Optimistic Optimisation of Composite Objective with Exponentiated Update

Shao Weijia[1*], Sivrikaya Fikret[2] and Albayrak Sahin[1,2]

[1*]Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Ernst-Reuter-Platz 7, Berlin, 10587, Germany.
[2] GT-ARC Gemeinnützige GmbH, Ernst-Reuter-Platz 7, Berlin, 10587, Country.

*Corresponding author(s). E-mail(s): weijia.shao@campus.tu-berlin.de;
Contributing authors: fikret.sivrikaya@gt-arc.com;
sahin.albayrak@dai-labor.de;

Abstract

This paper proposes a new family of algorithms for online optimisation of composite objectives. The algorithms can be interpreted as the combination of exponentiated gradient and p-norm algorithm. Combined with algorithmic ideas of adaptivity and optimism, the proposed algorithms achieve a sequence dependent regret upper bound, matching the best known bounds for sparse target decision variables. Furthermore, the algorithms have efficient implementations for popular composite objectives and constraints, and can be converted to stochastic optimisation algorithms with optimal accelerated rate for smooth objectives.

Keywords: Exponetiated Gradient, Composite Objective, Online Convex Optimisation, Sparsity

1 Introduction

In recent years, the minimisation of a high dimensional composite objective is involved in many machine learning problems Dhurandhar et al. (2018); Lu, Lin, and Yan (2014); Ribeiro, Singh, and Guestrin (2016); Xie, Bijral, and Ferres (2018). The additive gradient-based adaptive algorithms, such as adaptive subgradient methods (**Adagrad**) Duchi, Hazan, and Singer (2011) and its variants Alacaoglu, Malitsky, Mertikopoulos, and Cevher (2020); Joulani, Raj, Gyorgy, and Szepesvári (2020), which are often applied to estimating deep learning models, outperform standard online learning methods when the gradient vectors are sparse. However, such property can not be expected in every problem. In the task of explaining predictions of image classifier Dhurandhar et al. (2018); Ribeiro et al. (2016), we need to find a sparse perturbation explaining the prediction by solving the following constrained optimisation problems

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & l(x) + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|_2^2 \\ \text{s.t.} \quad & |x_i| \leq c_i \text{ for all } i = 1,\dots,d, \end{aligned}$$

where l is a function relating to the classifier. In such cases, the decision variables are expected to be in an implicitly defined ℓ_1 ball, but gradient vectors are usually very dense, for which the **Adagrad**-style algorithms do not have an optimal theoretical guarantee due to the sub-linear dependence of the regret on the dimension of the decision set.

The exponentiated gradient (**EG**) methods Arora, Hazan, and Kale (2012); Kivinen and Warmuth (1997), which are designed for estimating weights in the positive orthant, enjoy the regret bound growing logarithmically on the dimension. The $\mathbf{EG}^{\pm}$ algorithm generalises this idea to negative weights Kivinen and Warmuth (1997); Warmuth (2007). Given d dimensional problems with the maximum norm of the gradient bounded by G, the regret of performing $\mathbf{EG}^{\pm}$ is upper bounded by $\mathcal{O}(G\sqrt{T\ln d})$. As the performance of the $\mathbf{EG}^{\pm}$ algorithm depends strongly on the choice of hyperparameters, the p-norm algorithm Gentile (2003), which is insensitive to the tuning of hyperparameters, is introduced to approach the logarithmic behavior of $\mathbf{EG}^{\pm}$. Kakade, Shalev-Shwartz, and Tewari (2012) further extends the p-norm algorithm to learning with matrices. An adaptive version of the p-norm algorithm is analysed in Orabona, Crammer, and Cesa-Bianchi (2015), which has a regret upper bound proportional to $\sqrt{\sum_{t=1}^T \|g_t\|_{p_t}^2}$ for a given sequence of gradients $\{g_t\}$. Though this can be upper bounded by $\mathcal{O}(G\sqrt{T\ln d})$ for adaptively chosen p_t, it indicates the difficulty of extending the p-norm algorithm to obtain a regret depending on the sum of squared maximum norms of the gradient sequence, which is the target of this paper.

Recently, Ghai, Hazan, and Singer (2020) has introduced a hyperbolic regulariser for online mirror descent update (**HU**), which can be viewed as an interpolation between gradient descent and **EG**. It has a logarithmic behavior as in **EG** and a learning rate that can be flexibly scheduled as gradient descent. However, many optimisation problems with sparse target has a ℓ_1 or nuclear regulariser in the objective function. Otherwise, the optimisation algorithm has to pick decision variable from a compact decision set. Due to the hyperbolic regulariser, it is difficult to derive a closed form solution for either case. Ghai et al. (2020) has proposed a workaround by tuning a temperature-like hyperparameter so that the chosen decision variable can be normalised at each iteration, which is equivalent to the $\mathbf{EG}^{\pm}$ algorithm and makes the performance strongly depend on the tuning.

In this paper, we propose a family of algorithms for online optimisation of composite objectives. The algorithms employ an entropy-like regulariser combined with algorithmic ideas of adaptivity and optimism. Equipped with the regulariser, the online mirror descent (**OMD**) and the follow-the-regulariser-leader (**FTRL**) algorithms update the absolute value of the scalar components of the decision variable in the same way as **EG** in the positive orthant, while the direction of the decision variable are set in the same way as the p-norm algorithm. To derive the regret upper bound, we first show that the regulariser is strongly convex with respect to the ℓ_1-norm over the ℓ_1 ball. Then we analyse the algorithms in the comprehensive framework for optimistic algorithms with adaptive regularisers Joulani, György, and Szepesvári (2017). Given sequences of gradients $\{g_t\}$ and hints $\{h_t\}$, the proposed algorithms achieve a regret upper bound in the form of $\mathcal{O}(\sqrt{\ln d \sum_{t=1}^{T} \|g_t - h_t\|_\infty})$. With the techniques introduced in Ghai et al. (2020); Kakade et al. (2012), a spectral analogue of the entropy alike regulariser can be found, and proved to be strongly convex with respect to the nuclear norm over the nuclear ball, from which the best known regret upper bound depending on $\sqrt{\ln(\min\{m, n\})}$ for problems in $\mathbb{R}^{m,n}$ follows.

Furthermore, the algorithms have closed form solution for ℓ_1 and nuclear regularised objective function. For ℓ_2 and Frobenius regularised objectives, the update rules involve values of the principle branch of the *Lambert function*, which can be well approximated. For the ℓ_1 or nuclear ball constrained problems, we propose a sorting based procedure projecting the solution to the decision set. Finally, the proposed online algorithms can be converted into the algorithms for stochastic optimisation with the technique introduced in Joulani et al. (2020). We show that the converted algorithms guarantee an optimal accelerated convergence rate for smooth function. The convergence rate depends logarithmically on the dimension of the problem, which suggests its advantage compared to the accelerated **Adagrad**-Style algorithms Cutkosky (2019); Joulani et al. (2020); Levy, Yurtsever, and Cevher (2018).

The rest of the paper is organised as follows. Section 2 reviews the existing work. Section 3 introduces the notation and preliminary concepts. Next, we present and analyse our algorithms in Section 4. In Section 5, we derive efficient implementations for some popular choices of composite objectives, constraints and the stochastic optimisation. Section 6 demonstrates the experimental results using both synthetic and real world data. Finally, we conclude our work in Section 7.

2 Related Work

Our primary motivation is to solve the optimisation problems with an elastic net regulariser in their objective function, which are highly involved in the attacking (Cancela, Bolón-Canedo, and Alonso-Betanzos (2021); Carlini and Wagner (2017); Chen, Sharma, Zhang, Yi, and Hsieh (2018)) and explaining (Dhurandhar et al. (2018); Ribeiro et al. (2016)) deep neural networks. To solve the problem the proximal gradient method (**PGD**) Nesterov (2003) and its accelerated variants Beck and Teboulle (2009) are applied. However, these algorithms are not practical due to the tuning of the stepsize and the dependence of the convergence on the smoothness of the loss function.

In recent years, adaptive online learning algorithms, e.g., Duchi et al. (2011); Orabona et al. (2015); Orabona and Pál (2018), have become popular in the machine learning community. Given the gradient vectors $g_1, \ldots, g_t$ received at iteration t, the core idea of these algorithms is to set the stepsize proportional to $\frac{1}{\sqrt{\sum_{s=1}^{t-1} \|g_s\|_*^2}}$

to ensure a regret upper bounded by $\mathcal{O}(\sqrt{\sum_{t=1}^{T}\|g_t\|_*^2})$ after T iterations. Online learning algorithms with this adaptive regret can be directly applied to the stochastic optimisation problems (Alacaoglu et al. (2020); Li and Orabona (2019)) or can be easily converted into a stochastic algorithm Cesa-Bianchi and Gentile (2008) with convergence rate $\mathcal{O}(\frac{1}{\sqrt{T}})$. For unconstrained problems with smooth loss functions, this rate can be further improved to $\mathcal{O}(\frac{1}{T^2})$ by applying the acceleration techniques Cutkosky (2019); Kavis, Levy, Bach, and Cevher (2019); Levy et al. (2018). These acceleration techniques do not require prior knowledge on the smoothness of the loss function and guarantee convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ for non-smooth functions. For constrained problems with composite objectives, the combination of the adaptive optimistic optimisation Joulani et al. (2017) and the acceleration technique of proposed in Joulani et al. (2020) can be applied to achieve a convergence rate of $\mathcal{O}(\frac{1}{T^2})$.

Given a d-dimensional problem, the algorithms mentioned above have a regret upper bound depending (sub-) linearly on d. We are interested in a logarithmic dependence of the regret on the dimension, which can be attained by the **EG** family algorithms (Arora et al. (2012); Kivinen and Warmuth (1997); Warmuth (2007)) and their adaptive optimistic extension Steinhardt and Liang (2014). However, these algorithms work only for decision sets in the form of cross-polytopes, and require prior knowledge on the radius of the decision set for general convex optimisation problems. The p-norm algorithm Gentile (2003); Kakade et al. (2012) does not have the limitation mentioned above; however, it is impossible to obtain a adaptive regret upper bound depending on the gradient sequence Orabona et al. (2015), for which the acceleration techniques can not be applied.

The **HU** algorithm Ghai et al. (2020), which interpolates gradient descent and **EG**, can theoretically be applied to loss functions with elastic net regularisers and decision sets other than cross-polytopes. However, it is not practical due to the complex projection step. Following the idea of **HU**, we propose more practical algorithms interpolating **EG** and the p-norm algorithm. Our algorithms are based on an entropy-like regulariser, some similar variants of which are previously discussed in Cutkosky and Boahen (2017); Orabona (2013) for unconstrained optimisation.

3 Preliminary

The focus of this paper is the online convex optimisation (**OCO**) problems with the decision variable taken from a compact convex subset $\mathcal{K} \subseteq \mathbb{X}$ of finite dimensional vector space equipped with norm $\|\cdot\|$. Given a sequence of vectors $\{v_t\}$, we use the compressed-sum notation $v_{1:t} = \sum_{s=1}^{t} v_s$ for simplicity. We denote by $\mathbb{X}_*$ the dual space with dual norm $\|\cdot\|_*$. The bi-linear map combining vectors in $\mathbb{X}_*$ and $\mathbb{X}$ is denoted by

$$\langle \cdot, \cdot \rangle : \mathbb{X}_* \times \mathbb{X} \to \mathbb{R}, (\theta, x) \mapsto \theta x.$$

For $\mathbb{X} = \mathbb{R}^d$, we denote by $\|\cdot\|_1$ the ℓ_1 norm, the dual norm of which is the maximum norm denoted by $\|\cdot\|_\infty$. It is well known that the ℓ_2 norm denoted by $\|\cdot\|_2$ is self-dual. In case $\mathbb{X}$ is the space of the matrices, for simplicity, we also use $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ for the nuclear, Frobenius and spectral norm, respectively.

Let $\sigma : \mathbb{R}^{m,n} \to \mathbb{R}^{\min\{m,n\}}$ be the function mapping a matrix to its singular values. Define

$$\operatorname{diag} : \mathbb{R}^{\min\{m,n\}} \to \mathbb{R}^{m,n}, x \mapsto X$$

with

$$X_{ij} = \begin{cases} x_i, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, the singular value decomposition (**SVD**) of a matrix X can be expressed as

$$X = U \operatorname{diag}(\sigma(X)) V^\top.$$

Similarly, we write the eigendecomposition of a symmetric matrix X as

$$X = U \operatorname{diag}(\lambda(X)) U^\top,$$

where we denote by $\lambda : \mathbb{S}^d \mapsto \mathbb{R}^d$ the function mapping a symmetric matrix to its spectrum.

Given a convex set $\mathcal{K} \subseteq \mathbb{X}$ and a convex function $f : \mathcal{K} \to \mathbb{R}$ defined on $\mathcal{K}$, we denote by $\partial f(y) = \{g \in \mathbb{X}_* | \forall y \in \mathcal{K}. f(x) - f(y) \geq \langle g, x - y \rangle\}$ the subgradient of f at y. We refer to $\triangledown f(y)$ any element in $\partial f(y)$. A function is η-strongly convex w.r.t. $\|\cdot\|$ over $\mathcal{K}$ if

$$f(x) - f(y) \geq \langle \triangledown f(y), x - y \rangle + \frac{\eta}{2}\|x - y\|^2$$

holds for all $x, y \in \mathcal{K}$ and $\triangledown f(y) \in \partial f(y)$.

4 Algorithms and Analysis

The online convex optimisation (**OCO**) problem can be considered as an iterative game between a player and an adversary. In each round t of the game, the player makes a decision $x_t \in \mathcal{K}$. Next, the adversary selects and reveals a convex loss f_t to the player, who then suffers the loss $f_t(x_t)$. The target is to develop algorithms minimising the regret of not having chosen the best decision $x \in \mathcal{K}$

$$\mathcal{R}_{1:T} = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x).$$

In this paper, we consider the composite loss function $f_t(x) = l_t(x) + r_t(x)$, where $l_t : \mathcal{K} \to \mathbb{R}$ is a convex function revealed at each iteration and $r_t : \mathbb{X} \to \mathbb{R}_{\geq 0}$ a known closed convex function controlling the model complexity.

4.1 Generalised Entropy for Online Convex Optimisation

To provide an intuition of our algorithms, we begin with a short review on **EG** and p-norm algorithms for the case $f_t = l_t$. The **EG** algorithm can be considered as an instance of **OMD**, the update rules of which is given by

$$x_{t+1,i} \propto \exp(\ln(x_{t,i}) - \frac{1}{\eta} g_{t,i}),$$

where $g_t \in \partial f_t(x_t)$ is the subgradient, and $\eta > 0$ is the learning rate. Although the algorithm has the expected logarithmic dependence on the dimension, its update rule is applicable only to the decision variables on the standard simplex. For the problem with decision variables taken from an ℓ_1 ball $\{x | \|x\|_1 \leq D\}$, one can simply apply **EG** to the subgradient vector $[\frac{D}{2} g_t^\top, -\frac{D}{2} g_t^\top]^\top$ to update $[x_{t+1,+}^\top, x_{t+1,-}^\top]^\top$ at iteration t, and use $x_{t+1,+} - x_{t+1,-}$ for the prediction. However, if the decision set is implicitly given by a regularisation, the parameter D has to be tuned. Since applying an overestimated D increases regret, while using an underestimated D decreases the freedom of the model, the algorithm is sensitive to tuning. For composite objectives, **EG** is not practical due to its update rule.

Compared to **EG**, the p-norm algorithm, the update rule of which is given by

$$\begin{aligned} y_{t+1,i} =& \operatorname{sgn}(x_{t,i}) |x_{t,i}|^{p-1} \|x_t\|_p^{\frac{2}{p-1}} - \frac{1}{\eta} g_{t,i} \\ x_{t+1,i} =& \operatorname{sgn}(y_{t+1,i}) |y_{t+1,i}|^{q-1} \|y_{t+1}\|_q^{\frac{2}{q-1}}, \end{aligned}$$

is better applicable for unknown D. To combine the ideas of **EG** and the p-norm algorithm, we consider the following generalised entropy function.

$$\phi : \mathbb{R} \to \mathbb{R}, x \mapsto \alpha(|x| + \beta) \ln(\frac{|x|}{\beta} + 1) - \alpha|x|. \tag{1}$$

The function is twice differentiable and strictly convex, which is shown in the next lemma.

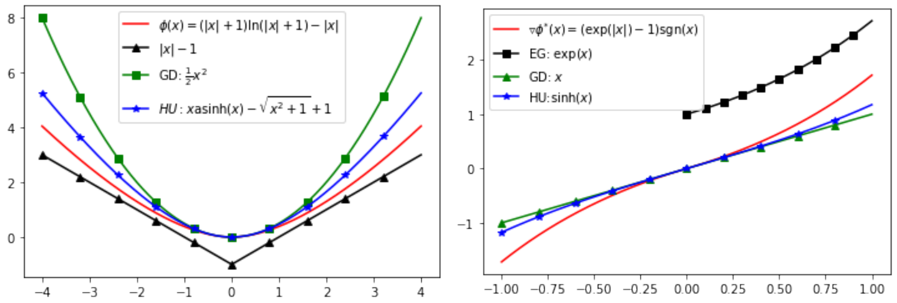
Lemma 1 *ϕ is twice continuous differentiable and strictly convex with*

1. $\phi'(x) = \alpha \ln(\frac{|x|}{\beta} + 1) \operatorname{sgn}(x)$
2. $\phi''(x) = \frac{\alpha}{|x| + \beta}$.

Furthermore, the convex conjugate given by $\phi^* : \mathbb{R} \to \mathbb{R}, \theta \mapsto \alpha\beta \exp \frac{|\theta|}{\alpha} - \beta|\theta| - \alpha\beta$ is also twice continuous differentiable with

1. $\phi^{*\prime}(\theta) = (\beta \exp \frac{|\theta|}{\alpha} - \beta) \operatorname{sgn}(\theta)$
2. $\phi^{*\prime\prime}(\theta) = \frac{\beta}{\alpha} \exp \frac{|\theta|}{\alpha}$.

Since we can expand the natural logarithm as $\ln(\frac{|x|}{\beta} + 1) = \frac{|x|}{\beta} - \frac{|x|^2}{2\beta^2} + \frac{|x|^3}{3\beta^3} - \ldots$, $\phi(x)$ can be intuitively considered as an interpolation between the absolute value and square. As observed in Figure 1a, it is closer to the absolute value compared to the hyperbolic entropy introduced in Ghai et al. (2020). Moreover, running **OMD** with



(a) Regularisers

(b) Mirror Maps

Fig. 1: Comparison of Convex Regularisers

regulariser $x \mapsto \sum_{i=1}^{d} \phi(x_i)$ yields an update rule

$$\begin{aligned} y_{t+1,i} &= \text{sgn}(x_{t,i}) \ln(\frac{|x_{t,i}|}{\beta} + 1) - \frac{1}{\alpha} g_{t,i} \\ x_{t+1,i} &= \text{sgn}(y_{t+1,i})(\beta \exp(|y_{t+1,i}|) - \beta), \end{aligned}$$

which sets the signs of coordinates like the p-norm algorithm and updates the scale similarly to **EG**. As illustrated in Figure 1b, the mirror map $\nabla\phi^*$ is close to the mirror map of **EG**, while the behavior of **HU** is more similar to the gradient descent update.

To obtain an adaptive and optimistic algorithm, we define the following time varying function

$$\phi_t : \mathbb{R}^d \to \mathbb{R}, x \mapsto \alpha_t \sum_{i=1}^{d} ((|x_i| + \beta) \ln(\frac{|x_i|}{\beta} + 1) - |x_i|), \tag{2}$$

and apply it to the adaptive optimistic **OMD** (**AO-OMD**) given by

$$x_{t+1} = \arg\min_{x \in \mathcal{K}} \langle g_t - h_t + h_{t+1}, x \rangle + r_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, x_t) \tag{3}$$

for the sequence of subgradients $\{g_t\}$ and hints $\{h_t\}$. In a bounded domain, ϕ_t is strongly convex w.r.t. $\|\cdot\|_1$, which is shown in the next lemma.

Lemma 2 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be convex and bounded such that $\|x\|_1 \leq D$ for all $x \in \mathcal{K}$. Then we have for all $x, y \in \mathcal{K}$*

$$\phi_t(x) - \phi_t(y) \geq \nabla\phi_t(y)^\top (x - y) + \frac{\alpha_t}{D + d\beta} \|x - y\|_1^2.$$

With the property of the strong convexity, the regret of **AO-OMD** with regulariser (2) can be analysed in the framework of optimistic algorithm Joulani et al. (2017) and is upper bounded by the following theorem.

Theorem 1 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a compact convex set. Assume that there is some $D > 0$ such that $\|x\|_1 \leq D$ holds for all $x \in \mathcal{K}$. Let $\{x_t\}$ be the sequence generated by update rule (3) with regulariser (2). Setting $\beta = \frac{1}{d}$, $\eta = \sqrt{\frac{1}{\ln(D+1)+\ln d}}$, and $\alpha_t = \eta\sqrt{\sum_{s=1}^{t-1}\|g_s - h_s\|_\infty^2}$, we obtain*

$$\mathcal{R}_{1:T} \leq r_1(x_1) + c(d,D)\sqrt{\sum_{t=1}^{T}\|g_t - h_t\|_\infty^2}$$

for some $c(d,D) \in \mathcal{O}(D\sqrt{\ln(D+1)+\ln d})$.

EG can also be considered as an instance of **FTRL** with constant learning rate. The update rule of the adaptive optimistic **FTRL** (**AO-FTRL**) is given by

$$x_{t+1} = \arg\min_{x \in \mathcal{K}} \langle g_{1:t} + h_{t+1}, x\rangle + r_{1:t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, x_1). \tag{4}$$

The regret of **AO-FTRL** is upper bounded by the following theorem.

Theorem 2 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a compact convex set with $d > e$. Assume that there is some $D \geq 1$ such that $\|x\|_1 \leq D$ holds for all $x \in \mathcal{K} \subseteq \mathbb{R}^d$. Let $\{x_t\}$ be the sequence generated by updating rule (4) with regulariser (2) at iteration t. Setting $\beta = \frac{1}{d}$, $\eta = \sqrt{\frac{1}{\ln(D+1)+\ln d}}$ and $\alpha_t = \eta\sqrt{\sum_{s=1}^{t-1}\|g_s - h_s\|_\infty^2}$ we obtain*

$$\mathcal{R}_{1:T} \leq c(d,D)\sqrt{\sum_{t=1}^{T}\|g_t - h_t\|_\infty^2}$$

for some $c(d,D) \in \mathcal{O}(D\sqrt{\ln(D+1)+\ln d})$.

4.2 Spectral Algorithm

We now consider the setting in which the decision variables are matrices taken from a compact convex set $\mathcal{K} \subseteq \mathbb{R}^{m,n}$. A direct attempt to solve this problem is to apply the updating rule (3) or (4) to the vectorised matrices. A regret bound of $\mathcal{O}(D\sqrt{T\ln(mn)})$ can be guaranteed if the ℓ_1 norm of the vectorised matrices from $\mathcal{K}$ are bounded by D, which is not optimal. In many applications, elements in $\mathcal{K}$ are assumed to have bounded nuclear norm, for which the regulariser

$$\Phi_t = \phi_t \circ \sigma \tag{5}$$

can be applied. The next theorem gives the strong convexity of Φ_t w.r.t. $\|\cdot\|_1$ over $\mathcal{K}$.

Theorem 3 *Let $\sigma : \mathbb{R}^{m,n} \to \mathbb{R}^d$ be the function mapping a matrix to its singular values. Then the function $\Phi_t = \phi_t \circ \sigma$ is $\frac{\alpha_t}{2(D+\min\{m,n\}\beta)}$-strongly convex w.r.t. the nuclear norm over the nuclear ball with radius D.*

The proof of Theorem 3 follows the idea introduced in Ghai et al. (2020). Define the operator

$$S : \mathbb{R}^{m,n} \to \mathbb{S}^{m+n}, X \mapsto \begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix}$$

The set $\mathcal{X} = \{S(X) | \in \mathbb{R}^{m,n}\}$ is a finite dimensional linear subspace of the space of symmetric matrices $\mathbb{S}^{m+n}$. Its dual space $\mathcal{X}_*$ determined by the Frobenius inner product can be represented by $\mathcal{X}$ itself. For any $S(X) \in \mathcal{X}$, the set of eigenvalues of $S(X)$ consists of the singular values and the negative singular values of X. Since ϕ is even, we have $\sum_{i=1}^d \phi(\sigma_i(X)) = \sum_{i=1}^d \phi(\lambda_i(X))$ for symmetric X. The next lemma shows that both $\Phi_t|_{\mathcal{X}}$ and $\Phi_t^*|_{\mathcal{X}}$ are twice differentiable.

Lemma 3 *Let $f : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable. Then the function given by*

$$F : \mathbb{S}^d \to \mathbb{R}, X \mapsto \sum_{i=1}^d f(\lambda_i(X))$$

is twice differentiable. Furthermore, let $X \in \mathbb{S}^d$ be a symmetric matrix with eigenvalue decomposition

$$X = U \operatorname{diag}(\lambda_1(X), \ldots, \lambda_d(X)) U^\top.$$

Define the matrix of the divided difference $\Gamma(f, X) = [\gamma(f, X)_{ij}]$ with

$$\gamma(f, X)_{ij} = \begin{cases} \frac{f(\lambda_i(X)) - f(\lambda_j(X))}{\lambda_i(X) - \lambda_j(X)}, & \text{if } \lambda_i(X) \neq \lambda_j(X) \\ f'(\lambda_i(X)), & \text{otherwise} \end{cases}$$

Then for any $G, H \in \mathbb{S}^d$, we have

$$D^2 F(X)(G, H) = \sum_{i.j} \gamma(f', X)_{ij} \tilde{g}_{ij} \tilde{h}_{ij},$$

where $\tilde{g}_{ij}$ and $\tilde{h}_{ij}$ are the elements of the i-th row and j-th column of the matrix $U^\top G U$ and $U^\top H U$, respectively.

Lemma 3 implies the unsurprising positive semidefiniteness of $D^2 F(X)$ for convex f. Furthermore, the exact expression of the second differential allows us to show the local smoothness of Φ_t^* using the local smoothness of ϕ^*. Together with Lemma 4, the locally strong convexity of $\Phi_t|_{\mathcal{X}}$ can be proved.

Lemma 4 *Let $\Phi : \mathbb{X} \to \mathbb{R}$ be a closed convex function such that Φ^* is twice differentiable at some $\theta \in \mathbb{X}_*$ with positive definite $D^2\Phi^*(\theta) \in \mathcal{L}(\mathbb{X}_*, \mathcal{L}(\mathbb{X}_*, \mathbb{R}))$. Suppose that $D^2\Phi^*(\theta)(v, v) \leq \|v\|_*^2$ holds for all $v \in \mathbb{X}_*$. Then we have $D^2\Phi(D\Phi^*(\theta))(x, x) \geq \|x\|^2$ for all $x \in \mathbb{X}$.*

Lemma 4 can be considered as a generalised version of the local duality of smoothness and convexity proved in Ghai et al. (2020). The required positive definiteness of $D^2\Phi_t^*(\theta)$ is guaranteed by the exact expression of the second differential described in Lemma 3 and the fact $\phi^{*\prime\prime}(\theta) > 0$ for all $\theta \in \mathbb{R}$. Finally, using the construction of $\mathcal{X}$, the locally strong convexity of $\Phi_t|_{\mathcal{X}}$ can be extended to Φ_t. The complete proofs of Theorem 3 and the technical lemmata can be found in Appendix B.1.

With the property of the strong convexity, the regret of applying (5) to **AO-OMD** and **AO-FTRL** can be upper bounded by the following theorems.

Theorem 4 *Let $\mathcal{K} \subseteq \mathbb{R}^{m,n}$ be a compact convex set. Assume that there is some $D > 0$ such that $\|x\|_1 \leq D$ holds for all $x \in \mathcal{K}$. Let $\langle x \rangle$ be the sequence generated by update rule (3) with regulariser (5) at iteration t. Setting $\beta = \frac{1}{\min\{m,n\}}$, $\eta = \sqrt{\frac{1}{\ln(D+1)+\ln\min\{m,n\}}}$, and $\alpha_t = \eta\sqrt{\sum_{s=1}^{t-1}\|g_s - h_s\|_\infty^2}$, we obtain*

$$\mathcal{R}_{1:T} \leq r_1(x_1) + c(m,n,D)\sqrt{\sum_{t=1}^{T}\|g_t - h_t\|_\infty^2}$$

with $c(m,n,D) \in \mathcal{O}(D\sqrt{\ln(D+1)+\ln\min\{m,n\}})$.

Theorem 5 *Let $\mathcal{K} \subseteq \mathbb{R}^{\min\{m,n\}}$ be a compact convex set with $\min\{m,n\} > e$. Assume that there is some $D \geq 1$ such that $\|x\|_1 \leq D$ holds for all $x \in \mathcal{K}$. Let $\langle x \rangle$ be the sequence generated by updating rule (4) with time varying regulariser (5). Setting $\beta = \frac{1}{\min\{m,n\}}$, $\eta = \sqrt{\frac{1}{\ln(D+1)+\ln\min\{m,n\}}}$ and $\alpha_t = \eta\sqrt{\sum_{s=1}^{t-1}\|g_s - h_s\|_\infty^2}$, we obtain*

$$\mathcal{R}_{1:T} \leq c(m,n,D)\sqrt{\sum_{t=1}^{T}\|g_t - h_t\|_\infty^2},$$

with $c(m,n,D) \in \mathcal{O}(D\sqrt{\ln(D+1)+\ln\min\{m,n\}})$.

With regulariser (5), both **AO-OMD** and **AO-FTRL** guarantee a regret upper bound proportional to $\sqrt{\ln\min\{m,n\}}$, which is the best known dependence on the size of the matrices.

5 Derived Algorithms

Given $z_{t+1} \in \mathbb{X}_*$ and a time varying convex function $R_{t+1} : \mathbb{X} \to \mathbb{R}$, we consider the following updating rule

$$\begin{aligned} y_{t+1} &= \nabla\phi_{t+1}^*(z_{t+1}) \\ x_{t+1} &= \arg\min_{x\in\mathcal{K}} R_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, y_{t+1}). \end{aligned} \tag{6}$$

It is easy to verify that (6) is equivalent to

$$\begin{aligned} x_{t+1} =& \arg\min_{x\in\mathcal{K}} R_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, y_{t+1}) \\ =& \arg\min_{x\in\mathcal{K}} R_{t+1}(x) + \phi_{t+1}(x) - \langle \nabla\phi_{t+1}(y_{t+1}), x\rangle \\ =& \arg\min_{x\in\mathcal{K}} R_{t+1}(x) + \phi_{t+1}(x) - \langle z_{t+1}, x\rangle \end{aligned}$$

Setting $z_{t+1} = \nabla\phi_{t+1}(x_t) - g_t + h_t - h_{t+1}$ and $R_{t+1} = r_{t+1}$, we obtain the **AO-OMD** update

$$\begin{aligned} x_{t+1} =& \arg\min_{x\in\mathcal{K}} \langle g_t - h_t + h_{t+1}, x\rangle - \langle \nabla\phi_{t+1}(x_t), x\rangle + \phi_{t+1}(x) + r_{t+1}(x) \\ =& \arg\min_{x\in\mathcal{K}} \langle g_t - h_t + h_{t+1}, x\rangle + r_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, x_t). \end{aligned}$$

Setting $z_{t+1} = -\nabla\phi_{t+1}(x_1) + g_{1:t} + h_{t+1}$ and $R_{t+1} = r_{1:t+1}$, we obtain the **AO-FTRL** update

$$x_{t+1} = \arg\min_{x\in\mathcal{K}} \langle g_{1:t} - \theta_1 + h_{t+1}, x\rangle + \phi_{t+1}(x) + r_{1:t+1}(x).$$

The rest of this section focuses on solving the second line of (6) for some popular choices of r and $\mathcal{K}$.

5.1 Elastic Net Regularisation

We first consider the setting of $\mathcal{K} = \mathbb{R}^d$ and $R_{t+1}(x) = \gamma_1\|x\|_1 + \frac{\gamma_2}{2}\|x\|_2^2$, which has countless applications in machine learning. It is easy to verify that the Bregman divergence associated with ψ_{t+1} is given by

$$\begin{aligned} \mathcal{B}_{\phi_{t+1}}(x, y) =& \alpha_{t+1}\sum_{i=1}^{d}((|x_i| + \beta)\ln(\frac{|x_i|}{\beta} + 1) - |x_i| \\ & - (\operatorname{sgn}(y_i)x_i + \beta)\ln(\frac{|y_i|}{\beta} + 1) + |y_i|). \end{aligned}$$

The minimiser of

$$R_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, y_{t+1})$$

in $\mathbb{R}^d$ can be simply obtained by setting the subgradient to 0. For $ln(\frac{|y_{i,t+1}|}{\beta} + 1) \leq \frac{\gamma_1}{\alpha_{t+1}}$, we set $x_{i,t+1} = 0$. Otherwise, the 0 subgradient implies $\operatorname{sgn}(x_{i,t+1}) = \operatorname{sgn}(y_{i,t+1})$ and $|x_{i,t+1}|$ given by the root of

$$\ln(\frac{|y_{i,t+1}|}{\beta} + 1) = \ln(\frac{|x_{i,t+1}|}{\beta} + 1) + \frac{\gamma_1}{\alpha_{t+1}} + \frac{\gamma_2}{\alpha_{t+1}}|x_{i,t+1}|$$

for $i = 1, \ldots, d$. For simplicity, we set $a = \beta$, $b = \frac{\gamma_2}{\alpha_{t+1}}$ and $c = \frac{\gamma_1}{\alpha_{t+1}} - \ln(\frac{|y_{i,t+1}|}{\beta} + 1)$. It can be verified that $|x_{i,t+1}|$ is given by

$$|x_{i,t+1}| = \frac{1}{b}W_0(ab\exp(ab - c)) - a, \tag{7}$$

where W_0 is the principle branch of the *Lambert function* and can be well approximated. For $\gamma_2 = 0$, i.e. the ℓ_1 regularised problem, $|x_{i,t+1}|$ has the closed form solution

$$|x_{i,t+1}| = \beta \exp(\ln(\frac{|y_{i,t+1}|}{\beta} + 1) - \frac{\gamma_1}{\alpha_{t+1}}) - \beta. \tag{8}$$

The implementation is described in Algorithm 1.

Algorithm 1 Solving $\min_{x \in \mathbb{R}^d} R_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, y_{t+1})$

```
for $i = 1, \dots, d$ do
    if $ln(\frac{|y_{i,t+1}|}{\beta} + 1) \leq \frac{\gamma_1}{\alpha_{t+1}}$ then
        $x_{t+1,i} \leftarrow 0$
    else
        $a \leftarrow \beta$
        $b \leftarrow \frac{\gamma_2}{\alpha_{t+1}}$
        $c \leftarrow \frac{\gamma_1}{\alpha_{t+1}} - \ln(\frac{y_{t+1,i}}{\beta} + 1)$
        $x_{t+1,i} \leftarrow \frac{1}{b} W_0(ab \exp(ab - c)) - a$
    end if
end for
Return $x_{t+1}$
```

5.2 Nuclear and Frobenius Regularisation

Similarly, we consider $\mathcal{K} = \mathbb{R}^{m,n}$ with a regulariser $R_{t+1}(x) = \gamma_1 \|x\|_1 + \frac{\gamma_2}{2} \|x\|_2^2$ mixed with nuclear and Frobenius norm. The second line of update rule (6) can be implemented as follows

$$\begin{aligned} \text{Compute SVD: } y_{t+1} =& U_{t+1} \operatorname{diag}(\tilde{y}_{t+1}) V_{t+1}^\top \\ \text{Apply Algorithm 1: } \tilde{x}_{t+1} =& \operatorname*{arg\,min}_{x \in \mathbb{R}^d} R_{t+1}(x) + \mathcal{B}_{\phi_{t+1}}(x, \tilde{y}_{t+1}) \\ \text{Construct: } x_{t+1} =& U_{t+1} \operatorname{diag}(\tilde{x}_{t+1}) V_{t+1}^\top. \end{aligned} \tag{9}$$

Let y_{t+1} and $\tilde{y}_{t+1}$ be as defined in (9). It is easy to verify

$$\begin{aligned} &\operatorname*{arg\,min}_{x \in \mathbb{R}^{m,n}} R_{t+1}(x) + \mathcal{B}_{\Phi_{t+1}}(x, y_{t+1}) \\ =& \operatorname*{arg\,min}_{x \in \mathbb{R}^{m,n}} R_{t+1}(x) + \Phi_{t+1}(x) - \langle U_{t+1} \operatorname{diag}(\nabla \phi_{t+1}(\tilde{y}_{t+1})) V_{t+1}^\top, x \rangle_F. \end{aligned} \tag{10}$$

From the characterisation of subgradient, it follows

$$\nabla R_{t+1}(x) = U \operatorname{diag}(\gamma_1 \operatorname{sgn}(\sigma(x)) + \gamma_2 \sigma(x)) V^\top,$$

and

$$\nabla \Phi_t(x) = U \operatorname{diag}(\nabla \phi_t(\sigma(x))) V^\top,$$

where $x = U \operatorname{diag}(\sigma(x))V^\top$ is **SVD** of x. Similar to the case in $\mathbb{R}^d$, $\tilde{x}_{t+1}$ is the root of

$$\gamma_1 \operatorname{sgn}(\sigma(x)) + \gamma_2\sigma(x) + \nabla\phi_t(\sigma(x)) = \nabla\phi_t(\tilde{y}_{t+1}).$$

The subgradient of the objective (10) at $x_{t+1} = U_{t+1} \operatorname{diag}(\tilde{x}_{t+1})V_{t+1}^\top$ is clearly 0.

5.3 Projection onto the Cross-Polytope

Next, we consider the setting where r_t is the zero function and $\mathcal{K}$ is the ℓ_1 ball with radius D. Clearly, we simply set $x_{t+1} = y_{t+1}$ for $\|y_{t+1}\|_1 \leq D$. Otherwise, Algorithm 2 describes a sorting based procedure projecting y_{t+1} onto the ℓ_1 ball with time complexity $\mathcal{O}(d \log d)$. The correctness of the algorithm is shown in the next lemma.

Algorithm 2 $\operatorname{project}(y, D, \beta)$

```
Sort $|y_i|$ to get the permutation $p$ such that $|y_{p(i)}| \leq |y_{p(i+1)}|$
Define $\theta(j) = |y_{p(j)}|(D + (d - j + 1)\beta) + \beta D - \beta \sum_{i \geq j} |y_{p(i)}|$
$\rho \leftarrow \min\{j | \theta(j) > 0\}$
$z \leftarrow \frac{\sum_{i=\rho}^{d}(|y_{p(i)}| + \beta)}{D + (d - \rho + 1)\beta}$
$x_i^* \leftarrow \max\{\frac{|y_i| + \beta}{z} - \beta, 0\} \operatorname{sgn}(y_i)$ for $i = 1 \dots d$
Return $x^*$
```

Lemma 5 *Let $y \in \mathbb{R}^d$ with $\|y\|_1 > D$ and x^* as returned by Algorithm 2, then we have*

$$x^* \in \arg\min_{x \in \mathcal{K}} \mathcal{B}_{\psi_{t+1}}(x, y).$$

For the case that $\mathcal{K} \subseteq \mathbb{R}^{m,n}$ is the nuclear ball with radius D and $\|y_{t+1}\|_1 > D$, we need to solve the problem

$$\min_{x \in \mathcal{K}} \Phi_{t+1}(x) - \langle U_{t+1} \operatorname{diag}(\nabla\phi_{t+1}(\tilde{y}_{t+1}))V_{t+1}^\top, x\rangle_F,$$

where the constant part of the Bregman divergence is removed. From the von Neumann's trace inequality, the Frobenius inner product is upper bounded by

$$\langle U_{t+1}\nabla\phi_{t+1}(\tilde{y}_{t+1})V_{t+1}^\top, x\rangle_F \leq \sigma(x)^\top \nabla\phi_{t+1}(\tilde{y}_{t+1}).$$

The equality holds when x and $U_{t+1}\nabla\phi_{t+1}(\tilde{y}_{t+1})V_{t+1}^\top$ share a simultaneous **SVD**, i.e. the minimiser has an **SVD** of the form

$$x = U_{t+1} \operatorname{diag}(\nabla\sigma(x))V_{t+1}^\top.$$

Thus the problem is reduced to

$$\begin{aligned} \min_{x \in \mathbb{R}^{\min\{m,n\}}} & \phi_{t+1}(x) - \nabla \phi_{t+1}(\tilde{y}_{t+1})^\top x \\ \text{s.t.} & \sum_{i=1}^{\min\{m,n\}} x_i \le D \\ & x_i \ge 0 \text{ for all } i = 1, \dots, \min\{m,n\}, \end{aligned}$$

which can be solved by Algorithm 2. Thus, the projection of update rule (6) can be implemented as follows

$$\begin{aligned} \text{Compute SVD: } y_{t+1} =& U_{t+1} \operatorname{diag}(\tilde{y}_{t+1}) V_{t+1}^\top \\ \text{Apply Algorithm 2: } \tilde{x}_{t+1} =& \operatorname{project}(\tilde{y}_{t+1}, D, \beta) \\ \text{Construct: } x_{t+1} =& U_{t+1} \operatorname{diag}(\tilde{x}_{t+1}) V_{t+1}^\top. \end{aligned} \tag{11}$$

5.4 Stochastic Acceleration

Finally, we consider the stochastic optimisation problem of the form

$$\min_{x \in \mathcal{K}} l(x) + r(x),$$

where $l : \mathbb{X} \to \mathbb{R}$ and $r : \mathcal{K} \to \mathbb{R}_{\ge 0}$ are closed convex functions. In the stochastic setting, instead of having a direct access to ∇l, we query a stochastic gradient g_t of l at z_t in each iteration t with $\mathbb{E}[g_t | z_t] \in \partial l(z_t)$. Algorithms with a regret bound of the form $\mathcal{O}(\sqrt{\sum_{t=1}^T \|g_t - h_t\|_*})$ can be easily converted into a stochastic optimisation algorithm by applying the update rule to the scaled stochastic gradient $a_t g_t$ and hint $a_{t+1} g_t$, which is described in Algorithm 3. Joulani et al. (2020) has shown the convergence of accelerating **Adagrad** for the problem in $\mathbb{R}^d$. We extend the result to any finite dimensional normed vector space in the following corollary.

Algorithm 3 Stochastic Acceleration

```
Input: optimistic algorithm $\mathcal{A}$, compact convex set $\mathcal{K}$ and closed convex function $r$
for $t = 1, \dots, T$ do
    $a_t \leftarrow t$
    $x_t$ from $\mathcal{A}$
    $z_t \leftarrow \frac{a_t}{a_{1:t}} x_t + (1 - \frac{a_t}{a_{1:t}}) z_{t-1}$
    Query $g_t$ such that $\mathbb{E}[g_t | z_t] \in \partial l(z_t)$
    Update $\mathcal{A}$ with $\mathcal{K}$, $\alpha_t r$, scaled subgradient $a_t g_t$ and hint $a_{t+1} g_t$
end for
Return $x_{t+1}$
```

Corollary 1 *Let* $(\mathbb{X}, \|\cdot\|)$ *be a finite dimensional normed vector space and* $\mathcal{K} \subseteq \mathbb{X}$ *a compact convex set. Denote by* $\mathcal{A}$ *be some optimistic algorithm generating* $x_t \in \mathcal{K}$ *at iteration* t*. Denote by*

$$\nu_t^2 = \mathbb{E}[\|g_t - \nabla l_t(z_t)\|_*^2 | z_t]$$

the variance. If $\mathcal{A}$ *has a regret upper bound in the form of*

$$c_1 + c_2 \sqrt{\sum_{t=1}^{T} \|a_t(g_t - g_{t-1})\|_*^2}$$

then there is some $L > 0$ *such that the error incurred by Algorithm 3 is upper bounded by*

$$\mathbb{E}[f(z_T) - f(x)] \leq \frac{c_1 + c_2\sqrt{8\sum_{t=1}^{T} a_t^2(\nu_t^2 + L^2)}}{a_{1:T}}.$$

Furthermore, if l *is* M*-smooth, then we have*

$$\mathbb{E}[f(z_T) - f(x)] \leq \frac{c_1 + c_2\sqrt{8\sum_{t=1}^{T} a_t^2\nu_t^2} + \sqrt{2}c_2 L + 2Mc_2^2}{a_{1:T}}.$$

Setting $\alpha_t = t$, we obtain a convergence of $\mathcal{O}(\frac{c_2}{\sqrt{T}})$ in general case, and $\mathcal{O}(\frac{c_2}{T^2} + \frac{c_2 \max_t \nu_t}{\sqrt{T}})$ for smooth loss function. Applying update rule (3) or (4) with regulariser (2) or (5) to Algorithm 3, the constant c_2 is proportional to $\sqrt{\ln d}$ and $\sqrt{\ln(\min\{m, n\})}$ for $\mathbb{X} = \mathbb{R}^d$ and $\mathbb{X} = \mathbb{R}^{m,n}$ respectively, while the accelerated **Adagrad** has a linear dependence on the dimension of the problem Joulani et al. (2020).

6 Experiments

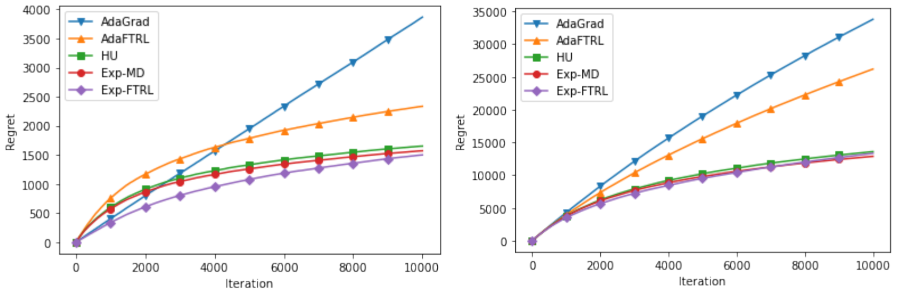
This section shows the empirical evaluation of the developed algorithms. We carry out experiments on both synthetic and real-world data and demonstrate the performances of the**OMD** (**EXP-MD**) and **FTRL** (**EXP-FTRL**) based on the exponentiated update.

6.1 Online Logistic Regression

For a sanity check, we simulate an d-dimensional online logistic regression problem, in which the values of the 1% ambient dimensions of the model parameter w^* are randomly and uniformly drawn from $[-1, 1]$. At each iteration t, we sample a random feature vector x_t from a uniform distribution over $[-1, 1]^d$ and generate a label $y_t \in \{-1, 1\}$ using a logit model, i.e. $\Pr[y_t = 1] = (1 + \exp(-w^\top x_t))^{-1}$. The goal is to minimise the cumulative regret

$$\mathcal{R}_{1:T} = \sum_{t=1}^{T} l_t(w_t) - \sum_{t=1}^{T} l_t(w^*)$$

with $l_t(w) = \ln(1 + \exp(-y_t w^\top x_t))$. We compare the performance of our algorithms with that of **AdaGrad**, **AdaFTRL** Duchi et al. (2011) and **HU** Ghai et al. (2020). For all of the candidates, we set the decision set $\mathcal{W} = \{w \in \mathbb{R}^d | \|w\|_1 \leq \|w^*\|_1\}$ according to prior knowledge on $\|w^*\|_1$. For both **AdaGrad** and **AdaFTRL**, we set



(a) Online Logistic Regression

(b) Online Multitask Learning

Fig. 2: Cumulative Regret of Online Learning Problems

the i-th entry of the proximal matrix H_t to $h_{ii} = 10^{-6} + \sum_{s=1}^{t-1} g_{s,i}^2$ as their theory suggested Duchi et al. (2011). The stepsize of **HU** is set to $\sqrt{\frac{1}{\sum_{s=1}^{t-1}\|g_s\|_\infty^2}}$, which leads to an adaptive regret upper bound. We set $d = 10.000$, run all algorithms for 10.000 iterations and average of the results over 20 trails for stability. The corresponding cumulative regret is plotted in Figure 2a. It is clear that **EXP-MD** and **EXP-FTRL** outperform **AdaGrad** and **AdaFTRL**. Our algorithms also have a small advantage over **HU**, which could be caused by the fact that decision variables selected by **HU**, which are projected onto the simplex through normalisation, are not really sparse, while the parameter generating the label actually is.

6.2 Online Multitask Learning

Next, we examine the performance of the developed spectral algorithms using a simulated online multi-task learning problem Kakade et al. (2012), in which we need to solve k highly correlated d-dimensional online prediction problems simultaneously. The data are generated as follows. We first randomly draw two orthogonal matrices $U \in \mathrm{GL}(d, \mathbb{R})$ and $V \in \mathrm{GL}(k, \mathbb{R})$. Then we generate a k-dimensional vector σ with r non-zero values randomly drawn from a uniform distribution over $[0, 10]$ and construct a low rank parameter matrix $W^* = U \operatorname{diag}(\sigma) V$. At each iteration t, k feature and label pairs $(x_{t,1}, y_{t,1}), \ldots, (x_{t,k}, y_{t,k})$ are generated using k logit models with the i-th parameters taken from the i-th rows of W. The loss function is given by $l_t(W) = \sum_{i=1}^{k} \ln(1 + \exp(-y_{t,i} w_i^\top x_{t,i}))$. We set $d = 100$, $k = 25$ and $r = 5$, take the nuclear ball $\{W \in \mathbb{R}^{d,k} | \|W\|_1 \leq \|W^*\|_1\}$ as the decision set and run the experiment as in subsection 6.1. The average of the results over 20 trials is shown in Figure 2b. Similar to the online logistic regression, our algorithms have a clear advantage over **AdaGrad** and **AdaFTRL** and slightly outperform **HU**.

6.3 Optimisation for Contrastive Explanations

Generating the contrastive explanation of a machine learning model Dhurandhar et al. (2018) is the most motivating application of our algorithm. Given a sample $x_0 \in \mathcal{X}$ and machine learning model $f : \mathcal{X} \to \mathbb{R}^K$, the constrastive explanation consists of a set of positive pertinent (**PP**) features and a set of pertinent negative (**PN**) features, which can be found by solving the following optimisation problem Dhurandhar et al. (2018)

$$\min_{x\in\mathcal{W}} \quad l_{x_0}(x)+\lambda_1\|x\|_1+\frac{\lambda_2}{2}\|x\|_2^2.$$

Let $\kappa\geq 0$ be a constant and define $k_0=\arg\max_i f(x_0)_i$. The loss function for finding **PP** is given by

$$l_{x_0}(x)=\max\{\max_{i\neq k_0} f(x)_i-f(x)_{k_0},-\kappa\},$$

which imposes a penalty on the features that do not justify the prediction. **PN** is the set of features altering the final classification, and is modeled by the following loss function

$$l_{x_0}(x)=\max\{f(x_0+x)_{k_0}-\max_{i\neq k_0} f(x_0+x)_i,-\kappa\}.$$

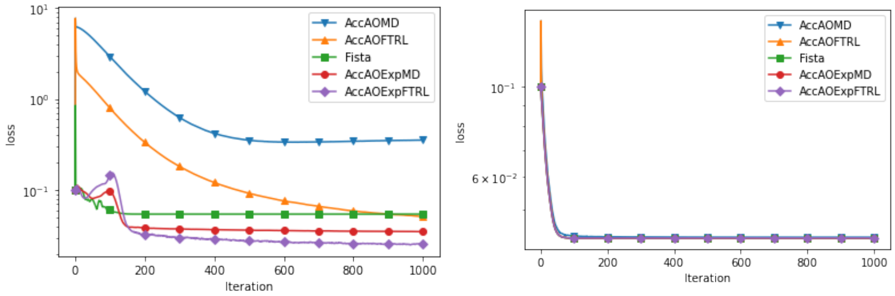
In the experiment, we first train a ResNet20 model He, Zhang, Ren, and Sun (2016) on the CIFAR-10 dataset Krizhevsky (2009), which attains a test accuracy of 91.49%. For each class of the images, we randomly pick 100 correctly classified images from the test dataset and generate **PP** and **PN** for them. For **PP**, we take the set of all feasible images as the decision set, while for **PN** we take the set of tensors x, such that x_0+x are feasible images. For both **PP** and **PN**, we set $\lambda_1=\lambda_2=0.5$.

We first consider the white-box setting, in which we have the access to ∇l_{x_0}. Our goal is to demonstrate the performance of the accelerated **AO-OMD** and **AO-FTRL** based on the exponentiated update (**AccAOExpMD** and **AccAOExpFTRL**). In Dhurandhar et al. (2018), the fast iterative shrinkage-thresholding algorithm (**FISTA**) Beck and Teboulle (2009) is applied to finding the **PP** and **PN**. Therefore, we take **FISTA** as our baseline. In addition, our algorithms are also compared with the accelerated **AO-OMD** and **AO-FTRL** with **AdaGrad**-style stepsizes (**AccAOMD** and **AccAOFTRL**) Joulani et al. (2020). All algorithms start from $x_1=0$. Figure 3 plots the convergence behavior of the five algorithms, averaged over the 1000 images. In the experiment for **PP**, our algorithms are obviously better than the **Adagrad**-style algorithms. Although **FISTA** converges faster at the first 100 iterations, it does not make further progress afterwards due to the tiny stepsize found by the backtracking rule. In the experiment for **PN**, all algorithms behave similarly. It is worth pointing out that the backtracking rule of **FISTA** requires multiple function evaluations, which are expensive for explaining deep neural networks.

Next, we consider the black-box setting, in which the gradient is estimated through the two-points estimation

$$\frac{1}{b}\sum_{i=1}^{b}\frac{d}{\mu}(f(x+\mu v_i)-f(x))v_i,$$

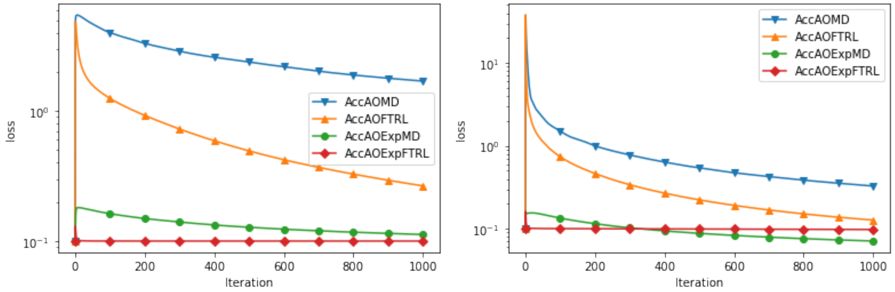
where v_i is sampled from the uniform distribution over the unit sphere. Here we set $b=10$ and $\mu=0.01$. Since the problem is stochastic, **FISTA** algorithm, which



(a) Convergence for Generating **PP**

(b) Convergence for Generating **PN**

Fig. 3: White Box Contrastive Explanations on CIFER-10



(a) Convergence for Generating **PP**

(b) Convergence for Generating **PN**

Fig. 4: Black Box Contrastive Explanations on CIFER-10

searches for the stepsize at each iteration, is not practical. Thus, we remove it from the comparison. It can be observed in Figure 4 that our algorithms outperform the **Adagrad**-style algorithms for generating both **PP** and **PN**.

7 Conclusion

This paper proposes and analyses a family of online optimisation algorithms based on an entropy-like regulariser combined with the ideas of optimism and adaptivity. The proposed algorithms have adaptive regret bounds depending logarithmically on the dimension of the problem, can handle popular composite objectives, and can be easily converted into stochastic optimisation algorithms with optimal accelerated convergence rates for smooth function. As a future research direction, we plan to analyse the convergence of the proposed algorithms together with variance reduction techniques for non-convex stochastic optimisation, and analyse their empirical performance for training deep neural networks.

Declarations

Funding

The research leading to these results received funding from the German Federal Ministry for Economic Affairs and Energy under Grant Agreement No. 01MK20002C.

Code availability

The implementation of the experiments and all algorithms involved in the experiments are available on GitHub https://github.com/mrdexteritas/exp_grad.

Availability of data and materials

The source code generating synthetic data, creating neural networks and model training are available on GitHub https://github.com/mrdexteritas/exp_grad. The CIFER-10 data are collected from https://www.cs.toronto.edu/~kriz/cifar.html.

Appendix A Missing Proofs of Section 3.1

A.1 Proof of Lemma 1

Proof It is straightforward that ϕ is differentiable at $x \neq 0$ with

$$\phi'(x) = \alpha \ln(\frac{|x|}{\beta} + 1) \operatorname{sgn}(x).$$

For any $h \in \mathbb{R}$, we have

$$\begin{aligned}\phi(0+h) - \phi(0) =&\alpha(|h| + \beta) \ln(\frac{|h|}{\beta} + 1) - \alpha|h| \\ \leq&\alpha(|h| + \beta)\frac{|h|}{\beta} - \alpha|h| \\ =&\frac{\alpha}{\beta}h^2,\end{aligned}$$

where the first inequality uses the fact $\ln x \leq x - 1$. Further more, we have

$$\begin{aligned}\phi(0+h) - \phi(0) =&\alpha(|h| + \beta) \ln(\frac{|h|}{\beta} + 1) - \alpha|h| \\ \geq&\alpha(|h| + \beta)(\frac{|h|}{|h| + \beta}) - \alpha|h| \\ \geq&0,\end{aligned}$$

where the first inequality uses the farc $\ln x \geq 1 - \frac{1}{x}$. Thus, we have

$$0 \leq \frac{\phi(0+h) - \phi(0)}{h} \leq \frac{\alpha}{\beta}h$$

for $h > 0$ and

$$\frac{\alpha}{\beta}h \leq \frac{\phi(0+h) - \phi(0)}{h} \leq 0.$$

for $h < 0$, from which it follows $\lim_{h\to 0} \frac{\phi(0+h)-\phi(0)}{h} = 0$. Similarly, we have for $x \neq 0$

$$\phi''(x) = \frac{\alpha}{|x|+\beta}.$$

Let $h \neq 0$, then we have

$$\frac{\phi'(0+h)-\phi'(0)}{h} = \frac{\alpha \ln(\frac{|h|}{\beta}+1)\operatorname{sgn}(h)}{h} = \frac{\alpha \ln(\frac{|h|}{\beta}+1)}{|h|}.$$

From the inequalities of the logarithm, it follows

$$\frac{\alpha}{|h|+\beta} \leq \frac{\phi'(0+h)-\phi'(0)}{h} \leq \frac{\alpha}{\beta}.$$

Thus, we obtain $\phi''(0) = \frac{\alpha}{\beta}$. By the definition of the convex conjugate we have

$$\phi^*(\theta) = \max_{x\in\mathbb{R}} \theta x - \phi(x), \tag{A1}$$

which is differentiable. The maximiser y satisfies

$$\ln(\frac{|y|}{\beta}+1)\operatorname{sgn}(y) = \theta.$$

Since $\ln(\frac{|y|}{\beta}+1) \geq 0$ holds, we have $\operatorname{sgn}(y) = \operatorname{sgn}(\theta)$ and

$$|y| = \beta \text{exp}(\frac{|\theta|}{\alpha}) - \beta.$$

Thus, we obtain the maximiser $y = \phi^{*\prime}(\theta)$ by setting

$$y = \operatorname{sgn}(\theta)(\beta \text{exp}(\frac{|\theta|}{\alpha}) - \beta). \tag{A2}$$

Combining (A1) and (A2), we obtain

$$\phi^*(\theta) = \alpha\beta \exp \frac{|\theta|}{\alpha} - \beta|\theta| - \alpha\beta.$$

To prove that ϕ^* is twice differentiable, it suffices to show that $\phi^{*\prime}$ is differentiable at 0. For any $h \neq 0$, we have

$$\frac{\phi^{*\prime}(0+h)-\phi^{*\prime}(0)}{h} = \frac{\operatorname{sgn}(h)(\beta\exp(\frac{|h|}{\alpha})-\beta)}{h}.$$

Applying the inequalities of the logarithm, we obtain

$$\frac{\beta}{\alpha} \leq \frac{\operatorname{sgn}(h)(\beta\exp(\frac{|h|}{\alpha})-\beta)}{h} \leq \frac{\beta}{\alpha}\exp(\frac{|h|}{\alpha}),$$

from which it follows ϕ^* is twice differentiable at 0 and

$$\phi^{*\prime\prime}(0) = \frac{\beta}{\alpha}.$$

□

A.2 Proof of Lemma 2

Proof Let $x \in \mathcal{K}$ be arbitrary. We have

$$\begin{aligned} v^\top \nabla^2 \phi_t(x) v =& \alpha_t \sum_{i=1}^d \frac{v_i^2}{|x_i| + \beta} \\ =& \alpha_t \sum_{i=1}^d \frac{v_i^2}{|x_i| + \beta} \sum_{i=1}^d (|x_i| + \beta) \frac{1}{\sum_{i=1}^d (|x_i| + \beta)} \\ \geq& \frac{\alpha_t}{\sum_{i=1}^d (|x_i| + \beta)} (\sum_{i=1}^d |v_i|)^2 \\ \geq& \frac{\alpha_t}{D + d\beta} (\sum_{i=1}^d |v_i|)^2 \\ =& \frac{\alpha_t}{D + d\beta} \|v\|_1^2 \end{aligned}$$

for all $v \in \mathbb{R}^d$, where the first inequality follows from the Cauchy-Schwarz inequality. This leads clearly to the strong convexity for a twice differentiable function. □

A.3 Proof of Theorem 1

Proposition 1 *Let $\mathcal{K} \subseteq \mathbb{X}$ be a convex set. Assume that $r_t : \mathcal{K} \to \mathbb{R}_{\geq 0}$ is closed convex function defined on $\mathcal{K}$ and $\psi_t : \mathcal{K} \mapsto \mathbb{R}$ is η_t-strongly convex w.r.t. $\|\cdot\|$ over $\mathcal{K}$. Then the sequence $\{x_t\}$ generated by (3) with regulariser $\{\psi_t\}$ guarantees*

$$\mathcal{R}_{1:T} \leq r_1(x_1) + \mathcal{B}_{\phi_1}(x, x_1) + \sum_{t=1}^T (\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_t}(x, x_t)) + \sum_{t=1}^T \frac{\|g_t - h_t\|_*^2}{2\eta_{t+1}}.$$

Proof From the optimality condition, it follows that for all $x \in \mathcal{K}$

$$\begin{aligned} &\langle g_t - h_t + h_{t+1} + \nabla r_{t+1}(x_{t+1}), x_{t+1} - x \rangle \\ \leq& \langle \nabla \phi_{t+1}(x_t) - \nabla \phi_{t+1}(x_{t+1}), x - x_{t+1} \rangle \\ =& \mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_{t+1}}(x, x_{t+1}) - \mathcal{B}_{\phi_{t+1}}(x_{t+1}, x_t), \end{aligned}$$

from which it follows

$$\begin{aligned} &\langle g_t, x_t - x \rangle + r_{t+1}(x_{t+1}) - r_{t+1}(x) \\ \leq& \langle g_t, x_t - x_{t+1} \rangle + \langle g_t - h_t + h_{t+1} + \nabla r_{t+1}(x_{t+1}), x_{t+1} - x \rangle + \langle h_t - h_{t+1}, x_{t+1} - x \rangle \\ \leq& \langle g_t - h_t, x_t - x_{t+1} \rangle + \langle h_t, x_t - x \rangle - \langle h_{t+1}, x_{t+1} - x \rangle \\ &+ \mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_{t+1}}(x, x_{t+1}) - \mathcal{B}_{\phi_{t+1}}(x_{t+1}, x_t) \end{aligned}$$

Adding up from 1 to T , we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}(\langle g_t, x_t - x\rangle + r_{t+1}(x_{t+1}) - r_{t+1}(x))\\
\leq&\sum_{t=1}^{T}\langle g_t - h_t, x_t - x_{t+1}\rangle + \sum_{t=1}^{T}(\langle h_t, x_t - x\rangle - \langle h_{t+1}, x_{t+1} - x\rangle)\\
&+\sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_{t+1}}(x, x_{t+1}) - \mathcal{B}_{\phi_{t+1}}(x_{t+1}, x_t))\\
\leq&\sum_{t=1}^{T}(\langle g_t - h_t, x_t - x_{t+1}\rangle - \mathcal{B}_{\phi_{t+1}}(x_{t+1}, x_t))\\
&+\langle h_1, x_1 - x\rangle - \langle h_{T+1}, x_{T+1} - x\rangle\\
&+\mathcal{B}_{\phi_1}(x, x_1) + \sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_t}(x, x_t))
\end{aligned}$$

h_1, h_{T+1} and x_{T+1}, which are artifacts of the analysis, can be set to 0. Then, we simply obtain

$$\begin{aligned}
&\sum_{t=1}^{T}(\langle g_t, x_t - x\rangle + r_t(x_t) - r_t(x))\\
=&\sum_{t=1}^{T}(\langle g_t, x_t - x\rangle + r_{t+1}(x_{t+1}) - r_{t+1}(x)) + r_1(x_1) - r_1(x) - r_{T+1}(x_{T+1}) + r_{T+1}(x)\\
\leq&\sum_{t=1}^{T}(\langle g_t, x_t - x\rangle + r_{t+1}(x_{t+1}) - r_{t+1}(x)) + r_1(x_1) - r_1(x) + r_{T+1}(x)\\
\leq&r_1(x_1) - r_1(x) + r_{T+1}(x) + \sum_{t=1}^{T}(\langle g_t - h_t, x_t - x_{t+1}\rangle - \mathcal{B}_{\phi_{t+1}}(x_{t+1}, x_t))\\
&+\mathcal{B}_{\phi_1}(x, x_1) + \sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_t}(x, x_t))
\end{aligned}$$

Since r_{T+1} is not involved in regret, we assume without loss of generality $r_1 = r_{T+1}$. From the η_t-strong convexity of ϕ_t we have

$$\begin{aligned}
&\langle g_t - h_t, x_t - x_{t+1}\rangle - \mathcal{B}_{\phi_{t+1}}(x_{t+1}, x_t)\\
\leq&\langle g_t - h_t, x_t - x_{t+1}\rangle - \frac{\eta_{t+1}}{2}\|x_t - x_{t+1}\|^2\\
\leq&\|g_t - h_t\|_*\|x_t - x_{t+1}\| - \frac{\eta_{t+1}}{2}\|x_t - x_{t+1}\|^2\\
\leq&\frac{\|g_t - h_t\|_*^2}{2\eta_{t+1}} + \frac{\eta_{t+1}}{2}\|x_t - x_{t+1}\|^2 - \frac{\eta_{t+1}}{2}\|x_t - x_{t+1}\|^2\\
=&\frac{\|g_t - h_t\|_*^2}{2\eta_{t+1}},
\end{aligned}$$

where the second inequality uses the definition of dual norm, the third inequality follows from the fact $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$. The claimed the result follows. □

Proof of Theorem 1 Proposition 1 can be directly applied, and we obtain

$$\begin{aligned}\mathcal{R}_{1:T} \leq & \sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x,x_t)-\mathcal{B}_{\phi_t}(x,x_t))+\sum_{t=1}^{T}\frac{D+d\beta}{2\alpha_t}\|g_t-h_t\|_\infty^2 \\ & +\mathcal{B}_{\phi_1}(x,x_1)+r(x_1).\end{aligned} \tag{A3}$$

Using Lemma 8, we bound the first term of (A3)

$$\begin{aligned}&\sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x,x_t)-\mathcal{B}_{\phi_t}(x,x_t)) \\ \leq & 4D(\ln(D+1)+\ln d)\sum_{t=2}^{T}(\alpha_{t+1}-\alpha_t) \\ \leq & 4D(\ln(D+1)+\ln d)\alpha_{T+1} \\ \leq & 4D(\ln(D+1)+\ln d)\eta\sqrt{\sum_{t=1}^{T}\|g_t-h_t\|_\infty^2}.\end{aligned}$$

Using Lemma 6, the second term of (A3) can be bounded as

$$\sum_{t=1}^{T}\frac{(D+1)\|g_t-h_t\|_\infty^2}{4\alpha_t}\leq\frac{D+1}{2\eta}\sqrt{\sum_{t=1}^{T}\|g_t-h_t\|_\infty^2}$$

The third term of (A3) is simply 0 since we set $\alpha_1=0$. Setting $\eta=\sqrt{\frac{1}{\ln(D+1)+\ln d}}$ and combining the results above, we obtain the claimed result. □

Proposition 2 *Let $\mathcal{K}\subseteq\mathbb{X}$ be a compact convex set such that $\|x\|\leq D$ holds for all $x\in\mathcal{K}$, $r_t:\mathcal{K}\to\mathbb{R}_{\geq 0}$ and $\phi_t:\mathcal{K}\mapsto\mathbb{R}$ closed convex function defined on $\mathcal{K}$. Assume ϕ_t is η_t-strongly convex w.r.t. $\|\cdot\|$ over $\mathcal{K}$ and $\phi_t\leq\phi_{t+1}$ for all $t=1,\ldots,T$. Then the sequence $\{x_t\}$ generated by* (4) *with guarantees*

$$\mathcal{R}_{1:T}\leq\phi_{T+1}(x)+\sum_{t=1}^{T}\frac{2D\|g_t-h_t\|_*^2}{\sqrt{16D^2\eta_t^2+\|g_t-h_t\|_*^2}}. \tag{A4}$$

Proof of Proposition 2 First, define $\psi_t=r_{1:t}+\phi_t$. Then, we have

$$\begin{aligned}&\sum_{t=1}^{T}\psi_{t+1}^*(\theta_{t+1}-h_{t+1})-\psi_t^*(\theta_t-h_t) \\ =&\psi_{T+1}^*(\theta_{T+1}-h_{T+1})-\psi_1^*(\theta_1-h_1) \\ \geq&\langle\theta_{T+1}-h_{T+1},x\rangle-\psi_{T+1}(x)-\psi_1^*(\theta_1-h_1) \\ \geq&\langle-\sum_{t=1}^{T}g_t-h_{T+1},x\rangle-\psi_{T+1}(x)-\psi_1^*(\theta_1-h_1)\end{aligned}$$

Setting the artifacts h_{T+1} to 0, rearranging and adding $\sum_{t=1}^{T}\langle g_t, w_t\rangle$ to both sides, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\langle g_t, x_t - x\rangle\\
\le&\psi_{T+1}(x) + \psi_1^*(\theta_1 - h_1) + \sum_{t=1}^{T}(\psi_{t+1}^*(\theta_{t+1} - h_{t+1}) - \psi_t^*(\theta_t - h_t) + \langle g_t, x_t\rangle)\\
=&\psi_{T+1}(x) - \langle h_1, x_1\rangle - r_1(x_1)\\
&+\sum_{t=1}^{T}(\psi_{t+1}^*(\theta_{t+1} - h_{t+1}) - \psi_t^*(\theta_{t+1}))\\
&+\sum_{t=1}^{T}(\psi_t^*(\theta_{t+1}) - \psi_t^*(\theta_t - h_t) + \langle \theta_t - \theta_{t+1}, \nabla\psi_t^*(\theta_t - h_t)\rangle)\\
\le&\psi_{T+1}(x) - \langle h_1, x_1\rangle - r_1(x_1)\\
&+\sum_{t=1}^{T}(\psi_{t+1}^*(\theta_{t+1} - h_{t+1}) - \psi_t^*(\theta_{t+1}))\\
&+\sum_{t=1}^{T}(\psi_t^*(\theta_{t+1}) - \psi_t^*(\theta_t - h_t) + \langle \theta_t - \theta_{t+1}, \nabla\psi_t^*(\theta_t - h_t)\rangle),
\end{aligned}$$

From the definition of ψ_t, it follows

$$\psi_{T+1}(x) = \phi_{T+1}(x) + r_{1:T+1}(x) = \phi_{T+1}(x) + r_{1:T}(x),$$

where we assumed $r_{T+1} \equiv 0$, since it is not involved in the regret. Furthermore, we have for $t \ge 1$

$$\begin{aligned}
&\psi_{t+1}^*(\theta_{t+1} - h_{t+1}) - \psi_t^*(\theta_{t+1})\\
\le&\langle \theta_{t+1} - h_{t+1}, x_{t+1}\rangle - \psi_{t+1}(x_{t+1}) - \langle \theta_{t+1}, x_{t+1}\rangle + \psi_t(x_{t+1})\\
=&-\langle h_{t+1}, x_{t+1}\rangle - \psi_{t+1}(x_{t+1}) + \psi_t(x_{t+1})\\
=&-\langle h_{t+1}, x_{t+1}\rangle - r_{1:t+1}(x_{t+1}) + r_{1:t}(x_{t+1}) - \phi_{t+1}(x_{t+1}) + \phi_t(x_{t+1})\\
\le&-\langle h_{t+1}, x_{t+1}\rangle - r_{t+1}(x_{t+1}),
\end{aligned}$$

where the first inequality uses the definition of convex conjugate and the second inequality follows from the fact $\phi_{t+1} \le \phi_t$. Adding up from 1 to T, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}(\psi_{t+1}^*(\theta_{t+1} - h_{t+1}) - \psi_t^*(\theta_{t+1}))\\
\le&-\sum_{t=1}^{T} r_{t+1}(x_{t+1}) - \sum_{t=1}^{T}\langle h_{t+1}, x_{t+1}\rangle\\
=&r_1(x_1) + \langle h_1, x_1\rangle - r_{T+1}(x_{t+1}) - \langle h_{T+1}, x_{T+1}\rangle - \sum_{t=1}^{T} r_t(x_t) - \sum_{t=1}^{T}\langle h_t, x_t\rangle\\
=&r_1(x_1) + \langle h_1, x_1\rangle - \sum_{t=1}^{T} r_t(x_t) - \sum_{t=1}^{T}\langle h_t, x_t\rangle,
\end{aligned}$$

where we use $r_{T+1} \equiv 0$ and $h_{T+1} = 0$. Combining the inequality above and rearranging, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(\langle g_t, x_t - x\rangle + r_t(x_t) - r_t(x)) \\
\le&\phi_{T+1}(x) + \sum_{t=1}^{T}(\psi_t^*(\theta_{t+1}) - \psi_t^*(\theta_t - h_t) + \langle \theta_t - h_t - \theta_{t+1}, \nabla\psi_t^*(\theta_t - h_t)\rangle) \\
\le&\phi_{T+1}(x) + \sum_{t=1}^{T}\mathcal{B}_{\psi_t^*}(\theta_{t+1}, \theta_t - h_t).
\end{aligned} \tag{A5}$$

Next, by the definition of the Bregman divergence, we have

$$\begin{aligned}
&\mathcal{B}_{\psi_t^*}(\theta_{t+1}, \theta_t - h_t) \\
\le&\langle \theta_{t+1}, \nabla\psi_t^*(\theta_{t+1})\rangle - \psi_t(\nabla\psi_t^*(\theta_{t+1})) - \langle \theta_t - h_t, x_t\rangle + \psi_t(x_t) + \langle g_t - h_t, x_t\rangle \\
=&\langle \theta_t - h_t, \nabla\psi_t^*(\theta_{t+1}) - x_t\rangle - \psi_t(\nabla\psi_t^*(\theta_{t+1})) + \psi_t(x_t) + \langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle \\
=&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t).
\end{aligned}$$

Since ϕ_t is η_t strongly convex, we have

$$\begin{aligned}
&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \\
\le&\frac{1}{2\eta_t}\|g_t - h_t\|_*^2 + \frac{\eta_t}{2}\|x_t - \nabla\psi_t^*(\theta_{t+1})\|^2 - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \\
\le&\frac{1}{2\eta_t}\|g_t - h_t\|_*^2
\end{aligned} \tag{A6}$$

We also have

$$\begin{aligned}
&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \\
\le&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle \\
\le&2D\|g_t - h_t\|_*.
\end{aligned} \tag{A7}$$

Putting (A6) and (A7) together, we have

$$\begin{aligned}
&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \\
\le&\min\{\frac{1}{2\eta_t}\|g_t - h_t\|_*^2, 2D\|g_t - h_t\|_*\} \\
\le&\frac{1}{\frac{2\eta_t}{\|g_t - h_t\|_*^2} + \frac{1}{2D\|g_t - h_t\|_*}} \\
\le&\frac{2D\|g_t - h_t\|_*^2}{4D\eta_t + \|g_t - h_t\|_*} \\
\le&\frac{2D\|g_t - h_t\|_*^2}{\sqrt{16D^2\eta_t^2 + \|g_t - h_t\|_*^2}}
\end{aligned}$$

Combining the inequalities above, we obtain

$$\mathcal{R}_{1:T} \le \phi_{T+1}(x) + \sum_{t=1}^{T}\frac{2D\|g_t - h_t\|_*^2}{\sqrt{16D^2\eta_t^2 + \|g_t - h_t\|_*^2}}$$

□

A.4 Proof of Theorem 2

Proof of Theorem 2 We take the Bregman divergence $\mathcal{B}_{\phi_t}(x, x_1)$ as the regulariser at iteration t. Since $\mathcal{B}_{\phi_t}(x, x_1)$ is non-negative, increasing with t and $\frac{2\alpha_t}{D+\beta d}$ strongly-convex w.r.t. $\|\cdot\|_1$, Proposition 2 can be directly applied, and we get

$$\begin{aligned}
\mathcal{R}_{1:T} \leq & \mathcal{B}_{\phi_{T+1}}(x, x_1) + \sum_{t=1}^{T} \frac{2D\|g_t - h_t\|_\infty^2}{\sqrt{\frac{64D^2\alpha_t^2}{(D+\beta d)^2} + \|g_t - h_t\|_\infty^2}} \\
= & \mathcal{B}_{\phi_{T+1}}(x, x_1) + \frac{2D}{\eta} \sum_{t=1}^{T} \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\frac{64D^2}{(D+\beta d)^2} \sum_{s=1}^{t-1} \|g_s - h_t\|_\infty^2 + \frac{1}{\eta^2}\|g_t - h_t\|_\infty^2}}
\end{aligned}$$

Setting $\beta = \frac{1}{d}$ and $\eta = \frac{1}{\sqrt{\ln(D+1)+\ln d}}$, we have

$$\begin{aligned}
& \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\frac{64D^2}{(D+\beta d)^2} \sum_{s=1}^{t-1} \|g_s - h_t\|_\infty^2 + \frac{1}{\eta^2}\|g_t - h_t\|_\infty^2}} \\
= & \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\frac{64D^2}{(D+1)^2} \sum_{s=1}^{t-1} \|g_s - h_t\|_\infty^2 + (\ln(D+1) + \ln d)\|g_t - h_t\|_\infty^2}} \\
\leq & \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\sum_{s=1}^{t-1} \|g_s - h_t\|_\infty^2 + \|g_t - h_t\|_\infty^2}} \\
= & \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\sum_{s=1}^{t} \|g_s - h_t\|_\infty^2}},
\end{aligned}$$

where the inequality uses the assumptions $D \geq 1$ and $d > e$. Adding up from 1 to T, we obtain

$$\begin{aligned}
\mathcal{R}_{1:T} \leq & \mathcal{B}_{\phi_{T+1}}(x, x_1) + 2D\sqrt{\ln(D+1) + \ln d} \sum_{t=1}^{T} \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\sum_{s=1}^{t} \|g_s - h_t\|_\infty^2}} \\
\leq & \mathcal{B}_{\phi_{T+1}}(x, x_1) + 4D\sqrt{\ln(D+1) + \ln d} \sqrt{\sum_{t=1}^{T} \|g_t - h_t\|_\infty^2}
\end{aligned}$$

The first term can be bounded by Lemma 8

$$\mathcal{B}_{\phi_{T+1}}(x, x_1) + \leq 4D\sqrt{\ln(D+1) + \ln d} \sqrt{\sum_{t=1}^{T} \|g_t - h_t\|_\infty^2}$$

Combining the inequality above, we obtain

$$\mathcal{R}_{1:T} \leq (D, d) \sqrt{\sum_{t=1}^{T} \|g_t - h_t\|_\infty^2},$$

with $c(D, d) \in \mathcal{O}(D\sqrt{\ln(D+1) + \ln d})$, which is the claimed result. □

Appendix B Missing Proofs of Section 3.2

B.1 Proof of Theorem 3

The Proof of Theorem 3 is based on the idea of Ghai et al. (2020). We first revise some technical lemmata.

Proof of Lemma 3 Define $\tilde{F} : \mathbb{S}^d \to \mathbb{S}^d, X \mapsto U \operatorname{diag}(f(\lambda_1(X)), \dots, f(\lambda_d(X)))U^\top$. Apparently, we have $F(X) = \operatorname{Tr} \tilde{F}(X)$. From the Theorem V.3.3 in Bhatia (2013), it follows that $\tilde{F}$ is differentiable and

$$D\tilde{F}(X)(H) = U(\Gamma(f, X) \odot U^\top HU)U^\top.$$

Using the linearity of the trace and the chain rule, F is differentiable and the directional derivative at X in H is given by

$$\begin{aligned}
D_H F(X) =& D \operatorname{Tr}(\tilde{F}(X)) \circ D\tilde{F}(X)(H) \\
=& \operatorname{Tr}(D\tilde{F}(X)(H)) \\
=& \operatorname{Tr}(U(\tilde{\Gamma}(f, X) \odot U^\top HU)U^\top) \\
=& \operatorname{Tr}(\tilde{\Gamma}(f, X) \odot U^\top HU) \\
=& \sum_{i=1}^{d} f'(\lambda_i(X))\tilde{h}_{ii} \\
=& \operatorname{Tr}(U \operatorname{diag}(f'(\lambda_1(X)), \dots, f'(\lambda_d(X)))U^\top H)
\end{aligned}$$

where $\tilde{h}_{ii}$ is the i-th element in the diagonal of the matrix $U^\top HU$. Next, define

$$\bar{F} : \mathbb{S}^d \to \mathbb{S}^d, X \mapsto U \operatorname{diag}(f'(\lambda_1(X)), \dots, f'(\lambda_d(X)))U^\top.$$

And we have

$$DF(X) = H \mapsto \operatorname{Tr}(\bar{F}(X)H)$$

Applying Theorem V.3.3 in Bhatia (2013) again, we obtain the differentiability of $\bar{F}$ and

$$D\bar{F}(X)(G) = U(\Gamma(f', X) \odot U^\top GU)U^\top.$$

Note that $X \mapsto \operatorname{Tr}(X(\cdot))$ is a linear map between finite dimensional spaces. Thus F is twice differentiable. From the linearity of the trace operator and matrix multiplication, it follows that $D_H F(X)$ is differentiable. Applying the chain rule, we obtain

$$\begin{aligned}
D^2F(X)(G, H) =& D_G(D_H F)(X) \\
=& D(D_H F)(X)(G) \\
=& \operatorname{Tr}((D\bar{F}(X)(G))H) \\
=& \operatorname{Tr}(U(\Gamma(f', X) \odot U^\top GU)U^\top H) \\
=& \operatorname{Tr}((\Gamma(f', X) \odot U^\top GU)U^\top HU) \\
=& \sum_{i.j} \gamma(f', X)_{ij}\tilde{g}_{ij}\tilde{h}_{ij},
\end{aligned}$$

which is the claimed result. □

Proof of Lemma 4 Since $D^2\Phi^*(\theta) \in \mathcal{L}(\mathbb{X}_*, \mathcal{L}(\mathbb{X}_*, \mathbb{R}))$ is positive definite and $\mathbb{X}$ is finite dimensional, the map

$$f_\theta : \mathbb{X}_* \to \mathbb{X}, v \mapsto D^2\Phi^*(\theta)(v, \cdot)$$

is invertible. Furthermore, defining $\psi_\theta : \mathbb{X}_* \to \mathbb{R}, v \mapsto \frac{1}{2}D^2\Phi^*(\theta)(v, v)$, we have

$$\begin{aligned}
D\psi_\theta(v) =& \frac{1}{2}D^2\Phi^*(\theta)(v, \cdot) + \frac{1}{2}D^2\Phi^*(\theta)(\cdot, v) \\
=& f_\theta(v).
\end{aligned}$$

Thus, we obtain the convex conjugate ψ_θ^*

$$\begin{aligned}\psi_\theta^*(x) = &\sup_{v\in\mathbb{X}_*}\langle v, x\rangle - \psi_\theta(v)\\ =&\langle f_\theta^{-1}(x), x\rangle - \psi_\theta(f_\theta^{-1}(x))\\ =&\langle f_\theta^{-1}(x), x\rangle - \frac{1}{2}\langle f_\theta^{-1}(x), D^2\Phi^*(\theta)(f_\theta^{-1}(x),\cdot)\rangle\\ =&\langle f_\theta^{-1}(x), x\rangle - \frac{1}{2}\langle f_\theta^{-1}(x), f_\theta(f_\theta^{-1}(x))\rangle\\ =&\frac{1}{2}\langle f_\theta^{-1}(x), x\rangle\end{aligned}$$

by setting $x = D\psi_\theta(v)$. Denote by $I : \mathbb{X}\to\mathbb{X}, x\mapsto x$ the identity function. From $D\Phi^* = D\Phi^{-1}$, it follows

$$\begin{aligned}I(x) =&DI(v)(x)\\ =&D(D\Phi^*\circ D\Phi)(v)(x)\\ =&D^2\Phi^*(D\Phi(v))\circ D^2\Phi(v)(x),\\ =&D^2\Phi^*(\theta)\circ D^2\Phi(D\Phi^*(\theta))(x)\end{aligned}$$

for $\theta = D\Phi(v)$ and all $x\in\mathbb{X}$. Thus, we have $f_\theta^{-1} = D^2\Phi(D\Phi^*(\theta))$ and

$$\begin{aligned}\psi_\theta^*(x) =&\frac{1}{2}\langle f_\theta^{-1}(x), x\rangle\\ =&\frac{1}{2}D^2\Phi(D\Phi^*(\theta))(x,x).\end{aligned}$$

Finally, since $\psi_\theta(v)\leq\frac{1}{2}\|v\|_*^2$ holds for all $v\in\mathbb{X}_*$, we can reverse the order by applying Proposition 2.19 in Barbu and Precupanu (2012) and obtain for all $x\in\mathbb{X}$

$$\frac{1}{2}D^2\Phi(D\Phi^*(\theta))(x,x) = \psi_\theta^*(x)\geq\frac{1}{2}\|x\|^2,$$

which is the claimed result. □

Finally, we can prove Theorem 3.

Proof of Theorem 3 We start the proof by introducing the required definitions. Define the operator

$$S:\mathbb{R}^{m,n}\to\mathbb{S}^{m+n}, X\mapsto\begin{bmatrix}0 & X\\ X^\top & 0\end{bmatrix}$$

The set $\mathcal{X} = \{S(X)|X\in\mathbb{R}^{m,n}\}$ is a finite dimensional linear subspace of the space of symmetric matrices $\mathbb{S}^{m+n}$, and thus $(\mathcal{X},\|\cdot\|_1)$ is a finite dimensional Banach space. Its dual space $\mathcal{X}_*$ determined by the Frobenius inner product can be represented by $\mathcal{X}$ itself. Denote by $\mathbb{B}(D) = \{X\in\mathbb{R}^{m,n}|\|X\|_1\leq D\}$ the nuclear ball with radius D. Then the set $\mathcal{K} = \{S(X)|X\in\mathbb{B}(D)\}$ is a nuclear ball in $\mathcal{X}$ with radius $2D$, since $\|S(X)\|_1 = 2\|X\|_1$ for all $X\in\mathbb{R}^{m,n}$.

Let $S(X)\in\mathcal{K}$ be arbitrary. Denote by $F_t = \Phi_t|_{\mathcal{X}}$ the restriction of Φ_t to $\mathcal{X}$. Next, we show the strong convexity of F_t over $\mathcal{K}$. From the conjugacy formula of Theorem 2.4 in Lewis (1995) and Lemma 1, it follows

$$F_t^*(S(X)) = \phi_t^*\circ\sigma(S(X)) = \phi_t^*\circ\lambda(S(X)),$$

where the second equality follows from the fact that Φ_t^* is absolutely symmetric. By Lemma 1 and Lemma 3, F_t^* is twice differentiable. Let $X \in \mathcal{K}$ be arbitrary and $\Theta = DF_t(X) \in \mathcal{X}_*$. For simplicity, we define

$$f_t : \mathbb{R} \to \mathbb{R}, x \mapsto \alpha_t \beta \exp \frac{|x|}{\alpha_t} - \beta|x| - \alpha_t \beta.$$

Then, for all $H \in \mathcal{X}$,

$$D^2 F_t^*(\Theta)(H,H) = \sum_{ij} \gamma(f_t', \Theta)_{ij} \tilde{h}_{ij}^2,$$

where $\Gamma(f_t', \Theta) = [\gamma(f_t', \Theta)_{ij}]$ is the matrix of the second divided difference with

$$\gamma(f_t', \Theta)_{ij} = \begin{cases} \frac{f_t'(\lambda_i(\Theta)) - f_t'(\lambda_j(\Theta))}{\lambda_i(\Theta) - \lambda_j(\Theta)}, & \text{if } \lambda_i(\Theta) \neq \lambda_j(\Theta) \\ f_t''(\lambda_i(\Theta)), & \text{otherwise.} \end{cases}$$

$D^2 F_t^*(\Theta)$ is clearly positive definite over $\mathbb{S}^{m+n}$, since $\gamma(f_t', \Theta)_{ij} > 0$ for all i and j. Furthermore, from the mean value theorem and the convexity of f_t'', there is a $c_{ij} \in (0,1)$ such that

$$\begin{aligned} \frac{f_t'(\lambda_i(\Theta)) - f_t'(\lambda_j(\Theta))}{\lambda_i(\Theta) - \lambda_j(\Theta)} \leq & f_t'''(c_{ij}\lambda_i(\Theta) + (1-c_{ij})\lambda_j(\Theta)) \\ \leq & c_{ij} f_t''(\lambda_i(\Theta)) + (1-c_{ij}) f_t''(\lambda_j(\Theta)) \\ \leq & f_t''(\lambda_i(\Theta)) + f_t''(\lambda_j(\Theta)) \end{aligned}$$

holds for all $\lambda_i(\Theta) \neq \lambda_j(\Theta)$. Thus, we obtain

$$\begin{aligned} D^2 F_t^*(\Theta)(H,H) = & \sum_{ij} \gamma(f_t, \Theta)_{ij} \tilde{h}_{ij}^2 \\ \leq & \sum_{ij} (f_t''(\lambda_i(\Theta)) + f_t''(\lambda_j(\Theta))) \tilde{h}_{ij}^2 \\ = & 2 \sum_{i=1}^{m+n} f_t''(\lambda_i(\Theta)) \sum_{j=1}^{m+m} \tilde{h}_{ij}^2 \\ = & 2 \operatorname{Tr}(UHU^\top \operatorname{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta)) UHU^\top) \\ = & 2 \operatorname{Tr}(H^2 \operatorname{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \\ \leq & 2 \sum_{i=1}^{2\min\{m,n\}} \sigma_i(H^2) \sigma_i(\operatorname{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \end{aligned} \tag{B8}$$

where the last line uses von Neumann's trace inequality and the fact that the rank of $H \in \mathcal{X}$ and Θ is at most $2\min\{m,n\}$. Since H^2 is positive semi-definite, $\sigma_i(H^2) = \sigma_i(H)^2$ holds for all i. Furthermore, $f_t''(x) \geq 0$ holds for all $x \in \mathbb{R}$. Thus, the last line of (B8) can be rewritten into

$$\begin{aligned} D^2 F_t^*(\Theta)(H,H) \leq & 2 \sum_{i=1}^{2\min\{m,n\}} \sigma_i(H)^2 \sigma_i(\operatorname{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \\ \leq & 2\|H\|_\infty^2 \sum_{i=1}^{2\min\{m,n\}} \sigma_i(\operatorname{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \\ \leq & 2\|H\|_\infty^2 \sum_{i=1}^{2\min\{m,n\}} f_t''(\lambda_i(\Theta)). \end{aligned} \tag{B9}$$

Recall $\Theta = DF_t(S(X))$ for $S(X) \in \mathcal{K}$. Together with Lemma 1, we obtain

$$\begin{aligned} f_t''(\lambda_i(\Theta)) =& \frac{\beta}{\alpha_t} \exp \frac{|\lambda_i(\Theta)|}{\alpha_t} \\ =& \frac{\beta}{\alpha_t} \exp \frac{|\alpha_t \ln(\frac{|\lambda_i(S(X))|}{\beta} + 1)|}{\alpha_t} \\ =& \frac{|\lambda_i(S(X))| + \beta}{\alpha_t}. \end{aligned}$$

By the construction of $\mathcal{K}$, it is clear that $\sum_{i=1}^{2\min\{m,n\}} |\lambda_i(S(X))| \leq 2D$. Thus, (B9) can be simply further upper bounded by

$$\begin{aligned} D^2 F_t^*(\Theta)(H,H) \leq& 2\|H\|_\infty^2 \sum_{i=1}^{2\min\{m,n\}} \frac{|\lambda_i(S(X))| + \beta}{\alpha_t} \\ \leq& 2\|H\|_\infty^2 \frac{2D + 2\min\{m,n\}\beta}{\alpha_t} \end{aligned}$$

Finally, applying Lemma 4, we obtain

$$D^2 F_t(S(X))(Y,Y) \geq \frac{\alpha_t}{4(D + \min\{m,n\}\beta)} \|Y\|_1^2,$$

which implies the $\frac{\alpha_t}{4(D+\min\{m,n\}\beta)}$-strong convexity of F_t over $\mathcal{K}$.

Finally, we prove the strongly convexity of Φ_t over $B(D) \in \mathbb{R}^{m+n}$. Let $X, Y \in B(D)$ be arbitrary matrices in the nuclear ball. The following inequality can be obtained

$$\begin{aligned} & 2\Phi_t(X) - 2\Phi_t(Y) \\ =& \Phi_t(S(X)) - \Phi_t(S(Y)) \\ \geq& \langle D\Phi_t(S(Y)), S(X) - S(Y)\rangle_F + \frac{\alpha_t}{8(D + \min\{m,n\}\beta)} \|S(X) - S(Y)\|_1^2 \\ =& 2\langle D\Phi_t(Y), X - Y\rangle_F + \frac{\alpha_t}{2(D + \min\{m,n\}\beta)} \|X - Y\|_1^2, \end{aligned}$$

which implies the $\frac{\alpha_t}{2(D+\min\{m,n\}\beta)}$-strong convexity of Φ_t as desired. □

B.2 Proof of Theorem 4

Proof The proof is almost identical to the proof of Theorem 1. From the strong convexity of Φ_t shown in Theorem 3 and the general upper bound in Proposition 1, we obtain

$$\mathcal{R}_{1:T} \leq r_1(x_1) + \mathcal{B}_{\phi_1}(x, x_1) + \sum_{t=1}^{T} (\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_t}(x, x_t)) + \sum_{t=1}^{T} \frac{\|g_t - h_t\|_*^2}{2\eta_{t+1}}. \tag{B10}$$

Using Lemma 8, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(\mathcal{B}_{\Phi_{t+1}}(x,x_t)-\mathcal{B}_{\Phi_t}(x,x_t))\\
\leq&4D(\ln(D+1)+\ln\min\{m,n\})\sum_{t=2}^{T}(\alpha_{t+1}-\alpha_t)\\
\leq&4D(\ln(D+1)+\ln\min\{m,n\})\alpha_{T+1}\\
=&4D(\ln(D+1)+\ln\min\{m,n\})\eta\sqrt{\sum_{t=1}^{T}\|g_t-h_t\|_\infty^2}\\
=&4D\sqrt{\ln(D+1)+\ln\min\{m,n\}}\sqrt{\sum_{t=1}^{T}\|g_t-h_t\|_\infty^2}
\end{aligned}$$

Furthermore, from Lemma 6, it follows

$$\sum_{t=1}^{T}\frac{(D+1)\|g_t-h_t\|_\infty^2}{4\alpha_t}\leq\frac{D+1}{2}\sqrt{\ln(D+1)+\ln\min\{m,n\}}\sqrt{\sum_{t=1}^{T}\|g_t-h_t\|_\infty^2}$$

The claimed result is obtained by combining the inequalities above. □

B.3 Proof of Theorem 5

Proof Since $\mathcal{B}_{\Phi_t}(x,x_1)$ is non-negative, increasing and $\frac{2\alpha_t}{D+\beta d}$ strongly-convex w.r.t. $\|\cdot\|_1$, Proposition 2 can be directly applied, and we get

$$\begin{aligned}
\mathcal{R}_{1:T}\leq&\mathcal{B}_{\Phi_t}(x,x_1)+\sum_{t=1}^{T}\frac{2D\|g_t-h_t\|_\infty^2}{\sqrt{\frac{64D^2\alpha_t^2}{(D+\beta d)^2}+\|g_t-h_t\|_\infty^2}}\\
=&\mathcal{B}_{\Phi_t}(x,x_1)+\frac{2D}{\eta}\sum_{t=1}^{T}\frac{\|g_t-h_t\|_\infty^2}{\sqrt{\frac{64D^2}{(D+\beta d)^2}\sum_{s=1}^{t-1}\|g_s-h_t\|_\infty^2+\frac{1}{\eta^2}\|g_t-h_t\|_\infty^2}}
\end{aligned}$$

Setting $\beta=\frac{1}{\min\{m,n\}}$ and $\eta=\frac{1}{\sqrt{\ln(D+1)+\ln\min\{m,n\}}}$, we have

$$\begin{aligned}
&\frac{\|g_t-h_t\|_\infty^2}{\sqrt{\frac{64D^2}{(D+\beta d)^2}\sum_{s=1}^{t-1}\|g_s-h_t\|_\infty^2+\frac{1}{\eta^2}\|g_t-h_t\|_\infty^2}}\\
=&\frac{\|g_t-h_t\|_\infty^2}{\sqrt{\frac{64D^2}{(D+1)^2}\sum_{s=1}^{t-1}\|g_s-h_t\|_\infty^2+(\ln(D+1)+\ln d)\|g_t-h_t\|_\infty^2}}\\
\leq&\frac{\|g_t-h_t\|_\infty^2}{\sqrt{\sum_{s=1}^{t-1}\|g_s-h_t\|_\infty^2+\|g_t-h_t\|_\infty^2}}\\
=&\frac{\|g_t-h_t\|_\infty^2}{\sqrt{\sum_{s=1}^{t}\|g_s-h_t\|_\infty^2}},
\end{aligned}$$

where the inequality uses the assumptions $D \geq 1$ and $\min\{m, n\} > e$. Adding up from 1 to T, we obtain

$$\begin{aligned}\mathcal{R}_{1:T} \leq & \mathcal{B}_{\Phi_t}(x, x_1) + 2D\sqrt{\ln(D+1) + \ln \min\{m, n\}} \sum_{t=1}^{T} \frac{\|g_t - h_t\|_\infty^2}{\sqrt{\sum_{s=1}^{t} \|g_s - h_t\|_\infty^2}} \\ \leq & \mathcal{B}_{\Phi_t}(x, x_1) + 4D\sqrt{\ln(D+1) + \ln \min\{m, n\}} \sqrt{\sum_{t=1}^{T} \|g_t - h_t\|_\infty^2}\end{aligned}$$

The first term can be bounded by Lemma 8

$$\mathcal{B}_{\Phi_{T+1}}(x, x_1) \leq 4D(\ln(D+1) + \ln \min\{m, n\}) \sqrt{\sum_{t=1}^{T} \|g_t - h_t\|_\infty^2}$$

Combining the inequality above, we obtain

$$\mathcal{R}_{1:T} \leq c(D, m, n) \sqrt{\sum_{t=1}^{T} \|g_t - h_t\|_\infty^2},$$

with $c(D, m, n) \in \mathcal{O}(D\sqrt{\ln(D+1) + \ln \min\{m, n\}})$, which is the claimed result. □

Appendix C Missing Proofs of Section 3.4

C.1 Proof of Lemma 5

Proof of Lemma 5 Let x^* be the minimiser of $\mathcal{B}_{\psi_{t+1}}(x, y_{t+1})$ in $\mathcal{K}$. Using the the fact $\ln a \geq 1 - \frac{1}{a}$, we obtain

$$\ln(\frac{|x_i^*|}{\beta} + 1) \geq \frac{|x_i^*|}{|x_i^* + \beta|}$$

and

$$((|x_i^*| + \beta) \ln(\frac{|x_i^*|}{\beta} + 1) - |x_i^*| \geq 0.$$

Thus, $y_i = 0$ implies $x_i^* = 0$. Furthermore $\operatorname{sgn}(x_i^*) = \operatorname{sgn}(y_i)$ must hold for all i with $y_i \neq 0$, since otherwise we can always flip the sign of x_i^* to obtain smaller objective value. So we assume without loss of generality that $y_i \geq 0$. We claim that $\sum_{i=1}^{d} x_i^* = D$ holds for the minimiser x^*. If it is not the case, there must be some i with $x_i^* < y_i$, and increasing x_i^* by a small enough amount can decrease the objective function. Thus minimising the Bregman divergence can be rewritten into

$$\begin{aligned}\min_{x \in \mathbb{R}^d} \quad & \sum_{i=1}^{d} ((x_i + \beta) \ln \frac{x_i + \beta}{y_i + \beta} - x_i) \\ \text{s.t.} \quad & \sum_{i=1}^{d} x_i = D \\ & x_i \geq 0 \text{ for all } i = 1, \ldots, d.\end{aligned} \tag{C11}$$

Using Lagrange multipliers for $x \in \mathbb{R}^d$, $\lambda \in \mathbb{R}$ and $\nu \in \mathbb{R}_+^d$

$$\mathcal{L}(x, \lambda, \nu) = \sum_{i=1}^{d} ((x_i + \beta) \ln \frac{x_i + \beta}{y_i + \beta} - x_i) - \nu^\top x - \lambda(D - \sum_{i=1}^{d} x_i).$$

Setting $\frac{\partial \mathcal{L}}{\partial x_i} = 0$, we obtain

$$\ln \frac{x_i + \beta}{y_i + \beta} = \nu_i - \lambda.$$

From the complementary slackness, we have $\nu_i = 0$ for $x_i \neq 0$, which implies

$$x_i + \beta = \frac{1}{z}(y_i + \beta),$$

where $z = \exp(\lambda)$. Let x^* be the minimiser and $\mathcal{I} = \{i : x_i^* > 0\}$ the support of x^*. Then we have

$$D + |\mathcal{I}|\beta = \frac{1}{z}(\sum_{i \in \mathcal{I}} y_i + |\mathcal{I}|\beta).$$

Let p be a permutation of $\{1, \ldots, d\}$ such that $y_{p(i)} \leq y_{p(i+1)}$. Define

$$\theta(j) = y_{p(j)}(D + (d - j + 1)\beta) + \beta D - \beta \sum_{i \geq j} y_{p(i)}.$$

It follows from

$$\theta(j+1) - \theta(j) = (y_{p(j+1)} - y_{p(j)})(D + (d - j + 1)\beta) \geq 0$$

that $\theta(j)$ is increasing in j. Let $\rho = \min\{i | \theta(i) > 0\}$. For all $j < \rho$, $p(j)$ is not in the support $\mathcal{I}$, since otherwise it would imply $x^*_{p(j)} \leq 0$. Thus the minimisation problem (C11) is equivalent to

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \sum_{i=\rho}^{d} (x_{p(i)} + \beta) \ln \frac{x_{p(i)} + \beta}{y_{p(i)} + \beta} \\ \text{s.t.} \quad & \sum_{i=\rho}^{d} x_{p(i)} = D \\ & x_{p(i)} > 0 \text{ for all } i = \rho, \ldots, d. \end{aligned} \tag{C12}$$

Define function $R : \mathbb{R}_{>0} \to \mathbb{R}, x \mapsto x \ln x$. It can be verified that R is convex. The objective function in (C12) can be further rewritten into

$$\begin{aligned} & \sum_{i=\rho}^{d} (x_{p(i)} + \beta) \ln \frac{x_{p(i)} + \beta}{y_{p(i)} + \beta} \\ = & \sum_{i=\rho}^{d} (y_{p(i)} + \beta) R(\frac{x_{p(i)} + \beta}{y_{p(i)} + \beta}) \\ \geq & \frac{1}{\sum_{i=\rho}^{d} (y_{p(i)} + \beta)} R(\frac{\sum_{i=\rho}^{d} (x_{p(i)} + \beta)}{\sum_{i=\rho}^{d} (y_{p(i)} + \beta)}) \\ = & \frac{1}{\sum_{i=\rho}^{d} (y_{p(i)} + \beta)} R(\frac{D + (d - \rho + 1)\beta}{\sum_{i=\rho}^{d} (y_{p(i)} + \beta)}), \end{aligned}$$

where the inequality follows from the Jensen's inequality. The minimum is attained if and only if $\frac{x_{p(i)} + \beta}{y_{p(i)} + \beta}$ are equal for all i. This is only possible when $\sigma(i)$ is in the support $\mathcal{I}$ for all $i \geq \rho$. Thus we can set $z = \frac{\sum_{i=\rho}^{d} (|y_{p(i)}| + \beta)}{D + (d - \rho + 1)\beta}$ and obtain $x_i^* = \max\{\frac{(|y_i| + \beta) - \beta}{z}, 0\} \operatorname{sgn}(y_i)$ for $i = 1 \ldots d$, which is the claimed result. □

C.2 Proof of Corollary 1

Proposition 3 *Let $\{x_t\}$ be any sequences and $\{y_t\}$ be the sequence produced by $y_{t+1} = \frac{a_t}{a_{1:t}} x_t + (1 - \frac{a_t}{a_{1:t}}) y_t$. Choosing $a_t > 0$, we have, for all $x \in \mathcal{W}$*

$$a_{1:T}\mathbb{E}[f(y_{T+1}) - f(x)] \leq \mathbb{E}[\mathcal{R}_{1:T}] - \sum_{t=1}^{T}(a_{1:t-1}\mathcal{B}_l(y_t, y_{t+1})),$$

with $\mathcal{R}_{1:T} = \sum_{t=1}^{T} a_t(\langle g_t, x_t - x\rangle + r(x_t) - r(x))$.

Proof It is interesting to see that the average scheme can be considered as an instance of the linear coupling introduced in Allen-Zhu and Orecchia (2017). For any sequence $\{x_t\}$, $\{y_t\}$ and $z_t = \frac{a_t}{a_{1:t}} x_t + (1 - \frac{a_t}{a_{1:t}}) y_t$, we start the proof by bounding $a_t(f(y_{t+1}) - f(x))$ as follows

$$\begin{aligned} & a_t(l(y_{t+1}) - l(x)) \\ =& a_t(l(y_{t+1}) - l(z_t) + l(z_t) - l(x)) \\ =& a_t(l(y_{t+1}) - l(z_t) + \langle \nabla l(z_t), z_t - x\rangle - \mathcal{B}_l(z_t, x)) \\ =& a_t(l(y_{t+1}) - l(z_t) + \langle \nabla l(z_t), z_t - x_t\rangle + \langle \nabla l(z_t), x_t - x\rangle - \mathcal{B}_l(z_t, x)) \end{aligned} \tag{C13}$$

Denote by $\tau_t = \frac{a_t}{a_{1:t}}$ the weight. The first term of the the inequality above can be further bounded by

$$\begin{aligned} & a_t(l(y_{t+1}) - l(z_t) + \langle \nabla l(z_t), z_t - x_t\rangle) \\ =& a_t(l(y_{t+1}) - l(z_t) + \frac{1-\tau_t}{\tau_t}\langle \nabla l(z_t), y_t - z_t\rangle) \\ =& a_t(l(y_{t+1}) - l(z_t) + (\frac{1}{\tau_t} - 1)(l(y_t) - l(z_t)) - (\frac{1}{\tau_t} - 1)\mathcal{B}_l(y_t, z_t)) \\ =& a_t(\frac{1}{\tau_t} - 1)(l(y_t) - l(y_{t+1})) + \frac{a_t}{\tau_t}(l(y_{t+1}) - l(z_t)) - a_{1:t-1}\mathcal{B}_l(y_t, z_t). \end{aligned} \tag{C14}$$

Next, we have

$$\begin{aligned} & \sum_{t=1}^{T} a_t(\frac{1}{\tau_t} - 1)(f(y_t) - f(y_{t+1})) \\ =& \sum_{t=2}^{T} a_{1:t-1}(f(y_t) - f(y_{t+1})) \\ =& \sum_{t=1}^{T-1} a_t f(y_{t+1}) - a_{1:T-1} f(y_{T+1}) \\ =& \sum_{t=1}^{T} a_t f(y_{t+1}) - a_{1:T} f(y_{T+1}) \\ =& \sum_{t=1}^{T} a_t(f(y_{t+1}) - f(y_{T+1})) \end{aligned} \tag{C15}$$

Combining (C13), (C14) and (C15), we have

$$\begin{aligned}a_{1:T}(f(y_{T+1})-f(x)) =&\sum_{t=1}^{T}\frac{a_t}{\tau_t}(l(y_{t+1})-l(z_t))\\&+\sum_{t=1}^{T}\langle\nabla l(z_t),x_t-x\rangle\\&-\sum_{t=1}^{T}(a_{1:t-1}\mathcal{B}_l(y_t,z_t)-a_t\mathcal{B}_l(z_t,x)),\end{aligned}$$

Simply setting $y_{t+1}:=z_t$ makes the first term above 0 and implies $z_t=\frac{\sum_{s=1}^{t}a_sx_s}{a_{1:t}}$. Furthermore it follows from the convexity of r

$$r(y_{T+1})=r(\frac{\sum_{s=1}^{T}a_sx_s}{a_{1:T}})\leq\sum_{t=1}^{T}\frac{a_tr(x_t)}{a_{1:T}}.$$

Combining the inequalities above and rearranging, we obtain

$$\begin{aligned}a_{1:T}(f(y_{T+1})-f(x)) \leq&\sum_{t=1}^{T}a_t(\langle\nabla l(z_t),x_t-x\rangle+r(x_t)-r(x))\\&-\sum_{t=1}^{T}(a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)+a_t\mathcal{B}_l(z_t,x))\\\leq&\sum_{t=1}^{T}a_t(\langle\nabla l(z_t),x_t-x\rangle+r(x_t)-r(x))\\&-\sum_{t=1}^{T}(a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t))\end{aligned}$$

Furthermore, we have

$$\begin{aligned}&\mathbb{E}[\sum_{t=1}^{T}\langle a_t\nabla l(z_t),x_t-x\rangle]\\=&\mathbb{E}[\sum_{t=1}^{T}\langle a_tg_t,x_t-x\rangle]+\mathbb{E}[\sum_{t=1}^{T}\langle a_t(\nabla l_t-g_t),x_t-x\rangle]\\=&\mathbb{E}[\sum_{t=1}^{T}\langle a_tg_t,x_t-x\rangle]+\sum_{t=1}^{T}\mathbb{E}[\langle a_t(\nabla l_t-g_t),x_t-x\rangle]\\=&\mathbb{E}[\sum_{t=1}^{T}\langle a_tg_t,x_t-x\rangle]+\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}[\langle a_t(\nabla l_t-g_t),x_t-x\rangle|z_t]]\\=&\mathbb{E}[\sum_{t=1}^{T}\langle a_tg_t,x_t-x\rangle].\end{aligned}$$

Finally, we we obtain

$$\begin{aligned}a_{1:T}\mathbb{E}[f(y_{T+1})-f(x)]\leq&\mathbb{E}[\sum_{t=1}^{T}a_t(\langle g_t,x_t-x\rangle+r(x_t)-r(x))]\\&-\sum_{t=1}^{T}(a_{1:t-1}\mathcal{B}_l(y_t,y_{t+1})),\end{aligned}$$

which is the claimed result. □

Proof of Corollary 1 . In general case, we have

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{1:T}] \leq & c_1 + c_2\mathbb{E}[\sqrt{\sum_{t=1}^{T}\|a_t(g_t - g_{t-1})\|_*^2}] \\ \leq & c_1 + c_2\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2]} \\ \leq & c_1 + c_2\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2 | z_t]}.\end{aligned} \tag{C16}$$

Foll all t, we have

$$\begin{aligned}\mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2 | z_t] \leq & 2a_t^2(\mathbb{E}[\|g_t - \nabla l(z_t) - g_{t-1} + \nabla l(z_{t-1})\|_*^2 | z_t]) \\ & + 2a_t^2(\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2).\end{aligned} \tag{C17}$$

Since z_{t-1} is fixed when z_t is given, the first term above can be bounded by

$$\begin{aligned}& 2a_t^2(\mathbb{E}[\|g_t - \nabla l(z_t) - g_{t-1} + \nabla l(z_{t-1})\|_*^2 | z_t]) \\ \leq & 4a_t^2(\mathbb{E}[\|g_t - \nabla l(z_t)\|_*^2 | z_t] + \mathbb{E}[\|g_{t-1} - \nabla l(z_{t-1})\|_*^2 | z_t]) \\ \leq & 4a_t^2(\mathbb{E}[\|g_t - \nabla l(z_t)\|_*^2 | z_t] + \mathbb{E}[\|g_{t-1} - \nabla l(z_{t-1})\|_*^2 | z_{t-1}]) \\ \leq & 4a_t^2(\nu_t^2 + \nu_{t-1}^2).\end{aligned}$$

Since $\mathcal{K}$ is compact, there is some $L > 0$ such that $\|\nabla l(z)\|_* \leq L$ for all $z \in \mathbb{X}$. Thus the second term of (C17) can be bounded by

$$2a_t^2\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2 \leq 8a_t^2 L^2 \tag{C18}$$

Combining (C16), (C17) and (C18), we have

$$\mathbb{E}[\mathcal{R}_{1:T}] \leq c1 + c2\sqrt{8\sum_{t=1}^{T} a_t^2(\nu_t^2 + L^2)},$$

and combining with Proposition 3, we obtain

$$\mathbb{E}[f(z_T) - f(x)] \leq \frac{c1 + c2\sqrt{8\sum_{t=1}^{T} a_t^2(\nu_t^2 + L^2)}}{a_{1:T}}.$$

If l is M-smooth, then for $t \geq 2$, we have

$$\begin{aligned}2a_t^2\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2 \leq & \frac{4Ma_t^2}{a_{1:t-1}} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t). \\ & 8Ma_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t).\end{aligned} \tag{C19}$$

Using fact $2ab - a^2 \leq b^2$, we have

$$\begin{aligned}& 2c_2\sqrt{2M\sum_{t=2}^{T} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t)} - \sum_{t=2}^{T} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t) \\ \leq & 2Mc_2^2.\end{aligned} \tag{C20}$$

Combining (C16), (C17) and (C20), we have

$$
\begin{aligned}
&\mathbb{E}[\mathcal{R}_{1:T}] - \sum_{t=1}^{T} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t) \\
\le& c_1 + c_2\sqrt{\sum_{t=1}^{T} \mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2 | z_t]} - \sum_{t=1}^{T} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t) \\
\le& c_1 1 + c_2\sqrt{8\sum_{t=1}^{T} a_t^2(\nu_t^2)} \\
&\quad + c_2\sqrt{\sum_{t=1}^{T} 2a_t^2\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2} - \sum_{t=1}^{T} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t) \\
\le& c_1 1 + c_2\sqrt{8\sum_{t=1}^{T} a_t^2(\nu_t^2)} + c_2\sqrt{2}\|\nabla l(z_1)\|_* \\
&\quad + c_2\sqrt{\sum_{t=2}^{T} 2a_t^2\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2} - \sum_{t=2}^{T} a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t) \\
\le& c_1 + c_2\sqrt{8\sum_{t=1}^{T} a_t^2(\nu_t^2)} + \sqrt{2}c_2 L + 2Mc_2^2,
\end{aligned}
$$

which implies

$$
\mathbb{E}[f(z_T) - f(x)] \le \frac{c_1 + c_2\sqrt{8\sum_{t=1}^{T} a_t^2\nu_t^2} + \sqrt{2}c_2 L + 2Mc_2^2}{a_{1:T}}.
$$

□

Appendix D Technical Lemmata

Lemma 6 *For positive values $a_1, \ldots, a_n$ the following holds:*

1.

$$
\sum_{i=1}^{n} \frac{a_i}{\sum_{k=1}^{i} a_k + 1} \le \log(\sum_{i=1}^{n} a_i + 1)
$$

2.

$$
\sqrt{\sum_{i=1}^{n} a_i} \le \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j^2}} \le 2\sqrt{\sum_{i=1}^{n} a_i}.
$$

Proof The proof of (1) can be found in Lemma A.2 in Levy et al. (2018) For (2), we define $A_0 = 1$ and $A_i = \sum_{k=1}^{i} a_i + 1$ for $i > 0$. Then we have

$$\begin{aligned}\sum_{i=1}^{n} \frac{a_i}{\sum_{k=1}^{i} a_k + 1} =& \sum_{i=1}^{n} \frac{A_i - A_{i-1}}{A_i}\\ =& \sum_{i=1}^{n} (1 - \frac{A_{i-1}}{A_i})\\ \leq& \sum_{i=1}^{n} \ln \frac{A_i}{A_{i-1}}\\ =& \ln A_n - \ln A_0\\ =& \ln \sum_{i=1}^{n} (a_i + 1),\end{aligned}$$

where the inequality follows from the concavity of log. □

Lemma 7 *Let l be convex and M-smooth over $\mathbb{X}$, i.e.*

$$l(x) \leq l(y) + \langle \nabla l(y), x - y \rangle + \frac{M}{2} \|x - y\|^2.$$

Then

$$\|\nabla l(x) - \nabla l(y)\|_*^2 \leq 2M\mathcal{B}_l(x, y)$$

holds for all $x, y \in \mathbb{X}$.

Proof Let $x, y \in \mathbb{X}$ be arbitrary. Define $h : \mathbb{X} \to \mathbb{R}, z \mapsto l(z) - \langle \nabla l(y), z \rangle$. Clearly, h is M-smooth and minimised at y. Thus we have

$$\begin{aligned}h(y) =& \min_{z \in \mathbb{X}} h(z)\\ \leq& \min_{z \in \mathbb{X}} h(x) + \langle \nabla h(x), z - x \rangle + \frac{M}{2} \|z - x\|^2\\ \leq& \min_{\gamma \geq 0} h(x) - \|\nabla h(x)\|_* \gamma + \frac{M}{2} \gamma^2\\ =& h(x) - \frac{1}{2M} \|\nabla h(x)\|_*^2,\end{aligned}$$

where the first inequality uses the M-smoothness of h, and the second uses $\langle \nabla h(x), z - x \rangle \geq -\|\nabla h(x)\|_* \|z - x\|$, for which we choose z such that the equality holds. This implies

$$\frac{1}{2M} \|\nabla l(x) - \nabla l(y)\|_*^2 \leq l(x) - l(y) - \langle \nabla l(y), x - y \rangle = \mathcal{B}_l(x, y),$$

and the desired result follows. □

Lemma 8 *Define $\psi : \mathbb{R}^d \to \mathbb{R}, x \mapsto \sum_{i=1}^{d} \phi(x_i)$ for ϕ be as defined in (1). Assume $\|x\|_1 \leq D$ for all $x \in \mathcal{K} \subseteq \mathbb{R}^d$. Setting $\beta = \frac{1}{d}$, we obtain for all $x, y \in \mathcal{K}$*

$$\mathcal{B}_\psi(x, y) \leq 4D(\ln(D + 1) + \ln d).$$

Similarly, we define $\Psi : \mathbb{R}^{m,n} \to \mathbb{R}, x \mapsto \psi \circ \sigma(x)$. Assume $\|x\|_1 \leq D$ for all $x \in \mathcal{K} \subseteq \mathbb{R}^{m,n}$. Setting $\beta = \frac{1}{\min\{m,n\}}$, we obtain for all $x, y \in \mathcal{K}$

$$\mathcal{B}_\Psi(x, y) \leq 4D(\ln(D + 1) + \ln \min\{m, n\}).$$

Proof From the definition of the Bregman divergence it follows for all $x, y \in \mathcal{K}$

$$\begin{aligned}\mathcal{B}_\psi(x,y) =&\psi(x) - \psi(y) - \langle \nabla \psi(y), x - y\rangle\\ \leq&\langle \nabla\psi(x) - \nabla\psi(y), x - y\rangle\\ \leq&\|\nabla\psi(x) - \nabla\psi(y)\|_\infty \|x - y\|_1\\ \leq&2D(\|\nabla\psi(x)\|_\infty + \|\nabla\psi(y)\|_\infty).\end{aligned}$$

Using the closed form of $\|\nabla\psi(x)\|_\infty$, we have for $x \in \mathcal{K}$

$$\begin{aligned}\|\nabla\psi(x)\|_\infty =&\max_i |\ln(\frac{|x_i|}{\beta} + 1)|\\ \leq&|\ln(D + \beta)| + |\ln(\frac{1}{\beta})|\\ \leq&\ln(D + 1) + \ln d.\end{aligned}$$

Combine the inequalities above and choose $\beta = \frac{1}{d}$, we obtain

$$\mathcal{B}_\psi(x,y) =4D(\ln(D + 1) + \ln d).$$

Using the same argument, we have for all $x, y \in \mathcal{K} \subseteq \mathbb{R}^{m,n}$

$$\mathcal{B}_\Psi(x,y) =2D(\|\nabla\Psi(x)\|_\infty + \|\nabla\Psi(y)\|_\infty).$$

From the characterisation of subgradient, it follows for $x \in \mathcal{K}$

$$\begin{aligned}\|\nabla\Psi(x)\|_\infty =&\|\nabla\phi(\sigma(x))\|_\infty\\ \leq&\ln(D + 1) + \ln\frac{1}{\beta}.\end{aligned}$$

Combine the inequalities above and choose $\beta = \frac{1}{\min\{m,n\}}$, we obtain

$$\mathcal{B}_\Psi(x,y) \leq 4D(\ln(D + 1) + \ln\min\{m, n\}).$$

□

References

Alacaoglu, A., Malitsky, Y., Mertikopoulos, P., Cevher, V. (2020). A new regret analysis for adam-type algorithms. *International conference on machine learning* (pp. 202–210).

Allen-Zhu, Z., & Orecchia, L. (2017). Linear coupling: An ultimate unification of gradient and mirror descent. *8th innovations in theoretical computer science conference (itcs 2017).*

Arora, S., Hazan, E., Kale, S. (2012). The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, *8*(1), 121–164.

Barbu, V., & Precupanu, T. (2012). *Convexity and optimization in banach spaces*. Springer Science & Business Media.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, *2*(1), 183–202.

Bhatia, R. (2013). *Matrix analysis* (Vol. 169). Springer Science & Business Media.

Cancela, B., Bolón-Canedo, V., Alonso-Betanzos, A. (2021). A delayed elastic-net approach for performing adversarial attacks. *2020 25th international conference on pattern recognition (icpr)* (p. 378-384). 10.1109/ICPR48806.2021.9413170

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 ieee symposium on security and privacy (sp)* (pp. 39–57).

Cesa-Bianchi, N., & Gentile, C. (2008). Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, *54*(1), 386–390.

Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.-J. (2018). Ead: elastic-net attacks to deep neural networks via adversarial examples. *Thirty-second aaai conference on artificial intelligence.*

Cutkosky, A. (2019). Anytime online-to-batch, optimism and acceleration. *International conference on machine learning* (pp. 1446–1454).

Cutkosky, A., & Boahen, K. (2017). Online learning without prior information. *Conference on learning theory* (pp. 643–677).

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.

Duchi, J., Hazan, E., Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.

Gentile, C. (2003). The robustness of the p-norm algorithms. *Machine Learning*, *53*(3), 265–299.

Ghai, U., Hazan, E., Singer, Y. (2020). Exponentiated gradient meets gradient descent. *Algorithmic learning theory* (pp. 386–407).

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *2016 ieee conference on computer vision and pattern recognition* *(cvpr)* (p. 770-778). 10.1109/CVPR.2016.90

Joulani, P., György, A., Szepesvári, C. (2017). A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, and variational bounds. *Journal of Machine Learning Research*, *1*, 40.

Joulani, P., Raj, A., Gyorgy, A., Szepesvári, C. (2020). A simpler approach to accelerated optimization: iterative averaging meets optimism. *International conference on machine learning* (pp. 4984–4993).

Kakade, S.M., Shalev-Shwartz, S., Tewari, A. (2012). Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, *13*(1), 1865–1890.

Kavis, A., Levy, K.Y., Bach, F., Cevher, V. (2019). Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in neural information processing systems* (pp. 6260–6269).

Kivinen, J., & Warmuth, M.K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, *132*(1), 1–63.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images.

Levy, Y.K., Yurtsever, A., Cevher, V. (2018). Online adaptive methods, universality and acceleration. *Advances in neural information processing systems* (pp. 6500–6509).

Lewis, A.S. (1995). The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, *2*(1), 173–183.

Li, X., & Orabona, F. (2019). On the convergence of stochastic gradient descent with adaptive stepsizes. *The 22nd international conference on artificial intelligence and statistics* (pp. 983–992).

Lu, C., Lin, Z., Yan, S. (2014). Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, *24*(2), 646–654.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course* (Vol. 87). Springer Science & Business Media.

Orabona, F. (2013). Dimension-free exponentiated gradient. *Nips* (pp. 1806–1814).

Orabona, F., Crammer, K., Cesa-Bianchi, N. (2015). A generalized online mirror descent with applications to classification and regression. *Machine Learning*, *99*(3), 411–435.

Orabona, F., & Pál, D. (2018). Scale-free online learning. *Theoretical Computer Science*, *716*, 50–69.

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Steinhardt, J., & Liang, P. (2014). Adaptivity and optimism: An improved exponentiated gradient algorithm. *International conference on machine learning* (pp. 1593–1601).

Warmuth, M.K. (2007). Winnowing subspaces. *Proceedings of the 24th international conference on machine learning* (pp. 999–1006).

Xie, C., Bijral, A., Ferres, J.L. (2018). Nonstop: A nonstationary online prediction method for time series. *IEEE Signal Processing Letters*, *25*(10), 1545–1549.