Analysis of Online Composite Mirror Descent Algorithm

Yunwen Lei[1] yunwen.lei@hotmail.com
Ding-Xuan Zhou[1] mazhou@cityu.edu.hk
[1]Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

Keywords: Composite Mirror Descent, Learning theory, Online learning, Error Analysis

Abstract

We study the convergence of the online composite mirror descent algorithm which involves a mirror map to reflect the geometry of the data and a convex objective function consisting of a loss and a regularizer possibly inducing sparsity. Our error analysis provides convergence rates in terms of properties of the strongly convex differentiable mirror map and the objective function. For a class of objective functions with Hölder continuous gradients, the convergence rates of the excess (regularized) risk under polynomially decaying step sizes have the order $O(T^{-\frac{1}{2}}\log T)$ after T iterates. Our results improve the existing error analysis for the online composite mirror descent algorithm by avoiding averaging and removing boundedness assumptions, and sharpen the existing convergence rates of the last iterate for online gradient descent without any boundedness assumptions. Our methodology mainly depends on a novel error decomposition in terms of an excess Bregman distance, refined analysis of self-bounding properties of the objective function, and the resulted one-step progress bounds.

1 Introduction

Gradient descent is a classical powerful method for optimization and numerical computation. To approximate a minimizer of a convex function f on the Euclidean space $\mathbb{R}^d$, it defines a sequence $\{w_t\}_{t\in\mathbb{N}}$ of points iteratively by $w_{t+1} = w_t - \eta_t f'(w_t)$, where $f'(w_t)$ is a subgradient of f at w_t and η_t is a step size. Gradient descent is even more powerful in the era of big data and has been extended along different directions in various ways. **Mirror descent** is such an extension by relaxing the Hilbert space structure (Nemirovsky & Yudin, 1983; Beck & Teboulle, 2003) and allowing a Banach space norm on $\mathbb{R}^d$ such as the ℓ_p-norm with $1 \leq p \leq 2$, where $f'(w_t)$ is used for performing the gradient descent in the dual of the primal space $(\mathbb{R}^d, \|\cdot\|_p)$.

As a first-order optimization procedure, mirror descent provides an efficient way to solve large-scale optimization problems in a Banach space $(\mathcal{W}, \|\cdot\|)$ by introducing a sequence of *primal-dual* variables $\{(w_t, v_t)\}_{t=1}^{\infty}$ in the primal-dual space to replace the sequence $\{w_t\}$ in the gradient descent algorithm, and it is induced by a *mirror map* $\Psi : \mathcal{W} \to \mathbb{R}$. We assume Ψ to be Fréchet differentiable meaning that at every $w \in \mathcal{W}$, there exists a bounded linear operator $A_w : \mathcal{W} \to \mathbb{R}$ such that $\lim_{\tilde{w}\to 0} \frac{|\Psi(w+\tilde{w})-\Psi(w)-A_w\tilde{w}|}{\|\tilde{w}\|} = 0$. Denote the operator A_w as the gradient $\nabla\Psi(w)$ of Ψ at $w \in \mathcal{W}$ which lies in the dual space $(\mathcal{W}^*, \|\cdot\|_*)$. So $\nabla\Psi : \mathcal{W} \to \mathcal{W}^*$ is a map from the primal space $\mathcal{W}$ to the dual space $\mathcal{W}^*$ and is used to express the relationship $v_t = \nabla\Psi(w_t)$ for the primal-dual pair (w_t, v_t). We also assume that Ψ is σ-strongly convex with respect to $\|\cdot\|$ for some $\sigma > 0$ meaning that

$$D_\Psi(w, \tilde{w}) := \Psi(w) - \Psi(\tilde{w}) - \langle w - \tilde{w}, \nabla\Psi(\tilde{w})\rangle \geq \frac{\sigma}{2}\|w - \tilde{w}\|^2, \quad \forall w, \tilde{w} \in \mathcal{W},$$

where $\langle w-\tilde{w}, \nabla\Psi(\tilde{w})\rangle$ is the dual element $\nabla\Psi(\tilde{w}) \in \mathcal{W}^*$ acting on the element $w-\tilde{w} \in \mathcal{W}$. We call $D_\Psi(w, \tilde{w})$ the Bregman distance between w and $\tilde{w}$. Then the mirror descent algorithm applied to $\min_{w\in\mathcal{D}} f(w)$ with a convex function f, an initial point $w_1 \in \mathcal{D}$ and a convex set $\mathcal{D} \subset \mathcal{W}$ produces a sequence $\{w_t\}_{t=1}^{\infty}$ of points iteratively as

$$\begin{cases} \nabla\Psi(w_{t+\frac{1}{2}}) := \nabla\Psi(w_t) - \eta_t f'(w_t), \\ w_{t+1} = \arg\min_{w\in\mathcal{D}} D_\Psi(w, w_{t+\frac{1}{2}}), \qquad t \in \mathbb{N}, \end{cases} \tag{1.1}$$

where $\{\eta_t\}_{t=1}^{\infty}$ is a sequence of step sizes. The strong convexity of Ψ implies the invertibility of the map $\nabla\Psi : \mathcal{W} \to \mathcal{W}^*$ making the point $w_{t+\frac{1}{2}} \in \mathcal{W}$ and the Bregman distance $D_\Psi(w, \tilde{w})$ well defined. An important property of the mirror descent algorithm rests on its flexibility in choosing a mirror map to capture the geometry of the problem at hand, which is appealing to solve problems of high dimensions. Below we provide a class of specific mirror maps to illustrate their influence on the behavior of the algorithm.

Example 1. Let $1 < p \leq 2$ and $(\mathcal{W}, \|\cdot\|) = (\mathbb{R}^d, \|\cdot\|_p)$ with the ℓ_p-norm $\|\cdot\| = \|\cdot\|_p$ defined by $\|w\|_p = [\sum_{i=1}^d |w(i)|^p]^{\frac{1}{p}}$ for $w = (w(i))_{i=1}^d \in \mathcal{W}$. Then its dual space is $(\mathbb{R}^d, \|\cdot\|_{\frac{p}{p-1}})$. Take the p-norm divergence $\Psi_p(x) = \frac{1}{2}\|x\|_p^2$ as the mirror map. This mirror map, as shown in (Ball et al., 1994), is $(p-1)$-strongly convex over $\mathcal{W}$ with respect to the norm $\|\cdot\|_p$. Take $\mathcal{D} = \mathcal{W}$. When $p = 2$, the primal and dual spaces coincide and the mirror descent reduces to the gradient descent. When a minimizer $w^* = \arg\min_{w\in\mathcal{W}} f(w)$ of a convex function f is sparse, the mirror descent method with the mirror map Ψ_p and the specific choice of $p = 1 + \frac{1}{\log d}$ yields a convergence bound with a logarithmic dependence on d as proved in (Duchi et al., 2010).

In many machine learning problems, the objective function f often takes a composite form: $f(w) = \ell(w) + r(w)$ with a data fitting convex (loss) function $\ell(w)$ and a regularization term $r(w)$, which arises naturally in regularization schemes. For these composite optimization problems, the mirror descent directly applied to f, involving subgradients of r, would destroy some desirable effects suggested by the regularizer r (Duchi & Singer, 2009), such as the ℓ_1-norm for promoting sparsity. Instead, a variant of mirror descent called the **composite mirror descent** was introduced in (Lions & Mercier, 1979; Duchi et al., 2010). At the t-th iteration, composite mirror descent updates w_{t+1} by approximating f with, instead of its first-order approximation at w_t used in the mirror descent scheme, the first-order approximation of $\ell(w)$ at w_t plus $r(w)$

$$w_{t+1} = \arg\min_{w\in\mathcal{D}} \eta_t\langle w - w_t, \ell'(w_t)\rangle + \eta_t r(w) + D_\Psi(w, w_t), \quad t \in \mathbb{N}. \tag{1.2}$$

When the term $r(w)$ vanishes, the above composite mirror descent method coincides with the mirror descent method (1.1), which can be seen from a reformulation of (1.2) in terms of two steps similar to (1.1) (Duchi et al., 2010). Another motivation to keep $r(w)$ intact in (1.2) is that the first-order approximation of $r(w)$ would slow down the convergence rate since $r(w)$ can be non-smooth while $\ell(w)$ can be smooth. If we take the specific mirror map $\Psi = \Psi_2$ and $\mathcal{W} = \mathcal{D} = \mathbb{R}^d$, the composite mirror descent recovers the proximal gradient method or *forward-backward splitting* $w_{t+1} = \text{Prox}_{\eta_t r}(w_t - \eta_t\ell'(w_t))$ dated back to (Lions & Mercier, 1979; Duchi & Singer, 2009), where $\text{Prox}_r(w) = \arg\min_{\tilde{w}}[r(\tilde{w}) + \frac{1}{2}\|w - \tilde{w}\|_2^2]$ is the proximal operator. A typical choice of $\ell(w)$ in machine learning is $\ell(w) = \frac{1}{T}\sum_{t=1}^T \phi(y_t, \langle w, x_t\rangle)$, where $\{(x_t, y_t)\}_{t=1}^T$ is a training sample and $\phi(y, \langle w, x\rangle)$ is a loss function used to measure the performance of the linear model $x \to \langle w, x\rangle$ on the example (x, y). When the sample size T is large, composite mirror descent in online and stochastic settings is studied in Duchi et al. (2010), where the fixed objective function $f(w) = \ell(w) + r(w)$ is replaced by a sequence $f_t(w) = \ell_t(w) + r(w)$ with $\ell_t(w)$ being either an instantaneous loss in the online setting or a stochastic estimate of the objective function in the stochastic setting.

In this paper, we study the **online composite mirror descent** algorithm with the aim of error analysis. Throughout the paper, the primal space is $\mathcal{W} = \mathbb{R}^d$ with the norm $\|\cdot\|$, the dual space is $\mathcal{W}^* = \mathbb{R}^d$ with the dual norm $\|\cdot\|_*$, and $\langle w, x\rangle$ denotes the action of the dual element $x \in \mathcal{W}^*$ on $w \in \mathcal{W}$. Take $\mathcal{D} = \mathcal{W}$ to be $\mathbb{R}^d$. We assume a sequence of examples $(x_t, y_t), t \in \mathbb{N}$, to be independently drawn from a Borel probability measure ρ defined over $\mathcal{X} \times \mathcal{Y} \subset \mathcal{W}^* \times \mathbb{R}$. We assume that $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_0^+$ is convex in the second argument, and $r : \mathcal{W} \to \mathbb{R}_0^+$ is convex. Then the online composite mirror descent updates the sequence $\{w_t\}_{t\in\mathbb{N}}$ with $w_1 = 0$ by

$$w_{t+1} = \arg\min_{w\in\mathcal{D}} \eta_t\langle w - w_t, \phi'_-(y_t, \langle w_t, x_t\rangle)x_t\rangle + \eta_t r(w) + D_\Psi(w, w_t), \quad t \in \mathbb{N}, \tag{1.3}$$

where $\phi'_-(y, \cdot)$ denotes the left-side derivative of ϕ with respect to the second argument. This strategy of processing each observation per iteration enjoys a great computational advantage when compared to the composite mirror descent (1.2). For example, for the typical choice $\ell(w) = \frac{1}{T}\sum_{t=1}^T \phi(y_t, \langle w, x_t\rangle)$ in a machine learning setting, evaluating one single gradient in (1.2) requires going through the whole data set. This gradient evaluation becomes prohibitively expensive in the big data era when faced with large amounts of data (Bach & Moulines, 2013). Below we list some examples covered in the framework of online composite mirror descent (1.3).

Example 2. If we take $\Psi = \Psi_2$ and $r(w) = 0$ in (1.3), the online composite mirror descent recovers the online gradient descent learning with the linear kernel

$$w_{t+1} = w_t - \eta_t \phi'_-(y_t, \langle w_t, x_t \rangle) x_t.$$

For the least squares loss $\phi(y,a) = \frac{1}{2}(y-a)^2$, it further translates to the Kaczmarz algorithm.

Example 3. If we take $\Psi = \Psi_2$, the online composite mirror descent (1.3) recovers the online proximal gradient descent algorithm

$$w_{t+1} = \text{Prox}_{\eta_t r}\big(w_t - \eta_t \phi'_-(y_t, \langle w_t, x_t \rangle) x_t\big).$$

Example 4. If we take $\Psi = \Psi_p, 1 < p \leq 2$ and $r(w) = \lambda \|w\|_1$, the online composite mirror descent recovers the *stochastic mirror descent algorithm made sparse* (SMIDAS) proposed in (Shalev-Shwartz & Tewari, 2011)

$$\nabla \Psi_p(w_{t+\frac{1}{2}}) = \nabla \Psi_p(w_t) - \eta_t \phi'_-(y_t, \langle w_t, x_t \rangle) x_t, \quad \nabla \Psi_p(w_{t+1}) = \text{Prox}_{\eta_t \lambda \|\cdot\|_1}\big(\nabla \Psi_p(w_{t+\frac{1}{2}})\big).$$

Actually (Duchi et al., 2010), the iterate w_{t+1} defined above is equivalent to $w_{t+1} = \arg\min_{w \in \mathcal{D}} D_{\Psi_p}(w, w_{t+\frac{1}{2}}) + \eta_t \lambda \|w\|_1$. Different realizations of online composite mirror descent with the sparsity-inducing regularizer $r(w) = \lambda \|w\|_1$ have also been proposed and theoretically studied in (Langford et al., 2009) and (Shalev-Shwartz & Tewari, 2011).

Our error analysis is carried out in terms of the generalization error (risk) of the linear function $x \to \langle w, x \rangle$ associated with the vector $w \in \mathcal{W}$ defined by

$$\mathcal{E}^\phi(w) = \int_{\mathcal{X} \times \mathcal{Y}} \phi(y, \langle w, x \rangle) d\rho.$$

We estimate the excess risk $\mathbb{E}[\mathcal{E}^\phi(w_T)] - \mathcal{E}^\phi(w^*)$ for the last iterate w_T produced by (1.3), where $w^* \in \mathcal{W}$ is a vector attaining the minimal risk

$$w^* = \arg\min_{w \in \mathcal{W}} \mathcal{E}^\phi(w).$$

The algorithm and our analysis include three main ingredients: the loss function ϕ, the regularizer r, and the mirror map Ψ. Our results are stated in terms of properties of the loss functions ϕ and the regularizer r in addition to the strong convexity of the differentiable mirror map Ψ and the boundedness of the probability measure ρ. To illustrate our ideas, we state learning rates, to be proved in Section 4, for the case when $r(w) = \lambda \|w\|_1$ is the (scaled) 1-norm.

Assumption 1. We assume that the input data are uniformly bounded in the sense $R := \sup_{x\in\mathcal{X}} \|x\|_* < \infty$ and $|\phi|_0 := \sup_{y\in\mathcal{Y}} \phi(y,0) < \infty, |\phi|'_0 := \sup_{y\in\mathcal{Y}} |\phi'_-(y,0)| < \infty$.

The involved properties of ϕ is measured by the Hölder continuity of ϕ'_-.

Assumption 2. We assume that the loss function ϕ is convex in the second argument and its (sub)gradient is q-Hölder continuous for some $0 \le q \le 1$, meaning that there exists a constant $L_q \ge 0$ such that

$$|\phi'_-(y,a) - \phi'_-(y,\tilde{a})| \le L_q |a-\tilde{a}|^q, \qquad \forall a,\tilde{a}\in\mathbb{R}, y\in\mathcal{Y}. \tag{1.4}$$

Example 5. In the two extreme cases $q=0$ and $q=1$, convex loss functions satisfying the condition (1.4) include the hinge loss $\phi(y,a)=\max(0,1-ya)$ with $\mathcal{Y}=\{1,-1\}$ for classification with $q=0$, the least square loss $\phi(y,a)=\frac{1}{2}(y-a)^2$ and the logistic function $\phi(y,a)=\log(1+\exp(-ya))$ with $q=1$. The intermediate case include $\tilde{q}$-norm hinge loss $\phi(y,a)=\max(0,1-ya)^{\tilde{q}}$ for classification (Chen et al., 2004) and the $\tilde{q}$-th power absolute distance loss $\phi(y,a)=|y-a|^{\tilde{q}}$ for regression (Steinwart & Christmann, 2008) with $\tilde{q}\in(1,2]$ and $q=\tilde{q}-1$.

Denote $\mathbf{1}$ the vector in $\mathbb{R}^d$ with all components being 1. A norm Ω on $\mathbb{R}^d$ is said to be monotonic if $\Omega(x)\le\Omega(\tilde{x})$ whenever $x,\tilde{x}\in\mathbb{R}^d$ satisfy $|x(i)|\le|\tilde{x}(i)|$ for $i=1,\ldots,d$.

Theorem 1. *Assume that the mirror map Ψ is differentiable and σ-strongly convex for some $\sigma>0$ and the norm $\|\cdot\|_*$ is monotonic. Suppose that Assumption 1 and Assumption 2 hold. Consider the regularizer $r(w)=\lambda\|w\|_1$ with $0\le\lambda\le\lambda_0$ for some $\lambda_0>0$. If the step size is $\eta_t=\eta_1 t^{-\frac{1}{2}}$ with*

$$0<\eta_1\le\begin{cases}4^{-1}\sigma\min\big((R|\phi|'_0)^{-2},(RL_q)^{-2},(\lambda\|\mathbf{1}\|_*)^{-2}\big), & \text{if } q=0,\\ 4^{-1}\sigma\min\big(R^{-2}L_q^{-\frac{2}{q+1}},(\lambda\|\mathbf{1}\|_*)^{-2}\big), & \text{if } 0<q\le 1,\end{cases} \tag{1.5}$$

then we have

$$\mathbb{E}[\mathcal{E}^\phi(w_T)-\mathcal{E}^\phi(w^*)]\le cT^{-\frac{1}{2}}\log(eT)+\|w^*\|_1\lambda,$$

where c is a constant independent of T or λ, and the expectation $\mathbb{E}$ is taken with respect to the sample $\{(x_t,y_t)\}_{t=1}^T$.

2 Main Results

This section presents our main results on error analysis of the online composite mirror descent algorithm (1.3) given in terms of the following properties of the regularizer $r:\mathcal{W}\to\mathbb{R}_0^+$ in addition to those of the loss function ϕ.

Assumption 3. We assume that the convex regularizer $r:\mathcal{W}\to\mathbb{R}_0^+$ satisfies $r(0)=0$ and its (sub)gradient r' is p-Hölder continuous for some $0\le p\le 1$, meaning that there exists a constant $L_p\ge 0$ such that

$$\|r'(w)-r'(\tilde{w})\|_*\le L_p\|w-\tilde{w}\|^p, \quad \forall w\ne\tilde{w}\in\mathcal{W}. \tag{2.1}$$

Example 6. The (scaled) regularizer $\lambda\|w\|_{\tilde{p}}^{\tilde{p}}$ with $\lambda \geq 0$ and $1 \leq \tilde{p} \leq 2$ satisfies the condition (2.1) with $p = \tilde{p} - 1$ and norm $\|\cdot\| = \|\cdot\|_{\tilde{p}}$ (see Lemma C.3 in Appendix C). In particular, the classical ℓ_1-regularizer with $\tilde{p} = 1$ satisfies (2.1) with $p = 0$, $\|\cdot\| = \|\cdot\|_1$ and $L_p = 2\lambda$.

Now we can state our main results, to be proved in Section 4, on convergence rates of the excess regularized risk $\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)]$ for the last iterate of (1.3), where $\mathcal{E}^{\phi,r}(w)$ denotes the regularized risk of the linear function associated to w with the minimizer w_r^* defined by

$$\mathcal{E}^{\phi,r}(w) = \int_{\mathcal{X}\times\mathcal{Y}} \phi(y, \langle w, x\rangle) d\rho + r(w), \quad w_r^* = \arg\min_{w\in\mathcal{W}} \mathcal{E}^{\phi,r}(w).$$

Theorem 2. *Assume that the mirror map $\Psi : \mathcal{W} \to \mathbb{R}$ is differentiable and σ-strongly convex for some $\sigma > 0$. Suppose that Assumptions 1, 2, 3 hold with $0 \leq p, q \leq 1$. Consider the step size $\eta_t = \eta_1 t^{-\theta}$ with $\frac{\max(p,q)}{1+\max(p,q)} \leq \theta < 1$ and*

$$\eta_t \leq \begin{cases} 4^{-1}\sigma \min\left(R^{-2} L_q^{-\frac{2}{q+1}}, L_p^{-\frac{2}{p+1}}\right), & \text{if } q > 0, \\ 4^{-1}\sigma \min\left(R^{-2}[\max(|\phi|_0', L_q)]^{-2}, L_p^{-\frac{2}{p+1}}\right), & \text{if } q = 0. \end{cases} \tag{2.2}$$

Then we have

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \leq c \max\left(T^{-\theta}\log(eT), T^{\theta-1}\right), \tag{2.3}$$

where c is a constant depending on $\eta_1, \mathcal{E}^{\phi,r}(w_r^*), D_\Psi(w_r^*, 0), \sigma^{-1}, L_p, L_q, p, q, \theta, R, |\phi|_0, |\phi|_0'$ (explicitly given in the proof). Specifically, if $\theta = \frac{1}{2}$ we get

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] = O\left(T^{-\frac{1}{2}}\log T\right). \tag{2.4}$$

The existing research work on the online (stochastic) composite mirror descent algorithm (1.3) gives bounds on the regularized regret defined by

$$R(T, w) = \sum_{t=1}^{T} \left[\phi(y_t, \langle w_t, x_t\rangle) + r(w_t) - \phi(y_t, \langle w, x_t\rangle) - r(w)\right]$$

or the closely related excess regularized risk for the average of the iterates $\bar{w}_T := \frac{1}{T}\sum_{t=1}^{T} w_t$ (Cesa-Bianchi et al., 2004; Duchi et al., 2010; Langford et al., 2009; Shalev-Shwartz & Tewari, 2011; Duchi & Singer, 2009; Srebro et al., 2011). However, as shown in (Rosasco et al., 2014; Shamir & Zhang, 2013; Rakhlin et al., 2012), averaging can have a detrimental effect in the sense that it can slow down the convergence rates when the objective function is strongly convex, or destroy the sparsity of the solution which is often crucial for proper interpretations in many applications. Instead of studying regret bounds or the associated convergence rates for the average of iterates, we consider here the more challenging problem of the convergence of the last iterate, which would imply the convergence of the averaging scheme and would not destroy the sparsity. Our main results show that the excess regularized risk enjoys the convergence rate $O(T^{-\frac{1}{2}}\log T)$ with the step size $\eta_t = \eta_1 t^{-\frac{1}{2}}$, matching (up to a logarithmic factor) the minimax rates of order $O(T^{-\frac{1}{2}})$ for stochastic approximation in the non-strongly convex case (Agarwal et al., 2012). These results are established for a general class of objective functions with Hölder continuous (sub)gradients including Lipschitz objective functions and smooth objective functions.

Furthermore, our analysis does not need any boundedness assumption on $\|w_t\|$ or $\mathbb{E}[\|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2]$ as imposed in the literature (Duchi et al., 2010; Shamir & Zhang, 2013). For example, stochastic projected gradient descent is studied in (Shamir & Zhang, 2013) for non-smooth optimization which gives the convergence rate $O(T^{-\frac{1}{2}}\log T)$ for the last iterate. But their discussion requires the assumption of the existence of a constant G such that $\|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_2 \leq G$ for all $t \in \mathbb{N}$ and $\sup_{w,\tilde{w}\in\mathcal{D}} \|w - \tilde{w}\|_2 \leq G$ for points on the projected domain $\mathcal{D}$, which only holds when $\mathcal{D}$ is compact and thereby their algorithm requires an additional projection onto $\mathcal{D}$ per iteration. More recently, convergence of the last iterate for stochastic proximal gradient algorithms with $\Psi = \Psi_2$ is studied in (Rosasco et al., 2014) presenting a non-asymptotic bound in expectation in the strongly convex case and the almost sure convergence in the general case, but their discussion still needs the assumption of the existence of a sequence $\{\alpha_t\}_{t\in\mathbb{N}}$ and a constant $\beta > 0$ satisfying $\mathbb{E}[\|\ell'_t(w_t) - \ell'(w_t)\|^2] \leq \beta^2(1 + \alpha_t\|\ell'(w_t)\|^2)$ for all $t \in \mathbb{N}$, where $\ell_t(w) = \phi(y_t, \langle w, x_t\rangle)$ and $\ell(w) = \mathbb{E}[\phi(y, \langle w, x\rangle)]$.

In deriving the almost optimal convergence rates, we also get the following convergence rate for $\mathbb{E}[\|\nabla\mathcal{E}^{\phi,r}(w_T)\|_*]$, to be proved in Appendix C.

Corollary 3. *Under the conditions of Theorem 2 with $0 < p, q \leq 1$, with the step size $\eta_t = \eta_1 t^{-\frac{1}{2}}$ satisfying (2.2), we have*

$$\mathbb{E}[\|\nabla\mathcal{E}^{\phi,r}(w_T)\|_*] = O\big((T^{-\frac{1}{2}}\log T)^{\min(\frac{p}{p+1},\frac{q}{q+1})}\big).$$

To demonstrate our main results stated in Theorem 2, we present explicit learning rates for some special cases in the following subsections. It would be interesting to extend our results to non-convex loss functions including those from the minimum error entropy principle (Hu et al., 2015).

2.1 Online gradient descent learning

The first special case corresponds to $\Psi = \Psi_2$ and $r(w) = 0$. In this case, the online composite mirror descent algorithm (1.3) recovers the unregularized online gradient descent algorithms for regression and classification by selecting concrete loss functions such as the $\tilde{q}$-norm hinge loss $\phi(y, a) = \max(0, 1 - ya)^{\tilde{q}}$, the logistic function $\phi(y, a) = \log(1 + \exp(-ya))$ and the $\tilde{q}$-th power absolute distance loss $\phi(y, a) = |y - a|^{\tilde{q}}$.

Convergence for the last iterate has been extensively studied for the online gradient descent algorithm in reproducing kernel Hilbert spaces in (Smale & Yao, 2006; Ying & Zhou, 2006; Smale & Zhou, 2009; Tarres & Yao, 2014; Ying & Zhou, 2016) where the regularizer is approximated by its first-order approximation when updating $\{w_t\}$. The unregularized least squares online gradient descent algorithm in reproducing kernel Hilbert spaces is studied in (Ying & Pontil, 2008) and convergence rates of order $O(T^{-\frac{1}{2}}\log T)$ are derived. For a class of loss functions with q-Hölder continuous gradients with $0 < q \leq 1$, the unregularized online gradient descent learning with $r(w) = 0, \Psi = \Psi_2$ is considered in (Ying & Zhou, 2015) which establishes the convergence rate

$$\mathbb{E}[\mathcal{E}^{\phi}(w_T) - \mathcal{E}^{\phi}(w^*)] \leq O(T^{-\min(2^{-1}q\theta,1-\theta)})$$

with the step size $\eta_t = \eta_1 t^{-\theta}$, which would be $O(T^{-\frac{q}{q+2}})$ by taking $\theta = 2/(q+2)$. This convergence rate can at most attain $O(T^{-\frac{1}{3}})$ when the loss function is smooth. Theorem 2 immediately implies the following convergence rate $O(T^{-\frac{1}{2}}\log T)$ of the excess risk $\mathbb{E}[\mathcal{E}^{\phi}(w_T)] - \mathcal{E}^{\phi}(w^*)$ for unregularized online gradient descent algorithms. It is a great improvement and thereby solves the open question whether the rate $O(T^{-\frac{1}{3}})$ without the boundedness assumption can be improved for the unregularized online gradient descent algorithm applied to general loss functions (Ying & Zhou, 2015).

Corollary 4. *Consider the mirror map $\Psi = \Psi_2$ and $r(w) = 0$. Suppose Assumptions 1 and 2 hold. Take $\|\cdot\| = \|\cdot\|_2$ and $\sigma = 1$. For the step size $\eta_t = \eta_1 t^{-\frac{1}{2}}$ satisfying* (2.2) *with $L_p = 0$, we have $\mathbb{E}[\mathcal{E}^{\phi}(w_T) - \mathcal{E}^{\phi}(w^*)] = O(T^{-\frac{1}{2}}\log T)$.*

2.2 Online learning with sparsity-inducing regularizer

The second special case is given by $\Psi = \Psi_{\tilde{p}}$ with $1 < \tilde{p} \leq 2$ and $r(w) = \lambda\|w\|_1$. In this case, the online composite mirror descent algorithm (1.3) recovers the SMIDAS proposed in (Shalev-Shwartz & Tewari, 2011), whose convergence follows as a direct corollary of Theorem 2 by noting the identity $D_{\Psi_{\tilde{p}}}(w_r^*, 0) = \frac{1}{2}\|w_r^*\|_{\tilde{p}}^2$ and the $(\tilde{p}-1)$-strong convexity of $\Psi_{\tilde{p}}$ w.r.t. $\|\cdot\|_{\tilde{p}}$. Note that the dual norm of $\|\cdot\|_{\tilde{p}}$ is $\|\cdot\|_{\frac{\tilde{p}}{\tilde{p}-1}}$.

Corollary 5. *Consider the mirror map $\Psi_{\tilde{p}}$ with $1 < \tilde{p} \leq 2$ and $r(w) = \lambda\|w\|_1$ with $\lambda \geq 0$. Suppose Assumptions 1 and 2 hold. Take $\|\cdot\| = \|\cdot\|_{\tilde{p}}$ and $\sigma = \tilde{p} - 1$. Then for the step size $\eta_t = \eta_1 t^{-\frac{1}{2}}$ satisfying* (1.5)*, we have*

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] = O\big(T^{-\frac{1}{2}}\log T\big). \tag{2.5}$$

Remark 1. Consider the case $0 \leq q < 1$. If we choose $\Psi = \Psi_{\tilde{p}}, r(w) = \lambda\|w\|_1$ and

$$\eta_1 = c'(\tilde{p}-1)\min\big\{\big[\sup_{x\in\mathcal{X}}\|x\|_{\frac{\tilde{p}}{\tilde{p}-1}}\big]^{-2}, \lambda^{-2}\|\mathbf{1}\|_{\frac{\tilde{p}}{\tilde{p}-1}}^{-2}\big\}$$

with c' depending only on L_q and $|\phi|_0'$, the constant c hidden in the big O notation in (2.5) takes the form

$$c = \bar{c}\Big[(\tilde{p}-1)^{-1}\max\big\{\sup_{x\in\mathcal{X}}\|x\|^2_{\frac{\tilde{p}}{\tilde{p}-1}}, \lambda^2\|\mathbf{1}\|^2_{\frac{\tilde{p}}{\tilde{p}-1}}\big\}\|w_r^*\|_{\tilde{p}}^2 + \mathcal{E}^{\phi,r}(w_r^*)\Big], \tag{2.6}$$

where $\bar{c}$ is a constant depending only on q, L_q, $|\phi|_0$ and $|\phi|_0'$. For the choice $\tilde{p} = 1+\frac{1}{\log d}$, the constant c defined by (2.6) satisfies (note that $\|x\|_{1+\log d} \leq e\|x\|_\infty$ for $x \in \mathcal{W}^*$)

$$c \leq \bar{c}e^2\Big[\max\big\{\sup_{x\in\mathcal{X}}\|x\|^2_\infty, \lambda^2\|\mathbf{1}\|^2_\infty\big\}\|w_r^*\|^2_{\frac{1+\log d}{\log d}}\log d + \mathcal{E}^{\phi,r}(w_r^*)\Big].$$

For the choice $\tilde{p} = 2$, the constant c in (2.6) translates to

$$c = \bar{c}\Big[\max\big\{\sup_{x\in\mathcal{X}}\|x\|_2^2, \lambda^2\|\mathbf{1}\|_2^2\big\}\|w_r^*\|_2^2 + \mathcal{E}^{\phi,r}(w_r^*)\Big].$$

Therefore, for learning problems where the features are dense (i.e., $\|x\|_2$ closed to $d^{\frac{1}{2}}\|x\|_\infty$) and w_r^* is very spare (i.e., $\|w_r^*\|_1$ closed to $\|w_r^*\|_2$), the online composite mirror descent with $\Psi = \Psi_{1+\frac{1}{\log d}}$ would enjoy a faster convergence rate compared to that for $\Psi = \Psi_2$, especially in high dimensional problems (Duchi et al., 2010).

2.3 Online smoothed linearized Bregman iteration

The last special case corresponds to the least squares loss $\phi(y,a) = \frac{1}{2}(y-a)^2$, $r(w) = 0$, and the mirror map Ψ_ϵ, with a parameter $\epsilon > 0$, defined by $\Psi_\epsilon(w) = \lambda J_\epsilon(w) + \frac{1}{2}\|w\|_2^2$, where J_ϵ is a componentwise regularizer for robustness smoothing the 1-norm given by

$$J_\epsilon(w) = \sum_{i=1}^d F_\epsilon(w(i)) \quad \text{and} \quad F_\epsilon(\xi) = \begin{cases} \frac{\xi^2}{2\epsilon}, & \text{if } |\xi| \leq \epsilon, \\ |\xi| - \frac{\epsilon}{2}, & \text{otherwise.}\end{cases}$$

In this case, with $w_1 = v_1 = 0$, the online composite mirror descent algorithm (1.3) can be reformulated as

$$\begin{cases} v_{t+1} = v_t - \eta_t(\langle w_t, x_t\rangle - y_t)x_t = \nabla\Psi_\epsilon(w_{t+1}), \\ w_{t+1} = \arg\min_{w\in\mathcal{D}}\eta_t\big\langle w - w_t, (\langle w_t, x_t\rangle - y_t)x_t\big\rangle + \lambda J_\epsilon(w) + \dfrac{1}{2}\|w\|_2^2 - \big\langle w - w_t, v_t\big\rangle \\ \qquad\quad = \arg\min_{w\in\mathcal{D}}\lambda J_\epsilon(w) + \dfrac{1}{2}\|w\|_2^2 - \langle w, v_{t+1}\rangle = \arg\min_{w\in\mathcal{D}}\lambda J_\epsilon(w) + \dfrac{1}{2}\|w - v_{t+1}\|_2^2. \end{cases} \tag{2.7}$$

This is the online version of the linearized Bregman iteration (Cai et al., 2009) modified by smoothing the 1-norm in a ϵ-neighborhood of the origin: the online version of the original linearized Bregman iteration proposed in (Yin et al., 2008) corresponds to $\epsilon = 0$ with $v_{t+1} \in \partial\Psi_\epsilon(w_{t+1})$ and $J_\epsilon(w) = \|w\|_1$. The convergence of the iterates (2.7) is established in the following direct corollary of Theorem 2.

Corollary 6. *Let $\phi(y,a) = \frac{1}{2}(y-a)^2$, $r(w) = 0$, $\|\cdot\| = \|\cdot\|_2$, $\Psi = \Psi_\epsilon$ with $\epsilon > 0$ and $\sigma = 1$. Under Assumption 1, with the step size $\eta_t = \eta_1 t^{-\frac{1}{2}}$ satisfying* (2.2) *with $L_q = 1, L_p = 0$, we have $\mathbb{E}[\mathcal{E}^\phi(w_T) - \mathcal{E}^\phi(w^*)] = O(T^{-\frac{1}{2}} \log T)$.*

It would be interesting to extend the above result to the convergence of the original online linearized Bregman iteration without smoothing.

3 Ideas and Novelty in the Analysis

This section outlines the ideas and novelty in the proof of our main results. Our first novel point is a one-step progress bound established in (3.1) below to be proved in the next section, showing that the excess regularized error $\eta_t[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w)]$ can be controlled by the excess Bregman distance $\mathbb{E}_t[D_\Psi(w,w_t) - D_\Psi(w,w_{t+1})]$ plus the term $\mathbb{E}_t[\mathcal{E}^{\phi,r}(w_t)]$. Here $\mathbb{E}_t = \mathbb{E}[X|\mathcal{A}_t]$ denotes the conditional expectation given $\mathcal{A}_t$, the σ-algebra generated by $\{(x_1,y_1),\ldots,(x_{t-1},y_{t-1})\}$. A notable property of the one-step progress bound (3.1) is that it involves the regularized error $\mathbb{E}_t[\mathcal{E}^{\phi,r}(w_t)]$, rather than the dual norm of gradients encountered during the iterations, whose "boundedness" in expectation is established in (3.2). This "boundedness" of $\mathbb{E}[\mathcal{E}^{\phi,r}(w_t)]$ allows us to avoid assumptions on the boundedness of gradients imposed in the literature (Shamir & Zhang, 2013; Duchi et al., 2010), and demonstrates the novelty of our analysis.

Lemma 7. *Under Assumptions 1, 2, 3, the sequence $\{w_t\}_{t=1}^\infty$ generated by* (1.3) *satisfies*

$$\eta_t[\mathcal{E}^{\phi,r}(w_t)-\mathcal{E}^{\phi,r}(w)] \le \mathbb{E}_t[D_\Psi(w,w_t)-D_\Psi(w,w_{t+1})]+\eta_t^2\big[c_1\mathbb{E}_t[\mathcal{E}^{\phi,r}(w_t)]+c_2\big],\ \forall w \in \mathcal{W}, \tag{3.1}$$

where c_1 and c_2 are two constants independent of t or w (explicitly given in the proof). If we take the step size $\eta_t = \eta_1 t^{-\theta}$ satisfying (2.2) *with $\frac{\max(p,q)}{1+\max(p,q)} \le \theta < 1$, then for any $T \in \mathbb{N}$ we have*

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T)] \le \begin{cases} c_3 \log(eT), & \textit{if } \theta = \frac{\max(p,q)}{1+\max(p,q)}, \\ c_3, & \textit{otherwise}, \end{cases} \tag{3.2}$$

where c_3 is a constant independent of T (explicitly given in the proof).

Our second novel point is to derive error bounds and convergence rates for the last iterate from the one-step progress measured by Bregman distance in Lemma 8. It refines the recent error decomposition method for gradient descent schemes in (Lin et al., 2015b,a; Shamir & Zhang, 2013) reformulating $\eta_T\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)]$ as a summation of the *weighted average errors* and *moving weighted average errors* (see (B.1)), and is proved in Appendix B.

Lemma 8. *Let $\{\eta_t\}$ be a non-increasing sequence. Let $\{A_t\}_{t\in\mathbb{N}}$ be a sequence of random variables such that A_t is measurable with respect to $\mathcal{A}_t$. If*

$$\eta_t[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w)] \le \mathbb{E}_t[D_\Psi(w,w_t) - D_\Psi(w,w_{t+1})] + A_t, \quad \forall w \in \mathcal{W} \tag{3.3}$$

for every $t \in \mathbb{N}$, then we have

$$\eta_T \mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \leq \frac{1}{T} D_\Psi(w_r^*, 0) + \sum_{t=1}^{T-1} \frac{\mathbb{E}[A_t]}{T-t} + \mathbb{E}[A_T]. \tag{3.4}$$

Our last novel point is to get the "boundedness" of $\mathbb{E}[\mathcal{E}^{\phi,r}(w_t)]$ stated in (3.2) by applying Lemma 8 to the following one-step progress bound in terms of the excess Bregman distance and the dual norms of gradients, which can be controlled in terms of step sizes (Lemma 13). Lemma 9 improves Lemma 1 in (Duchi et al., 2010) in our situation. Unlike Lemma 1 in (Duchi et al., 2010) involving $-\phi(y_t, \langle w_t, x_t \rangle) - r(w_{t+1})$ in the associated one-step progress bound, (3.6) in Lemma 9 involves $-\phi(y_t, \langle w_t, x_t \rangle) - r(w_t)$ instead, which matches the form of (3.3) in Lemma 8 and is thereby crucial for applying Lemma 8 to get (3.2). As a comparison, Lemma 1 in (Duchi et al., 2010) could not yield a one-step progress bound of the form (3.3). The proof of Lemma 9 is given in the Appendix A.

Lemma 9. *For any $w \in \mathcal{W}$, the sequence $\{w_t\}_{t=1}^\infty$ generated by* (1.3) *satisfies*

$$\eta_t[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w)] \leq \mathbb{E}_t[D_\Psi(w, w_t) - D_\Psi(w, w_{t+1})] + \\ \sigma^{-1}\eta_t^2 \mathbb{E}_t[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t \rangle)x_t\|_*^2], \tag{3.5}$$

and

$$D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) \leq \eta_t\big[\phi(y_t, \langle w, x_t \rangle) + r(w) - \phi(y_t, \langle w_t, x_t \rangle) - r(w_t)\big] \\ + \sigma^{-1}\eta_t^2[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t \rangle)x_t\|_*^2]. \tag{3.6}$$

Remark 2. It should be emphasized that, a single application of Lemma 8 with the one-step progress bound given in (3.5) can only yield the convergence rate $O(T^{\frac{-1}{\max(p,q)+2}} \log T)$ with the step size $\eta_t = \eta_1 t^{-\frac{\max(p,q)+1}{\max(p,q)+2}}$. For the specific case $r(w) = 0$ and $q = 1$, this convergence rate translates to $O(T^{-\frac{1}{3}} \log T)$, matching the rate $O(T^{-\frac{1}{3}})$ established in (Ying & Zhou, 2015) within a logarithmic factor. The way we achieve the improvement from $O(T^{-\frac{1}{3}})$ to $O(T^{-\frac{1}{2}} \log T)$ rests on the following key observation due to a self-bounding property (see Lemmas 10, 11 below): although the iterates w_t can only be shown to lie in a ball with the asymptotically diverging radius $O([\sum_{\tilde{t}=1}^t \eta_{\tilde{t}}]^{\frac{1}{2}})$ (see Lemma 13 below), the expected norm of the associated gradient is always bounded since it is dominated by the regularized risk.

4 Proving Main Results

This section presents the proof of Theorem 2, which yields the conclusion of Theorem 1. Our proof consists of two parts. The first part applies Lemma 8 and the one-step progress bound (3.5) to establish a crude bound on the regularized risk (3.2), based on which the second part applies Lemma 8 and the one-step progress bound (3.1) to derive the convergence rate (2.3) for the last iterate of the online composite mirror descent.

We first provide some technical lemmas and inequalities used throughout the proof. It is clear that loss functions satisfying Assumption 2 always enjoy the following growth behavior

$$|\phi'_-(y,a)| \leq |\phi'_-(y,0)|+L_q|a|^q \leq \bar{c}_q(1+|a|^q), \quad \bar{c}_q := \max(|\phi|'_0, L_q), \ \ \forall a \in \mathbb{R}, y \in \mathcal{Y}. \tag{4.1}$$

Also, the regularizer $r(w)$ satisfying Assumption 3 meets the following growth condition

$$\|r'(w)\|_* = \|r'(w)-0\|_* \leq L_p\|w-0\|^p = L_p\|w\|^p, \quad \forall w \in \mathcal{W}, \tag{4.2}$$

where we have used the fact $0 \in \partial r(0)$ followed from the convexity of r and $0 \in \arg\min_{w\in\mathcal{W}} r(w)$. For $q \in (0,1]$, denote $c_q = 2L_q^{\frac{1}{q+1}}$ and $\tau_q = 4(1-q)^{-1}$ if $q<1$ and $\tau_q = 4$ if $q = 1$. Denote $p \vee q = \max(p,q)$ and $\theta^* = 2\theta - (1-\theta)(p\vee q)$.

The following two lemmas establish the self-bounding property for functions with Hölder continuous gradients (Srebro et al., 2010; Ying & Zhou, 2015), meaning that the gradients can be controlled by the function values. This self-bounding property allows us to transfer the one-step progress bound (3.5) in terms of gradients to the one-step progress bound (3.1) in terms of the regularized risk, and is essential for us to avoid the boundedness assumptions imposed in the literature (Shamir & Zhang, 2013; Duchi et al., 2010). Lemma 10 can be found in (Ying & Zhou, 2015), while Lemma 11 will be proved as a consequence of Lemma C.2 in Appendix C.

Lemma 10. *If the non-negative loss function $\phi(y,a)$ satisfies* (1.4) *with $q \in (0,1]$, then for $c_q = 2L_q^{\frac{1}{q+1}}$ we have*

$$|\phi'_-(y,a)| \leq c_q\phi(y,a)^{\frac{q}{1+q}}, \quad \forall y, a \in \mathbb{R}. \tag{4.3}$$

Lemma 11. *If the gradient of the non-negative regularizer $r(w)$ satisfies* (2.1) *with $p \in (0,1]$, then for $c_p = 2L_p^{\frac{1}{p+1}}$ we have*

$$\|r'(w)\|_* \leq c_p r(w)^{\frac{p}{1+p}}, \quad \forall w \in \mathcal{W}. \tag{4.4}$$

Lemma 12. *For any $\lambda \in (0,2]$, we have the following inequalities*

$$\sum_{t=1}^{T} t^{-\lambda} \leq 1 + \sum_{t=2}^{T}\int_{t-1}^{t} x^{-\lambda}dx = \begin{cases} \frac{T^{1-\lambda}-\lambda}{1-\lambda}, & \textit{if } \lambda < 1, \\ \log(eT), & \textit{if } \lambda = 1, \\ \frac{\lambda}{\lambda-1}, & \textit{if } \lambda > 1, \end{cases} \tag{4.5}$$

and

$$\sum_{t=1}^{T-1}\frac{t^{-\lambda}}{T-t} \leq \begin{cases} \tau_\lambda T^{-\lambda}\log(eT), & \textit{if } \lambda \leq 1, \\ 8(\lambda-1)^{-1}T^{-1}, & \textit{if } \lambda > 1. \end{cases} \tag{4.6}$$

To apply Lemmas 8 and 9, we need to estimate the growth behavior of $\|r'(w_t)\|_*$ and $\|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*$. This is achieved in the following lemma by showing that $\{w_t\}_{t\in\mathbb{N}}$ always lie inside a ball, under the Bregman divergence, with a controllable radius. The proof of Lemma 13 is given in Appendix D.

Lemma 13. *Suppose that Assumptions 1, 2 and 3 hold. If the step sizes satisfy* (2.2)*, then the sequence* $\{w_t\}_{t=1}^{\infty}$ *generated by* (1.3) *satisfies*

$$D_\Psi(0, w_t) \leq c_{p,q}\sum_{k=1}^{t-1}\eta_k, \tag{4.7}$$

and

$$\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2 \leq c_4\sigma \max\big(1, [\sum_{k=1}^{t-1}\eta_k]^{p\vee q}\big), \tag{4.8}$$

where $c_{p,q}$ *and* c_4 *are constants given by*

$$\begin{aligned} c_{p,q} &= |\phi|_0 + (1-p)(1+p)^{-1} + (1-q)(1+q)^{-1}, \\ c_4 &= \sigma^{-1}\big[2R^2\bar{c}_q^2 + 2\bar{c}_q^2R^{2q+2}[2c_{p,q}\sigma^{-1}]^q + L_p^2[2c_{p,q}\sigma^{-1}]^p\big]. \end{aligned}$$

We are now in a position to prove Lemma 7. The proof of (3.1) requires the one-step progress bound (3.5) and the self-bounding property established in Lemmas 10, 11, while the proof of (3.2) requires to apply Lemma 8 with the one-step progress bound (3.5) coupled with the bounds on the gradients established in Lemma 13.

Proof of Lemma 7. We first use the self-bounding property established in Lemmas 10, 11 to control $\sigma^{-1}\eta_t^2[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2]$.

For the case $0 < q < 1$, by $\frac{q+1}{2q} > 1$ and (4.3) the term $\sigma^{-1}\eta_t^2\|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2$ can be controlled as

$$\begin{aligned} \sigma^{-1}\eta_t^2\|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2 &\leq \sigma^{-1}\eta_t^2R^2|\phi'_-(y_t, \langle w_t, x_t\rangle)|^2 \leq \sigma^{-1}\eta_t^2R^2c_q^2\phi(y_t, \langle w_t, x_t\rangle)^{\frac{2q}{q+1}} \\ &\leq \sigma^{-1}\eta_t^2R^2c_q^2\left[\frac{\big[\phi(y_t, \langle w_t, x_t\rangle)^{\frac{2q}{q+1}}\big]^{\frac{q+1}{2q}}}{\frac{q+1}{2q}} + \frac{1}{\frac{q+1}{1-q}}\right] \\ &\leq \sigma^{-1}\eta_t^2R^2c_q^2(q+1)^{-1}\big[2q\phi(y_t, \langle w_t, x_t\rangle) + 1 - q\big], \end{aligned}$$

where we have used the Young's inequality

$$ab \leq \frac{a^s}{s} + \frac{b^{\tilde{s}}}{\tilde{s}}, \quad \forall a, b, s, \tilde{s} > 0 \text{ with } \frac{1}{s} + \frac{1}{\tilde{s}} = 1. \tag{4.9}$$

The above inequality holds obviously when $q = 1$. Moreover, according to (4.1) we have

$$\sigma^{-1}\eta_t^2R^2|\phi'_-(y_t, \langle w_t, x_t\rangle)|^2 \leq \sigma^{-1}\eta_t^2R^2 4\bar{c}_q^2, \quad \text{if } q = 0.$$

For the case $0 < p < 1$, by $\frac{1+p}{2p} > 1$ and (4.4) the term $\sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2$ can be controlled similarly by

$$\begin{aligned}\sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2 \leq \sigma^{-1}\eta_t^2 c_p^2 r(w_t)^{\frac{2p}{1+p}} &\leq \sigma^{-1}\eta_t^2 c_p^2\Bigg[\frac{\big[r(w_t)^{\frac{2p}{1+p}}\big]^{\frac{1+p}{2p}}}{\frac{1+p}{2p}} + \frac{1}{\frac{1+p}{1-p}}\Bigg]\\ &\leq \sigma^{-1}\eta_t^2 c_p^2(1+p)^{-1}[2pr(w_t)+1-p].\end{aligned}$$

The above inequality holds obviously when $p = 1$. From (4.2) we also have that $\sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2 \leq \sigma^{-1}\eta_t^2 L_p^2$ if $p = 0$.

Putting the above discussions together, we derive the following inequality

$$\begin{aligned}&\sigma^{-1}\eta_t^2\big[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t,\langle w_t,x_t\rangle)x_t\|_*^2\big]\\ &\quad\leq \eta_t^2\sigma^{-1}\big[R^2c_q^2(2q)(q+1)^{-1}\phi(y_t,\langle w_t,x_t\rangle) + 2pc_p^2(1+p)^{-1}r(w_t)\big]\\ &\qquad + \sigma^{-1}\eta_t^2\big[R^2c_q^2(1-q)(1+q)^{-1} + c_p^2(1-p)(1+p)^{-1}\\ &\qquad\qquad + 4(1-q)R^2\bar{c}_q^2 + (1-p)L_p^2\big].\end{aligned}$$

Plugging the above inequality into (3.5) yields the following one-step progress bound for the online composite mirror descent

$$\eta_t[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w)] \leq \mathbb{E}_t[D_\Psi(w,w_t) - D_\Psi(w,w_{t+1})] + \eta_t^2\big[c_1\mathbb{E}_t[\mathcal{E}^{\phi,r}(w_t)] + c_2\big],\ \forall w \in \mathcal{W},$$

where the constants c_1 and c_2 are given explicitly as

$$\begin{aligned}c_1 &= \sigma^{-1}\max(R^2c_q^2(2q)(q+1)^{-1}, 2pc_p^2(1+p)^{-1}),\\ c_2 &= \sigma^{-1}\big[R^2c_q^2(1-q)(1+q)^{-1} + c_p^2(1-p)(1+p)^{-1} + 4(1-q)R^2\bar{c}_q^2 + (1-p)L_p^2\big].\end{aligned}$$

This proves the first desired estimate (3.1).

We turn to the second desired estimate (3.2). Plugging (4.8) into (3.5) immediately yields that

$$\eta_t[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w)] \leq \mathbb{E}_t[D_\Psi(w,w_t) - D_\Psi(w,w_{t+1})] + c_4\eta_t^2\max\Big\{1, \big[\sum_{k=1}^{t-1}\eta_k\big]^{p\vee q}\Big\},$$

which implies (3.3) with $A_t = c_4\eta_t^2\max\Big\{1, \big[\sum_{k=1}^{t-1}\eta_k\big]^{p\vee q}\Big\}$. Therefore, we can apply Lemma 8 to obtain

$$\begin{aligned}\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \leq \frac{1}{T\eta_T}D_\Psi(w_r^*,0) + c_4\eta_T^{-1}\sum_{t=1}^{T-1}\frac{\eta_t^2\max\Big\{1, \big[\sum_{k=1}^{t-1}\eta_k\big]^{p\vee q}\Big\}}{T-t}\\ + c_4\eta_T\max\Big\{1, \big[\sum_{k=1}^{T-1}\eta_k\big]^{p\vee q}\Big\}. \quad (4.10)\end{aligned}$$

According to Lemma 12, the definition of θ^* and the step size $\eta_t = \eta_1 t^{-\theta}$ with $(p \vee q)(1 + p \vee q)^{-1} \le \theta < 1$ we have

$$\begin{aligned}
\eta_T^{-1} \sum_{t=1}^{T-1} \frac{\eta_t^2 \max\big(1, \big[\sum_{k=1}^{t-1} \eta_k\big]^{p\vee q}\big)}{T-t} &\le \eta_T^{-1} \sum_{t=1}^{T-1} \frac{\eta_t^2 \max\big(1, [\eta_1(1-\theta)^{-1}]^{p\vee q} t^{(1-\theta)(p\vee q)}\big)}{T-t} \\
&\le \eta_1^2 \eta_T^{-1} \big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) \sum_{t=1}^{T-1} \frac{t^{(1-\theta)(p\vee q)-2\theta}}{T-t} \\
&= \eta_1^2 \eta_T^{-1} \big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) \sum_{t=1}^{T-1} \frac{t^{-\theta^*}}{T-t} \\
&\le \eta_1 \big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) \begin{cases} \tau_{\theta^*} T^{\theta-\theta^*} \log(eT), & \text{if } \theta^* \le 1, \\ 8(\theta^*-1)^{-1} T^{\theta-1}, & \text{if } \theta^* > 1. \end{cases}
\end{aligned}$$

Furthermore, it follows from (4.5) that

$$\begin{aligned}
\eta_T \max\Big\{1, \big[\sum_{k=1}^{T-1} \eta_k\big]^{p\vee q}\Big\} &\le \eta_1 T^{-\theta} \max\Big\{1, [\eta_1(1-\theta)^{-1}]^{p\vee q} T^{(1-\theta)(p\vee q)}\Big\} \\
&\le \eta_1 \big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) T^{\theta-\theta^*}.
\end{aligned}$$

Plugging the above two bounds into (4.10), we see

$$\begin{aligned}
\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] &\le \eta_1^{-1} T^{\theta-1} D_\Psi(w_r^*, 0) + c_4 \eta_1 \big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) T^{\theta-\theta^*} \\
&\quad + c_4 \eta_1 \big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) \begin{cases} \tau_{\theta^*} T^{\theta-\theta^*} \log(eT), & \text{if } \theta \le \frac{1+p\vee q}{2+p\vee q}, \\ 8(\theta^*-1)^{-1} T^{\theta-1}, & \text{if } \theta > \frac{1+p\vee q}{2+p\vee q}, \end{cases}
\end{aligned}$$

where we observe that $\theta^* \le 1$ if and only if $\theta \le [1 + (p \vee q)]\big[2 + (p \vee q)\big]^{-1}$. Note that $\theta \le \theta^*$ can be equivalently written as $\theta \ge (p \vee q)\big(1 + p \vee q\big)^{-1}$.

When $\theta = (p \vee q)(1 + p \vee q)^{-1}$, we have $T^{\theta-\theta^*} \log(eT) = \log(eT)$.

When $(p \vee q)(1 + p \vee q)^{-1} < \theta \le (1 + p \vee q)(2 + p \vee q)^{-1}$, then $\theta - \theta^* < 0$ and the elementary inequality

$$\max_{x>0} \big\{x^{-\tau} \log x\big\} \le \frac{1}{e\tau}, \quad \forall \tau > 0 \tag{4.11}$$

imply that

$$\begin{aligned}
T^{\theta-\theta^*} \log(eT) &= (eT)^{\theta-\theta^*} \log(eT) e^{\theta^*-\theta} \le e^{-1} (\theta^* - \theta)^{-1} e^{\theta^*-\theta} \\
&= [(1 + p \vee q)\theta - p \vee q]^{-1} e^{\theta+(p\vee q)\theta - p\vee q - 1} \le [(1 + p \vee q)\theta - p \vee q]^{-1}.
\end{aligned}$$

When $\theta > (1 + p \vee q)(2 + p \vee q)^{-1}$, $T^{\theta-1}$ is bounded by 1.

Combining the above three cases together, we know that

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \le \eta_1^{-1} D_\Psi(w_r^*, 0) +$$
$$c_4\eta_1\big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) \times \begin{cases} [1 + \tau_{\theta^*}\log(eT)], & \text{if } \theta = \frac{p\vee q}{1+p\vee q}, \\ [1 + 8(\theta^* - 1)^{-1}], & \text{if } \theta > \frac{1+p\vee q}{2+p\vee q}, \\ \big[1 + \tau_{\theta^*}[(1+p\vee q)\theta - p\vee q]^{-1}\big], & \text{otherwise.} \end{cases}$$

The above inequality verifies the desired estimate (3.2) with the constant c_3 given by

$$c_3 = \mathcal{E}^{\phi,r}(w_r^*) + \eta_1^{-1} D_\Psi(w_r^*, 0) + c_5,$$

where

$$c_5 = c_4\eta_1\big(1 \vee [\eta_1(1-\theta)^{-1}]^{p\vee q}\big) \times \begin{cases} [1 + \tau_{\theta^*}], & \text{if } \theta = \frac{p\vee q}{1+p\vee q}, \\ [1 + 8(\theta^* - 1)^{-1}], & \text{if } \theta > \frac{1+p\vee q}{2+p\vee q}, \\ \big[1 + \tau_{\theta^*}[(1+p\vee q)\theta - p\vee q]^{-1}\big], & \text{otherwise.} \end{cases}$$

The proof of Lemma 7 is complete. □

We are in a position to prove our main results.

Proof of Theorem 2. We prove our conclusion in two cases according to different values of θ.

If $\theta > (p \vee q)(1 + p \vee q)^{-1}$, then (3.2) implies $\mathbb{E}[\mathcal{E}^{\phi,r}(w_t)] \le c_3$. According to (3.1), we can apply Lemma 8 with $A_t = (c_1\mathbb{E}_t[\mathcal{E}^{\phi,r}(w_t)] + c_2)\eta_t^2$ and use the inequality $\mathbb{E}[A_t] \le (c_1c_3 + c_2)\eta_t^2$ to obtain

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \le (T\eta_T)^{-1} D_\Psi(w_r^*, 0) + (c_1c_3 + c_2)\eta_T^{-1}\sum_{t=1}^{T-1}\frac{\eta_t^2}{T-t} + (c_1c_3 + c_2)\eta_T$$
$$= \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*, 0) + (c_1c_3 + c_2)\eta_1 T^\theta \sum_{t=1}^{T-1}\frac{t^{-2\theta}}{T-t} + (c_1c_3 + c_2)\eta_1 T^{-\theta}$$
$$\le \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*, 0) + (c_1c_3 + c_2)\eta_1 T^{-\theta} + (c_1c_3 + c_2)\eta_1 \begin{cases} \tau_{2\theta}T^{-\theta}\log(eT), & \text{if } \theta \le \frac{1}{2}, \\ 8(2\theta - 1)^{-1}T^{\theta-1}, & \text{if } \theta > \frac{1}{2}, \end{cases}$$

where we have used (4.6) in the last inequality.

If $\theta = (p \vee q)(1 + p \vee q)^{-1}$, then (3.2) implies

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_t)] \le c_3 \log(eT), \quad \forall t \le T.$$

Analyzing analogously to the case $\theta > (p \vee q)(1 + p \vee q)^{-1}$ yields

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \le \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*, 0) + (c_1c_3\log(eT) + c_2)\eta_1\big[T^\theta \sum_{t=1}^{T-1}\frac{t^{-2\theta}}{T-t} + T^{-\theta}\big]$$
$$\le \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*, 0) + (c_1c_3\log(eT) + c_2)\eta_1\big[\tau_{2\theta}T^{-\theta}\log(eT) + T^{-\theta}\big]$$
$$\le \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*, 0) + (c_1c_3 + c_2)\eta_1\big[\tau_{2\theta}T^{-\theta}\log^2(eT) + T^{-\theta}\log(eT)\big]$$
$$\le \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*, 0) + (c_1c_3 + c_2)\eta_1\big[4\tau_{2\theta}e^{\theta-2}\theta^{-2} + e^{\theta-1}\theta^{-1}\big],$$

where we have used the fact that $2\theta = 2(p\vee q)(1+p\vee q)^{-1} \leq 1$ followed from $p\vee q \leq 1$, and the last step uses the following inequalities due to (4.11)

$$T^{-\theta}\log^2(eT) = [(eT)^{-\frac{\theta}{2}}\log(eT)]^2 e^{\theta} \leq 4e^{\theta-2}\theta^{-2}, \quad T^{-\theta}\log(eT) \leq e^{\theta-1}\theta^{-1}.$$

The following inequality follows for any $T \in \mathbb{N}$

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T)] \leq c_6 := \mathcal{E}^{\phi,r}(w_r^*) + \eta_1^{-1}D_\Psi(w_r^*,0) + (c_1c_3+c_2)\eta_1\big[4\tau_{2\theta}e^{\theta-2}\theta^{-2} + e^{\theta-1}\theta^{-1}\big].$$

Analyzing analogously to the case $\theta > (p\vee q)(1+p\vee q)^{-1}$ by applying Lemma 8 and (3.1) with $\mathbb{E}[\mathcal{E}^{\phi,r}(w_T)]$ bounded above and noting $\theta \leq \frac{1}{2}$ yields

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \leq \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*,0) + (c_1c_6+c_2)\eta_1T^{-\theta} + (c_1c_6+c_2)\eta_1\tau_{2\theta}T^{-\theta}\log(eT).$$

Combining the above discussions in two different cases together, we get

$$\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \leq \begin{cases} \eta_1^{-1}T^{\theta-1}D_\Psi(w_r^*,0) + c_7T^{-\theta}[1+\tau_{2\theta}\log(eT)], & \text{if } \theta \leq \frac{1}{2}, \\ [\eta_1^{-1}D_\Psi(w_r^*,0) + 8c_7(2\theta-1)^{-1}]T^{\theta-1} + c_7T^{-\theta}, & \text{if } \theta > \frac{1}{2}, \end{cases} \tag{4.12}$$

where c_7 is the constant defined by

$$c_7 := \begin{cases} (c_1c_6+c_2)\eta_1, & \text{if } \theta = \frac{p\vee q}{1+p\vee q}, \\ (c_1c_3+c_2)\eta_1, & \text{otherwise.} \end{cases}$$

This verifies the desired error bound (2.3) with the constant

$$c = \begin{cases} \eta_1^{-1}D_\Psi(w_r^*,0) + c_7(1+\tau_{2\theta}), & \text{if } \theta \leq \frac{1}{2}, \\ \eta_1^{-1}D_\Psi(w_r^*,0) + 8c_7(2\theta-1)^{-1} + c_7, & \text{otherwise.} \end{cases}$$

The proof of Theorem 2 is complete. □

Proof of Theorem 1. According to the definition of w_r^* we have

$$\mathcal{E}^{\phi}(w_r^*) + \lambda\|w_r^*\|_1 \leq \mathcal{E}^{\phi}(w^*) + \lambda\|w^*\|_1. \tag{4.13}$$

For any $w \in \mathcal{W}$, the monotonic property of $\|\cdot\|_*$ implies $\|r'(w)\|_* \leq \lambda\|\mathbf{1}\|_*$ and therefore $r(w) = \lambda\|w\|_1$ satisfies (2.1) with $p = 0, \|\cdot\| = \|\cdot\|_1$ and $L_p = 2\lambda\|\mathbf{1}\|_*$. It then follows from Theorem 2 and (4.13) that

$$\begin{aligned} \mathbb{E}[\mathcal{E}^{\phi}(w_T) + \lambda\|w_T\|_1] &\leq \mathcal{E}^{\phi}(w_r^*) + \lambda\|w_r^*\|_1 + cT^{-\frac{1}{2}}\log(eT) \\ &\leq \mathcal{E}^{\phi}(w^*) + \lambda\|w^*\|_1 + cT^{-\frac{1}{2}}\log(eT), \end{aligned} \tag{4.14}$$

where c is a constant depending on $\eta_1, \mathcal{E}^{\phi,r}(w_r^*), D_\Psi(w_r^*,0), \sigma^{-1}, q, L_q, R, |\phi|_0, |\phi|_0'$. Eq. (4.13), together with the inequality $\mathcal{E}^{\phi}(w^*) \leq \mathcal{E}^{\phi}(w_r^*)$ due to the definition of w^*, implies $\|w_r^*\|_1 \leq \|w^*\|_1$ and then $D_\Psi(w_r^*,0) \leq \sup_{\|w\|_1\leq\|w^*\|_1} D_\Psi(w,0)$. Furthermore, (4.13) and the assumption $\lambda \leq \lambda_0$ imply

$$\mathcal{E}^{\phi}(w_r^*) \leq \mathcal{E}^{\phi}(w^*) + \lambda\|w^*\|_1 \leq \mathcal{E}^{\phi}(w^*) + \lambda_0\|w^*\|_1.$$

That is, both $D_\Psi(w_r^*,0)$ and $\mathcal{E}^{\phi}(w_r^*)$ can be upper bounded by constants independent of λ or T. Therefore, the constant c in (4.14) is independent of T or λ. The proof of Theorem 1 is complete. □

Acknowledgements

We are grateful to the anonymous referees for their constructive comments. The work described in this paper is partially supported by the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006].

Appendix

In this appendix, we prove our claimed technical lemmas.

A Proof of Lemma 9

Proof of Lemma 9. The first-order optimality condition for the minimization problem (1.3) implies the existence of an $r'(w_{t+1}) \in \partial r(w_{t+1})$ such that

$$\eta_t \phi'_-(y_t, \langle w_t, x_t \rangle) x_t + \eta_t r'(w_{t+1}) + \nabla\Psi(w_{t+1}) - \nabla\Psi(w_t) = 0.$$

Combining this with

$$D_\Psi(u,v) + D_\Psi(v,w) - D_\Psi(u,w) = \langle u - v, \nabla\Psi(w) - \nabla\Psi(v) \rangle$$

yields

$$\begin{aligned} D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) &= D_\Psi(w, w_{t+1}) + D_\Psi(w_{t+1}, w_t) - D_\Psi(w, w_t) - D_\Psi(w_{t+1}, w_t) \\ &= \langle w - w_{t+1}, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1}) \rangle - D_\Psi(w_{t+1}, w_t) \\ &= \langle w - w_{t+1}, \eta_t \phi'_-(y_t, \langle w_t, x_t \rangle) x_t + \eta_t r'(w_{t+1}) \rangle - D_\Psi(w_{t+1}, w_t) \\ &= \eta_t \langle w - w_t, \phi'_-(y_t, \langle w_t, x_t \rangle) x_t \rangle + \eta_t \langle w_t - w_{t+1}, \phi'_-(y_t, \langle w_t, x_t \rangle) x_t \rangle \\ &\quad + \eta_t \langle w - w_{t+1}, r'(w_{t+1}) \rangle - D_\Psi(w_{t+1}, w_t). \end{aligned}$$

This together with the convexity of ϕ and r, and the σ-strong convexity of Ψ gives

$$\begin{aligned} D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) &\le \eta_t \big[\phi(y_t, \langle w, x_t \rangle) - \phi(y_t, \langle w_t, x_t \rangle) \big] \\ &\quad + \eta_t [r(w) - r(w_{t+1})] + \eta_t \langle w_t - w_{t+1}, \phi'_-(y_t, \langle w_t, x_t \rangle) x_t \rangle - 2^{-1}\sigma \|w_{t+1} - w_t\|^2 \\ &\le \eta_t \big[\phi(y_t, \langle w, x_t \rangle) - \phi(y_t, \langle w_t, x_t \rangle) \big] + \eta_t [r(w) - r(w_t)] \\ &\quad + \eta_t [r(w_t) - r(w_{t+1})] + \eta_t \|w_t - w_{t+1}\| \|\phi'_-(y_t, \langle w_t, x_t \rangle) x_t\|_* - 2^{-1}\sigma \|w_{t+1} - w_t\|^2. \end{aligned} \tag{A.1}$$

The convexity of r yields

$$\eta_t [r(w_t) - r(w_{t+1})] \le \eta_t \langle w_t - w_{t+1}, r'(w_t) \rangle \le \eta_t \|w_t - w_{t+1}\| \|r'(w_t)\|_*.$$

This together with the elementary inequality $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ gives

$$\begin{aligned}
&\eta_t[r(w_t) - r(w_{t+1})] + \eta_t\|w_t - w_{t+1}\|\|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_* - 2^{-1}\sigma\|w_{t+1} - w_t\|^2 \\
&\leq \eta_t\|w_t - w_{t+1}\|\left[\|r'(w_t)\|_* + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*\right] - 2^{-1}\sigma\|w_{t+1} - w_t\|^2 \\
&\leq \frac{\sigma}{2}\|w_{t+1} - w_t\|^2 + \frac{\eta_t^2}{2\sigma}[\|r'(w_t)\|_* + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*]^2 - 2^{-1}\sigma\|w_{t+1} - w_t\|^2 \\
&\leq \sigma^{-1}\eta_t^2[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2].
\end{aligned}$$

Plugging this estimate into (A.1) gives

$$\begin{aligned}
D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) \leq \eta_t\left[\phi(y_t, \langle w, x_t\rangle) + r(w) - \phi(y_t, \langle w_t, x_t\rangle) - r(w_t)\right] \\
+ \sigma^{-1}\eta_t^2[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2].
\end{aligned}$$

This establishes (3.6). Reformulation followed with conditional expectation with given $\mathcal{A}_t$ (note that w_t is measurable w.r.t $\mathcal{A}_t$) on both sides yields

$$\begin{aligned}
\eta_t[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w)] \leq \mathbb{E}_t[D_\Psi(w, w_t) - D_\Psi(w, w_{t+1})]+ \\
\sigma^{-1}\eta_t^2\mathbb{E}_t[\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2].
\end{aligned}$$

The proof of Lemma 9 is complete. □

B Proof of Lemma 8

We use our ideas from (Lin & Zhou, 2015; Lin et al., 2015b,a) to prove Lemma 8.

Proof of Lemma 8. We proceed with the proof in four steps.

Step 1: Error decomposition. The following identity (Shamir & Zhang, 2013; Lin et al., 2015b) holds for any sequence $\{s_t\}_{t\in\mathbb{N}}$

$$s_T = \frac{1}{T}\sum_{t=1}^{T} s_t + \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}(s_t - s_{T-k}).$$

Applying this to $s_t = \eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_r^*)]$ yields

$$\begin{aligned}
&\eta_T\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] = \frac{1}{T}\sum_{t=1}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_r^*)]+ \\
&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\Big(\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_r^*)] - \eta_{T-k}\mathbb{E}[\mathcal{E}^{\phi,r}(w_{T-k}) - \mathcal{E}^{\phi,r}(w_r^*)]\Big),
\end{aligned}$$

from which we derive

$$\begin{aligned}
&\eta_T\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] = \frac{1}{T}\sum_{t=1}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_r^*)]+ \\
&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\Big(\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_{T-k})] + (\eta_t - \eta_{T-k})\mathbb{E}[\mathcal{E}^{\phi,r}(w_{T-k}) - \mathcal{E}^{\phi,r}(w_r^*)]\Big).
\end{aligned}$$

The definition of w_r^* implies $\mathbb{E}[\mathcal{E}^{\phi,r}(w_{T-k})] \geq \mathbb{E}[\mathcal{E}^{\phi,r}(w_r^*)]$, which, coupled with the fact that $\{\eta_t\}_{t\in\mathbb{N}}$ is non-increasing, guarantees the non-positivity of the last term in the above inequality and thereby implies

$$\eta_T\mathbb{E}[\mathcal{E}^{\phi,r}(w_T) - \mathcal{E}^{\phi,r}(w_r^*)] \leq \frac{1}{T}\sum_{t=1}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_r^*)] \\ + \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_{T-k})]. \quad \text{(B.1)}$$

The first and second term in the right-hand side of the above inequality are called the *weighted average errors* and *moving weighted average errors*, respectively.

Step 2: Controlling Weighted Average Errors. Applying the assumption (3.3) with $w = w_r^*$ and taking expectation over remaining random variables imply

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_r^*)] &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[D_\Psi(w_r^*, w_t) - D_\Psi(w_r^*, w_{t+1})] + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[A_t] \\ &= \frac{1}{T}\mathbb{E}[D_\Psi(w_r^*, w_1) - D_\Psi(w_r^*, w_{T+1})] + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[A_t] \\ &\leq \frac{1}{T}D_\Psi(w_r^*, 0) + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[A_t].\end{aligned}$$

Step 3: Controlling Moving Weighted Average Errors. Applying (3.3) with $w = w_{T-k}$ (note w_{T-k} is measurable with respect to $\mathcal{A}_t$ for any $t \geq T-k$) followed with expectations over remaining random variables implies

$$\begin{aligned}&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_{T-k})] \\ &= \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\eta_t\mathbb{E}[\mathcal{E}^{\phi,r}(w_t) - \mathcal{E}^{\phi,r}(w_{T-k})] \\ &\leq \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\Big[\sum_{t=T-k}^{T}\mathbb{E}\big[D_\Psi(w_{T-k}, w_t) - D_\Psi(w_{T-k}, w_{t+1})\big] + \sum_{t=T-k}^{T}\mathbb{E}[A_t]\Big] \\ &\leq \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\Big[\mathbb{E}[D_\Psi(w_{T-k}, w_{T-k}) - D_\Psi(w_{T-k}, w_{T+1})] + \sum_{t=T-k}^{T}\mathbb{E}[A_t]\Big] \\ &\leq \sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\mathbb{E}[A_t].\end{aligned}$$

Step 4: Combining the above results. Plugging the error bounds in the above two steps into (B.1) yields

$$\begin{aligned}\eta_T\mathbb{E}[\mathcal{E}^{\phi,r}(w_T)-\mathcal{E}^{\phi,r}(w_r^*)] &\leq \frac{1}{T}D_\Psi(w_r^*,0)+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[A_t]+\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\mathbb{E}[A_t]\\ &=\frac{1}{T}D_\Psi(w_r^*,0)+\sum_{t=1}^{T-1}\frac{\mathbb{E}[A_t]}{T-t}+\mathbb{E}[A_T],\end{aligned}$$

where the last inequality uses the following identity

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}A_t+\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}A_t &= \frac{1}{T}\sum_{t=1}^{T}A_t+\sum_{t=1}^{T-1}A_t\sum_{k=T-t}^{T-1}\frac{1}{k(k+1)}+A_T\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\\ &=\frac{1}{T}\sum_{t=1}^{T}A_t+\sum_{t=1}^{T-1}A_t\Big(\frac{1}{T-t}-\frac{1}{T}\Big)+[1-T^{-1}]A_T\\ &=\sum_{t=1}^{T-1}\frac{A_t}{T-t}+A_T.\end{aligned}$$

The proof of Lemma 8 is complete. □

C Proving Self-bounding Properties

The following lemma is an extension of Proposition 1 in (Ying & Zhou, 2015), which considers the case $\alpha=\beta$.

Lemma C.1. *Let $\phi:\mathbb{R}\to\mathbb{R}_0^+$ be a differentiable function. Suppose that there exist constants $0<\alpha\leq\beta\leq 1$ and $L>0$ such that*

$$|\phi'(s)-\phi'(\tilde{s})|\leq L\max(|s-\tilde{s}|^\alpha,|s-\tilde{s}|^\beta),\quad \forall s,\tilde{s}\in\mathbb{R}. \tag{C.1}$$

Then, for any $s\in\mathbb{R}$ we have

$$|\phi'(s)|\leq 2\max\big(L^{\frac{1}{\alpha+1}}\phi(s)^{\frac{\alpha}{\alpha+1}},L^{\frac{1}{\beta+1}}\phi(s)^{\frac{\beta}{\beta+1}}\big). \tag{C.2}$$

Proof. Let $s\in\mathbb{R}$ be any real number. It suffices to consider the case $\phi'(s)\neq 0$. We proceed with the discussion by considering two cases according to the value of $\phi'(s)$.

If $|\phi'(s)|\leq 2L$, we take $r=s-(2^{-1}L^{-1}|\phi'(s)|)^{\frac{1}{\alpha}}\frac{\phi'(s)}{|\phi'(s)|}$. Then $|r-s|\leq 1$. According to the mean-value theorem, there exists ξ between s and r such that $\phi(r)=\phi(s)+\phi'(\xi)(r-s)$. Therefore, by (C.1) and the condition $0<\alpha\leq\beta$,

$$\begin{aligned}0\leq\phi(r)&=\phi(s)+\phi'(s)(r-s)+(\phi'(\xi)-\phi'(s))(r-s)\\ &\leq\phi(s)+\phi'(s)(r-s)+L|r-s|\max(|\xi-s|^\alpha,|\xi-s|^\beta)\\ &\leq\phi(s)+\phi'(s)(r-s)+L|r-s|\max(|r-s|^\alpha,|r-s|^\beta)\\ &=\phi(s)+\phi'(s)(r-s)+L|r-s|^{\alpha+1}\\ &=\phi(s)-(2^{-1}L^{-1})^{\frac{1}{\alpha}}|\phi'(s)|^{\frac{1}{\alpha}+1}+L2^{-\frac{\alpha+1}{\alpha}}L^{-\frac{\alpha+1}{\alpha}}|\phi'(s)|^{\frac{\alpha+1}{\alpha}}\\ &=\phi(s)-2^{-\frac{\alpha+1}{\alpha}}L^{-\frac{1}{\alpha}}|\phi'(s)|^{\frac{1}{\alpha}+1}.\end{aligned}$$

Hence

$$|\phi'(s)| \leq 2L^{\frac{1}{\alpha+1}}\phi(s)^{\frac{\alpha}{\alpha+1}}.$$

If $|\phi'(s)| > 2L$, we take $r = s - (2^{-1}L^{-1}|\phi'(s)|)^{\frac{1}{\beta}}\frac{\phi'(s)}{|\phi'(s)|}$. Then $|r - s| > 1$. Analyzing analogously to the first case we get

$$\begin{aligned}
0 &\leq \phi(s) + \phi'(s)(r-s) + L\max(|r-s|^{\alpha+1}, |r-s|^{\beta+1}) \\
&= \phi(s) + \phi'(s)(r-s) + L|r-s|^{\beta+1} \\
&= \phi(s) - |\phi'(s)|^{\frac{1}{\beta}+1}(2^{-1}L^{-1})^{\frac{1}{\beta}} + L(2^{-1}L^{-1})^{\frac{\beta+1}{\beta}}|\phi'(s)|^{\frac{\beta+1}{\beta}} \\
&= \phi(s) - 2^{-\frac{\beta+1}{\beta}}L^{-\frac{1}{\beta}}|\phi'(s)|^{\frac{1}{\beta}+1}.
\end{aligned}$$

Hence,

$$|\phi'(s)| \leq 2L^{\frac{1}{\beta+1}}\phi(s)^{\frac{\beta}{\beta+1}}.$$

Combining the above discussion together yields the inequality (C.2). □

Lemma C.2 with the case $\alpha = \beta = 1$ was considered in (Srebro et al., 2010).

Lemma C.2. *Let* $r : \mathcal{W} \to \mathbb{R}_0^+$ *be differentiable. Suppose that there exists constants* $0 < \alpha \leq \beta \leq 1$ *and* $L > 0$ *such that*

$$\|r'(w) - r'(\tilde{w})\|_* \leq L\max(\|w - \tilde{w}\|^\alpha, \|w - \tilde{w}\|^\beta), \quad \forall w, \tilde{w} \in \mathcal{W}. \tag{C.3}$$

Then we have

$$\|r'(w)\|_* \leq 2\max(L^{\frac{1}{\alpha+1}}r(w)^{\frac{\alpha}{\alpha+1}}, L^{\frac{1}{\beta+1}}r(w)^{\frac{\beta}{\beta+1}}), \quad \forall w \in \mathcal{W}.$$

Proof. Fix $w \in \mathcal{W}$. For any $\bar{w}$ such that $\|w - \bar{w}\| \leq 1$ we define a function $f_{\bar{w}} : \mathbb{R} \to \mathbb{R}_0^+$ by

$$f_{\bar{w}}(t) = r(\bar{w} + t(w - \bar{w})).$$

It can be directly checked that

$$f'_{\bar{w}}(t) = \langle w - \bar{w}, r'(\bar{w} + t(w - \bar{w}))\rangle$$

and by (C.3)

$$\begin{aligned}
|f'_{\bar{w}}(t) - f'_{\bar{w}}(\tilde{t})| &= \big|\big\langle w - \bar{w}, r'(\bar{w} + t(w - \bar{w})) - r'(\bar{w} + \tilde{t}(w - \bar{w}))\big\rangle\big| \\
&\leq \big\|r'(\bar{w} + t(w - \bar{w})) - r'(\bar{w} + \tilde{t}(w - \bar{w}))\big\|_* \|w - \bar{w}\| \\
&\leq L\max\big(\|w - \tilde{w}\|^\alpha|t - \tilde{t}|^\alpha, \|w - \tilde{w}\|^\beta|t - \tilde{t}|^\beta\big)\|w - \tilde{w}\| \\
&\leq L\max(|t - \tilde{t}|^\alpha, |t - \tilde{t}|^\beta).
\end{aligned}$$

where in the last inequality we have used the inequality $\|w - \bar{w}\| \leq 1$. So the function $f_{\bar{w}}$ satisfies the condition (C.1) and thereby Lemma C.1 can be applied here to obtain

$$|f'_{\bar{w}}(t)| \leq 2\max\big(L^{\frac{1}{\alpha+1}}f_{\bar{w}}(t)^{\frac{\alpha}{\alpha+1}}, L^{\frac{1}{\beta+1}}f_{\bar{w}}(t)^{\frac{\beta}{\beta+1}}\big), \quad \forall t \in \mathbb{R},$$

from which it immediately follows that

$$\begin{aligned}\|r'(w)\|_* &= \sup_{\bar{w}:\|w-\bar{w}\|\le 1}\langle w-\bar{w}, r'(w)\rangle = \sup_{\bar{w}:\|w-\bar{w}\|\le 1} f'_{\bar{w}}(1)\\ &\le 2\max\big(L^{\frac{1}{\alpha+1}} f_{\bar{w}}(1)^{\frac{\alpha}{\alpha+1}}, L^{\frac{1}{\beta+1}} f_{\bar{w}}(1)^{\frac{\beta}{\beta+1}}\big)\\ &= 2\max\big(L^{\frac{1}{\alpha+1}} r(w)^{\frac{\alpha}{\alpha+1}}, L^{\frac{1}{\beta+1}} r(w)^{\frac{\beta}{\beta+1}}\big).\end{aligned}$$

The proof of Lemma C.2 is complete. □

Proof of Corollary 3. Define a map $g:\mathcal{W}\to\mathbb{R}$ by $g(w)=\mathcal{E}^{\phi,r}(w)-\mathcal{E}^{\phi,r}(w_r^*)$. The definition of w_r^* implies $g(w)\ge 0$. For any $w,\tilde{w}\in\mathcal{W}$, we have

$$g(w)-g(\tilde{w})=\mathbb{E}[\phi(y,\langle w,x\rangle)-\phi(y,\langle\tilde{w},x\rangle)]+r(w)-r(\tilde{w}),$$

from which it follows that

$$\begin{aligned}\|\nabla g(w)-\nabla g(\tilde{w})\| &= \big\|\mathbb{E}\big[\phi'(y,\langle w,x\rangle)x-\phi'(y,\langle\tilde{w},x\rangle)x\big]+r'(w)-r'(\tilde{w})\big\|_*\\ &\le \mathbb{E}\big[\big\|\phi'(y,\langle w,x\rangle)x-\phi'(y,\langle\tilde{w},x\rangle)x\big\|_*\big]+\|r'(w)-r'(\tilde{w})\|_*\\ &= \mathbb{E}\big[|\phi'(y,\langle w,x\rangle)-\phi'(y,\langle\tilde{w},x\rangle)|\|x\|_*\big]+\|r'(w)-r'(\tilde{w})\|_*.\end{aligned}$$

Applying (1.4), (2.1) yields

$$\begin{aligned}\|\nabla g(w)-\nabla g(\tilde{w})\| &\le L_q\mathbb{E}[|\langle w-\tilde{w},x\rangle|^q\|x\|_*]+L_p\|w-\tilde{w}\|^p\\ &\le L_q\mathbb{E}[\|w-\tilde{w}\|^q\|x\|_*^{q+1}]+L_p\|w-\tilde{w}\|^p\\ &\le L\max(\|w-\tilde{w}\|^p,\|w-\tilde{w}\|^q),\end{aligned}$$

where $L:=L_p+L_q\mathbb{E}[\|x\|_*^{q+1}]$. So the condition (C.3) is satisfied and we can apply Lemma C.2 to get

$$\|\nabla g(w)\|_*\le 2\max\big(L^{\frac{1}{p+1}}g(w)^{\frac{p}{p+1}}, L^{\frac{1}{q+1}}g(w)^{\frac{q}{q+1}}\big),\quad \forall w\in\mathcal{W}.$$

Setting $w=w_T$ and taking expectations on both sides we find

$$\begin{aligned}\mathbb{E}[\|\nabla\mathcal{E}^{\phi,r}(w_T)\|_*]=\mathbb{E}[\|\nabla g(w_T)\|_*] &\le 2\max\big(L^{\frac{1}{p+1}}\mathbb{E}[g(w_T)^{\frac{p}{p+1}}], L^{\frac{1}{q+1}}\mathbb{E}[g(w_T)^{\frac{q}{q+1}}]\big)\\ &\le 2\max\big(L^{\frac{1}{p+1}}\big[\mathbb{E}[g(w_T)]\big]^{\frac{p}{p+1}}, L^{\frac{1}{q+1}}\big[\mathbb{E}[g(w_T)]\big]^{\frac{q}{q+1}}\big),\end{aligned}$$

where we have used the Jensen inequality. Applying (2.4) of Theorem 2 gives

$$\mathbb{E}[g(w_T)]=\mathbb{E}[\mathcal{E}^{\phi,r}(w_T)-\mathcal{E}^{\phi,r}(w_r^*)]=O(T^{-\frac{1}{2}}\log T)$$

and thereby

$$\mathbb{E}[\|\nabla\mathcal{E}^{\phi,r}(w_T)\|_*]=O\big((T^{-\frac{1}{2}}\log T)^{\min(\frac{p}{p+1},\frac{q}{q+1})}\big).$$

The proof of Corollary 3 is complete. □

The following lemma provides a class of regularizers satisfying the condition (2.1). For $a \in \mathbb{R}$, denote by $\text{sgn}(a)$ the sign of a, i.e., $\text{sgn}(a) = 1$ if $a > 0$, $\text{sgn}(a) = -1$ if $a < 0$ and $\text{sgn}(a) = 0$ if $a = 0$.

Lemma C.3. *The function $r_q(w) = \|w\|_q^q$ with $1 \leq q \leq 2$ defined on $\mathcal{W}$ satisfies*

$$\|r_q'(w) - r_q'(\tilde{w})\|_{q^*} \leq 2q\|w - \tilde{w}\|_q^{q-1}, \quad \forall w, \tilde{w} \in \mathcal{W},$$

where $q^* = \frac{q}{q-1}$ is the conjugate exponent of q.

Proof. If $q = 1$, then for any $w \in \mathcal{W}$ the associated subgradient $r_1'(w)$ would satisfy $\|r_1'(w)\|_\infty \leq 1$, from which we immediately derive

$$\|r_1'(w) - r_1'(\tilde{w})\|_\infty \leq 2\|w - \tilde{w}\|_1^0.$$

If $q > 1$, then the gradient of r_q at w can be calculated by $\nabla r_q(w) = q\big(\text{sgn}(w(i))|w(i)|^{q-1}\big)_{i=1}^d$, from which we have

$$\begin{aligned}\|\nabla r_q(w) - \nabla r_q(\tilde{w})\|_{q^*} &= q\Big[\sum_{i=1}^d \big|\text{sgn}(w(i))|w(i)|^{q-1} - \text{sgn}(\tilde{w}(i))|\tilde{w}(i)|^{q-1}\big|^{q^*}\Big]^{\frac{1}{q^*}} \\ &\leq q\Big[\sum_{i=1}^d 2|w(i) - \tilde{w}(i)|^{(q-1)q^*}\Big]^{\frac{1}{q^*}} \\ &= q\Big[\sum_{i=1}^d 2|w(i) - \tilde{w}(i)|^q\Big]^{\frac{1}{q^*}} = q2^{\frac{q-1}{q}}\|w - \tilde{w}\|_q^{q-1},\end{aligned}$$

where we use the following inequality stated in (Lei et al., 2015)

$$|\text{sgn}(a)|a|^\alpha - \text{sgn}(b)|b|^\alpha| \leq 2|a - b|^\alpha, \quad \forall a, b \in \mathbb{R}, \alpha \in (0, 1].$$

The proof of Lemma C.3 is complete. □

D Proof of Lemma 13

Proof of Lemma 13. We first prove (4.7). Applying (3.6) of Lemma 9 with $w = 0$ shows

$$\begin{aligned}D_\Psi(0, w_{t+1}) - D_\Psi(0, w_t) &\leq \eta_t\big[\phi(y_t, \langle 0, x_t\rangle) - \phi(y_t, \langle w_t, x_t\rangle) + r(0) - r(w_t)\big] \\ &\quad + \sigma^{-1}\eta_t^2\big[\|r'(w_t)\|_*^2 + \|x_t\|_*^2|\phi_-'(y_t, \langle w_t, x_t\rangle)|^2\big] \\ &\leq \eta_t\big[\phi(y_t, \langle 0, x_t\rangle) - \phi(y_t, \langle w_t, x_t\rangle) - r(w_t)\big] \\ &\quad + \sigma^{-1}\eta_t^2\big[\|r'(w_t)\|_*^2 + R^2|\phi_-'(y_t, \langle w_t, x_t\rangle)|^2\big]. \quad \text{(D.1)}\end{aligned}$$

We now tackle the terms $-\eta_t\phi(y_t, \langle w_t, x_t\rangle) + \sigma^{-1}\eta_t^2R^2|\phi_-'(y_t, \langle w_t, x_t\rangle)|^2$ and $-\eta_t r(w_t) + \sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2$, separately. We perform the deduction in three steps.

Step 1. We first bound $-\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2|\phi'_-(y_t,\langle w_t,x_t\rangle)|^2$ according to different values of q.

If $q=1$, applying Lemma 10 shows that

$$\begin{aligned}
&-\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2|\phi'_-(y_t,\langle w_t,x_t\rangle)|^2\\
&\leq -\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2c_q^2\phi(y_t,\langle w_t,x_t\rangle)\\
&= -\eta_t\phi(y_t,\langle w_t,x_t\rangle)\big[1-\sigma^{-1}\eta_tR^2c_q^2\big]\leq 0,
\end{aligned}$$

where in the last step we have used (2.2).

If $0<q<1$, applying Lemma 10 and Young's inequality (4.9) implies

$$\begin{aligned}
&-\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2|\phi'_-(y_t,\langle w_t,x_t\rangle)|^2\\
&\leq -\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2c_q^2\phi(y_t,\langle w_t,x_t\rangle)^{\frac{2q}{q+1}}\\
&\leq -\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\eta_t(1+q)^{-1}\Big[2q\phi(y_t,\langle w_t,x_t\rangle)+(1-q)[\sigma^{-1}R^2c_q^2\eta_t]^{\frac{1+q}{1-q}}\Big]\\
&= -\eta_t(1-2q(1+q)^{-1})\phi(y_t,\langle w_t,x_t\rangle)+\eta_t(1-q)(1+q)^{-1}[\sigma^{-1}R^2c_q^2\eta_t]^{\frac{1+q}{1-q}}.
\end{aligned}$$

Since $2q(1+q)^{-1}<1$, this is bounded by

$$\eta_t(1-q)(1+q)^{-1}[\sigma^{-1}R^2c_q^2\eta_t]^{\frac{1+q}{1-q}}\leq\eta_t\big[(1-q)(1+q)^{-1}\big],$$

where in the last step we have used (2.2) and $c_q=2L_q^{\frac{1}{q+1}}$.

If $q=0$, then from (4.1) we have

$$-\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2|\phi'_-(y_t,\langle w_t,x_t\rangle)|^2\leq\sigma^{-1}\eta_t^2R^24\bar{c}_q^2\leq\eta_t,$$

where the last inequality follows from the assumption $\eta_t\leq\sigma(2R\bar{c}_q)^{-2}$.

Combining the above discussions together we have that for any $q\in[0,1]$

$$-\eta_t\phi(y_t,\langle w_t,x_t\rangle)+\sigma^{-1}\eta_t^2R^2|\phi'_-(y_t,\langle w_t,x_t\rangle)|^2\leq(1-q)(1+q)^{-1}\eta_t. \tag{D.2}$$

Step 2. We now bound $-\eta_tr(w_t)+\sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2$ in three cases according to the value of p.

If $p=1$, from Lemma 11 and the assumption (2.2) we have

$$\begin{aligned}
-\eta_tr(w_t)+\sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2&\leq-\eta_tr(w_t)+\sigma^{-1}\eta_t^2c_p^2r(w_t)\\
&=-\eta_tr(w_t)[1-\sigma^{-1}\eta_tc_p^2]\leq 0.
\end{aligned}$$

If $0<p<1$, Lemma 11 and Young's inequality imply

$$\begin{aligned}
-\eta_tr(w_t)+\sigma^{-1}\eta_t^2\|r'(w_t)\|_*^2&\leq\eta_t[-r(w_t)+\sigma^{-1}\eta_tc_p^2r(w_t)^{\frac{2p}{p+1}}]\\
&\leq\eta_t\Big[-r(w_t)+(1+p)^{-1}\big[2pr(w_t)+(1-p)[\sigma^{-1}\eta_tc_p^2]^{\frac{p+1}{1-p}}\big]\Big]\\
&\leq\eta_tr(w_t)[-1+2p(1+p)^{-1}]+(1+p)^{-1}(1-p)[\sigma^{-1}\eta_tc_p^2]^{\frac{p+1}{1-p}}\eta_t\\
&\leq(1-p)(1+p)^{-1}\eta_t,
\end{aligned}$$

where in the last step we have used (2.2).

If $p = 0$, the assumption $\eta_t \leq \sigma L_p^{-2}$ implies

$$-\eta_t r(w_t) + \sigma^{-1}\eta_t^2 \|r'(w_t)\|_*^2 \leq -\eta_t r(w_t) + \sigma^{-1}\eta_t^2 L_p^2 \leq \eta_t.$$

According to the above deductions we derive for any $p \in [0,1]$

$$-\eta_t r(w_t) + \sigma^{-1}\eta_t^2 \|r'(w_t)\|_*^2 \leq (1-p)(1+p)^{-1}\eta_t. \tag{D.3}$$

Step 3. Plugging (D.2) and (D.3) back into (D.1) we get

$$D_\Psi(0, w_{t+1}) - D_\Psi(0, w_t) \leq \eta_t\big[\phi(y_t, \langle 0, x_t\rangle) + (1-p)(1+p)^{-1} + (1-q)(1+q)^{-1}\big].$$

Taking a summation from $t = 1$ to T yields (4.7) as

$$\begin{aligned} D_\Psi(0, w_{T+1}) &= \sum_{t=1}^{T} [D_\Psi(0, w_{t+1}) - D_\Psi(0, w_t)] + D_\Psi(0, 0) \\ &\leq \sum_{t=1}^{T} \eta_t\big[|\phi|_0 + (1-p)(1+p)^{-1} + (1-q)(1+q)^{-1}\big]. \end{aligned}$$

We then prove (4.8). The σ-strong convexity of Ψ, coupled with the inequality $D_\Psi(0, w_t) \leq c_{p,q}\sum_{k=1}^{t-1}\eta_k$ given by (4.7), implies

$$\frac{\sigma}{2}\|w_t\|^2 \leq D_\Psi(0, w_t) \leq c_{p,q}\sum_{k=1}^{t-1}\eta_k,$$

from which we have

$$\|w_t\|^{2q} \leq \Big[2c_{p,q}\sigma^{-1}\sum_{k=1}^{t-1}\eta_k\Big]^q, \qquad \|w_t\|^{2p} \leq \Big[2c_{p,q}\sigma^{-1}\sum_{k=1}^{t-1}\eta_k\Big]^p \tag{D.4}$$

and by (4.1)

$$\begin{aligned} \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2 &\leq \|x_t\|_*^2 \bar{c}_q^2[1 + |\langle w_t, x_t\rangle|^q]^2 \leq 2R^2\bar{c}_q^2[1 + \|w_t\|^{2q}R^{2q}] \\ &\leq 2R^2\bar{c}_q^2\Big[1 + \Big[2c_{p,q}\sigma^{-1}\sum_{k=1}^{t-1}\eta_k\Big]^q R^{2q}\Big]. \end{aligned} \tag{D.5}$$

Also, it follows from the growth condition (4.2) and (D.4) that

$$\|r'(w_t)\|_*^2 \leq L_p^2\|w_t\|^{2p} \leq L_p^2\Big[2c_{p,q}\sigma^{-1}\sum_{k=1}^{t-1}\eta_k\Big]^p. \tag{D.6}$$

Combining (D.5) and (D.6) together yields

$$\begin{aligned} &\|r'(w_t)\|_*^2 + \|\phi'_-(y_t, \langle w_t, x_t\rangle)x_t\|_*^2 \\ &\leq 2R^2\bar{c}_q^2 + 2\bar{c}_q^2 R^{2q+2}\big[2c_{p,q}\sigma^{-1}\big]^q\Big[\sum_{k=1}^{t-1}\eta_k\Big]^q + L_p^2\big[2c_{p,q}\sigma^{-1}\big]^p\Big[\sum_{k=1}^{t-1}\eta_k\Big]^p \\ &\leq c_4\sigma\max\Big\{1, \Big[\sum_{k=1}^{t-1}\eta_k\Big]^{p\vee q}\Big\}. \end{aligned}$$

This proves (4.8) and completes the proof of Lemma 13. □

E Proof of Lemma 12

The inequality (4.5) is obvious. The inequality (4.6) is a slight modification of Lemma 2.6 in (Lin et al., 2015a).

Proof of Lemma 12. We only prove (4.6) here. We split the sum in two parts as follows (we denote by $\lfloor a \rfloor$ the largest integer not larger than a)

$$\sum_{t=1}^{T-1} \frac{t^{-\lambda}}{T-t} = \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} \frac{t^{-\lambda}}{T-t} + \sum_{t=\lfloor \frac{T}{2} \rfloor + 1}^{T-1} \frac{t^{-\lambda}}{T-t} \leq 2T^{-1} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-\lambda} + \left(2T^{-1}\right)^{\lambda} \sum_{t=\lfloor \frac{T}{2} \rfloor + 1}^{T-1} (T-t)^{-1}$$

$$\leq 2T^{-1} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-\lambda} + (2T^{-1})^{\lambda} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-1} \leq 2T^{-1} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-\lambda} + (2T^{-1})^{\lambda} \log(eT).$$

If $\lambda < 1$, we have

$$\begin{aligned} 2T^{-1} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-\lambda} + (2T^{-1})^{\lambda} \log(eT) &\leq 2T^{-1}(1-\lambda)^{-1}(2^{-1}T)^{1-\lambda} + (2T^{-1})^{\lambda} \log(eT) \\ &\leq 2^{\lambda} T^{-\lambda} (1-\lambda)^{-1} + 2^{\lambda} T^{-\lambda} \log(eT) \\ &\leq 2^{\lambda} T^{-\lambda} \log(eT) \left[1 + (1-\lambda)^{-1}\right] \\ &\leq 2^{\lambda+1} (1-\lambda)^{-1} T^{-\lambda} \log(eT). \end{aligned}$$

If $\lambda = 1$, we have

$$2T^{-1} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-\lambda} + (2T^{-1})^{\lambda} \log(eT) \leq 2T^{-1} \log(eT) + 2T^{-1} \log(eT) = 4T^{-1} \log(eT).$$

If $1 < \lambda \leq 2$, we have

$$\begin{aligned} 2T^{-1} \sum_{t=1}^{\lfloor \frac{T}{2} \rfloor} t^{-\lambda} + (2T^{-1})^{\lambda} \log(eT) &\leq 2T^{-1} \lambda (\lambda-1)^{-1} + (2T^{-1})^{\lambda} \log(eT) \\ &= 2T^{-1} \lambda (\lambda-1)^{-1} + 2^{\lambda} e^{\lambda-1} T^{-1} (eT)^{1-\lambda} \log(eT) \\ &\leq 2T^{-1} \lambda (\lambda-1)^{-1} + 2^{\lambda} e^{\lambda-2} T^{-1} (\lambda-1)^{-1} \\ &\leq 8(\lambda-1)^{-1} T^{-1}, \end{aligned}$$

where we have used (4.11) in the second inequality.

The above bounds together can be written as (4.6). This proves Lemma 12. □

References

Agarwal, A., Bartlett, P. L., Ravikumar, P., & Wainwright, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 5(58), 3235–3249.

Bach, F. & Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, (pp. 773–781).

Ball, K., Carlen, E. A., & Lieb, E. H. (1994). Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1), 463–482.

Beck, A. & Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3), 167–175.

Cai, J.-F., Osher, S., & Shen, Z. (2009). Linearized bregman iterations for compressed sensing. *Mathematics of Computation*, 78(267), 1515–1536.

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.

Chen, D.-R., Wu, Q., Ying, Y., & Zhou, D.-X. (2004). Support vector machine soft margin classifiers: error analysis. *The Journal of Machine Learning Research*, 5, 1143–1175.

Deimling, K. (1985). *Nonlinear functional analysis*. Springer.

Duchi, J. & Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. In *Advances in Neural Information Processing Systems*, (pp. 495–503).

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., & Tewari, A. (2010). Composite objective mirror descent. In *COLT*, (pp. 14–26). Citeseer.

Hu, T., Fan, J., Wu, Q., & Zhou, D.-X. (2015). Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(04), 437–455.

Langford, J., Li, L., & Zhang, T. (2009). Sparse online learning via truncated gradient. In *Advances in neural information processing systems*, (pp. 905–912).

Lei, Y., Ding, L., & Zhang, W. (2015). Generalization performance of radial basis function networks. *IEEE Transactions on Neural Networks and Learning Systems*, 26(3), 551–564.

Lin, J., Rosasco, L., Villa, S., & Zhou, D.-X. (2015a). Modified Fejér sequences and applications. *arXiv preprint arXiv:1510.04641*.

Lin, J., Rosasco, L., & Zhou, D.-X. (2015b). Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, To appear.

Lin, J. & Zhou, D.-X. (2015). Learning theory of randomized Kaczmarz algorithm. *Journal of Machine Learning Research*, 16, 3341–3365.

Lions, P.-L. & Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6), 964–979.

Nemirovsky, A.-S. & Yudin, D.-B. (1983). *Problem complexity and method efficiency in optimization.* John Wiley & Sons.

Rakhlin, A., Shamir, O., & Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, (pp. 449–456).

Rosasco, L., Villa, S., & Vũ, B. C. (2014). Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*.

Shalev-Shwartz, S. & Tewari, A. (2011). Stochastic methods for ℓ_1-regularized loss minimization. *The Journal of Machine Learning Research*, 12, 1865–1892.

Shamir, O. & Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *Proceedings of The 30th International Conference on Machine Learning*, (pp. 71–79).

Smale, S. & Yao, Y. (2006). Online learning algorithms. *Foundations of computational mathematics*, 6(2), 145–170.

Smale, S. & Zhou, D.-X. (2009). Online learning with markov sampling. *Analysis and Applications*, 7(01), 87–113.

Srebro, N., Sridharan, K., & Tewari, A. (2010). Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, (pp. 2199–2207).

Srebro, N., Sridharan, K., & Tewari, A. (2011). On the universality of online mirror descent. In *Advances in neural information processing systems*, (pp. 2645–2653).

Steinwart, I. & Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Tarres, P. & Yao, Y. (2014). Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9), 5716–5735.

Yin, W., Osher, S., Goldfarb, D., & Darbon, J. (2008). Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging sciences*, 1(1), 143–168.

Ying, Y. & Pontil, M. (2008). Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5), 561–596.

Ying, Y. & Zhou, D.-X. (2006). Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11), 4775–4788.

Ying, Y. & Zhou, D.-X. (2015). Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, To appear.

Ying, Y. & Zhou, D.-X. (2016). Online pairwise learning algorithms. *Neural computation*, 28(4), 743–777.