Neurocomputing 595 (2024) 127789

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Regularized online exponentially concave optimization

Xu Yang [a], Peng Tian [b], Xiao Cheng [c], Yuanyu Wan [a,*], Mingli Song [d]

[a] School of Software Technology, Zhejiang University, Ningbo 315048, China
[b] State Grid Chongqing Electric Power Company, Chongqing 400015, China
[c] State Grid Chongqing Electric Power Research Institute, Chongqing 401123, China
[d] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Communicated by T. Yang

Keywords:
Online learning
Exponential concavity
Regularization
Proximal mapping
Regret bound

ABSTRACT

In this paper, we investigate regularized online exponentially concave (abbr. exp-concave) optimization, in which each loss function consists of a time-varying exp-concave function and a fixed convex regularization. If the whole loss function is exp-concave, a classical method called online Newton step (ONS) enjoys an $O(d \log T)$ regret bound, where d is the dimensionality and T is the time horizon. However, in the regularized setting, the sum of an exp-concave function and a convex regularization is not necessarily an exp-concave function, which implies that ONS is not applicable. To address this problem, we propose the proximal online Newton step (ProxONS), and show that it can attain the same $O(d \log T)$ regret bound for any convex regularization. The main idea is to first perform an iteration of ONS with the exp-concave part in each loss function and then perform a proximal mapping with the regularization part. Furthermore, we demonstrate that by utilizing the standard online-to-batch conversion, our ProxONS can be extended to solve stochastic optimization with a regularized exp-concave objective, and enjoy an $O(d \log T/T)$ convergence rate with high probability. Experimental results on two real datasets verify the effectiveness of our ProxONS.

1. Introduction

Regularized online optimization [1–3] is a general learning framework that can find many applications in machine learning, such as ℓ_1-norm regularized logistic regression, support vector machine, and least squares [4–6]. Specifically, it is formulated as a game between an online algorithm and an adversary, which is iteratively performed in T consecutive rounds. In each round $t \in [T]$, the online algorithm is required to select a decision $\mathbf{x}_t$ from a convex set $\mathcal{K} \subseteq \mathbb{R}^d$, and then suffers a loss of the form

$$F_t(\mathbf{x}) = f_t(\mathbf{x}) + r(\mathbf{x}), \tag{1}$$

where $f_t(\cdot) : \mathcal{K} \mapsto \mathbb{R}$ is a convex function selected by the adversary and $r(\cdot) : \mathcal{K} \mapsto \mathbb{R}$ is a convex regularization function. The performance of the online algorithm is commonly measured by the gap between the cumulative loss of its decisions and the optimal fixed decision, i.e.,

$$R_T = \sum_{t=1}^{T} F_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} F_t(\mathbf{x}), \tag{2}$$

which is referred to as the regret.

In the literature, there exist plenty of methods [7–18] that can minimize the regret for specific types of loss function $F_t(\mathbf{x})$ by simply utilizing the gradient of $F_t(\mathbf{x})$ and ignoring the composite structure shown in (1). For example, online gradient descent (OGD) [7] attains $O(\sqrt{T})$ and $O(\log T)$ regret bounds for convex and strongly convex $F_t(\mathbf{x})$ respectively, and online Newton step (ONS) [8] attains an $O(d \log T)$ regret bound for exp-concave $F_t(\mathbf{x})$. Moreover, these regret bounds are known to be minimax optimal [19,20]. However, ignoring the composite structure in (1) can result in undesirable effects. For example, a major reason for using the ℓ_1-norm regularization $r(\mathbf{x}) = \|\mathbf{x}\|_1$ is to promote the sparsity of decisions [21], but directly applying OGD on the loss function $F_t(\mathbf{x})$ actually cannot promote the sparse decisions.

To address the above problem, previous studies [1–3] have proposed improved methods that update the decision by utilizing the gradient of $f_t(\mathbf{x})$ and exploiting the regularization $r(\mathbf{x})$ in an explicit way such as implementing a proximal mapping with $r(\mathbf{x})$. In this way, these methods not only have the ability to minimize the regret for a specific type of loss function $F_t(\mathbf{x})$, but also can benefit from the presence of the regularization function $r(\mathbf{x})$. However, they can only utilize the convexity and strong convexity of the loss function $F_t(\mathbf{x})$ to recover the $O(\sqrt{T})$ and $O(\log T)$ regret bounds obtained by OGD. It is still an open problem whether the exponential concavity (abbr. exp-concavity) of the loss function $F_t(\mathbf{x})$ can be utilized to recover the $O(d \log T)$ regret bound obtained by ONS without ignoring the presence of the regularization

* Corresponding author.
E-mail address: wanyy@zju.edu.cn (Y. Wan).

https://doi.org/10.1016/j.neucom.2024.127789
Received 22 September 2023; Accepted 27 April 2024
Available online 22 May 2024

function $r(\mathbf{x})$. It is worth noting that the exp-concavity is a weaker property than the strong convexity [22].

In this paper, instead of focusing on exp-concave $F_t(\mathbf{x})$, we consider a more general scenario, in which only the time-varying function $f_t(\mathbf{x})$ is exp-concave, and the regularization $r(\mathbf{x})$ can be any convex function. As proved by Yang et al. [23], the sum of an exp-concave function and a convex regularization is not necessarily an exp-concave function, which implies that ONS actually is not applicable in this general scenario even ignoring the composite structure in (1). To answer the above open problem and address the limitation of ONS, we propose a proximal variant of ONS, namely ProxONS, by combining it with the standard proximal mapping technique [1]. Specifically, in each round of our ProxONS, the main idea is to first perform an iteration of ONS with the exp-concave function $f_t(\mathbf{x})$ and then perform a proximal mapping with the regularization function $r(\mathbf{x})$. Our theoretical analysis reveals that ProxONS can attain the $O(d \log T)$ regret bound for any convex regularization function $r(\mathbf{x})$, which is more general than ONS.

Furthermore, we demonstrate that by utilizing the standard online-to-batch conversion [24], our ProxONS can be extended to solve stochastic optimization with a regularized exp-concave objective, and enjoy an $O(d \log T/T)$ convergence rate with high probability. Notice that when the whole objective is exp-concave, Mahdavi et al. [25] have proved that the combination of ONS with the standard online-to-batch conversion can enjoy the $O(d \log T/T)$ convergence rate with high probability. By comparison, our theoretical result for stochastic optimization actually generalizes that of Mahdavi et al. [25] to the case with any additional regularization. Finally, we conduct experiments on two real datasets to verify the effectiveness of our ProxONS.

Organization. The remainder of this paper is organized as follows. In Section 2, we review several relevant regularized online optimization methods, and their applications to regularized stochastic optimization. In Section 3, we introduce the procedures of our ProxONS and its stochastic extension, and provide corresponding theoretical guarantees. We present the proofs for our theoretical results in Section 4. In Section 5, we validate the effectiveness of the proposed algorithms through experiments. In Section 6, we summarize our proposed algorithms and discuss potential future work.

Notation. We use lowercase italic letters to represent scalars, such as the regularization parameter λ, and lowercase bold letters to represent vectors, such as the decision vector $\mathbf{x}$. Matrices are in uppercase bold letters, such as Q for a positive definite matrix, I_d for the identity matrix of $d \times d$. Let $\|\mathbf{x}\|_Q = \sqrt{\mathbf{x}^\top Q \mathbf{x}}$ denote the norm of a vector $\mathbf{x}$ induced by the matrix Q. Let $\nabla f(\mathbf{x})$ denote the gradient of $f(\mathbf{x})$ at the point $\mathbf{x}$ and $\operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$ denote a minimizer of $f(\mathbf{x})$ over the convex set $\mathcal{K}$. Let $\operatorname{prox}_r^Q(\mathbf{x})$ denote the proximal mapping of $r(\mathbf{x})$ on a point $\mathbf{x} \in \mathbb{R}^d$ with respect to a positive definite matrix $Q \in \mathbb{R}^{d\times d}$. We use $\mathbb{E}_{\xi\sim P}[f(\mathbf{x};\xi)]$ to denote the expectation of $f(\mathbf{x};\xi)$ where ξ denotes a random variable drawn from the distribution P. We summarize the notations used in our paper in Table 1.

Table 1
Notations used in this paper.

Notation	Meaning
λ	Scalar
$\mathbf{x}$	Vector
Q	Positive definite matrix
I_d	Identity matrix of $d \times d$
$\|\mathbf{x}\|_Q$	Vector norm induced by the matrix Q
$\nabla f(\mathbf{x})$	Gradient of $f(\mathbf{x})$ at the point $\mathbf{x}$
$\operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$	Minimizer of $f(\mathbf{x})$ over the convex set $\mathcal{K}$
$\operatorname{prox}_r^Q(\mathbf{x})$	Proximal mapping of $r(\mathbf{x})$ with respect to the matrix Q
$\mathbb{E}_{\xi\sim P}[f(\mathbf{x};\xi)]$	Expectation of $f(\mathbf{x};\xi)$

2. Related work

In this section, we briefly review related work on regularized online optimization, and discuss their applications to regularized stochastic optimization.

2.1. Regularized online optimization

Forward backward splitting (FOBOS) [1] is the first method for minimizing the regret of regularized online optimization, while explicitly exploiting the regularization structure.[1] In each round, it updates as follows

$$\begin{aligned} \mathbf{x}'_t &= \mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \left\{ \frac{1}{2} \left\|\mathbf{x} - \mathbf{x}'_t\right\|_2^2 + \eta_t r(\mathbf{x}) \right\}, \end{aligned} \tag{3}$$

where η_t is a learning rate. The first step of FOBOS is a gradient descent step with respect to the function $f_t(\mathbf{x})$, which is commonly utilized to minimize the regret. The second step of FOBOS introduces a regularization function $r(\mathbf{x})$, which can be rewritten to

$$\mathbf{x}_{t+1} = \operatorname{prox}_r^{\eta_t^{-1} I_d}(\mathbf{x}'_t), \tag{4}$$

where

$$\operatorname{prox}_r^Q(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{K}} \left\{ r(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_Q^2 \right\} \tag{5}$$

denotes the proximal mapping of $r(\cdot)$ on a point $\mathbf{x} \in \mathbb{R}^d$ with respect to a positive definite matrix $Q \in \mathbb{R}^{d\times d}$. This step is critical for exploiting the regularization structure. Moreover, as proved by Duchi and Singer [1], FOBOS attains $O(\sqrt{T})$ and $O(\log T)$ regret bounds for convex and strongly convex $F_t(\mathbf{x})$ respectively. Later, Duchi et al. [3] propose a generalized version of FOBOS, namely composite mirror descent, by replacing the Euclidean distance $\frac{1}{2}\left\|\mathbf{x} - \mathbf{x}'_t\right\|_2^2$ in (3) with the Bregman divergence.

Besides, Xiao [2] proposes regularized dual averaging (RDA), which is an online extension of the primal–dual subgradient method [26] and updates as follows

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \left\{ \langle \bar{\mathbf{g}}_t, \mathbf{x}\rangle + r(\mathbf{x}) + \frac{1}{t\eta_t} h(\mathbf{x}) \right\}, \tag{6}$$

where $\bar{\mathbf{g}}_t = \frac{1}{t}\sum_{i=1}^t \nabla f_i(\mathbf{x}_i)$ denotes the average gradient and $h(\mathbf{x})$ is a strongly convex auxiliary function. As discussed by Xiao [2], RDA can also achieve the $O(\sqrt{T})$ and $O(\log T)$ regret bounds for convex and strongly convex $F_t(\mathbf{x})$ respectively, and is able to generate significantly more sparse decisions than FOBOS by using $r(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ for some $\lambda > 0$. We notice that by setting $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, the update of RDA in (6) can be rewritten to the following proximal mapping

$$\mathbf{x}_{t+1} = \operatorname{prox}_r^{(t\eta_t)^{-1} I_d}\left(-t\eta_t \bar{\mathbf{g}}_t\right). \tag{7}$$

One limitation of the above methods is that they cannot utilize the exp-concavity of loss functions, which is a property weaker than strong convexity [22]. It is well-known that if the composite structure in (1) can be ignored, there exists a classical online Newton step (ONS) method for exp-concave functions [8], which updates as follows

$$\begin{aligned} \mathbf{y}_{t+1} &= \mathbf{x}_t - \frac{1}{\beta} A_t^{-1} \nabla F_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}_{t+1}\|_{A_t}^2 \end{aligned} \tag{8}$$

in each round, where $A_t = \sum_{i=1}^t \nabla F_i(\mathbf{x}_i)\nabla F_i(\mathbf{x}_i)^\top + \epsilon I_d$ approximates the second order information of loss functions, $\epsilon > 0$ is a parameter such that A_t is positive definite, and β is a parameter for adjusting the learning rate. It can achieve a regret bound of $O(d \log T)$ by utilizing the exp-concavity of $F_t(\mathbf{x})$. However, it remains unclear whether the

[1] The abbreviation of this method follows the original paper of Duchi and Singer [1], though there does not exist the second 'O' in its full name.

exp-concavity of $F_t(\mathbf{x})$ can be exploited without ignoring the composite structure. This paper provides an affirmative answer by proposing ProxONS, which attains the $O(d \log T)$ regret bound for exp-concave $f_t(\mathbf{x})$ with any convex regularization function $r(\mathbf{x})$.

2.2. Regularized stochastic optimization

One significant application of methods for regularized online optimization is to solve stochastic optimization with a regularized objective, which can be formulated as

$$\min_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x}) \equiv \mathbb{E}_{\xi\sim P}[f(\mathbf{x};\xi)] + r(\mathbf{x}), \tag{9}$$

where P is an underlying distribution and $f(\mathbf{x};\xi) : \mathbb{R}^d \mapsto \mathbb{R}$ is a loss function depending on ξ. If independent and identically distributed (i.i.d.) samples $\xi_1, \ldots, \xi_T \sim P$ are given, there exists a standard online-to-batch conversion [24], which can extend online methods to this stochastic setting. To be precise, it runs the online methods with the loss function $F_t(\mathbf{x}) = f_t(\mathbf{x}) + r(\mathbf{x})$ where $f_t(\mathbf{x}) = f(\mathbf{x};\xi_t)$, and utilize the average decision

$$\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \tag{10}$$

as the solution of stochastic optimization. By applying the online-to-batch conversion to an online method with a regret bound of R_T, one can simply achieve the following convergence rate

$$\mathbb{E}_{\xi_1,\ldots,\xi_T}\left[F(\bar{\mathbf{x}}_T) - F(\mathbf{x}^*)\right] = O\left(R_T/T\right), \tag{11}$$

where $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$. However, this rate only holds in expectation, which cannot make precise the fluctuations of the convergence. To address this problem, previous studies not only combine the online-to-batch conversion with FOBOS [1] and RDA [2], but also establish a high-probability convergence rate of $O(1/\sqrt{T})$ for convex objective and a high-probability convergence rate of $O(\log T/T)$ for strongly convex objective. Inspired by these studies, in this paper, we similarly extend our ProxONS to the problem of regularized stochastic optimization and achieve a high-probability convergence rate of $O(d \log T/T)$ for exp-concave $f(\mathbf{x};\xi)$ and any convex $r(\mathbf{x})$. We notice that Mahdavi et al. [25] have extended ONS to achieve a high-probability convergence rate of $O(d \log T/T)$ in the special case with exp-concave $f(\mathbf{x};\xi)$ and $r(\mathbf{x}) = 0$. Our result can be regarded as a significant generalization of that in Mahdavi et al. [25].

3. Main results

In this section, we first present our ProxONS for regularized online exp-concave optimization. Then, we introduce its theoretical guarantee on the regret bound and its application in stochastic optimization.

3.1. ProxONS

Our method is a proximal variant of ONS. In the beginning, since $r(\mathbf{x})$ is the only available information, we set

$$\mathbf{x}_1 = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} r(\mathbf{x}). \tag{12}$$

Then, in each round t, similar to ONS, our method exploits the exp-concavity of $f_t(\mathbf{x})$ by utilizing

$$A_t = \sum_{i=1}^{t} \nabla f_i(\mathbf{x}_i)\nabla f_i(\mathbf{x}_i)^\top + \epsilon I_d \tag{13}$$

with $\epsilon > 0$ to approximate the second order information of $f_t(\mathbf{x})$. Subsequently, we use the negative direction of the product of the inverse of matrix A_t and the gradient vector $\nabla f_t(\mathbf{x}_t)$ to make the following update

$$\mathbf{y}_{t+1} = \mathbf{x}_t - (\beta A_t)^{-1}\nabla f_t(\mathbf{x}_t), \tag{14}$$

Algorithm 1 Proximal Online Newton Step

Input: parameters β and ϵ
$\mathbf{x}_1 = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} r(\mathbf{x})$, $A_0 = \epsilon I_d$
for $t = 1, 2, \ldots, T$ **do**
  $A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top$
  $\mathbf{y}_{t+1} = \mathbf{x}_t - (\beta A_t)^{-1}\nabla f_t(\mathbf{x}_t)$
  $\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \left\{ \frac{1}{2}\left\|\mathbf{x} - \mathbf{y}_{t+1}\right\|_{\beta A_t}^2 + r(\mathbf{x}) \right\}$
end for

where the parameter β is also inherited from ONS. This step ensures that the optimization process moves in the direction of decreasing values of the objective function. Moreover, the matrix A_t can be incrementally updated by setting $A_0 = \epsilon I_d$ and computing

$$A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top \tag{15}$$

in each round t.

Furthermore, to explicitly exploiting the regularization $r(\mathbf{x})$, inspired by (3) and (7), we compute the decision $\mathbf{x}_{t+1}$ by using the following proximal mapping of $r(\mathbf{x})$ on the point $\mathbf{y}_{t+1}$ with respect to βA_t, i.e,

$$\mathbf{x}_{t+1} = \operatorname{prox}_r^{\beta A_t}(\mathbf{y}_{t+1}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \left\{ \frac{1}{2}\left\|\mathbf{x} - \mathbf{y}_{t+1}\right\|_{\beta A_t}^2 + r(\mathbf{x}) \right\}. \tag{16}$$

The detailed procedures of our method are summarized in Algorithm 1, which is named as proximal online Newton step (ProxONS).

Remark 1. First, our ProxONS can reduce to the original ONS when $r(\mathbf{x}) = 0$. Second, although our proximal mapping in (16) is inspired by (3) and (7), the positive definite matrix utilized in our ProxONS is βA_t, which is carefully designed according to the ONS step in (14) and significantly differs from that utilized in (3) and (7). Third, we notice that the proximal mapping in (16) may not have a closed-form solution. In that case, we can use existing proximal methods for offline optimization such as the proximal gradient descent method [27] to iteratively find a solution.

3.2. Regret bound

Following Hazan et al. [8], we first introduce some necessary assumptions.

Assumption 1. All functions $f_1(\cdot), \ldots, f_T(\cdot)$ are α-exp-concave over $\mathcal{K}$, i.e., $\exp(-\alpha f_t(\cdot))$ is concave over $\mathcal{K}$ for any $t \in [T]$.

Assumption 2. All gradients of $f_1(\cdot), \ldots, f_T(\cdot)$ are bounded by G, i.e., it holds that

$$\|\nabla f_t(\mathbf{x})\|_2 \le G \tag{17}$$

for any $\mathbf{x} \in \mathcal{K}$ and $t \in [T]$.

Assumption 3. The diameter of the decision set $\mathcal{K}$ is bounded by D, i.e., it holds that

$$\|\mathbf{x} - \mathbf{y}\|_2 \le D \tag{18}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$.

Then, we present the regret bound of ProxONS in the following theorem.

Theorem 1. *Under Assumptions 1, 2 and 3, for any $\mathbf{x} \in \mathcal{K}$, Algorithm 1 with $\beta = \frac{1}{2}\min\left\{\frac{1}{4GD}, \alpha\right\}$ and $\epsilon = \frac{1}{\beta^2 D^2}$ ensures*

$$\sum_{t=1}^{T} F_t(\mathbf{x}_t) - \sum_{t=1}^{T} F_t(\mathbf{x}) \le \frac{1}{2\beta} + \frac{d}{2\beta}\ln\left(\beta^2 G^2 D^2 T + 1\right). \tag{19}$$

Remark 2. Theorem 1 implies that for any convex regularization $r(\mathbf{x})$, our ProxONS has the capability to exploit the exp-concavity of $f_t(\mathbf{x})$ to achieve a regret bound of $O(d \log T)$, which matches the regret bound of ONS in the special case with $r(\mathbf{x}) = 0$. Moreover, in terms of T, our regret bound is better than the $O(\sqrt{T})$ regret bound for convex loss functions achieved by existing methods such as FOBOS and RDA.

3.3. Application to stochastic optimization

Furthermore, we extend our ProxONS to solve stochastic optimization with a regularized objective, which has been formulated in (9). Specifically, we assume that i.i.d. samples $\xi_1, \ldots, \xi_T \sim P$ are given. According to the standard online-to-batch conversion [24], we first generate decisions $\mathbf{x}_1, \ldots, \mathbf{x}_T$ by running our ProxONS with $f_t(\mathbf{x}) = f(\mathbf{x}; \xi_t)$, and then utilize the average decision

$$\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \tag{20}$$

as the solution.

In this case, we introduce the following two assumptions on each individual function $f(\mathbf{x}; \xi)$, which are analogue to Assumptions 1 and 2 required in the online setting.

Assumption 4. For any $\xi \in P$, the function $f(\mathbf{x}; \xi)$ is α-exp-concave over $\mathcal{K}$, i.e., $\exp(-\alpha f(\mathbf{x}; \xi))$ is concave over $\mathcal{K}$.

Assumption 5. For any $\xi \in P$, all gradients of the function $f(\mathbf{x}; \xi)$ are bounded by G, i.e., it holds that

$$\|\nabla f(\mathbf{x}; \xi)\|_2 \leq G \tag{21}$$

for any $\mathbf{x} \in \mathcal{K}$.

Under Assumptions 3, 4, and 5, we establish the following theoretical guarantee on the solution $\bar{\mathbf{x}}_T$.

Theorem 2. *Let $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$ and $\hat{\beta} = \frac{1}{2} \min\left\{ \frac{1}{4GD}, \alpha \right\}$. Under Assumptions 3, 4, and 5, by applying Algorithm 1 with $\beta = \frac{\hat{\beta}}{3}$ and $\epsilon = \frac{1}{\hat{\beta}^2 D^2}$ to functions $\{f_t(\mathbf{x}) = f(\mathbf{x}; \xi_t)\}_{t=1}^T$, with a probability at least $1 - 2\delta$, the average decision $\bar{\mathbf{x}}$ of Algorithm 1 ensures*

$$F(\bar{\mathbf{x}}_T) - F(\mathbf{x}^*) \leq \frac{1}{6\hat{\beta}T} + \frac{3}{2\beta T} \ln\left(\hat{\beta}^2 G^2 D^2 T + 1\right) + \frac{C_{T,1}}{T} + \frac{\beta C_{T,2}}{3T}, \tag{22}$$

where $C_{T,1} = \frac{24}{\hat{\beta}} \log \frac{\sqrt{2T+1}}{\delta} + 2GD\sqrt{\log \frac{2T+1}{\delta^2}}$ and $C_{T,2} = 8G^2D^2 \log \frac{\sqrt{2T+1}}{\delta} + G^2D^2\sqrt{\log \frac{2T+1}{\delta^2}}$.

Remark 3. From Theorem 2, the combination of our ProxONS with the standard online-to-batch conversion can achieve a high-probability convergence rate of $O(d \log T/T)$ for stochastic optimization with a regularized exp-concave objective. This rate matches the convergence rate of Mahdavi et al. [25], which is also derived by exploiting the exp-concavity of $f_t(\mathbf{x})$, but limits to the special case with $r(\mathbf{x}) = 0$. Moreover, in terms of T, our rate is faster than the $O(1/\sqrt{T})$ rate achieved by only utilizing the convexity of the objective.

4. Theoretical analysis

In this section, we first introduce some lemmas to support our analysis, and then prove Theorems 1 and 2.

4.1. Supporting results

The following results are used throughout our analysis.

Lemma 1. *For any $\mathbf{x} \in \mathcal{K}$, Algorithm 1 ensures*

$$\begin{aligned} &\nabla f_t(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}) + \nabla r(\mathbf{x}_{t+1})^\top (\mathbf{x}_{t+1} - \mathbf{x}) \\ \leq &\frac{\beta}{2}\|\mathbf{x}_t - \mathbf{x}\|_{A_t}^2 - \frac{\beta}{2}\|\mathbf{x}_{t+1} - \mathbf{x}\|_{A_t}^2 + \frac{1}{2\beta}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2. \end{aligned} \tag{23}$$

Proof. Let $\mathbf{y}_{t+1} = \mathbf{x}_t - (\beta A_t)^{-1} \nabla f_t(\mathbf{x}_t)$. Notice that

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \left\{ \frac{1}{2} \left\|\mathbf{x} - \mathbf{y}_{t+1}\right\|_{\beta A_t}^2 + r(\mathbf{x}) \right\}. \tag{24}$$

According to the first order optimality condition [28], we have

$$\langle \mathbf{x} - \mathbf{x}_{t+1}, \beta A_t(\mathbf{x}_{t+1} - \mathbf{x}_t) + \nabla f_t(\mathbf{x}_t) + \nabla r(\mathbf{x}_{t+1}) \rangle \geq 0. \tag{25}$$

Then, we have

$$\begin{aligned} &\frac{\beta}{2}\|\mathbf{x}_t - \mathbf{x}\|_{A_t}^2 - \frac{\beta}{2}\|\mathbf{x}_{t+1} - \mathbf{x}\|_{A_t}^2 \\ =&\frac{\beta}{2}\mathbf{x}_t^\top A_t \mathbf{x}_t - \frac{\beta}{2}\mathbf{x}_{t+1}^\top A_t \mathbf{x}_{t+1} + \langle \beta A_t(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x} \rangle \\ \geq&\frac{\beta}{2}\mathbf{x}_t^\top A_t \mathbf{x}_t - \frac{\beta}{2}\mathbf{x}_{t+1}^\top A_t \mathbf{x}_{t+1} + \langle \beta A_t(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_{t+1} \rangle \\ &- \langle \mathbf{x} - \mathbf{x}_{t+1}, \nabla f_t(\mathbf{x}_t) + \nabla r(\mathbf{x}_{t+1}) \rangle \\ =&\frac{\beta}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{A_t}^2 + \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x} + \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ &+ \langle \nabla r(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x} \rangle \\ \geq&\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x} \rangle + \langle \nabla r(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x} \rangle \\ &+ \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ \frac{\beta}{2}\|\mathbf{x}\|_{A_t}^2 + \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle \right\} \\ \geq&\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x} \rangle + \langle \nabla r(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x} \rangle - \frac{1}{2\beta}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2, \end{aligned} \tag{26}$$

where the first inequality is due to (25).

By rearranging terms in the above inequality, we complete the proof.

Lemma 2 (*Lemma 3 of Hazan et al. [8]*). *If a function $f(\mathbf{x}) : \mathcal{K} \mapsto \mathbb{R}$ is α-exp-concave, $\|\nabla f(\mathbf{x})\|_2 \leq G$ for any $\mathbf{x} \in \mathcal{K}$, and Assumption 3 holds, we have*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\hat{\beta}}{2}(\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \tag{27}$$

for $\hat{\beta} \leq \frac{1}{2} \min\left\{ \frac{1}{4GD}, \alpha \right\}$ and any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$.

Lemma 3 (*Lemma 11 of Hazan et al. [8]*). *Let $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathbb{R}^d$ be a sequence of vector such that $\|\mathbf{u}_t\|_2 \leq c$ for any $t \in [T]$ and some $c > 0$. Define*

$$V_t = \sum_{\tau=1}^{t} \mathbf{u}_\tau \mathbf{u}_\tau^\top + \epsilon I_d, \tag{28}$$

where $\epsilon > 0$ is some constant. Then, we have

$$\sum_{t=1}^{T} \mathbf{u}_t^\top V_t^{-1} \mathbf{u}_t \leq d \ln\left(\frac{c^2 T}{\epsilon} + 1 \right). \tag{29}$$

Lemma 4 (*Lemma 4 of Mahdavi et al. [25]*). *Let $f_t(\mathbf{x}) = f(\mathbf{x}; \xi_t)$, $L(\mathbf{x}) = \mathbb{E}_{\xi \sim P}[f(\mathbf{x}; \xi)]$. Under Assumptions 3 and 5, with a probability at least $1 - \delta$, for all $T > 0$, any $\mathbf{x}^* \in \mathcal{K}$, and any $\hat{\beta} > 0$, we have*

$$\begin{aligned} &\sum_{t=1}^{T} \langle \nabla L(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_t), (\mathbf{x}_t - \mathbf{x}^*) \rangle \\ \leq &\frac{\hat{\beta}}{12} \sum_{t=1}^{T} \left(|(\mathbf{x}_t - \mathbf{x}^*)^\top \nabla f_t(\mathbf{x}_t)|^2 + \mathbb{E}_{t-1}\left[|(\mathbf{x}_t - \mathbf{x}^*)^\top \nabla f_t(\mathbf{x}_t)|^2 \right] \right) + C_{T,1}, \end{aligned} \tag{30}$$

where $C_{T,1} = \frac{24}{\hat{\beta}} \log \frac{\sqrt{2T+1}}{\delta} + 2GD\sqrt{\log \frac{2T+1}{\delta^2}}$.

Lemma 5 (*Lemma 5 of Mahdavi et al. [25]*). *Let $f_t(\mathbf{x}) = f(\mathbf{x};\xi_t)$. Under Assumptions 3 and 5, with a probability at least $1-\delta$, for all $T>0$ and any $\mathbf{x}^* \in \mathcal{K}$, we have*

$$\begin{aligned}&3\sum_{t=1}^{T}|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2-5\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]\\ \le&4\left(8G^2D^2\log\frac{\sqrt{2T+1}}{\delta}+G^2D^2\sqrt{\log\frac{2T+1}{\delta^2}}\right).\end{aligned}\tag{31}$$

4.2. Proof of Theorem 1

Let $\hat{\beta}=\frac{1}{2}\min\left\{\frac{1}{4GD},\alpha\right\}$. For any $t\in[T]$, define

$$c_t=\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x})-\frac{\hat{\beta}}{2}(\mathbf{x}-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}-\mathbf{x}_t).\tag{32}$$

Since $f_t(\mathbf{x})$ is exp-concave and $r(\mathbf{x})$ is convex, we have

$$\begin{aligned}&f_t(\mathbf{x}_t)+r(\mathbf{x}_{t+1})-f_t(\mathbf{x})-r(\mathbf{x})\\ \le&c_t+r(\mathbf{x}_{t+1})-r(\mathbf{x})\\ \le&c_t+\nabla r(\mathbf{x}_{t+1})^\top(\mathbf{x}_{t+1}-\mathbf{x}),\end{aligned}\tag{33}$$

where the first inequality is due to Lemma 2 and the definition of c_t, the second inequality is due to the convexity of $r(\mathbf{x})$.

Combining (33) with Lemma 1, we have

$$\begin{aligned}&f_t(\mathbf{x}_t)+r(\mathbf{x}_{t+1})-f_t(\mathbf{x})-r(\mathbf{x})\\ \le&\frac{\beta}{2}\|\mathbf{x}_t-\mathbf{x}\|_{A_t}^2-\frac{\beta}{2}\|\mathbf{x}_{t+1}-\mathbf{x}\|_{A_t}^2+\frac{1}{2\beta}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2\\ &-\frac{\hat{\beta}}{2}|(\mathbf{x}-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2.\end{aligned}\tag{34}$$

Since $\|\mathbf{x}_t-\mathbf{x}\|_{A_t}=\sqrt{(\mathbf{x}_t-\mathbf{x})^\top A_t(\mathbf{x}_t-\mathbf{x})}$, we have

$$\begin{aligned}&\sum_{t=1}^{T}\frac{\beta}{2}\|\mathbf{x}_t-\mathbf{x}\|_{A_t}^2-\sum_{t=1}^{T}\frac{\beta}{2}\|\mathbf{x}_{t+1}-\mathbf{x}\|_{A_t}^2\\ =&\frac{\beta}{2}\|\mathbf{x}_1-\mathbf{x}\|_{A_1}^2+\sum_{t=2}^{T}\frac{\beta}{2}(\mathbf{x}_t-\mathbf{x})^\top A_t(\mathbf{x}_t-\mathbf{x})\\ &-\sum_{t=1}^{T-1}\frac{\beta}{2}(\mathbf{x}_{t+1}-\mathbf{x})^\top A_t(\mathbf{x}_{t+1}-\mathbf{x})-\frac{\beta}{2}\|\mathbf{x}_{T+1}-\mathbf{x}\|_{A_T}^2\\ =&\frac{\beta}{2}\|\mathbf{x}_1-\mathbf{x}\|_{A_1}^2-\frac{\beta}{2}\|\mathbf{x}_{T+1}-\mathbf{x}\|_{A_T}^2\\ &+\sum_{t=2}^{T}\frac{\beta}{2}(\mathbf{x}_t-\mathbf{x})^\top(A_t-A_{t-1})(\mathbf{x}_t-\mathbf{x}).\end{aligned}\tag{35}$$

Then, combining (34) and (35), we have

$$\begin{aligned}&\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t)+r(\mathbf{x}_{t+1})\right)-\sum_{t=1}^{T}\left(f_t(\mathbf{x})+r(\mathbf{x})\right)\\ \le&\frac{\beta}{2}\|\mathbf{x}_1-\mathbf{x}\|_{A_1}^2-\frac{\beta}{2}\|\mathbf{x}_{T+1}-\mathbf{x}\|_{A_T}^2\\ &+\sum_{t=2}^{T}\frac{\beta}{2}(\mathbf{x}_t-\mathbf{x})^\top(A_t-A_{t-1})(\mathbf{x}_t-\mathbf{x})\\ &+\sum_{t=1}^{T}\frac{1}{2\beta}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2-\sum_{t=1}^{T}\frac{\hat{\beta}}{2}|(\mathbf{x}-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\\ \le&\frac{\beta}{2}(\mathbf{x}_1-\mathbf{x})^\top A_0(\mathbf{x}_1-\mathbf{x})+\sum_{t=1}^{T}\frac{1}{2\beta}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2\\ \le&\frac{\beta\epsilon\|\mathbf{x}_1-\mathbf{x}\|_2^2}{2}+\frac{d}{2\beta}\ln\left(\frac{G^2T}{\epsilon}+1\right)\\ \le&\frac{1}{2\beta}+\frac{d}{2\beta}\ln\left(\beta^2G^2D^2T+1\right),\end{aligned}\tag{36}$$

where the second inequality is due to $\beta=\hat{\beta}$, $A_t-A_{t-1}=\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)$ and the fact that the matrix A_t is positive definite, the third inequality is due to $A_0=\epsilon I_d$, $\|\nabla f_t(\mathbf{x}_t)\|_2\le G$ and Lemma 3, and the last inequality is due to $\epsilon=\frac{1}{\beta^2D^2}$ and $\|\mathbf{x}_1-\mathbf{x}\|_2\le D$.

Finally, since $\mathbf{x}_1=\operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}r(\mathbf{x})$ implies that $r(\mathbf{x}_1)\le r(\mathbf{x}_{T+1})$, we have

$$\begin{aligned}&\sum_{t=1}^{T}F_t(\mathbf{x}_t)-\sum_{t=1}^{T}F_t(\mathbf{x})\\ =&\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t)+r(\mathbf{x}_t)\right)-\sum_{t=1}^{T}\left(f_t(\mathbf{x})+r(\mathbf{x})\right)\\ \le&\sum_{t=1}^{T}f_t(\mathbf{x}_t)+\sum_{t=1}^{T}r(\mathbf{x}_{t+1})-\sum_{t=1}^{T}\left(f_t(\mathbf{x})+r(\mathbf{x})\right)\\ \le&\frac{1}{2\beta}+\frac{d}{2\beta}\ln\left(\beta^2G^2D^2T+1\right),\end{aligned}\tag{37}$$

where the first inequality is derived by relaxing $r(\mathbf{x}_1)$ to $r(\mathbf{x}_{T+1})$ and the last inequality is due to (36).

4.3. Proof of Theorem 2

Let $f_t(\mathbf{x})=f(\mathbf{x};\xi_t)$ and $L(\mathbf{x})=\mathbb{E}_{\xi\sim P}[f(\mathbf{x};\xi)]$. Because of Lemma 2, we have

$$f_t(\mathbf{x}_t)-f_t(\mathbf{x}^*)\le\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^*)-\frac{\hat{\beta}}{2}|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2.\tag{38}$$

Then, let $\mathbb{E}_{t-1}[\mathbf{x}]$ denote the expectation conditioned on the randomness until round $t-1$. By taking the expectation of both sides in the above inequality, we have

$$\begin{aligned}&L(\mathbf{x}_t)-L(\mathbf{x}^*)\le\nabla L(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^*)-\frac{\hat{\beta}}{2}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]\\ =&\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^*)-\frac{\hat{\beta}}{2}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]\\ &+\langle\nabla L(\mathbf{x}_t)-\nabla f_t(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^*\rangle.\end{aligned}\tag{39}$$

According to Lemma 1 and $\beta=\frac{\hat{\beta}}{3}$, we have

$$\begin{aligned}&\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^*)+\nabla r(\mathbf{x}_{t+1})^\top(\mathbf{x}_{t+1}-\mathbf{x}^*)\\ \le&\frac{\hat{\beta}}{6}\|\mathbf{x}_t-\mathbf{x}^*\|_{A_t}^2-\frac{\hat{\beta}}{6}\|\mathbf{x}_{t+1}-\mathbf{x}^*\|_{A_t}^2+\frac{3}{2\hat{\beta}}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2.\end{aligned}\tag{40}$$

Combining the above two inequalities, we have

$$\begin{aligned}&(L(\mathbf{x}_t)+r(\mathbf{x}_{t+1}))-(L(\mathbf{x}^*)+r(\mathbf{x}^*))\\ \le&\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^*)+\nabla r(\mathbf{x}_{t+1})^\top(\mathbf{x}_{t+1}-\mathbf{x}^*)\\ &-\frac{\hat{\beta}}{2}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]+\langle\nabla L(\mathbf{x}_t)-\nabla f_t(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^*\rangle\\ \le&\frac{\hat{\beta}}{6}\|(\mathbf{x}_t-\mathbf{x}^*)\|_{A_t}^2-\frac{\hat{\beta}}{6}\|(\mathbf{x}_{t+1}-\mathbf{x}^*)\|_{A_t}^2+\frac{3}{2\hat{\beta}}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2\\ &-\frac{\hat{\beta}}{2}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]+\langle\nabla L(\mathbf{x}_t)-\nabla f_t(\mathbf{x}_t),(\mathbf{x}_t-\mathbf{x}^*)\rangle\\ =&\frac{\hat{\beta}}{6}\|(\mathbf{x}_t-\mathbf{x}^*)\|_{A_{t-1}}^2-\frac{\hat{\beta}}{6}\|(\mathbf{x}_{t+1}-\mathbf{x}^*)\|_{A_t}^2+\frac{3}{2\hat{\beta}}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2\\ &-\frac{\hat{\beta}}{2}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]+\langle\nabla L(\mathbf{x}_t)-\nabla f_t(\mathbf{x}_t),(\mathbf{x}_t-\mathbf{x}^*)\rangle\\ &+\frac{\hat{\beta}}{6}|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2.\end{aligned}\tag{41}$$

Summing the two sides of the above inequality over $t=1,\dots,T$, we have

$$\begin{aligned}&\sum_{t=1}^{T}(L(\mathbf{x}_t)+r(\mathbf{x}_{t+1}))-\sum_{t=1}^{T}(L(\mathbf{x}^*)+r(\mathbf{x}^*))\\ \le&\frac{\hat{\beta}}{6}\|(\mathbf{x}_1-\mathbf{x}^*)\|_{A_0}^2-\frac{\hat{\beta}}{6}\|(\mathbf{x}_{T+1}-\mathbf{x}^*)\|_{A_T}^2+\frac{3}{2\hat{\beta}}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2\end{aligned}$$

Table 2
Summary of datasets.

Datasets	# Training	# Testing	# Features
a9a	32 561	16 281	123
mushrooms	5687	2437	112

$$+\frac{\hat{\beta}}{6}\sum_{t=1}^{T}|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2-\frac{\hat{\beta}}{2}\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[|(\mathbf{x}^*-\mathbf{x}_t)^\top\nabla f_t(\mathbf{x}_t)|^2\right]$$
$$+\sum_{t=1}^{T}\langle\nabla L(\mathbf{x}_t)-\nabla f_t(\mathbf{x}_t),(\mathbf{x}_t-\mathbf{x}^*)\rangle. \tag{42}$$

Combining (42) with Lemma 4, with a probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(L(\mathbf{x}_t)+r(\mathbf{x}_{t+1}))-\sum_{t=1}^{T}(L(\mathbf{x}^*)+r(\mathbf{x}^*))\\
\le&\frac{\hat{\beta}}{6}\|(\mathbf{x}_1-\mathbf{x}^*)\|_{A_0}^2-\frac{\hat{\beta}}{6}\|(\mathbf{x}_{T+1}-\mathbf{x}^*)\|_{A_T}^2+\frac{3}{2\hat{\beta}}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2+C_{T,1}\\
&+\frac{\hat{\beta}}{12}\left(3\sum_{t=1}^{T}|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2-5\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2\right]\right)\\
\le&\frac{1}{6\hat{\beta}}+\frac{3}{2\hat{\beta}}\ln\left(\hat{\beta}^2G^2D^2T+1\right)+C_{T,1}\\
&+\frac{\hat{\beta}}{12}\left(3\sum_{t=1}^{T}|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2-5\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[|(\mathbf{x}_t-\mathbf{x}^*)^\top\nabla f_t(\mathbf{x}_t)|^2\right]\right),
\end{aligned} \tag{43}$$

where the second inequality is due to $A_0=\epsilon I_d$, $\epsilon=\frac{1}{\hat{\beta}^2D^2}$, $\|\mathbf{x}_1-\mathbf{x}^*\|_2\le D$, $\|\nabla f_t(\mathbf{x}_t)\|_2\le G$, and Lemma 3.

Then, combining Lemma 5 with (43), with a probability at least $1-2\delta$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(L(\mathbf{x}_t)+r(\mathbf{x}_{t+1}))-\sum_{t=1}^{T}(L(\mathbf{x}^*)+r(\mathbf{x}^*))\\
\le&\frac{1}{6\hat{\beta}}+\frac{3}{2\hat{\beta}}\ln\left(\hat{\beta}^2G^2D^2T+1\right)+C_{T,1}+\frac{\hat{\beta}C_{T,2}}{3},
\end{aligned} \tag{44}$$

where $C_{T,2}=8G^2D^2\log\frac{\sqrt{2T+1}}{\delta}+G^2D^2\sqrt{\log\frac{2T+1}{\delta^2}}$.

Moreover, because of $\mathbf{x}_1=\operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}r(\mathbf{x})$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(L(\mathbf{x}_t)+r(\mathbf{x}_{t+1}))-\sum_{t=1}^{T}(L(\mathbf{x}^*)+r(\mathbf{x}^*))\\
\ge&\sum_{t=1}^{T}(L(\mathbf{x}_t)+r(\mathbf{x}_t))-\sum_{t=1}^{T}(L(\mathbf{x}^*)+r(\mathbf{x}^*))\\
\ge&T(F(\bar{\mathbf{x}}_T)-F(\mathbf{x}^*)),
\end{aligned} \tag{45}$$

where the last inequality is due to $\bar{x}_T=\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$ and Jensen's inequality. Finally, by combining the above two inequalities, we complete this proof.

5. Experiments

In this section, we perform numerical experiments to verify the effectiveness of our ProxONS. First, we compare ProxONS against FOBOS [1] and RDA [2] to show its ability to exploit the exp-concavity. Moreover, we compare ProxONS with ONS [8] to show its ability to exploit the regularization.

5.1. Experimental settings

We consider online ℓ_1-norm regularized logistic regression for binary classification on two real datasets—a9a and mushrooms from the LIBSVM repository [29]. Both datasets are divided into the training part and the testing part, and their details are summarized in Table 2.[2] In each round t, the learning algorithm first selects a decision $\mathbf{x}_t$ from $\mathcal{K}=\mathbb{R}^d$. Then, it receives a single training example $(\mathbf{w}_t,y_t)$ where $\mathbf{w}_t\in\mathbb{R}^d$ and $y_t\in\{+1,-1\}$, and suffer the ℓ_1 norm regularized logistic loss which is the sum of

$$f_t(\mathbf{x}_t)=\log\left(1+\exp\left(-y_t\mathbf{w}_t^\top\mathbf{x}_t\right)\right) \tag{46}$$

and $r(\mathbf{x}_t)=\lambda\|\mathbf{x}_t\|_1$, where $\lambda=0.001$. It is well-known that the function $f_t(\mathbf{x})$ is convex and exp-concave [30]. Furthermore, the initial decision for all methods is set to the zero vector. For FOBOS and RDA, we set the learning rate as $\eta_t=c/\sqrt{t}$ by searching the constant c from $\{1e{-}3,1e{-}2,\dots,100\}$. For ONS and ProxONS, the parameters β and ϵ are also searched from $\{1e{-}3,1e{-}2,\dots,100\}$.

5.2. Experimental results

We adopt the average loss, the test accuracy, and the sparsity of the final decision as the performance metrics. Fig. 1 shows the comparison of the average loss and the test accuracy among different algorithms on a9a and mushrooms. From the first line of Fig. 1, we find that during the training process, our ProxONS outperforms FOBOS and RDA in terms of the average loss, which verifies the advantage of ProxONS in exploiting the exp-concavity. Moreover, we find that ONS and our ProxONS perform similarly in terms of the average loss, which is reasonable because the procedures of these two methods are very similar except for the treatment of regularization. To further verify the performance of different algorithms, we report the test accuracy as shown in the second line of Fig. 1. It is evident that when the number of training data increases, the overall test accuracy shows an increasing trend, implying that all algorithms are effectively learning. Moreover, ONS and our ProxONS perform better in terms of the test accuracy compared to FOBOS and RDA, which is consistent with their theoretical advantage on the regret bound. In summary, our algorithm's effectiveness is thoroughly demonstrated through comparisons of average loss and test accuracy with other algorithms.

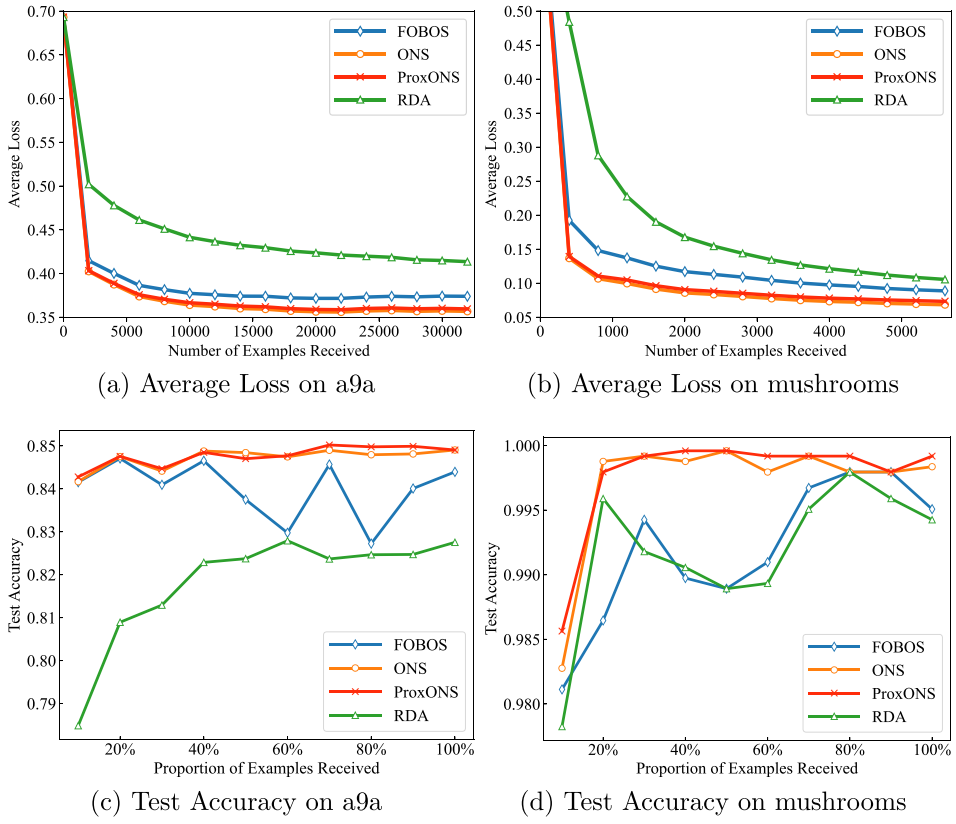
Furthermore, to verify the effect of regularization, we show the sparsity of the final decision generated by different algorithms in Fig. 2. We find that the final decision of ONS is dense, which is caused by ignoring the composite structure. By contrast, the final decision of our ProxONS is much more sparse than ONS, which verifies its ability to exploit the regularization. Probably due to the sparsity of our ProxONS, its test accuracy is slightly better than that of ONS, as shown in the second line of Fig. 1. Additionally, we also notice that the final decisions of RDA and FOBOS are more sparse than our ProxONS. In the future, we will investigate how to further improve the sparsity of our method.

6. Conclusions

In this paper, we propose the ProxONS method for the regularized online exp-concave optimization. According to our analysis, it obtains a regret bound of $O(d\log T)$ for any convex regularization, which is more general than ONS for the non-regularized case, and better than existing methods with an $(\sqrt{T})$ regret bound for convex loss functions. Furthermore, we demonstrate that our ProxONS with the standard online-to-batch conversion can attain a high-probability $O(d\log T/T)$ convergence rate for stochastic optimization with a regularized exp-concave objective. Finally, numerical experiments on two real datasets verify the ability of our ProxONS to exploit the exp-concavity and the regularization.

Notice that both ONS and our ProxONS need $O(d^2)$ time to compute the inverse of A_t in each round. Existing studies [31,32] have combined ONS with matrix sketching [33,34] to address this problem. In the future, it is appealing to apply this idea to accelerate ProxONS.

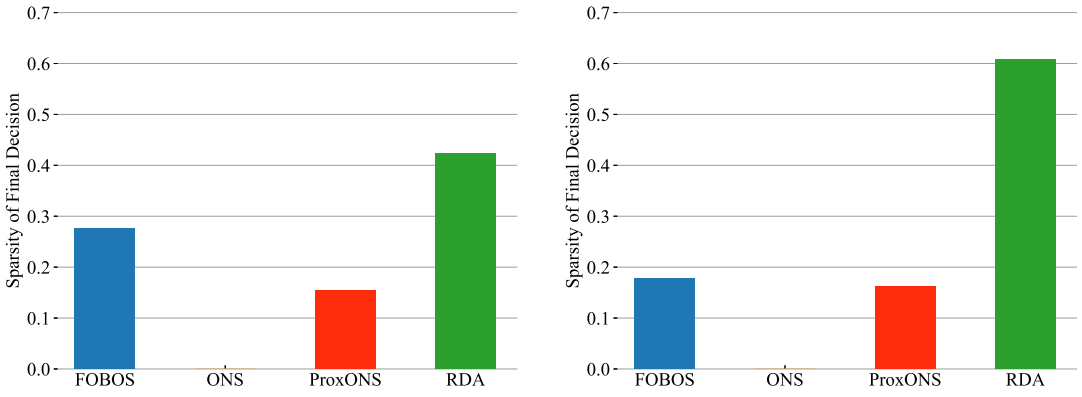
[2] The original mushrooms dataset does not split the training part and the testing part. We randomly select 70% data for training.



(a) Average Loss on a9a (b) Average Loss on mushrooms

(c) Test Accuracy on a9a (d) Test Accuracy on mushrooms

Fig. 1. Comparison of the average loss and the test accuracy among different algorithms.



(a) Sparsity of Final Decision on a9a (b) Sparsity of Final Decision on mushrooms

Fig. 2. Comparison of the sparsity of the final decision among different algorithms.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by State Grid science and technology project, China (5700-202327286A-1-1-ZN).

References

[1] J.C. Duchi, Y. Singer, Efficient online and batch learning using forward backward splitting, J. Mach. Learn. Res. 10 (99) (2009) 2899–2934.

[2] L. Xiao, Dual averaging method for regularized stochastic learning and online optimization, in: Advances in Neural Information Processing Systems 22, 2009, pp. 2543–2596.

[3] J.C. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari, Composite objective mirror descent, in: Proceedings of the 23rd Conference on Learning Theory, 2010, pp. 14–26.

[4] J. Langford, L. Li, T. Zhang, Sparse online learning via truncated gradient, J. Mach. Learn. Res. 10 (3) (2009).

[5] S. Shalev-Shwartz, A. Tewari, Stochastic methods for ℓ_1-regularized loss minimization, J. Mach. Learn. Res. 12 (52) (2011) 1865–1892.

[6] G.-X. Yuan, C.-H. Ho, C.-J. Lin, An improved GLMNET for L1-regularized logistic regression, J. Mach. Learn. Res. 13 (64) (2012) 1999–2030.

[7] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in: Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 928–936.

[8] E. Hazan, A. Agarwal, S. Kale, Logarithmic regret algorithms for online convex optimization, Mach. Learn. 69 (2) (2007) 169–192.

[9] S. Shalev-Shwartz, Y. Singer, A primal-dual perspective of online learning algorithms, Mach. Learn. 69 (2–3) (2007) 115–142.
[10] S. Shalev-Shwartz, Online learning and online convex optimization, Found. Trends Mach. Learn. 4 (2) (2011) 107–194.
[11] S.C. Hoi, D. Sahoo, J. Lu, P. Zhao, Online learning: A comprehensive survey, Neurocomputing 459 (2021) 249–289.
[12] V. Losing, B. Hammer, H. Wersing, Incremental on-line learning: A review and comparison of state of the art algorithms, Neurocomputing 275 (2018) 1261–1274.
[13] C. Wang, S. Xu, D. Yuan, B. Zhang, Z. Zhang, Distributed online convex optimization with a bandit primal-dual mirror descent push-sum algorithm, Neurocomputing 497 (2022) 204–215.
[14] J. Li, C. Gu, Z. Wu, Online distributed stochastic learning algorithm for convex optimization in time-varying directed networks, Neurocomputing 416 (2020) 85–94.
[15] Y. Wan, W.-W. Tu, L. Zhang, Strongly adaptive online learning over partial intervals, Sci. China Inf. Sci. 65 (2022) 202101.
[16] F. Orabona, N. Cesa-Bianchi, C. Gentile, Beyond logarithmic bounds in online learning, in: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, 2012, pp. 823–831.
[17] E. Hazan, S. Kale, Projection-free online learning, in: Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 1843–1850.
[18] Y. Wan, L. Zhang, Projection-free online learning over strongly convex sets, in: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021, pp. 10076–10084.
[19] J.D. Abernethy, P. Bartlett, A. Rakhlin, A. Tewari, Optimal strategies and minimax lower bounds for online convex games, in: Proceedings of the 21st Annual Conference on Learning Theory, 2008, pp. 415–424.
[20] E. Hazan, S. Kale, Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization, J. Mach. Learn. Res. 15 (71) (2014) 2489–2512.
[21] J. Zeng, J. Fang, Z. Xu, Sparse SAR imaging based on $L_{1/2}$ regularization, Sci. China Inf. Sci. 55 (2012) 1755–1775.
[22] E. Hazan, Introduction to online convex optimization, Found. Trends Optim. 2 (3–4) (2016) 157–325.
[23] T. Yang, Z. Li, L. Zhang, A simple analysis for exp-concave empirical minimization with arbitrary convex regularizer, in: Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, 2018, pp. 445–453.
[24] N. Cesa-Bianchi, A. Conconi, C. Gentile, On the generalization ability of on-line learning algorithms, IEEE Trans. Inform. Theory 50 (9) (2004) 2050–2057.
[25] M. Mahdavi, L. Zhang, R. Jin, Lower and upper bounds on the generalization of stochastic exponentially concave optimization, in: Proceedings of the 28th Conference on Learning Theory, 2015, pp. 1305–1320.
[26] Y. Nesterov, Primal-dual subgradient methods for convex problems, Math. Program. 120 (1) (2009) 221–259.
[27] M. Schmidt, N. Roux, F. Bach, Convergence rates of inexact proximal-gradient methods for convex optimization, in: Advances in Neural Information Processing Systems 24, 2011, pp. 1458–1466.
[28] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[29] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (27) (2011) 1–27.
[30] E. Hazan, T. Koren, K.Y. Levy, Logistic regression: Tight bounds for stochastic and online optimization, in: Proceedings of the 27th Conference on Learning Theory, 2014, pp. 197–209.
[31] H. Luo, A. Agarwal, N. Cesa-Bianchi, J. Langford, Efficient second order online learning by sketching, in: Advances in Neural Information Processing Systems 29, 2016, pp. 902–910.
[32] L. Luo, C. Chen, Z. Zhang, W.-J. Li, T. Zhang, Robust frequent directions with application in online learning, J. Mach. Learn. Res. 20 (45) (2019) 1–41.
[33] E. Liberty, Simple and deterministic matrix sketching, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 581—588.
[34] M. Ghashami, E. Liberty, J.M. Phillips, D.P. Woodruff, Frequent directions: Simple and deterministic matrix sketching, SIAM J. Comput. 45 (5) (2016) 1762–1792.