Machine Learning (2022) 111:4719–4764
https://doi.org/10.1007/s10994-022-06229-1

Optimistic optimisation of composite objective with exponentiated update

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

Received: 3 February 2022 / Revised: 31 May 2022 / Accepted: 22 July 2022 /
Published online: 22 August 2022

Abstract
This paper proposes a new family of algorithms for the online optimisation of composite objectives. The algorithms can be interpreted as the combination of the 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 the optimal accelerated rate for smooth objectives.

Keywords Exponetiated gradient · Composite objective · Online convex optimisation · Sparsity

1 Introduction

Many machine learning problems involve minimising high dimensional composite objectives (Dhurandhar et al., 2018; Lu et al., 2014; Ribeiro et al., 2016; Xie et al., 2018). For example, in the task of explaining predictions of an image classifier (Dhurandhar et al., 2018; Ribeiro et al., 2016), we need to find a sufficiently small set of features explaining the prediction by solving the following constrained optimisation problem

Editors: Krzysztof Dembczynski and Emilie Devijver.

✉ Weijia Shao
weijia.shao@campus.tu-berlin.de

Fikret Sivrikaya
fikret.sivrikaya@gt-arc.com

Sahin Albayrak
sahin.albayrak@dai-labor.de

[1] Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

[2] GT-ARC Gemeinnützige GmbH, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

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

where l is a function relating to the classifier, λ_1 controls the sparsity of the feature set, λ_2 controls the complexity of the feature set, and $c_1, \dots, c_d$ are the ranges of the features. For l with a complicated structure and large d, it is practical to solve the problem by optimising the first-order approximation of the objective function (Lan, 2020). However, the first-order methods can not attain optimal performance due to the non-smooth component $\lambda_1\|\cdot\|_1$. Furthermore, the purpose of introducing the ℓ_1 regularisation is to ensure the sparsity of the decision variable. Applying the first-order algorithms directly on the subgradient of $\lambda_1\|\cdot\|_1$ does not lead to sparse updates (Duchi et al., 2010). We refer to the objective function consisting of a loss with a complicated structure and a simple (possibly non-smooth) convex regularisation term as a composite objective.

This paper focuses on the more general online convex optimisation (*OCO*), which 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 l_t to the player, who then suffers the composite loss $f_t(x) = l_t(x) + r_t(x)$, where $l_t : \mathcal{K} \to \mathbb{R}$ is a convex function revealed at each iteration and $r_t : \mathbb{X} \to \mathbb{R}_{\ge 0}$ is a known closed convex function. The target is to develop algorithms minimising the regret of not choosing the best decision $x \in \mathcal{K}$

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

An online optimisation algorithm can be converted into a stochastic optimisation algorithm using the online-to-batch conversion technique (Cesa-Bianchi et al., 2004), which is our primary motivation. In addition to that, online optimisation also has many direct applications, such as recommender systems (Song et al., 2014) and time series prediction (Anava et al., 2013).

Given a sequence of subgradients $\{g_t\}$ of $\{l_t\}$, we are interested in the so-called adaptive algorithms ensuring regret bounds of the form $\mathcal{O}(\sqrt{\sum_{t=1}^{T}\|g_t\|_*^2})$. The adaptive algorithms are worst-case optimal in the online setting (McMahan & Streeter, 2010) and can be converted into stochastic optimisation algorithms with optimal convergence rates (Cutkosky, 2019; Joulani et al., 2020; Kavis et al., 2019; Levy et al., 2018). The adaptive subgradient methods (*AdaGrad*) (Duchi et al., 2011) and their variants (Alacaoglu et al., 2020; Duchi et al., 2011; Orabona & Pál, 2018; Orabona et al., 2015) have become the most popular adaptive algorithms in recent years. They are often applied to estimating deep learning models and outperform standard optimisation algorithms when the gradient vectors are sparse. However, such property can not be expected in every problem. If the decision variables are in an ℓ_1 ball and gradient vectors are dense, the *Adagrad*-style algorithms do not have an optimal theoretical guarantee due to the sub-linear regret dependence on the dimensionality.

The exponentiated gradient (*EG*) methods (Arora et al., 2012; Kivinen & Warmuth, 1997), which are designed for estimating weights in the positive orthant, enjoy the regret bound growing logarithmically with the dimensionality. The $EG^{\pm}$ algorithm generalises this idea to negative weights (Kivinen & Warmuth, 1997; Warmuth, 2007). Given d dimensional problems with the maximum norm of the gradient bounded by G, the regret of $EG^{\pm}$

is upper bounded by $\mathcal{O}(G\sqrt{T \ln d})$. As the performance of the $EG^{\pm}$ algorithm depends strongly on the choice of hyperparameters, the p-norm algorithm (Gentile, 2003), which is less sensitive to the tuning of hyperparameters, is introduced to approach the logarithmic behaviour of $EG^{\pm}$. Kakade et al. (2012) further extends the p-norm algorithm to learning with matrices. An adaptive version of the p-norm algorithm is analysed in Orabona et al. (2015), which has a regret upper bound proportional to $\|x\|_{p,*}^2 \sqrt{\sum_{t=1}^T \|g_t\|_p^2}$ for a given sequence of gradients $\{g_t\}$. By choosing $p = 2 \ln d$, a regret upper bound $\mathcal{O}(\|x\|_1^2 \sqrt{\ln d \sum_{t=1}^T \|g_t\|_\infty^2})$ can be achieved. However, tuning hyperparameters is still required to attain the optimal regret $\mathcal{O}(\|x\|_1 \sqrt{\ln d \sum_{t=1}^T \|g_t\|_\infty^2})$.

Recently, Ghai et al. (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 behaviour as in *EG* and a stepsize that can be flexibly scheduled as gradient descent. However, many optimisation problems with sparse targets have an ℓ_1 or nuclear regulariser in the objective function. Otherwise, the optimisation algorithm has to pick a 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 to normalise the decision variable at each iteration, which is equivalent to the $EG^{\pm}$ algorithm and leads to a performance dependence on the tuning.

This paper proposes a family of algorithms for the 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. The directions of the decision variables 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 et al., 2017). Given the radius of decision set D, sequences of gradients $\{g_t\}$ and hints $\{h_t\}$, the proposed algorithms achieve a regret upper bound in the form of $\mathcal{O}(D\sqrt{\ln d \sum_{t=1}^T \|g_t - h_t\|_\infty^2})$. With the techniques introduced in Ghai et al. (2020), a spectral analogue of the entropy-like 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 solutions for the ℓ_1 and nuclear regularised objective functions. For the ℓ_2 and Frobenius regularised objectives, the update rules involve values of the principal branch of the *Lambert function*, which can be well approximated. We propose a sorting based procedure projecting the solution to the decision set for the ℓ_1 or nuclear ball constrained problems. Finally, the proposed online algorithms can be converted into 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 objective functions. The convergence rate depends logarithmically on the dimensionality of the problem, which suggests its advantage compared to the accelerated *AdaGrad*-Style algorithms (Cutkosky, 2019; Joulani et al., 2020; Levy et al., 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 Sect. 4. In Sect. 5, we derive efficient implementations for some popular choices of composite objectives, constraints and stochastic optimisation. Section 6 demonstrates the empirical evaluations using both synthetic and real-world data. Finally, we conclude our work in Sect. 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 attacking (Cancela et al., 2021; Carlini & Wagner, 2017; Chen et al., 2018) and explaining (Dhurandhar et al., 2018; Ribeiro et al., 2016) deep neural networks. The proximal gradient method (*PGD*) (Nesterov, 2003) and its accelerated variants (Beck & Teboulle, 2009) are usually applied to solving the problem. However, these algorithms are not practical since they require prior knowledge about the smoothness of the objective function to ensure their convergence.

The *AdaGrad*-style algorithms (Alacaoglu et al., 2020; Duchi et al., 2011; Orabona & Pál, 2018; Orabona et al., 2015) have become popular in the machine learning community in recent years. Given the gradient vectors $g_1, \dots, g_t$ received at iteration t, the core idea of these algorithms is to set the stepsizes 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 & Orabona, 2019) or can be converted into a stochastic algorithm (Cesa-Bianchi & Gentile, 2008) with a convergence rate $\mathcal{O}(\frac{1}{\sqrt{T}})$. This rate can be further improved to $\mathcal{O}(\frac{1}{T^2})$ for unconstrained problems with smooth loss functions by applying the acceleration techniques (Cutkosky, 2019; Kavis et al., 2019; Levy et al., 2018). These acceleration techniques do not require prior knowledge about the smoothness of the loss function and a guarantee convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ for non-smooth functions. Joulani et al. (2020) has proposed a simple approach to accelerate optimistic online optimisation algorithms with adaptive regret bound.

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 regret dependence on the dimensionality, which can be attained by the *EG* family algorithms (Arora et al., 2012; Kivinen & Warmuth, 1997; Warmuth, 2007) and their adaptive optimistic extension (Steinhardt & Liang, 2014). However, these algorithms work only for decision sets in the form of cross-polytopes and require prior knowledge about 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 still requires prior knowledge about the problem to attain optimal performance (Orabona et al., 2015). 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. The core of our algorithm is a symmetric logarithmic function. Orabona (2013) first introduced the idea of composing the single-dimensional symmetric logarithmic function and a norm to generalise *EG* to the infinite-dimensional space. It has become popular for parameter-free optimisation (Cutkosky & Boahen, 2016, 2017a, b; Kempka et al., 2019)

since one can easily construct an adaptive regulariser with this composition (Cutkosky & Boahen, 2017a). In this paper, instead of using the composition, we apply the symmetric logarithmic function directly to each entry of a vector to construct a symmetric entropy-like function that is strongly convex with respect to the ℓ_1 norm. We analyse *MD* and *FTRL* with the entropy-like function in the framework developed in Joulani et al. (2017). The analysis of the spectral analogue of the entropy-like function follows the idea proposed in Ghai et al. (2020).

3 Preliminary

The focus of this paper is *OCO* with the decision variable taken from a compact convex subset $\mathcal{K} \subseteq \mathbb{X}$ of finite dimensional vector space equipped with a 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 the 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

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

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

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

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

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

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 $\nabla f(y)$ any element in $\partial f(y)$. A function is η-strongly convex with respect to $\|\cdot\|$ over $\mathcal{K}$ if

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

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

4 Algorithms and analysis

In this section, we present and analyse our algorithms, which begins with a short review on *EG* and the *p*-norm algorithm 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

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

where $g_t \in \partial f_t(x_t)$ is the subgradient, and $\eta > 0$ is the stepsize. Although the algorithm has the expected logarithmic dependence on the dimensionality, 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 apply the $EG^{\pm}$ trick, i.e. use the 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 choose the decision variable $x_{t+1,+} - x_{t+1,-}$. However, if the decision set is implicitly given by a regularisation term, 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

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

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

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

In the next lemma, we show the twice differentiability and strict convexity of ϕ, based on which a strongly convex potential function for *OMD* in a compact decision set can be constructed.

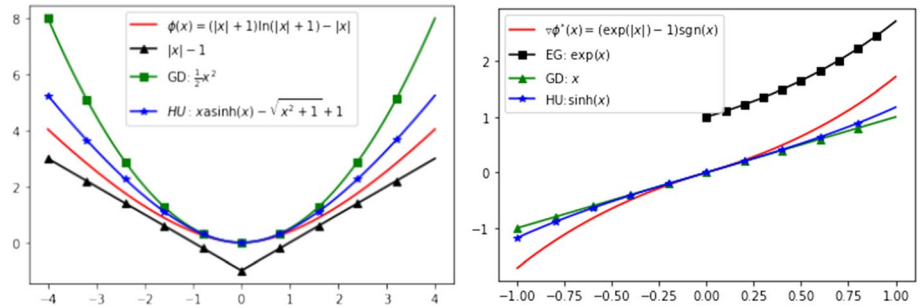
Lemma 1 *ϕ is twice continuous differentiable and strictly convex with*

1. $\phi'(x) = \alpha \ln\left(\frac{|x|}{\beta} + 1\right)\operatorname{sgn}(x)$
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. $\phi^{*\prime}(\theta) = \left(\beta \exp\frac{|\theta|}{\alpha} - \beta\right)\operatorname{sgn}(\theta)$
2. $\phi^{*\prime\prime}(\theta) = \frac{\beta}{\alpha} \exp\frac{|\theta|}{\alpha}$.

Since we can expand the natural logarithm as $\ln(\frac{|x|}{\beta} + 1) = \frac{|x|}{\beta} - \frac{|x|^2}{2\beta^2} + \frac{|x|^3}{3\beta^3} - \cdots$, $\phi(x)$ can be intuitively considered as an interpolation between the absolute value and square. As



(a) Regularisers **(b)** Mirror Maps

Fig. 1 Comparison of convex regularisers

observed in Fig. 1a, it is closer to the absolute value compared to the hyperbolic entropy introduced in Ghai et al. (2020). Moreover, running *OMD* with regulariser $x \mapsto \sum_{i=1}^{d} \phi(x_i)$ yields an update rule

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

which sets the signs of coordinates like the p-norm algorithm and updates the scale similarly to *EG*. As illustrated in Fig. 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.

4.1 Algorithms in the Euclidean space

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

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

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 with respect to $\|\cdot\|_1$, 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 a constant stepsize. 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 algorithms

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 with respect to $\|\cdot\|_1$ over $\mathcal{K}$, which allows us to use $\{\Phi_t\}$ as the potential functions in *OMD* and *FTRL*.

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 with respect to 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

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

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 2.1".

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 $\{x_t\}$ 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 $\{x_t\}$ 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 closed convex function $R_{t+1} : \mathcal{K} \to \mathbb{R}$, 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

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

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

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

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 principal 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

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

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 $i = 1, \ldots, d$ **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 the nuclear and Frobenius norm. The second line of update rule (6) can be implemented as follows

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

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

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

From the characterisation of subgradient, it follows

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

and

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

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

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

The subgradient of the objective (10) at $x_{t+1} = U_{t+1}\text{diag}(\tilde{x}_{t+1})V_{t+1}^\top$ 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 $\|y_{t+1}\|_1 \le D$. 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 $\text{project}(y, D, \beta)$

Sort $|y_i|$ to get the permutation p such that $|y_{p(i)}| \le |y_{p(i+1)}|$
Define $\theta(j) = |y_{p(j)}|(D + (d - j + 1)\beta) + \beta D - \beta \sum_{i \ge j} |y_{p(i)}|$
$\rho \leftarrow \min\{j | \theta(j) > 0\}$
$z \leftarrow \frac{\sum_{i=\rho}^{d}(|y_{p(i)}| + \beta)}{D + (d - \rho + 1)\beta}$
$x_i^* \leftarrow \max\{\frac{|y_i| + \beta}{z} - \beta, 0\} \text{sgn}(y_i)$ 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

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

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

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

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

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

Thus the problem is reduced to

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

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

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

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 $\mathcal{O}(\sqrt{\sum_{t=1}^T \|g_t - h_t\|_*^2})$ 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 $\mathcal{K}$, $\alpha_{t+1}r$, 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*

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

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

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

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

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

Setting $\alpha_t = t$, we obtain a convergence of $\mathcal{O}(\frac{c_2}{\sqrt{T}})$ in general case, and $\mathcal{O}(\frac{c_2}{T^2} + \frac{c_2 \max_t v_t}{\sqrt{T}})$ 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 dimensionality (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 model parameter w^* has a 99% sparsity and the non-zero values are randomly drawn from the uniform distribution over $[-1, 1]$. 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 choose $d = 10{,}000$ and compare our algorithms with *AdaGrad*, *AdaFTRL* (Duchi et al., 2011) and *HU* (Ghai et al., 2020). For both *AdaGrad* and *AdaFTRL*, we set the i-th entry of the proximal matrix H_t to $h_{ii} = 10^{-6} + \sum_{s=1}^{t-1} g_{s,i}^2$

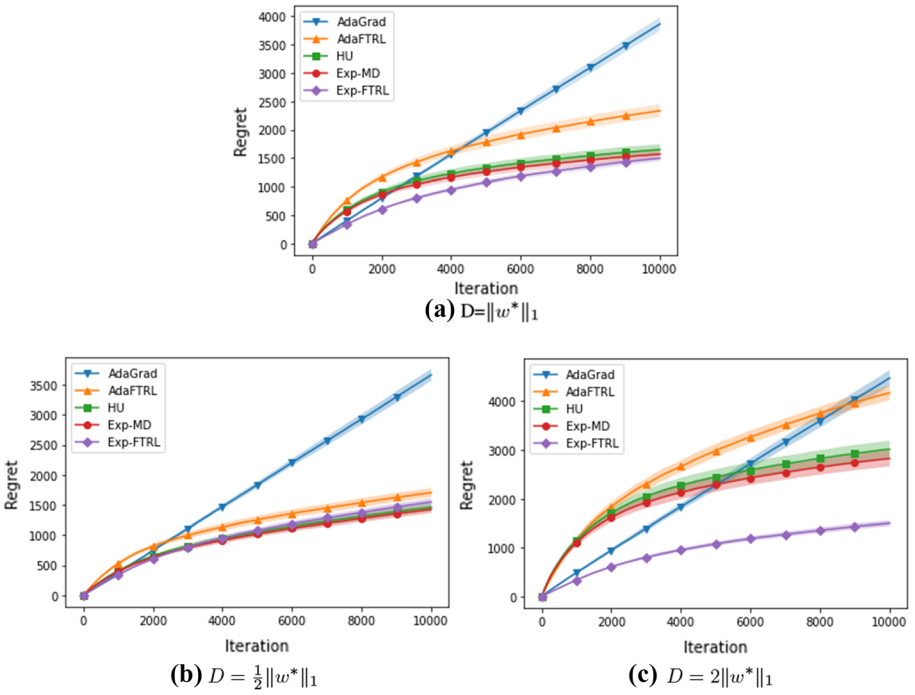


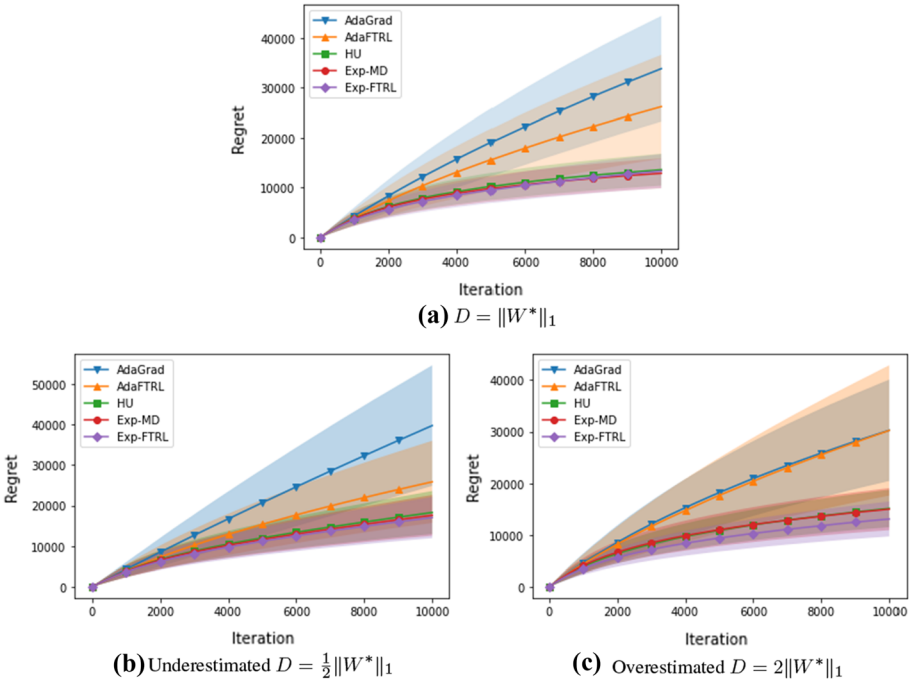
Fig. 2 Online logistic regression

as their theory suggested (Duchi et al., 2011). The stepsize of *HU* is set to $\sqrt{\frac{1}{\sum_{s=1}^{t-1}\|g_s\|_\infty^2}}$ leading to an adaptive regret upper bound. All algorithms take decision variables from an ℓ_1 ball $\{w \in \mathbb{R}^d | \|w\|_1 \leq D\}$, which is the ideal case for *HU*. We examine the performances of the algorithms with known, underestimated and overestimated $\|w^*\|_1$ by setting $D = \|w^*\|_1$, $D = \frac{1}{2}\|w^*\|_1$ and $D = 2\|w^*\|_1$, respectively. For each choice of D, we simulate the online process of each algorithm for 10,000 iterations and repeat the experiments for 20 trials.

Figure 2 plots the curves of the average cumulative regret with the ranges of standard deviation as shaded regions. As can be observed, our algorithms have a clear and stable advantage over the *AdaGrad*-style algorithms and slightly outperform *HU* in the experiments with known $\|w^*\|_1$. As the combination of the entropy-like regulariser and *FTRL* can also be used for parameter-free optimisation (Cutkosky & Boahen, 2017a), overestimating $\|w^*\|_1$ does not have a tangible impact on the performance of *Exp-FTRL*, which leads to its clear advantage over the rest.

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



(a) $D = \|W^*\|_1$

(b) Underestimated $D = \frac{1}{2}\|W^*\|_1$

(c) Overestimated $D = 2\|W^*\|_1$

Fig. 3 Online multitask learning

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 $W^* = U\mathrm{diag}(\sigma)V$. 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 $\{W \in \mathbb{R}^{d,k} | \|W\|_1 \leq D\}$ as the decision set and run the experiment as in Sect. 6.1. The average and standard deviation of the results over 20 trials are shown in Fig. 3.

Similar to the online logistic regression, our algorithms have a clear advantage over *AdaGrad* and *AdaFTRL* and slightly outperform *HU* in all settings. While the regret of the *AdaGrad*-style algorithms spread over a wider range, our algorithms yield relatively stabler results. The superiority of *Exp-FTRL* for the overestimated $\|W^*\|_1$ can also be observed from Fig. 3c.

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 this paper. Given a sample $x_0 \in \mathcal{X}$ and machine learning model $f : \mathcal{X} \to \mathbb{R}^K$, the contrastive explanation consists of a set of

pertinent positive (*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

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

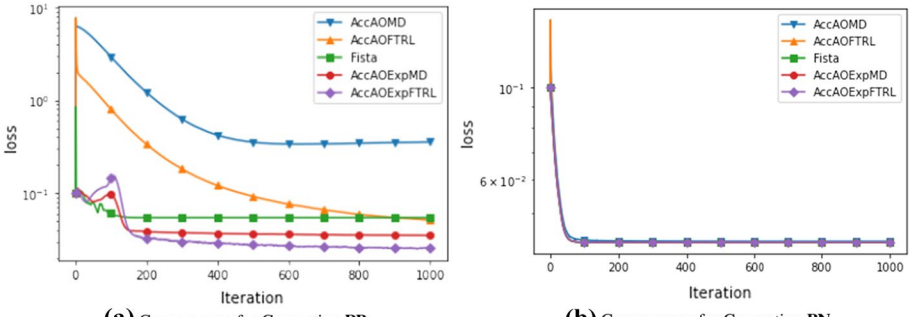
which imposes a penalty on the features that do not justify the prediction. *PN* is the set of features altering the final classification and is modelled by the following loss function

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

In the experiment, we first train a ResNet20 model (He et al., 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 $x_0 + x$ is a feasible image.

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 & 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).

We pick $\lambda_1 = \lambda_2 = \frac{1}{2}$, which is the largest value from the set $\{2^{-i}|i \in \mathbb{N}\}$ allowing *FISTA* to attain a negative loss l_{x_0} for 10 randomly selected images. All algorithms start from $x_1 = 0$. Figure 4 plots the convergence behaviour 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



(a) Convergence for Generating **PP** **(b)** Convergence for Generating **PN**

Fig. 4 White box contrastive explanations on CIFAR-10

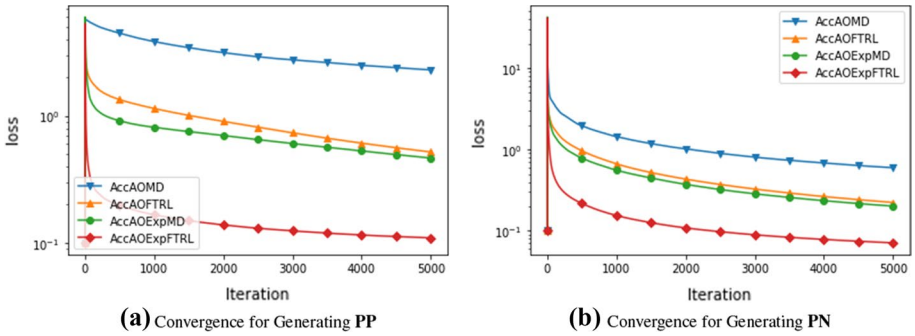


Fig. 5 Black box contrastive explanations: high variance setting

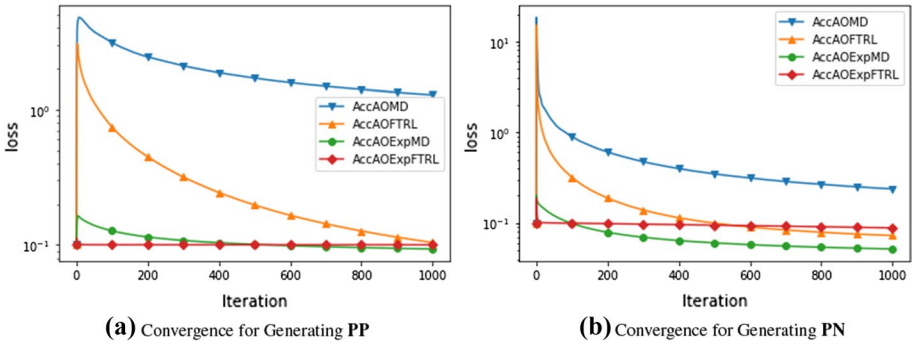


Fig. 6 Black box contrastive explanations: low variance setting

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

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

where δ, μ are constants and v_i is a random vector. Following Chen et al. (2019), we set $\delta = d$ and sample v_i independently from the uniform distribution over the unit sphere for *AdaGrad*-style algorithms. Since the convergence of our algorithms depends on the variance of the gradient estimation in $(\mathbb{R}^d, \|\cdot\|_\infty)$, we set $\delta = 1$ and sample $v_{i,1}, \ldots, v_{i,d}$ independently from Rademacher distribution according to Corollary 3 in Duchi et al. (2015). To ensure a small bias of the gradient estimation, we set $\mu = \frac{1}{\sqrt{dT}}$, which is the recommended value for non-convex and constrained optimisation in Chen et al. (2019). The performances of the algorithms are examined in the high and low variance settings with $b = 1$

and $b = \sqrt{T}$, respectively. Since the problem is stochastic, *FISTA*, which searches for the stepsize at each iteration, is not practical. Thus, we remove it from the comparison.

Figure 5 plots the convergence behaviour of the algorithms in the high variance setting. Our algorithms outperform the *AdaGrad*-style algorithms for generating both *PP* and *PN*. Furthermore, the *FTRL* based algorithms have higher convergence rates than the *MD* based ones at the first few iterations, leading to overall better performance. The experimental results of the low variance setting are plotted in Fig. 6. Though *AccAOExpFTRL* yields the smallest objective value at the beginning of the experiments, it gets stuck in the local minimum around 0 and is outperformed by *AccAOExpMD* and *AccAOFTRL* at the later iterations. Overall, the algorithms based on the exponentiated update have an advantage over the *AdaGrad*-style algorithms for both high and low variance settings.

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 dimensionality 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.

Appendix 1: Missing proofs of section 3.1

Appendix 1.1: Proof of Lemma 1

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

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

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

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

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

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

where the first inequality uses the fact $\ln x \geq 1-\frac{1}{x}$. 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

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

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

$$\phi^*(\theta) = \max_{x\in\mathbb{R}} \theta x - \phi(x), \tag{12}$$

which is differentiable. The maximiser y satisfies

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

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

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

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

$$y = \mathrm{sgn}(\theta)\left(\beta\exp\left(\frac{|\theta|}{\alpha}\right)-\beta\right). \tag{13}$$

Combining (12) and (13), 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

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

Applying the inequalities of the logarithm, we obtain

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

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

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

□

Appendix 1.2: Proof of Lemma 2

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

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

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. □

Appendix 1.3: 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 $x \in \mathcal{K}$

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

Then, we have

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

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 the 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

$$
\begin{aligned}
\mathcal{R}_{1:T} \leq& \sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_t}(x, x_t)) + \sum_{t=1}^{T}\frac{D + d\beta}{2\alpha_t}\|g_t - h_t\|_\infty^2 \\
&+ \mathcal{B}_{\phi_1}(x, x_1) + r(x_1).
\end{aligned} \tag{14}
$$

Using Lemma 8, we bound the first term of (14)

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

Using Lemma 6, the second term of (14) 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 (14) is simply 0 since we set $\alpha_1=0$. Setting $\eta=\sqrt{\frac{1}{\ln(D+1)+\ln d}}$ and combining the inequalities above, we obtain the claimed result. □

Appendix 1.4: Proof of Theorem 2

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*

$$\mathcal{R}_{1:T}\leq\phi_{T+1}(x)+\sum_{t=1}^{T}\frac{2D\|g_t-h_t\|_*^2}{\sqrt{16D^2\eta_t^2+\|g_t-h_t\|_*^2}}. \tag{15}$$

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

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

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

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

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

$$\begin{aligned}
&\sum_{t=1}^{T}(\langle g_t, x_t - x\rangle + r_t(x_t) - r_t(x)) \\
&\leq \phi_{T+1}(x) + \sum_{t=1}^{T}(\psi_t^*(\theta_{t+1}) - \psi_t^*(\theta_t - h_t) + \langle \theta_t - h_t - \theta_{t+1}, \nabla\psi_t^*(\theta_t - h_t)\rangle) \qquad (16)\\
&\leq \phi_{T+1}(x) + \sum_{t=1}^{T}\mathcal{B}_{\psi_t^*}(\theta_{t+1}, \theta_t - h_t).
\end{aligned}$$

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

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

Since ϕ_t is η_t strongly convex, we have

$$\begin{aligned}
&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \\
&\leq \frac{1}{2\eta_t}\|g_t - h_t\|_*^2 + \frac{\eta_t}{2}\|x_t - \nabla\psi_t^*(\theta_{t+1})\|^2 - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \qquad (17)\\
&\leq \frac{1}{2\eta_t}\|g_t - h_t\|_*^2
\end{aligned}$$

We also have

$$\begin{aligned}
&\langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle - \mathcal{B}_{\psi_t}(\nabla\psi_t^*(\theta_{t+1}), x_t) \\
&\leq \langle g_t - h_t, x_t - \nabla\psi_t^*(\theta_{t+1})\rangle \qquad (18)\\
&\leq 2D\|g_t - h_t\|_*.
\end{aligned}$$

Putting (17) and (18) together, we have

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

Combining the inequalities above, we obtain

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

□

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

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

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

Appendix 2: Missing Proofs of section 3.2

Appendix 2.1: 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 $\tilde{F} : \mathbb{S}^d \to \mathbb{S}^d, X \mapsto U\text{diag}(f(\lambda_1(X)), \dots, f(\lambda_d(X)))U^\top$. Apparently, we have $F(X) = \text{Tr}\tilde{F}(X)$. 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

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

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

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

And we have

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

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 $X \mapsto \text{Tr}(X(\cdot))$ 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

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

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 Lemmas 1 and 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 $H \in \mathcal{X}$,

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

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

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

$D^2F_t^*(\Theta)$ 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

$$\begin{aligned}D^2 F_t^*(\Theta)(H,H) &= \sum_{ij} \gamma(f_t,\Theta)_{ij} \tilde{h}_{ij}^2 \\ &\leq \sum_{ij} (f_t''(\lambda_i(\Theta)) + f_t''(\lambda_j(\Theta)))\tilde{h}_{ij}^2 \\ &= 2\sum_{i=1}^{m+n} f_t''(\lambda_i(\Theta)) \sum_{j=1}^{m+m} \tilde{h}_{ij}^2 \\ &= 2\mathrm{Tr}(UHU^\top \mathrm{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta)) UHU^\top) \\ &= 2\mathrm{Tr}(H^2 \mathrm{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \\ &\leq 2 \sum_{i=1}^{2\min\{m,n\}} \sigma_i(H^2)\sigma_i(\mathrm{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta)))\end{aligned} \tag{19}$$

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 (19) can be rewritten into

$$\begin{aligned}D^2 F_t^*(\Theta)(H,H) &\leq 2 \sum_{i=1}^{2\min\{m,n\}} \sigma_i(H)^2 \sigma_i(\mathrm{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \\ &\leq 2\|H\|_\infty^2 \sum_{i=1}^{2\min\{m,n\}} \sigma_i(\mathrm{diag}(f_t''(\lambda_1(\Theta)), \ldots, f_t''(\lambda_{m+n}(\Theta))) \\ &\leq 2\|H\|_\infty^2 \sum_{i=1}^{2\min\{m,n\}} f_t''(\lambda_i(\Theta)).\end{aligned} \tag{20}$$

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

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

By the construction of $\mathcal{K}$, it is clear that $\sum_{i=1}^{2\min\{m,n\}} |\lambda_i(S(X))| \leq 2D$. Thus, (20) can be simply further upper bounded by

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

Finally, applying Lemma 4, we obtain

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

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

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

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

Appendix 2.2: 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

$$\mathcal{R}_{1:T} \le r_1(x_1) + \mathcal{B}_{\phi_1}(x, x_1) + \sum_{t=1}^{T}(\mathcal{B}_{\phi_{t+1}}(x, x_t) - \mathcal{B}_{\phi_t}(x, x_t)) + \sum_{t=1}^{T}\frac{\|g_t - h_t\|_*^2}{2\eta_{t+1}}. \tag{21}$$

Using Lemma 8, we have

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

Furthermore, from Lemma 6, it follows

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

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

Appendix 2.3: 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

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

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

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

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

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

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 inequalities 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 3: Missing Proofs of section 3.4

Appendix 3.1: 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

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

and

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

Thus, $y_i = 0$ implies $x_i^* = 0$. Furthermore $\mathrm{sgn}(x_i^*) = \mathrm{sgn}(y_i)$ 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

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & \sum_{i=1}^{d}\left((x_i+\beta)\ln\frac{x_i+\beta}{y_i+\beta} - x_i\right) \\ \text{s.t.} \quad & \sum_{i=1}^{d} x_i = D \\ & x_i \geq 0 \text{ for all } i = 1,\ldots,d. \end{aligned} \tag{22}$$

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

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

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

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

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 (22) is equivalent to

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \sum_{i=\rho}^{d} (x_{p(i)} + \beta) \ln \frac{x_{p(i)} + \beta}{y_{p(i)} + \beta} \\ \text{s.t.} \quad & \sum_{i=\rho}^{d} x_{p(i)} = D \\ & x_{p(i)} > 0 \text{ for all } i = \rho, \ldots, d. \end{aligned} \tag{23}$$

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 (23) can be further rewritten into

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

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 $x_i^*=\max\{\frac{(|y_i|+\beta)-\beta}{z},0\}\mathrm{sgn}(y_i)$ for $i=1\dots d$, which is the claimed result. □

Appendix 3.2: 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

$$
\begin{aligned}
&a_t(l(y_{t+1})-l(x))\\
&=a_t(l(y_{t+1})-l(z_t)+l(z_t)-l(x))\\
&=a_t(l(y_{t+1})-l(z_t)+\langle\nabla l(z_t),z_t-x\rangle-\mathcal{B}_l(z_t,x))\\
&=a_t(l(y_{t+1})-l(z_t)+\langle\nabla l(z_t),z_t-x_t\rangle+\langle\nabla l(z_t),x_t-x\rangle-\mathcal{B}_l(z_t,x))
\end{aligned}
\tag{24}
$$

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

$$
\begin{aligned}
& a_t(l(y_{t+1}) - l(z_t) + \langle \nabla l(z_t), z_t - x_t \rangle) \\
&= a_t(l(y_{t+1}) - l(z_t) + \frac{1-\tau_t}{\tau_t} \langle \nabla l(z_t), y_t - z_t \rangle) \\
&= a_t(l(y_{t+1}) - l(z_t) + \left(\frac{1}{\tau_t} - 1\right)(l(y_t) - l(z_t)) - \left(\frac{1}{\tau_t} - 1\right)\mathcal{B}_l(y_t, z_t)) \\
&= a_t\left(\frac{1}{\tau_t} - 1\right)(l(y_t) - l(y_{t+1})) + \frac{a_t}{\tau_t}(l(y_{t+1}) - l(z_t)) - a_{1:t-1}\mathcal{B}_l(y_t, z_t).
\end{aligned} \tag{25}
$$

Next, we have

$$
\begin{aligned}
& \sum_{t=1}^{T} a_t\left(\frac{1}{\tau_t} - 1\right)(f(y_t) - f(y_{t+1})) \\
&= \sum_{t=2}^{T} a_{1:t-1}(f(y_t) - f(y_{t+1})) \\
&= \sum_{t=1}^{T-1} a_t f(y_{t+1}) - a_{1:T-1} f(y_{T+1}) \\
&= \sum_{t=1}^{T} a_t f(y_{t+1}) - a_{1:T} f(y_{T+1}) \\
&= \sum_{t=1}^{T} a_t(f(y_{t+1}) - f(y_{T+1}))
\end{aligned} \tag{26}
$$

Combining (24), (25) and (26), 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 $z_t = \frac{\sum_{s=1}^{t} a_s x_s}{a_{1:t}}$. Furthermore it follows from the convexity of r

$$
r(y_{T+1}) = r\left(\frac{\sum_{s=1}^{T} a_s x_s}{a_{1:T}}\right) \leq \sum_{t=1}^{T} \frac{a_t r(x_t)}{a_{1:T}}.
$$

Combining the inequalities above and rearranging, we obtain

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

Furthermore, we have

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

Finally, we we obtain

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

which is the claimed result. □

Proof of Corollary 1 First of all, we have

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{1:T}] &\le c_1 + c_2\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|a_t(g_t - g_{t-1})\|_*^2}\right] \\ &\le c_1 + c_2\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2]} \\ &\le c_1 + c_2\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2|z_t]}.\end{aligned} \tag{27}$$

For all t, we have

$$\begin{aligned}\mathbb{E}[\|a_t(g_t - g_{t-1})\|_*^2|z_t] \le{}& 2a_t^2(\mathbb{E}[\|g_t - \nabla l(z_t) - g_{t-1} + \nabla l(z_{t-1})\|_*^2|z_t]) \\ &+ 2a_t^2(\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2).\end{aligned} \tag{28}$$

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

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

Since $\mathcal{K}$ is compact, there is some $L > 0$ such that $\|\nabla l(z)\|_* \le L$ for all $z \in \mathbb{X}$. Thus the second term of (28) can be bounded by

$$2a_t^2\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2 \le 8a_t^2L^2 \tag{29}$$

Combining (27), (28) and (29), we have

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

and combining with Proposition 3, we obtain

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

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

$$\begin{aligned}2a_t^2\|\nabla l(z_t) - \nabla l(z_{t-1})\|_*^2 \le{}& \frac{4Ma_t^2}{a_{1:t-1}}a_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t). \\ & 8Ma_{1:t-1}\mathcal{B}_l(z_{t-1}, z_t).\end{aligned} \tag{30}$$

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

$$2c_2\sqrt{2M\sum_{t=2}^{T}a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)-\sum_{t=2}^{T}a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)} \leq 2Mc_2^2. \tag{31}$$

Combining (27), (28) and (31), we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{R}_{1:T}]-\sum_{t=1}^{T}a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)\\
&\leq c_1+c_2\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|a_t(g_t-g_{t-1})\|_*^2|z_t]}-\sum_{t=1}^{T}a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)\\
&\leq c_1 1+c_2\sqrt{8\sum_{t=1}^{T}a_t^2(v_t^2)}\\
&\quad+c_2\sqrt{\sum_{t=1}^{T}2a_t^2\|\nabla l(z_t)-\nabla l(z_{t-1})\|_*^2}-\sum_{t=1}^{T}a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)\\
&\leq c_1 1+c_2\sqrt{8\sum_{t=1}^{T}a_t^2(v_t^2)}+c_2\sqrt{2}\|\nabla l(z_1)\|_*\\
&\quad+c_2\sqrt{\sum_{t=2}^{T}2a_t^2\|\nabla l(z_t)-\nabla l(z_{t-1})\|_*^2}-\sum_{t=2}^{T}a_{1:t-1}\mathcal{B}_l(z_{t-1},z_t)\\
&\leq c_1+c_2\sqrt{8\sum_{t=1}^{T}a_t^2(v_t^2)}+\sqrt{2}c_2L+2Mc_2^2,
\end{aligned}$$

which implies

$$\mathbb{E}[f(z_T)-f(x)]\leq\frac{c_1+c_2\sqrt{8\sum_{t=1}^{T}a_t^2v_t^2}+\sqrt{2}c_2L+2Mc_2^2}{a_{1:T}}.$$

□

Appendix 4: Technical lemmata

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

1. $$\sum_{i=1}^{n}\frac{a_i}{\sum_{k=1}^{i}a_k+1}\leq\log\left(\sum_{i=1}^{n}a_i+1\right)$$

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

2.

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

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

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 $x \in \mathcal{K}$

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

Combining the inequalities above and choosing $\beta = \frac{1}{d}$, 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

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

□

Author contributions Conceptualization: WS; Methodology: WS; Formal analysis and investigation: WS; Software: WS; Validation: WS, FS; Visualization: WS; Writing - original draft preparation: WS; Writing - review and editing: WS, FS; Funding acquisition: SA; Resources: SA; Supervision: FS, SA.

Funding Open Access funding enabled and organized by Projekt DEAL. The research leading to these results received funding from the German Federal Ministry for Economic Affairs and Climate Action under Grant Agreement No. 01MK20002C.

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 CIFAR-10 data are collected from https://www.cs.toronto.edu/~kriz/cifar.html.

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

Declarations

Conflict of interest The authors declare that they have no conflict of interest or competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

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

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

Anava, O., Hazan, E., Mannor, S., & Shamir, O. (2013). Online learning for time series prediction. In *Conference on learning theory* (pp. 172–184).

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*. Berlin: Springer.

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). Berlin: Springer.

Cancela, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2021). A delayed elastic-net approach for performing adversarial attacks. In *2020 25th International conference on pattern recognition (ICPR)* (pp. 378–384). https://doi.org/10.1109/ICPR48806.2021.9413170.

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

Cesa-Bianchi, N., Conconi, A., & Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory, 50*(9), 2050–2057.

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. In *Thirty-second AAAI conference on artificial intelligence*.

Chen, X., Liu, S., Xu, K., Li, X., Lin, X., Hong, M., & Cox, D. (2019). Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Berlin: Curran Associates, Inc.

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

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

Cutkosky, A., & Boahen, K. A. (2016). Online convex optimization with unconstrained domains and losses. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Berlin: Curran Associates, Inc.

Cutkosky, A., & Boahen, K. A. (2017b). Stochastic and adversarial online learning without hyperparameters. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Berlin: Curran Associates, Inc.

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. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems.* (Vol. 31). Berlin: 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*, 2121–2159.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., & Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory, 61*(5), 2788–2806.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., & Tewari, A. (2010). Composite objective mirror descent. In A. T. Kalai & M. Mohri (Eds.), *COLT 2010—The 23rd conference on learning theory, Haifa, Israel, June 27–29, 2010* (pp. 14–26). Omnipress.

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. In *Algorithmic learning theory* (pp. 386–407).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770-778). https://doi.org/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. In *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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 6260–6269). Berlin: Curran Associates Inc.

Kempka, M., Kotlowski, W., & Warmuth, M. K. (2019). Adaptive scale-invariant online algorithms for learning linear models. In *International conference on machine learning* (pp. 3321–3330).

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. Master's thesis, University of Tront.

Lan, G. (2020). *First-order and stochastic optimization methods for machine learning*. Berlin: Springer.

Levy, Y. K., Yurtsever, A., & Cevher, V. (2018). Online adaptive methods, universality and acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 6500–6509). Berlin: Curran Associates Inc.

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. In *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.

McMahan, H. B., & Streeter, M. J. (2010). Adaptive bound optimization for online convex optimization. In A. T. Kalai & M. Mohri (Eds.), *COLT 2010—The 23rd conference on learning theory, Haifa, Israel, June 27–29, 2010* (pp. 244–256). Omnipress.

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

Orabona, F. (2013). Dimension-free exponentiated gradient. In *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. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Song, L., Tekin, C., & Van Der Schaar, M. (2014). Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing, 9*(3), 433–445.

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

Warmuth, M. K. (2007). Winnowing subspaces. In *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.