that, for general convex, \( \alpha \)-exp-concave and \( \lambda \)-strongly convex loss functions, UFO attains the minimax \( O(\sqrt{T}) \), \( O \left( \frac{\alpha}{h} \log (hL_\ast) \right) \), and \( O \left( \frac{1}{h} \log (hL_\ast) \right) \) regret bounds in the worst case, and achieves tighter bounds whenever \( \sum_{t=1}^T \|g_t\|^2 = o(T) \). Under the assumption of \( h \)-smoothness, UFO enjoys the state-of-the-art small-loss regret bounds for the three types of loss functions.

**Conclusion and Future Work**

In this paper, we propose a more universal algorithm for OCO, which can not only adapt to general convex, exp-concave and strongly convex functions, but also automatically exploit smoothness to achieve tighter results. The proposed algorithm, named UFO, follows the LEF to deal with the uncertainty of the loss function type, and utilizes carefully designed surrogate loss functions and expert algorithms to make use of the smoothness. We show that, similarly to the existing state-of-the-art universal method, UFO achieves minimax optimality for general convex, exp-concave and strongly convex functions simultaneously. Moreover, under the assumption of smoothness, the regret bound of UFO for the three types of loss functions attain \( O \left( \sqrt{L_\ast} \right) \), \( O(d \log L_\ast) \), and \( O \left( \frac{1}{\sqrt{L_\ast}} \right) \), which are much tighter as long as \( L_\ast = o(T) \). Finally, we note it is the first time that the \( O(\log L_\ast) \) regret bound is attained for strongly convex and smooth functions. In the future, we will investigate our algorithm to achieve broader universality.

---

3Due to the page limitation, we only provide the order of regret bounds. The exact regret bound can be found in the Appendix.
References


Proof of Main Theorems

Proof of Theorem 1

Let $x_{t+1} = x_t - \alpha_t g_t$. We have $\forall x \in \mathcal{X}$, 

$$f_t(x_t) - f_t(x) \leq (x_t - x)^T g_t - \frac{\lambda}{2} \|x_t - x\|^2$$

For the first term, we have 

$$(x_t - x^T) = \|x_t - x\|^2 + (x - x^T)^T (x_t - x)$$

$$= \|x_t - x\|^2 - \|x^T - x\|^2 - (x_t - x^T)^T (x_t - x)$$

$$= \|x_t - x\|^2 - \|x^T - x\|^2 + \|x_t - x^T\|^2$$

which implies 

$$\frac{1}{2\alpha_t} ((x_t - x)^T (x_t - x)$$

Plugging (17) into (16), we have 

$$f_t(x_t) - f_t(x) \leq \frac{1}{2\alpha_t} ((x_t - x)^T - (x^T - x)^T + \|x_t - x^T\|^2$$

where the last inequality is due to the following lemma, which implies that the weighted projection procedure is non-expensive.

Lemma 7 ((McMahan and Streeter, 2010), Lemma 3). Let $A \in \mathbb{R}^{d \times d}$ be a positive definite matrix and $\mathcal{X}$ be a convex set. Then, we have $\forall x_1, x_2 \in \mathbb{R}^d$, 

$$\|\Pi_A^T x_1 - \Pi_A^T x_2\|_A \leq \|x_1 - x_2\|_A.$$ 

Summing (18) over 1 to $T$, we have 

$$\sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x)$$ 

$$\leq \frac{1}{2\alpha_1} \|x_1 - x\|^2 + \sum_{t=2}^T \left( \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \lambda \right) \frac{\|x_t - x\|^2}{2}$$

$$+ \frac{1}{2} \sum_{t=1}^T \alpha_t \|g_t\|^2.$$ 

(19) 

For the second term, we have 

$$\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \lambda = \frac{\lambda \|g_t\|^2}{G^2} - \lambda \leq 0.$$ 

Thus 

$$\sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x)$$ 

$$\leq \lambda D^2 + \sum_{t=1}^T \frac{\|g_t\|^2}{G^2}$$ 

$$\leq \lambda D^2 + \frac{G^2}{2\lambda} \log \left( \frac{1}{G^2} \sum_{t=1}^T \|g_t\|^2 + 1 \right)$$ 

(20) 

where the last inequality is based on the following lemma:

Lemma 8. Let $\ell_1, \ldots, \ell_T$ and $\delta$ be non-negative real numbers. Then 

$$\sum_{t=1}^T \ell_t^2 + \frac{\ell_t^2}{\ell_t^2} \leq \ln \left( \frac{1}{\delta} \sum_{t=1}^T \ell_t^2 + 1 \right).$$ 

To obtain the small-loss regret bound, we introduce the following lemma.

Lemma 9 ((Orabona et al., 2012), Corollary 5). Let $a, b, c, d > 0$ satisfy 

$$x - d \leq a \ln(bx + c).$$ 

Then 

$$x - d \leq a \ln \left( 2 \left( \frac{ab}{e} \ln \left( \frac{2ab}{e} + db + c \right) \right) \right).$$ 

(21) 

Remark 4. We note that, Lemma 9 requires $x > 0$. If $x = 0$, from $-d \leq a \ln c$, we can easily get 

$$-d \leq a \ln \left( 2 \left( \frac{ab}{e} \ln \left( \frac{2ab}{e} + db + c \right) \right) \right)$$ 

due to the fact that the function is non-decreasing and $\ln \frac{2ab+e}{e} > 0$. Thus, overall, for nonnegative $x$, we have 

$$x - d \leq a \ln \left( 2 \left( \frac{ab}{e} \ln \left( \frac{2ab}{e} + db + c \right) \right) \right).$$ 

From (20), we have 

$$\sum_{t=1}^T f_t(x_t) - \lambda D^2 + \sum_{t=1}^T f_t(x)$$ 

$$\leq \frac{G^2}{2\lambda} \log \left( \frac{1}{G^2} \sum_{t=1}^T \|g_t\|^2 + 1 \right)$$ 

$$\leq \frac{G^2}{2\lambda} \log \left( \frac{4h}{G^2} \sum_{t=1}^T f_t(x_t) + 1 \right)$$ 

$$\leq \frac{G^2}{2\lambda} \ln \left( \frac{8h}{G^2} \sum_{t=1}^T f_t(x) + \frac{8\lambda h D^2}{G^2} + \frac{4h}{\lambda} \ln \left( \frac{4h + \lambda \epsilon}{\epsilon \lambda} + 2 \right) \right).$$ 

(21)
Proof of Theorem 2

We first prove Lemma 3 and Lemma 4, then provide the proof of Theorem 2. For the sake of clarity, we restate the two Lemmas as follows.

Lemma 3. Suppose Assumptions 1 and 2 hold. Then, we have \( \forall \eta \in (0, \frac{1}{\sqrt{2}T}) \), \( x \in X \),

\[ \| \nabla \hat{s}^T_t(x) \|^2 \leq 2 \eta^2 \| g_t \|^2. \]

Proof. By the definition of \( \hat{s}^T_t \), we have \( \forall x \in X \),

\[ \| \nabla \hat{s}^T_t(x) \|^2 = (\eta g_t + 2\eta^2 \| g_t \|^2 (x - x_t))^2 \]

\[ = \eta^2 \| g_t \|^2 + 4 \eta^2 \| g_t \|^2 (x - x_t) \top g_t \]

\[ + 4 \eta^4 \| g_t \|^4 \| x - x_t \|^2 \]

\[ \leq \eta^2 \| g_t \|^2 + 4 \eta^2 \| g_t \|^2 + \frac{4}{\delta^2} \| g_t \|^2 \]

\[ \leq 2 \eta^2 \| g_t \|^2 \]

where the first inequality is derived from \( \eta \leq \frac{1}{\sqrt{2}T} \). \( \Box \)

Lemma 4. Let \( \lambda_t^\eta = 2 \eta^2 \| g_t \|^2 \). We have \( \forall x_1, x_2 \in X \),

\[ \hat{s}^T_t(x_1) \geq \hat{s}^T_t(x_2) + \nabla \hat{s}^T_t(x_2) \top (x_1 - x_2) + \frac{\lambda^\eta_t}{2} \| x_1 - x_2 \|^2. \]

Proof. When \( g_t \neq 0 \), \( \hat{s}^T_t \) is \( 2 \eta^2 \| g_t \|^2 \)-strongly convex, and the inequality holds. When \( g_t = 0 \), by definition, \( \forall x \in X \), we have \( \hat{s}^T_t(x) = \lambda^\eta_t = 0 \), and \( \nabla \hat{s}^T_t(x) = 0 \), thus the inequality still holds. \( \Box \)

Now we are ready to prove Theorem 2. Let \( x_{t+1}^\eta = x_{t+1}^{\eta, \hat{s}} - \alpha^\eta_t \nabla \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) \). By Lemma 4, we have \( \forall x \in X \),

\[ \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) \geq \hat{s}^T_t(x) + \nabla \hat{s}^T_t(x) \top (x_{t+1}^{\eta, \hat{s}} - x) + \frac{\lambda^\eta_t}{2} \| x_{t+1}^{\eta, \hat{s}} - x \|^2. \]

For the first term, following similar arguments as in the proof of Theorem 1, we have

\[ (x_{t+1}^{\eta, \hat{s}} - x_t)^\top \nabla \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) - \frac{\lambda^\eta_t}{2} \| x_{t+1}^{\eta, \hat{s}} - x \|^2 \]

\[ = \frac{1}{\alpha^\eta_t} (x_{t+1}^{\eta, \hat{s}} - x_t)^\top (x_{t+1}^{\eta, \hat{s}} - x) - \frac{\lambda^\eta_t}{2} \| x_{t+1}^{\eta, \hat{s}} - x \|^2. \]

and thus

\[ \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) - \hat{s}^T_t(x) \]

\[ \leq \frac{1}{2\alpha^\eta_t} \| x_{t+1}^{\eta, \hat{s}} - x \|^2 - \| x_{t+1}^{\eta, \hat{s}} - x \|^2 + \| x_{t+1}^{\eta, \hat{s}} - x \|^2 \]

\[ + \frac{\lambda^\eta_t}{2} \| \nabla \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) \|^2 \]

\[ - \frac{\lambda^\eta_t}{2} \| x_{t+1}^{\eta, \hat{s}} - x \|^2. \]

Summing the above inequality over 1 to \( T \), we have

\[ \sum_{t=1}^T \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) - \sum_{t=1}^T \hat{s}^T_t(x) \]

\[ \leq \frac{1}{2\alpha^\eta_t} \| x_{t+1}^{\eta, \hat{s}} - x \|^2 + \sum_{t=2}^T \left( \frac{1}{\alpha^\eta_t} - \frac{1}{\alpha^\eta_{t-1}} \right) \| x_{t+1}^{\eta, \hat{s}} - x \|^2 \]

\[ + \frac{1}{2} \sum_{t=1}^T \alpha^\eta_t \| \nabla \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) \|^2 \]

\[ = \frac{\| x_{t+1}^{\eta, \hat{s}} - x \|^2}{2\alpha^\eta_t} + \frac{1}{2} \sum_{t=1}^T \alpha^\eta_t \| \nabla \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) \|^2. \]

By Lemma 3, we have

\[ \sum_{t=1}^T \alpha^\eta_t \| \nabla \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) \|^2 \leq \sum_{t=1}^T \| g_t \|^2 \]

\[ \leq \sum_{t=1}^T \| g_t \|^2 \]

\[ \leq \sum_{t=1}^T \| g_t \|^2. \]

The proof can be finished by using Lemma 8.

Proof of Theorem 3

We have

\[ R(T) = \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x_*) \leq \sum_{t=1}^T g_t(x_t - x_*) \]

\[ \leq \frac{1}{\eta} \sum_{t=1}^T (\hat{s}^T_t(x_t) - \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}})) + \sum_{t=1}^T \eta \| g_t \|^2 \| x_t - x_* \|^2 \]

\[ \leq \frac{1}{\eta} \sum_{t=1}^T (\hat{s}^T_t(x_t) - \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}})) + \sum_{t=1}^T \eta \| g_t \|^2 \| x_t - x_* \|^2 \]

\[ + \frac{1}{\eta} \sum_{t=1}^T (\hat{s}^T_t(x_{t+1}^{\eta, \hat{s}}) - \hat{s}^T_t(x_*)). \]

Note that \( \sum_{t=1}^T \hat{s}^T_t(x_t) = 0 \). To proceed, we introduce the following lemma.

Lemma 10. For every grid point \( \eta \), we have

\[ \sum_{t=1}^T (\hat{s}^T_t(x_t) - \hat{s}^T_t(x_{t+1}^{\eta, \hat{s}})) \leq \ln(\log T + 3). \]

By the above lemma and Theorem 2, we have

\[ R(T) \leq \frac{\ln(\log T + 3) + \ln(\log T + 3) + 1}{\eta} \]

\[ + \eta V_f^\eta. \]

Note that

\[ \ln(\log T + 3) + \ln(\log T + 3) + 1 \geq 1. \]
Thus, the optimal $\eta^*$

$$
\eta^* = \sqrt{\frac{\ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1}{V_T^2}}.
$$

If $\eta^* \leq \frac{1}{5GD}$, there exists a grid point $\eta$ such that $\eta^* \in [\frac{3}{2}, \eta]$. Thus, for this $\eta$, we have

$$
R(T) \leq 3 \sqrt{V_T^2 \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right)}.
$$

If $\eta^* > \frac{1}{5GD}$,

$$
V_T^2 \leq 25G^2D^2 \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right)
$$

and for grid point $\eta = \frac{1}{5GD}$,

$$
R(T) \leq 10GD \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right).
$$

Overall, we have

$$
R(T) \leq \sum_{t=1}^{T} g_t^\top (x_t - x_*)
$$

$$
\leq 3 \sqrt{V_T^2 \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right)} + 10GD \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right).
$$

If the loss functions are $\lambda$-strongly convex, we have

$$
R(T) \leq \sum_{t=1}^{T} g_t^\top (x_t - x_*) - \sum_{t=1}^{T} \frac{\lambda}{2} \|x_t - x_*\|^2
$$

$$
\leq 3 \sqrt{V_T^2 \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right)} + 10GD \left( \ln(\log T + 3) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right)
$$

$$
- \sum_{t=1}^{T} \frac{\lambda}{2} \|x_t - x_*\|^2
$$

$$
\leq \sum_{t=1}^{T} \frac{\lambda}{2} \|x_t - x_*\|^2 + \left( 10GD + \frac{9G^2}{2\lambda} \right) \left( \ln(\log T + 3) + 1 \right)
$$

$$
+ \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1
$$

$$
- \sum_{t=1}^{T} \frac{\lambda}{2} \|x_t - x_*\|^2
$$

$$
= \left( 10GD + \frac{9G^2}{2\lambda} \right) \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right)
$$

$$
+ \left( 10GD + \frac{9G^2}{2\lambda} \right) \left( \ln(\log T + 3) + 1 \right)
$$

(25)

where the last inequality is derived by the arithmetic-geometric mean inequality. Finally, for smooth and nonnegative strongly convex functions, define

$$
m = \left( 10GD + \frac{9G^2}{2\lambda} \right)
$$

and

$$
n = \left( 10GD + \frac{9G^2}{2\lambda} \right) \left( \ln(\log T + 3) + 1 \right).
$$

By (25) and Lemma 9, we have

$$
\sum_{t=1}^{T} f_t(x_t) - \left( \sum_{t=1}^{T} f_t(x_*) + n \right)
$$

$$
\leq m \log \left( \frac{4h}{G^2} \sum_{t=1}^{T} f_t(x_t) + 1 \right)
$$

$$
\leq m \ln \left( \frac{8h}{G^2} \sum_{t=1}^{T} f_t(x_t) + \frac{8hn}{G^2}
$$

$$
+ 8mh \ln \left( \frac{8n + eG^2}{eG^2} \right) + 2 \right).
$$

(26)

**Proof of Theorem 4**

To begin with, we introduce the following lemma that bounds the meta-regret, i.e., the difference between the cumulative loss of the meta-algorithm and that of an expert algorithm.

**Lemma 11.** For every grid point $\eta$, we have

$$
\sum_{t=1}^{T} \hat{s}_t^\eta(x_t) - \sum_{t=1}^{T} s_t^\eta(x_t^\eta, \eta) \leq \ln \left( 2 \log T + 6 \right)
$$

(27)

$$
\sum_{t=1}^{T} \ell_t^\eta(x_t) - \sum_{t=1}^{T} \ell_t^\eta(x_t^\eta, \eta) \leq \ln \left( 2 \log T + 6 \right)
$$

(28)

and

$$
\sum_{t=1}^{T} c_t^\eta(x_t) - \sum_{t=1}^{T} c_t^\eta(x_t^\eta, \eta) \leq \sum_{t=1}^{T} \eta^2 \|g_t\|^2 D^2 + \ln \left( 2 \log T + 6 \right).
$$

(29)
Now, we are ready to prove the main theorem. By the definition of \( c_i^\eta \), for every grid point \( \eta \), we have

\[
R(T) = \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^\ast) 
\]

\[
\leq \sum_{t=1}^{T} \sum_{\eta \in \mathcal{X}} \left| c_i^\eta(x_t) - c_i^\eta(x_n^\eta) \right| + \sum_{t=1}^{T} \left( c_i^\eta(x^\ast) - c_i^\eta(x_t) \right) 
\]

\[
= \sum_{t=1}^{T} (c_i^\eta(x_t) - c_i^\eta(x_n^\eta)) + \sum_{t=1}^{T} (c_i^\eta(x^\ast) - c_i^\eta(x_t)) \tag{13}
\]

Next, we introduce the following lemma (Zhang et al., 2019).

**Lemma 12.** We have \( \forall x \in \mathcal{X} \)

\[
\sum_{t=1}^{T} c_i^\eta(x_t^\eta) - \sum_{t=1}^{T} c_i^\eta(x) \leq 2\sqrt{D^2} \sqrt{\delta + \sum_{t=1}^{T} \| \nabla c_i^\eta(x_t^\eta) \|^2} \tag{2}
\]

\[
= \sqrt{2D^2} \left( \delta + \eta \sum_{t=1}^{T} \| g_t \|^2 \right). \tag{30}
\]

Combining (29), (30) and (31), we get

\[
R(T) \leq \frac{\ln(2 \log T + 6) + \sqrt{2D^2} \sqrt{\delta + \eta \sum_{t=1}^{T} \| g_t \|^2}}{\eta} 
\]

\[
+ \frac{\sum_{t=1}^{T} \eta \| g_t \|^2 D^2}{2} \tag{32}
\]

Note that \( \delta = \eta^2 \). Let \( v = \ln(2 \log T + 6) \). The optimal \( \eta^* \)

\[
\eta^* = \sqrt{\frac{v}{D^2 \sum_{t=1}^{T} \| g_t \|^2}} \geq \frac{1}{5GD\sqrt{T}}. \tag{33}
\]

If \( \eta^* \geq \frac{1}{5GD} \), then there exists a grid point \( \eta \) such that \( \eta^* \in [\frac{1}{2}, \eta] \), and

\[
R(T) \leq \sqrt{2D^2} \left( 1 + \sum_{t=1}^{T} \| g_t \|^2 + 3 \sqrt{vD^2 \sum_{t=1}^{T} \| g_t \|^2} \right). \tag{34}
\]

Therefore,

\[
R(T) \leq \sqrt{2D^2} \left( 1 + \sum_{t=1}^{T} \| g_t \|^2 + 3 \sqrt{vD^2 \sum_{t=1}^{T} \| g_t \|^2} + \frac{1}{\sqrt{v}} \right) \tag{35}
\]

We introduce the following lemma.

**Lemma 13 (Lemma 19 of (Shalev-Shwartz, 2007)).** Let \( x, b, c \) be nonnegative numbers. Then,

\[
x - b \leq c \]

implies

\[
x - c \leq b \sqrt{x} + c. \tag{36}
\]

By the lemma above, we have

\[
\left( \frac{1}{4h} + \sum_{t=1}^{T} f_t(x_t) \right) \text{ and } \left( \frac{1}{4h} + \sum_{t=1}^{T} f_t(x^\ast) + r \right)
\]

\[
\leq 2h \left( \sqrt{2D + 3D^2v} \right) \left( \frac{1}{4h} + \sum_{t=1}^{T} f_t(x_t) \right) \tag{37}
\]

Next, we upper bound the regret by using \( s_i^\eta \). By similar arguments as in the proof of Theorem 3, we can easily get

\[
R(T) \leq \sum_{t=1}^{T} \| g_t \|^2 \leq 25G^2v \tag{38}
\]

and \( \eta = \frac{1}{5GD} \) yields

\[
R(T) \leq \sqrt{2D^2} \sqrt{1 + 25G^2v + 10DGv}. \tag{39}
\]
For $\lambda$-strongly convex functions,
\[
R(T) \leq \sum_{t=1}^{T} g^\top_t (x_t - x_*) - \sum_{t=1}^{T} \frac{\lambda}{2} \|x_t - x_*\|^2 \\
\leq 3 \sqrt{V_T \left( \ln(2 \log T + 6) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right)} \\
+ 10GD \left( \ln(2 \log T + 6) + \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 1 \right) \\
- \sum_{t=1}^{T} \frac{\lambda}{2} \|x_t - x_*\|^2 \\
\leq m \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + n^\beta
\]
where
\[m = \left( 10GD + \frac{9G^2}{2\lambda} \right)\]
\[n^\beta = \left( 10GD + \frac{9G^2}{2\lambda} \right) \ln(2 \log T + 6) + 1).\]

Finally, if the loss functions are smooth and nonnegative, we have
\[
\sum_{t=1}^{T} f_t(x_t) - \left( \sum_{t=1}^{T} f_t(x_*) + n^\beta \right) \\
\leq m \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) \\
\leq m \ln \left( \frac{8b}{G^2} \sum_{t=1}^{T} f_t(x_*) \right) \\
+ \frac{8hn^\beta}{G^2} + \frac{8mh}{G^2} \ln \left( \frac{8mh + eG^2}{eG^2} \right) + 2 \right) .
\]

Finally, we utilize $\ell_t^\eta$ to upper bound the regret.
\[
R(T) \leq \sum_{t=1}^{T} g^\top_t (x_t - x_*) \\
\leq \sum_{t=1}^{T} -\ell_t^\eta(x_*) + \frac{\eta^2}{\eta} (g^\top_t (x_t - x_*))^2 \\
+ \sum_{t=1}^{T} \left( \ell_t^\eta(x_t) - \ell_t^\eta(x_*) \right)^2 \\
+ \frac{\eta}{\eta} \sum_{t=1}^{T} \left( g^\top_t (x_t - x_*) \right)^2 \\
+ \frac{\eta}{\eta} \sum_{t=1}^{T} \left( \ell_t^\eta(x_t) - \ell_t^\eta(x_*) \right)^2 \\
+ \frac{\eta}{\eta} \sum_{t=1}^{T} \left( \ell_t^\eta(x_t^\eta) - \ell_t^\eta(x_*) \right)^2.
\]

For the last term, we have the following lemma.

**Lemma 14.** For every grid point $\eta$, we have $\forall x \in \mathcal{X}$,
\[
\sum_{t=1}^{T} \ell_t^\eta(x_t^\eta) - \sum_{t=1}^{T} \ell_t(x) \leq 2d \ln \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 2.
\]

**Proof.** The proof is similar to that of Theorem 1 in (Orabona et al., 2012). Note that $\ell_t^\eta$ is 1-exp-concave for $\eta \in (0, \frac{1}{2\beta^2})$. Thus, based on the proof of Theorem 2 in (Hazan et al., 2007), we have
\[
\sum_{t=1}^{T} \ell_t^\eta(x_t^\eta) - \sum_{t=1}^{T} \ell_t(x) \\
\leq \frac{1}{2\beta} \sum_{t=1}^{T} \left( \nabla \ell_t^\eta(x_t) \right)^\top \left[ \frac{1}{t+1} \sum_{t=1}^{T} \nabla \ell_t^\eta(x_t) \right] + \frac{1}{2\beta} \\
\leq \frac{d}{2\beta} \ln \left( \frac{\beta^2 D^2}{d} \sum_{t=1}^{T} \|\nabla \ell_t^\eta(x_t)\|^2 + 1 \right) + \frac{1}{2\beta} \\
\leq \frac{d}{2\beta} \ln \left( \frac{2\eta^2 \beta^2 D^2}{d} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + \frac{1}{2\beta} \\
\leq 2d \ln \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 2.
\]

where the third inequality is due to Lemma 6. Combining (28), (34) and Lemma 14, and using similar arguments as in the proof of Theorem 3, we have
\[
R(T) \leq \sum_{t=1}^{T} g^\top_t (x_t - x_*) \\
\leq \ln \left( 2 \log T + 6 \right) + 2d \ln \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 2 \\
+ \eta \sqrt{\left( \ln(2 \log T + 6) + 2d \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 2 \right)} \\
\leq 3 \sqrt{V_T \left( \ln(2 \log T + 6) + 2d \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 2 \right)} \\
\leq 10GD \left( \ln(2 \log T + 6) \\
+ 2d \log \left( \frac{1}{G^2} \sum_{t=1}^{T} \|g_t\|^2 + 1 \right) + 2 \right)
\]

For $\alpha$-exp-concave functions, combining (36), Lemma 1 and
Lemma 13, we have
\[ R(T) \leq \left( 20GD + \frac{9}{\beta} \right) d \log \left( \frac{1}{G^2} \sum_{t=1}^{T} ||g_t||^2 + 1 \right) \]
\[ + \left( 10GD + \frac{9}{\beta} \right) (2 + \ln(2 \log T + 6)) \] (37)
\[ = m^d d \log \left( \frac{1}{G^2} \sum_{t=1}^{T} ||g_t||^2 + 1 \right) + n^d \]
where
\[ m^d = \left( 20GD + \frac{9}{\beta} \right) \]
and
\[ n^d = \left( 10GD + \frac{9}{2\beta} \right) (2 + \ln(2 \log T + 6)). \]

If the loss functions are nonnegative and smooth, we have
\[ \sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x_*) + n^d \]
(38)
\[ \leq m^d d \log \left( \frac{1}{G^2} \sum_{t=1}^{T} ||g_t||^2 + 1 \right) + n^d \]
\[ \leq m^d \ln \left( \frac{8h}{G^2} \sum_{t=1}^{T} f_t(x_*) + \frac{8hn^d}{G^2} \right) \]
\[ + \frac{8m^d dh}{G^2} \ln \left( \frac{8m^d dh + 4eG^2}{eG^2} + 2 \right). \]

**Proof of Lemmas**

**Proof of Lemma 6**

We have \( \forall x \in \mathcal{X} \),
\[ \| \nabla f_t^\eta(x) \|^2 = \| \eta g_t + 2\eta^2 g_t g_t^\top (x - x_t) \|^2 \]
\[ = \| \eta g_t + 4\eta^2 g_t g_t^\top (x - x_t) \|^2 \]
\[ + 4\eta^4 \| (x - x_t) \|^2 \| g_t \|^2 \]
\[ \leq 4\eta^2 \| g_t \|^2 + 4\eta^4 \| (x - x_t) \|^2 \]
\[ \leq 2\eta^2 \| g_t \|^2. \]

Note that \( \eta \in (0, \frac{1}{5GD}] \).

**Proof of Lemma 10**

We first introduce the following inequality. For every grid point \( \eta \),
\[ e^{-\eta^2 (x_t - x_t^\eta)^\top g_t} \leq e^{-\eta (x_t - x_t^\eta)^\top g_t} - \eta^2 \| g_t \|^2 \| x_t - x_t^\eta \|^2 \]
\[ \leq e^{-\eta (x_t - x_t^\eta)^\top g_t} - (\eta (x_t - x_t^\eta)^\top g_t)^2 \]
\[ \leq 1 + \eta (x_t - x_t^\eta)^\top g_t \]
(40)

where the first inequality comes from the Holder’s inequality, and the second inequality is because \( e^{x - x^2} \leq 1 + x \) for any \( x \geq -\frac{2}{3} \) (van Erven and Koolen, 2016). Note that \( \eta \in (0, \frac{1}{5GD}] \). Define
\[ \Phi_T = \sum_{t=1}^{T} \eta^\eta (x_t - x_t^\eta)^\top g_t. \]

We have
\[ \Phi_{T+1} - \Phi_T \]
\[ = \sum_{\eta} \pi_1^\eta (x_t^\eta)^\top g_t \] (41)
\[ \leq \sum_{\eta} \pi_1^\eta (x_t^\eta)^\top g_t \]
\[ \leq \sum_{\eta} \pi_1^\eta (x_t^\eta)^\top g_t. \]

On the other hand, by the update rule of \( x_t \) and \( \pi_t^\eta \), we have
\[ x_{T+1} = \frac{\sum_{\eta} \pi_1^\eta (x_t^\eta)^\top g_t}{\sum_{\eta} \pi_1^\eta (x_t^\eta)^\top g_t}. \]

Note that all \( \pi_1^\eta \) shares the same denominator. Combining (41) and (42), we have \( \Phi_{T+1} - \Phi_T \leq 0 \), which implies that \( \Phi_t \) is non-increasing, i.e.,
\[ \Phi_T \leq \Phi_{T-1} \leq \cdots \leq \Phi_0 = 0. \]

Thus,
\[ 0 \leq - \ln \left( \prod_{t=1}^{T} \pi_1^\eta (x_t^\eta)^\top g_t + \ln \prod_{t=1}^{T} \pi_1^\eta (x_t^\eta)^\top g_t \right) \]
\[ \leq \sum_{t=1}^{T} \eta (x_t^\eta)^\top g_t + \ln \left( \prod_{t=1}^{T} \pi_1^\eta (x_t^\eta)^\top g_t \right) \]
\[ \leq \ln (\log T + 3). \]

**Proof of Lemma 11**

The proof is similar to that of Lemma 1 in (Wang et al., 2019), but with different prior weights of experts as well as the surrogate losses. We first introduce some useful inequalities. For every gradient point \( \eta \),
\[ e^{\eta (x_t - x_t^\eta)^\top g_t} \leq e^{\eta (x_t - x_t^\eta)^\top g_t} \]
\[ \leq 1 + \eta (x_t - x_t^\eta)^\top g_t \]
(43)
\[
\begin{align*}
\sum_{\eta} e^{-c_{i}^{\eta}(x_{i})} & \leq e^{\eta(x_{i}-x_{i}^{\eta})^{\top} g_{i}} - (\eta\|g_{i}\|D)^{2} \\
& \leq e^{\eta(x_{i}-x_{i}^{\eta})^{\top} g_{i}} - (\eta\|x_{i}-x_{i}^{\eta}\|^{\top} g_{i})^{2} \\
& \leq 1 + \eta(x_{i} - x_{i}^{\eta})^{\top} g_{i}.
\end{align*}
\]

Define
\[
\Phi_{T} = \sum_{\eta} \left( \pi_{1}^{\eta, x} e^{-\sum_{t=1}^{T} \delta_{t}^{\eta}(x_{t}^{\eta, x})} + \pi_{1}^{\eta, c} e^{-\sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c})} \right).
\]

We have
\[
\begin{align*}
\Phi_{T+1} - \Phi_{T} &= \sum_{\eta} \pi_{1}^{\eta, x} e^{-\sum_{t=1}^{T} \delta_{t}^{\eta}(x_{t}^{\eta, x})} \left( e^{-\delta_{T+1}^{\eta}(x_{T+1}^{\eta, x})} - 1 \right) \\
& \quad + \sum_{\eta} \pi_{1}^{\eta, c} e^{-\sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c})} \left( e^{-\ell_{T+1}^{\eta}(x_{T+1}^{\eta, c})} - 1 \right) \\
& \leq \sum_{\eta} \pi_{1}^{\eta, x} e^{-\sum_{t=1}^{T} \delta_{t}^{\eta}(x_{t}^{\eta, x})} \eta(x_{T+1} - x_{T+1}^{\eta, x})^{\top} g_{T+1} \\
& \quad + \sum_{\eta} \pi_{1}^{\eta, c} e^{-\sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c})} \eta(x_{T+1} - x_{T+1}^{\eta, c})^{\top} g_{T+1} \\
& \leq 0
\end{align*}
\]

where first inequality is due to (43), (44) and (45), and the second inequality can be derived from the update rule of \(x_{i}\) and weights. (47) implies that \(\Phi_{t}\) is non-increasing. Thus,
\[
\begin{align*}
0 & \leq - \ln \left( \pi_{1}^{\eta, x} e^{-\sum_{t=1}^{T} \delta_{t}^{\eta}(x_{t}^{\eta, x})} \right) = \sum_{t=1}^{T} \delta_{t}^{\eta}(x_{t}^{\eta, x}) + \ln 3C \\
0 & \leq - \ln \left( \pi_{1}^{\eta, c} e^{-\sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c})} \right) = \sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c}) + \ln 3C
\end{align*}
\]

and
\[
0 \leq - \ln \left( \pi_{1}^{\eta, c} e^{-\sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c})} \right) = \sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}^{\eta, c}) + \ln 3C.
\]

The proof is finished by noticing \(\sum_{t=1}^{T} \delta_{t}^{\eta}(x_{t}) = \sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}) = 0\), and
\[
\sum_{t=1}^{T} \ell_{t}^{\eta}(x_{t}) = \sum_{t=1}^{T} \eta^{2} \|g_{t}\| D^{2}.
\]