Supplementary Materials

Huang Fang [1] **Nicholas J. A. Harvey** [1] **Victor S. Portella** [1] **Michael P. Friedlander** [1]

A. Facts and propositions

A.1. Scalar inequalities

Fact A.1. For any $a > 0$ and $b, x \in \mathbb{R}$, we have $-ax^2 + bx \leq b^2/4a$.

Fact A.2. $e^{-x} \leq 1 - x + \frac{x^2}{2}$ for $x \geq 0$.

Fact A.3. $\sum_{i=1}^{t} \frac{1}{\sqrt{i}} \leq 2\sqrt{t} - 1$ for $t \geq 1$.

Fact A.4. $\log(x) \leq x - 1$ for $x \geq 0$.

The following proposition is a variant of an inequality that is frequently used in online learning; see, e.g., (Auer et al., 2002, Lemma 3.5), (McMahan, 2017, Lemma 4).

Proposition A.5. Let $u > 0$ and $a_1, a_2, \ldots, a_T \in [0, u]$. Then

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{u + \sum_{i<t} a_i}} \;\leq\; 2\sqrt{\sum_{t=1}^{T} a_t}.$$

Although it is easy to prove this inequality by induction, the following proof may provide more intuition. The proof is based on a generic lemma on approximating sums by integrals.

Lemma A.6 (Sums with chain rule). Let $S \subseteq \mathbb{R}$ be an interval. Let $F : S \to \mathbb{R}$ be concave and differentiable on the interior of S. Let $u \geq 0$ and let $A : \{0, \ldots, T\} \to S$ satisfy $A(i) - A(i-1) \in [0, u]$ for each $1 \leq i \leq T$. Then

$$\sum_{i=1}^{T} F'\big(u + A(i-1)\big) \cdot (A(i) - A(i-1)) \;\leq\; F(A(T)) - F(A(0)).$$

As $u \to 0$, the left-hand side becomes comparable to $\int_0^T F'(A(x))A'(x)\,dx$, an expression that has no formal meaning since A is only defined on integers. If this expression existed, it would equal the right-hand side by the chain rule.

Proof of Lemma A.6. Since F is concave, $f := F'$ is non-increasing. Fix any $1 \leq i \leq T$ and observe that $f(x) \geq f(A(i)) \geq f(u + A(i-1))$ for all $x \leq A(i)$. Thus

$$f(u + A(i-1)) \cdot (A(i) - A(i-1)) \;\leq\; \int_{A(i-1)}^{A(i)} f(x)\,dx \;=\; F(A(i)) - F(A(i-1)).$$

Summing over i, the right-hand side telescopes, which yields the result. □

Proof of Proposition A.5. Apply Lemma A.6 with $S = \mathbb{R}_{\geq 0}$, $F(x) = 2\sqrt{x}$ and $A(i) = \sum_{1 \leq j \leq i} a_j$. □

Proposition A.7. Let $x, y, \alpha, \beta > 0$.

$$\begin{aligned} \text{If} \quad & x - y \;\leq\; \alpha\sqrt{x} + \beta \\ \text{then} \quad & x - y \;\leq\; \alpha\sqrt{y} + \beta + \alpha\sqrt{\beta} + \alpha^2. \end{aligned}$$

Proof. The proposition's hypothesis yields

$$y + \beta + \frac{\alpha^2}{4} \;\geq\; x - \alpha\sqrt{x} + \frac{\alpha^2}{4} \;=\; \Big(\sqrt{x} - \frac{\alpha}{2}\Big)^2.$$

Taking the square root and rearranging,

$$\sqrt{x} \;\leq\; \sqrt{y + \beta + \frac{\alpha^2}{4}} + \frac{\alpha}{2}.$$

Squaring both sides and rearranging,

$$\begin{aligned} x \;&\leq\; y + \alpha\sqrt{y + \beta + \frac{\alpha^2}{4}} + \beta + \frac{\alpha^2}{2} \\ &\leq\; y + \alpha\sqrt{y} + \alpha\sqrt{\beta} + \beta + \alpha^2, \end{aligned}$$

by subadditivity of the square root. □

A.2. Bregman divergence properties

The following lemma collects basic facts regarding the Bregman divergence induced by a mirror map of the Legendre type. See (Zhang, n.d.) and (Rockafellar, 1970, Chapter 26). For a detailed discussion on the necessity (or the lack thereof) of mirror maps of Legendre type, see (Bubeck, 2011, Chapter 5)

Lemma A.8. The Bregman divergence induced by Φ satisfies the following properties:

- $D_\Phi(x, y)$ is convex in x;
- $\nabla\Phi(\nabla\Phi^*(z)) = z$ and $\nabla^*\Phi(\nabla\Phi(x)) = x$ for all x and z;
- $D_\Phi(x, y) = D_{\Phi^*}(\nabla\Phi(y), \nabla\Phi(x))$ for all x and y.

Proposition A.9. If Φ is ρ-strongly convex with respect to $\|\cdot\|$ then $D_\Phi(x, y) \geq \frac{\rho}{2}\|x - y\|^2$.

A.2.1. Differences of Bregman divergences

Recall that in Section 3 we defined the notation

$$D_\Phi\big(\substack{a\\b}; c\big) \;\coloneqq\; D_\Phi(a, c) - D_\Phi(b, c) \;=\; \Phi(a) - \Phi(b) - \langle\, \nabla\Phi(c),\, a - b\,\rangle.$$

This has several useful properties, which we now discuss.

Proposition A.10. $D_\Phi\big(\substack{a\\b}; p\big)$ is linear in $\hat{p}$. In particular,

$$D_\Phi\big(\substack{a\\b}; \nabla\Phi^*(\hat{p} - \hat{q})\big) \;=\; D_\Phi\big(\substack{a\\b}; p\big) + \langle\, \hat{q},\, a - b\,\rangle \qquad \forall \hat{q} \in \mathbb{R}^n.$$

Proof. Immediate from the definition. □

Proposition A.11. For all $a, b, c, d \in \mathcal{D}$,

$$D_\Phi\big(\substack{a\\b}; d\big) - D_\Phi\big(\substack{a\\b}; c\big) \;=\; \langle\, \hat{c} - \hat{d},\, a - b\,\rangle \;=\; D_\Phi\big(\substack{a\\b}; d\big) + D_\Phi\big(\substack{b\\a}; c\big).$$

Proof. The first equality holds from Proposition A.10 with $\hat{p} = \hat{c}$ and $\hat{q} = \hat{c} - \hat{d}$. The second equality holds since $D_\Phi\big(\substack{b\\a}; c\big) = -D_\Phi\big(\substack{a\\b}; c\big)$. □

An immediate consequence is the "generalized triangle inequality for Bregman divergence". See Bubeck (Bubeck, 2015, Eq. (4.1)), Beck and Teboulle (Beck & Teboulle, 2003, Lemma 4.1) or Zhang (Zhang, n.d., Eq. (3)).

Proposition A.12. For all $a, b, d \in \mathcal{D}$,

$$D_\Phi(a, d) - D_\Phi(b, d) + D_\Phi(b, a) \;=\; \langle\, \hat{a} - \hat{d},\, a - b\,\rangle$$

Proof. Apply Proposition A.11 with $c = a$ and use $D_\Phi(a, a) = 0$. □

Proposition A.13. Let $a, b, c, u, v \in \mathbb{R}^n$ satisfy $\gamma\hat{a} + (1-\gamma)\hat{b} = \hat{c}$ for some $\gamma \in \mathbb{R}$. Then

$$\gamma D_\Phi\big(\genfrac{}{}{0pt}{1}{u}{v}; a\big) \;+\; (1-\gamma) D_\Phi\big(\genfrac{}{}{0pt}{1}{u}{v}; b\big) \;=\; D_\Phi\big(\genfrac{}{}{0pt}{1}{u}{v}; c\big).$$

Proof. By definition of D_Φ, the claimed identity is equivalent to

$$(1-\gamma)\big(\Phi(u) - \Phi(v) - \langle\, \nabla\Phi(a),\, u - v \,\rangle\big) \;+\; \gamma\big(\Phi(u) - \Phi(v) - \langle\, \nabla\Phi(b),\, u - v \,\rangle\big) \;=\; \big(\Phi(u) - \Phi(v) - \langle\, \nabla\Phi(c),\, u - v \,\rangle\big).$$

This equality holds by canceling $\Phi(u) - \Phi(v)$ and by the assumption that $\nabla\Phi(c) = (1-\gamma)\nabla\Phi(a) + \gamma\nabla\Phi(b)$. □

The following proposition is the "Pythagorean theorem for Bregman divergence". Recall that $\Pi^\Phi_{\mathcal{X}}(y) = \arg\min_{u \in \mathcal{X}} D_\Phi(u, y)$. Proofs may be found in (Bubeck, 2015, Lemma 4.1) or (Zhang, n.d., Eq. (17)).

Proposition A.14. Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set. Let $p \in \mathbb{R}^n$ and $\pi = \Pi^\Phi_{\mathcal{X}}(p)$. Then

$$D_\Phi\big(\genfrac{}{}{0pt}{1}{z}{\pi}; p\big) \;\geq\; D_\Phi\big(\genfrac{}{}{0pt}{1}{z}{\pi}; \pi\big) \;=\; D_\Phi(z, \pi) \qquad \forall z \in \mathcal{X}.$$

A generalization of the previous proposition can be obtained by using the linearity property.

Proposition A.15. Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set. Let $p \in \mathbb{R}^n$ and $\pi = \Pi^\Phi_{\mathcal{X}}(p)$. Then

$$D_\Phi\big(\genfrac{}{}{0pt}{1}{v}{\pi}; \nabla\Phi^*(\hat{p} - \hat{q})\big) \;\geq\; D_\Phi\big(\genfrac{}{}{0pt}{1}{v}{\pi}; \Phi^*(\hat{\pi} - \hat{q})\big) \qquad \forall v \in \mathcal{X},\; \hat{q} \in \mathbb{R}^n.$$

Proof.

$$\begin{aligned}
D_\Phi\big(\genfrac{}{}{0pt}{1}{v}{\pi}; \nabla\Phi^*(\hat{p} - \hat{q})\big) &= D_\Phi\big(\genfrac{}{}{0pt}{1}{v}{\pi}; p\big) + \langle\, \hat{q},\, v - \pi \,\rangle && \text{(by Proposition A.10)}\\
&\geq D_\Phi\big(\genfrac{}{}{0pt}{1}{v}{\pi}; \pi\big) + \langle\, \hat{q},\, v - \pi \,\rangle && \text{(by Proposition A.14)}\\
&= D_\Phi\big(\genfrac{}{}{0pt}{1}{v}{\pi}; \nabla\Phi^*(\hat{\pi} - \hat{q})\big) && \text{(by Proposition A.10)}
\end{aligned}$$

□

B. Missing details from the proof of Theorem 4.1

Due to space constraints we had to omit some calculations from the proof of Theorem 4.1 the main body of the paper. In particular, we claimed that summing the expression from Claim 4.2 over $t \in [T]$ yields Theorem 4.1. For the sake of completeness we present the missing calculations.

Summing (4.8) over t and using Claim 4.2 leads to the desired telescoping sum.

$$\begin{aligned}
\sum_{t=1}^{T} \big(f_t(x_t) - f_t(z)\big) &\leq \sum_{t=1}^{T} \left(\frac{D_\Phi\big(\genfrac{}{}{0pt}{1}{x_t}{x_{t+1}}; w_{t+1}\big)}{\eta_t} + \Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big) D_\Phi(z, x_1) + \frac{D_\Phi(z, x_t)}{\eta_t} - \frac{D_\Phi(z, x_{t+1})}{\eta_{t+1}} \right)\\
&\leq \sum_{t=1}^{T} \frac{D_\Phi\big(\genfrac{}{}{0pt}{1}{x_t}{x_{t+1}}; w_{t+1}\big)}{\eta_t} + \left(\frac{1}{\eta_1} + \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \right) D_\Phi(z, x_1)\\
&= \sum_{t=1}^{T} \frac{D_\Phi\big(\genfrac{}{}{0pt}{1}{x_t}{x_{t+1}}; w_{t+1}\big)}{\eta_t} + \frac{D_\Phi(z, x_1)}{\eta_{T+1}}.
\end{aligned}$$

C. Proof of Theorem 4.3

Proof (of Theorem 4.3).

Fix $z \in \mathcal{X}$. The first step is the same as in Theorem 4.1.

$$\begin{aligned}
f_t(x_t) - f_t(z) &\overset{\text{(i)}}{\leq} \langle \hat{g}_t, x_t - z \rangle\\
&= \frac{1}{\eta_t} \langle \hat{x}_t - \hat{w}_{t+1}, x_t - z \rangle
\end{aligned}$$

$$\overset{\text{(ii)}}{=} \frac{1}{\eta_t}\Big(D_\Phi(x_t, w_{t+1}) + D_\Phi(z, x_t) - D_\Phi(z, w_{t+1})\Big) \tag{C.1}$$

Here (i) is the subgradient inequality. For (ii), recall our notation $\hat{x}_t = \nabla\Phi(x_t)$ and $\hat{w}_{t+1} = \nabla\Phi(w_{t+1})$, then use the generalized triangle inequality for Bregman divergences (Proposition A.12).

From Claim 4.4 and the definition $\gamma_t = \eta_{t+1}/\eta_t$, we obtain

$$D_\Phi(z, w_{t+1}) \;\geq\; \frac{\eta_t}{\eta_{t+1}} D_\Phi(z, x_{t+1}) - \Big(\frac{\eta_t}{\eta_{t+1}} - 1\Big) D_\Phi(z, x_1) + D_\Phi(y_{t+1}, w_{t+1}). \tag{C.2}$$

The remainder is very similar to Theorem 4.1. Plugging (C.2) into (C.1), we obtain

$$\begin{aligned}
\text{(C.1)} \;&\leq\; \frac{1}{\eta_t}\Bigg(D_\Phi(x_t, w_{t+1}) + D_\Phi(z, x_t) - \frac{\eta_t}{\eta_{t+1}} D_\Phi(z, x_{t+1}) + \Big(\frac{\eta_t}{\eta_{t+1}} - 1\Big) D_\Phi(z, x_1) - D_\Phi(y_{t+1}, w_{t+1})\Bigg) \\
&= \frac{D_\Phi(x_t, w_{t+1}) - D_\Phi(y_{t+1}, w_{t+1})}{\eta_t} + \frac{D_\Phi(z, x_t)}{\eta_t} - \frac{D_\Phi(z, x_{t+1})}{\eta_{t+1}} + \Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big) D_\Phi(z, x_1). \qquad \text{(C.3)}
\end{aligned}$$

Summing (C.3) over t, the $D_\Phi(z, x_t)$ terms telescope, and we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \big(f_t(x_t) - f_t(z)\big) \\
&\leq \sum_{t=1}^{T} \Bigg(\frac{D_\Phi(x_t, w_{t+1}) - D_\Phi(y_{t+1}, w_{t+1})}{\eta_t} + \frac{D_\Phi(z, x_t)}{\eta_t} - \frac{D_\Phi(z, x_{t+1})}{\eta_{t+1}} + \Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big) D_\Phi(z, x_1)\Bigg) \\
&\leq \sum_{t=1}^{T} \frac{D_\Phi(x_t, w_{t+1}) - D_\Phi(y_{t+1}, w_{t+1})}{\eta_t} + \Bigg(\frac{1}{\eta_1} + \sum_{t=1}^{T}\Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)\Bigg) D_\Phi(z, x_1) \\
&= \sum_{t=1}^{T} \frac{D_\Phi(x_t, w_{t+1}) - D_\Phi(y_{t+1}, w_{t+1})}{\eta_t} + \frac{D_\Phi(z, x_1)}{\eta_{T+1}}, \qquad \text{(C.4)}
\end{aligned}$$

as desired.

□

D. Proof of Theorem 4.6

Proof (of Theorem 4.6). As previously mentioned, this proof parallels the proof of Theorem 4.1.

Fix $z \in \mathcal{X}$. The first step is very similar to the proof of Theorem 4.1.

$$\begin{aligned}
f_t(x_t) - f_t(z) \;&\leq\; \langle\, \hat{g}_t,\, x_t - z \,\rangle \qquad \text{(subgradient inequality)} \\
&= \frac{1}{\eta_t}\langle\, \hat{y}_t - \hat{w}_{t+1},\, x_t - z \,\rangle \qquad \text{(by (4.15))} \\
&= \frac{1}{\eta_t}\Big(D_\Phi(x_t, w_{t+1}) - D_\Phi(z, w_{t+1}) + D_\Phi\big(\genfrac{}{}{0pt}{}{z}{x_t}; y_t\big)\Big). \qquad \text{(D.1)}
\end{aligned}$$

where we have used Proposition A.11 instead of Proposition A.12.

As in the proof of Theorem 4.1, the next step is to relate $D_\Phi(z, w_{t+1})$ to $D_\Phi(z, y_{t+1})$ so that (D.1) can be bounded using a telescoping sum. The following claim is similar to Claim 4.2.

Claim D.1. Assume that $\gamma_t = \eta_{t+1}/\eta_t \in (0, 1]$. Then

$$\text{(D.1)} \;\leq\; \frac{D_\Phi\big(\genfrac{}{}{0pt}{}{x_t}{x_{t+1}}; w_{t+1}\big)}{\eta_t} + \underbrace{\Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)}_{\text{telescopes}} D_\Phi(z, x_1) + \underbrace{\frac{D_\Phi\big(\genfrac{}{}{0pt}{}{z}{x_t}; y_t\big)}{\eta_t} - \frac{D_\Phi\big(\genfrac{}{}{0pt}{}{z}{x_{t+1}}; y_{t+1}\big)}{\eta_{t+1}}}_{\text{telescopes}}.$$

Proof. The first two steps are identical to the proof of Claim 4.2.

$$\begin{aligned}
&\gamma_t\big(D_\Phi(z,w_{t+1}) - D_\Phi(x_{t+1},w_{t+1})\big) \;+\; (1-\gamma_t)D_\Phi(z,x_1) \\
\geq\;& \gamma_t D_\Phi\big(\substack{z\\x_{t+1}};w_{t+1}\big) \;+\; (1-\gamma_t)D_\Phi\big(\substack{z\\x_{t+1}};x_1\big) && \text{(since } D_\Phi(x_{t+1},x_1)\geq 0 \text{ and } \gamma_t \leq 1) \\
=\;& D_\Phi\big(\substack{z\\x_{t+1}};y_{t+1}\big) && \text{(by Proposition A.13 and (4.3)).}
\end{aligned}$$

Rearranging and using $\gamma_t > 0$ yields

$$D_\Phi(z,w_{t+1}) \;\geq\; D_\Phi(x_{t+1},w_{t+1}) \;-\; \Big(\frac{1}{\gamma_t}-1\Big)D_\Phi(z,x_1) \;+\; \frac{D_\Phi\big(\substack{z\\x_{t+1}};y_{t+1}\big)}{\gamma_t}. \tag{D.2}$$

Plugging this into (D.1) yields

$$\begin{aligned}
\text{(D.1)} \;=\;& \frac{1}{\eta_t}\Big(D_\Phi(x_t,w_{t+1}) - D_\Phi(z,w_{t+1}) + D_\Phi\big(\substack{z\\x_t};y_t\big)\Big) \\
\leq\;& \frac{1}{\eta_t}\left(D_\Phi(x_t,w_{t+1}) - D_\Phi(x_{t+1},w_{t+1}) + \Big(\frac{1}{\gamma_t}-1\Big)D_\Phi(z,x_1) - \frac{D_\Phi\big(\substack{z\\x_{t+1}};y_{t+1}\big)}{\gamma_t} + D_\Phi\big(\substack{z\\x_t};y_t\big)\right),
\end{aligned}$$

by (D.2). The claim follows by the definition of γ_t. □

The final step is very similar to the proof of Theorem 4.1. Summing (D.1) over t and using Claim D.1 leads to the desired telescoping sum.

$$\begin{aligned}
\sum_{t=1}^{T}\big(f_t(x_t)-f_t(z)\big) \;\leq\;& \sum_{t=1}^{T}\left(\frac{D_\Phi\big(\substack{x_t\\x_{t+1}};w_{t+1}\big)}{\eta_t} + \Big(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\Big)D_\Phi(z,x_1) + \frac{D_\Phi\big(\substack{z\\x_t};y_t\big)}{\eta_t} - \frac{D_\Phi\big(\substack{z\\x_{t+1}};y_{t+1}\big)}{\eta_{t+1}}\right) \\
\leq\;& \sum_{t=1}^{T}\frac{D_\Phi\big(\substack{x_t\\x_{t+1}};w_{t+1}\big)}{\eta_t} \;+\; \left(\frac{1}{\eta_1}+\sum_{t=1}^{T}\Big(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\Big)\right)D_\Phi(z,x_1) \\
=\;& \sum_{t=1}^{T}\frac{D_\Phi\big(\substack{x_t\\x_{t+1}};w_{t+1}\big)}{\eta_t} \;+\; \frac{D_\Phi(z,x_1)}{\eta_{T+1}}.
\end{aligned}$$

For the second inequality we have also used that $D_\Phi\big(\substack{z\\x_1};y_1\big) = D_\Phi(z,x_1)$ since $x_1 = y_1$. Thus, the above shows that

$$\operatorname{Regret}(T,z) \;\leq\; \sum_{t=1}^{T}\frac{D_\Phi\big(\substack{x_t\\x_{t+1}};w_{t+1}\big)}{\eta_t} + \frac{D_\Phi(z,x_1)}{\eta_{T+1}} \qquad \forall T>0. \tag{D.3}$$

Notice that (D.3) is syntactically identical to (4.7); the only difference is the definition of w_{t+1} in these two settings. However the bound (D.3) requires further development because, curiously, this proof has not yet used the definition of x_t. To conclude the theorem, we will provide an upper bound on (D.3) that incorporates the definition $x_t = \Pi_{\mathcal{X}}^{\Phi}(y_t)$. Specifically, we will control $D_\Phi\big(\substack{x_t\\x_{t+1}};w_{t+1}\big)$ by applying Proposition A.15 as follows. Taking $p = y_t$, $\pi = x_t = \Pi_{\mathcal{X}}^{\Phi}(y_t)$, $v = x_{t+1}$ and $\hat{q} = \eta_t\hat{g}_t$, we obtain

$$\begin{aligned}
D_\Phi\big(\substack{x_t\\x_{t+1}};w_{t+1}\big) \;=\;& -D_\Phi\big(\substack{v\\\pi};\nabla\Phi^*(\hat{p}-\hat{q})\big) && \text{(since } \hat{w}_{t+1} = \hat{y}_t - \eta_t\hat{g}_t = \hat{p}-\hat{q}) \\
\leq\;& -D_\Phi\big(\substack{v\\\pi};\nabla\Phi^*(\hat{\pi}-\hat{q})\big) && \text{(by Proposition A.15)} \\
=\;& D_\Phi\big(\substack{x_t\\x_{t+1}};\nabla\Phi^*(\hat{x}_t-\eta_t\hat{g}_t)\big).
\end{aligned}$$

Plugging this into (D.3) completes the proof. □

E. Additional proofs for Section 5.1

Proof (of Proposition 5.2). First we apply Proposition A.12 with $a = x$, $b = x'$ and $d = \nabla\Phi^*(\hat{x}-\hat{q})$ to obtain

$$D_\Phi\big(\substack{x\\x'};w\big) \;=\; \langle\, \hat{x}-\hat{d},\, x-x'\,\rangle - D_\Phi(x',x)$$

$$
\begin{aligned}
&= \langle\, \hat{q},\, x - x' \,\rangle - D_\Phi(x', x) \\
&\leq \|\hat{q}\|_* \|x - x'\| - \frac{\rho}{2}\|x - x'\|^2 && \text{(definition of dual norm and Proposition A.9)} \\
&\leq \|\hat{q}\|_*^2 / 2\rho && \text{(by Fact A.1).}
\end{aligned}
$$
□

F. Additional proofs for Section 5.2

An initial observation shows that Λ is non-negative in the experts' setting.

Proposition F.1. $\Lambda(a, b) \geq 0$ for all $a \in \mathcal{X}$, $b \in \mathcal{D}$.

Proof. Let us write $\Lambda(a,b) = -\sum_{i=1}^n a_i \ln \frac{b_i}{a_i} + \ln\left(\sum_{i=1}^n b_i\right)$. Since a is a probability distribution, we may apply Jensen's inequality to show that this expression is non-negative. □

Proof (of Proposition 5.4). Since $a, b \in \mathcal{X}$ we have $\|a\|_1 = \|b\|_1 = 1$. Then

$$
\begin{aligned}
D_\Phi\left(\tbinom{a}{b}; c\right) &= D_{\mathrm{KL}}(a, c) - D_{\mathrm{KL}}(b, c) \\
&= \big(D_{\mathrm{KL}}(a, c) + 1 - \|c\|_1 + \ln\|c\|_1\big) - \big(D_{\mathrm{KL}}(b, c) + 1 - \|c\|_1 + \ln\|c\|_1\big) \\
&= \Lambda(a, c) - \Lambda(b, c) \qquad \text{(by definition of } \Lambda) \\
&\leq \Lambda(a, c) \qquad \text{(by Proposition F.1).}
\end{aligned}
$$
□

Proof (of Proposition 5.8). Let $b = \nabla\Phi^*(\hat{a} - \eta\hat{q})$. By (5.1), $b_i = a_i \exp(-\eta\hat{q}_i)$. Then

$$
\begin{aligned}
\Lambda(a, \nabla\Phi^*(\hat{a} - \eta\hat{q})) &= \sum_{i=1}^n a_i \ln(a_i/b_i) + \ln\|b\|_1 \\
&= \sum_{i=1}^n \eta a_i \hat{q}_i + \ln\left(\sum_{i=1}^n a_i \exp(-\eta\hat{q}_i)\right) \\
&\leq \sum_{i=1}^n \eta a_i \hat{q}_i + \sum_{i=1}^n a_i \exp(-\eta\hat{q}_i) - 1 \qquad \text{(by Fact A.4)} \\
&\leq \sum_{i=1}^n \eta a_i \hat{q}_i + \sum_{i=1}^n a_i \left(1 - \eta\hat{q}_i + \frac{\eta^2\hat{q}_i^2}{2}\right) - 1 \qquad \text{(by Fact A.2)} \\
&\leq \eta^2 \sum_{i=1}^n a_i \hat{q}_i / 2,
\end{aligned}
$$

using $\sum_{i=1}^n a_i = 1$ (since $a \in \mathcal{X}$) and $\hat{q}_i^2 \leq \hat{q}_i$ (since $\hat{q} \in [0,1]^n$). □

G. Remarks on lower bounds fo the expert's problem

We have seen that dual averaging achieves regret $\sqrt{T \ln n}$ for all T. Here we present a lower bound analysis for DA showing that this is the best one can hope for.

Theorem G.1. There exists a value of n such that, for every $T > 0$, there exists a sequence of vectors $\{c_i \mid c_i \in \{0,1\}^n\}_{i=1}^T$, such that

$$
\lim_{t \to \infty} \frac{\mathrm{Regret}_{\mathrm{DA}}(t)}{\sqrt{t \ln n}} \geq 1,
$$

where $\mathrm{Regret}_{\mathrm{DA}}(T)$ denotes the worst-case regret (that is, taking the supremum of the comparison point over the simples) of the dual averaging algorithm used in (Bubeck, 2011, Theorem 2.4).

It is known in the literature (Cesa-Bianchi & Lugosi, 2006, §3.7) that no algorithm can achieve a regret bound better than $\sqrt{T/2 \, \ln n}$ for the problem of learning with expert advice (as $(T, n) \to \infty$). Thus, there is still a $\sqrt{2}$ gap between the best upper and lower bounds (that hold for all T) for prediction with expert advice. This gap was previously pointed out by Gerchinovitz (2011, pp. 52).

Algorithm 1 Adaptive randomized weighted majority based on DA (Cesa-Bianchi & Lugosi, 2006)

```
Input: $\eta : \mathbb{N} \to \mathbb{R}$
$x_1 = [1/n, 1/n, ...]$
for $t = 1, 2, \dots$ do
    Incur cost $f(x_t)$ and receive $\hat{g}_t \in \partial f_t(x_t)$
    for j=1,2,...,n do
        $y_{t+1,j} = x_{1,j} \exp\left(-\eta_t \sum_{k=1}^{t} \hat{g}_k(j)\right)$
    end for
    $x_{t+1} = y_{t+1}/\|y_{t+1}\|_1$
end for
```

Proof. The detailed algorithm described in (Bubeck, 2011, Theorem 2.4) is shown in Algorithm 1, where η_t is set as $\sqrt{4\ln n/t}$, we consider the case when $n = 2$, and construct the following cost vectors:

$$c_t = \begin{cases} [1,0]^\intercal & 1 \leq t < \tau, t \text{ is odd} \\ [0,1]^\intercal & 1 \leq t < \tau, t \text{ is even} \\ [1,0]^\intercal & \tau \leq t \leq T, \end{cases} \qquad \forall t \geq 1,$$

where $\tau := \lfloor T - \log(T)\sqrt{T} \rfloor$. Without loss of generality, we assume that τ is an odd number.

Throughout the remainder of this proof, denote $\mathrm{Regret}(T)$ as the worts-case regret on T rounds, that is,

$$\mathrm{Regret}(T) := \sup_{z \in \Delta_n} \mathrm{Regret}(T, z).$$

It is obvious that the second expert is the best one and our regret at time T is

$$\mathrm{Regret}(T) = \sum_{1 \leq t < \tau} c_t^\intercal x_t - \frac{\tau - 1}{2} + \sum_{\tau \leq t \leq T} c_t^\intercal x_t$$

It is also easy to check that

$$x_t = \begin{cases} [1/2, 1/2]^\intercal & 1 \leq t < \tau, t \text{ is odd} \\ \left[\dfrac{1}{1+\exp(\eta_{t-1})}, \dfrac{1}{1+\exp(-\eta_{t-1})}\right]^\intercal & 1 \leq t < \tau, t \text{ is even} \\ \left[\dfrac{1}{1+\exp(\eta_{t-1}(t-\tau))}, \dfrac{1}{1+\exp(-\eta_{t-1}(t-\tau))}\right]^\intercal & \tau \leq t \leq T \end{cases}$$

$$\mathrm{Regret}(T) = \underbrace{\sum_{1 \leq t < \tau, t \text{ is even}} \left(\frac{1}{1+\exp(-\eta_{t-1})} - \frac{1}{2}\right)}_{\text{Term 1}} + \underbrace{\sum_{t=\tau}^{T} \frac{1}{1+\exp((t-\tau)\eta_{t-1})}}_{\text{Term 2}}$$

For Term 1,

$$\begin{aligned} \text{Term 1} &\overset{\text{(i)}}{\geq} \sum_{1 \leq t < \tau, t \text{ is even}} \frac{1 - \exp(-\eta_{t-1})}{4} \\ &= \sum_{1 \leq t < \tau, t \text{ is even}} \frac{\eta_{i-1}}{4} + O(\eta_{i-1}^2) \\ &\overset{\text{(ii)}}{\geq} \frac{\sqrt{4\ln n}}{4}\sqrt{\tau} + o(\tau) \end{aligned}$$

where (i) is true by $\frac{1}{2-x} - \frac{1}{2} \geq \frac{x}{4}\ \forall x \in (0,1)$ and (ii) is true by using the fact that $\sum_{t=1}^{\tau} \frac{1}{\sqrt{t}} \geq 2\sqrt{\tau} - 2$

By the definition of τ, we have $\lim_{T\to\infty} \frac{\tau}{T} = 1$, thus

$$\lim_{T\to\infty} \frac{\text{Term 1}}{\sqrt{T\ln n}} = \frac{1}{2}. \tag{G.1}$$

For Term 2,

$$\begin{aligned}
\text{Term 2} &= \sum_{t=\tau}^{T} \frac{1}{1+\exp((t-\tau)\eta_{t-1})} \\
&\geq \sum_{t=\tau}^{T} \frac{1}{1+\exp((t-\tau)\sqrt{\frac{4\ln n}{t-1}})} \\
&\geq \sum_{t=\tau}^{T} \frac{1}{1+\exp((t-\tau)\sqrt{\frac{4\ln n}{\tau-1}})} \\
&\geq \int_{t=\tau}^{T} \frac{1}{1+\exp((t-\tau)\sqrt{\frac{4\ln n}{\tau-1}})} dt \\
&= \int_{y=0}^{\log(T)\sqrt{T}} \frac{1}{1+\exp(y\sqrt{\frac{4\ln n}{\tau-1}})} dy
\end{aligned} \tag{G.2}$$

Note that

$$\int \frac{1}{1+\exp(\beta y)} dy = y - \frac{\ln\left(1+e^{\beta y}\right)}{\beta}$$

Set $\beta = \sqrt{\frac{4\ln n}{\tau-1}}$ and plug the above result to Eq G.2, we get the following,

$$\text{Term 2} \geq \log(T)\sqrt{T} - \frac{\ln\left(1+\exp\left(\sqrt{\frac{4\ln n}{\tau-1}}\log(T)\sqrt{T}\right)\right)}{\sqrt{\frac{4\ln n}{\tau-1}}} + \frac{\ln 2}{\sqrt{\frac{4\ln n}{\tau-1}}}$$

Using the fact that $\ln(1+e^x) = x + o(x)$,

$$\text{Term 2} \geq \log(T)\sqrt{T} - \log(T)\sqrt{T} + o\left(\sqrt{\frac{4\ln n}{\tau-1}}\log(T)\sqrt{T}\right) + \ln 2\sqrt{\frac{(\tau-1)}{4\ln n}}$$

Note that $n = 2$, thus

$$\lim_{T\to\infty} \frac{\text{Term 2}}{\sqrt{T\ln n}} = \lim_{T\to\infty} \frac{\ln 2\sqrt{\frac{(\tau-1)}{4\ln 2}}}{\sqrt{T\ln 2}} = \frac{1}{2} \tag{G.3}$$

Combining Eq. G.1 and Eq. G.3, we conclude that

$$\lim_{T\to\infty} \frac{\text{Regret(T)}}{\sqrt{T\ln n}} = \lim_{T\to\infty} \frac{\text{Term 1}}{\sqrt{T\ln n}} + \lim_{T\to\infty} \frac{\text{Term 2}}{\sqrt{T\ln n}} \geq \frac{1}{2} + \frac{1}{2} = 1. \qquad \square$$

H. Aditional proofs for Section 6

At many points throughout this section we will need to talk about optimality condition for problems where we minimize a convex function over a convex set. Such conditions depend on the *normal cone* of the set on which the optimization is taking place.

Definition H.1. The normal cone to $C \subseteq \mathbb{R}^n$ at $x \in \mathbb{R}^n$ is the set $N_C(x) := \{s \in \mathbb{R}^n \mid \langle s, y-x\rangle \leq 0 \ \forall y \in C\}$.

Lemma H.2 ((Rockafellar, 1970, Theorem 27.4)). Let $h\colon \mathcal{C} \to \mathbb{R}$ be a closed convex function such that $(\operatorname{ri}\mathcal{C}) \cap (\operatorname{ri}\mathcal{X}) \neq \varnothing$. Then, $x \in \arg\min_{z\in\mathcal{X}} h(z)$ if and only if there is $\hat{g} \in \partial h(x)$ such that $-\hat{g} \in N_{\mathcal{X}}(x)$.

Using the above result allows us to derive a useful characterization of points that realize the Bregman projections.

Lemma H.3. Let $y \in \mathcal{D}$ and $x \in \bar{\mathcal{D}}$. Then $x = \Pi_{\mathcal{X}}^{\Phi}(y)$ if and only if $x \in \mathcal{D} \cap \mathcal{X}$ and $\nabla\Phi(y) - \nabla\Phi(x) \in N_{\mathcal{X}}(x)$.

Proof. Suppose $x \in \mathcal{D} \cap \mathcal{X}$ and $\nabla\Phi(y) - \nabla\Phi(x) \in N_{\mathcal{X}}(x)$. Since $\nabla\Phi(y) - \nabla\Phi(x) = -\nabla(D_{\Phi}(\cdot, y))(x)$, by Lemma H.2 we conclude that $x \in \arg\min_{z\in\mathcal{X}} D(z, y)$. Now suppose $x = \Pi_{\mathcal{X}}^{\Phi}(y)$. By Lemma H.2 together with the definition of Bregman divergence, this is the case if and only if there is $-g \in \partial\Phi(x)$ such that $-(g - \nabla\Phi(y)) \in N_{\mathcal{X}}(x)$. Since Φ is of Legendre type we have $\partial\Phi(z) = \varnothing$ for any $z \notin \mathcal{D}$ (see (Rockafellar, 1970, Theorem 26.1)). Thus, $x \in \mathcal{D}$ and $g = \nabla\Phi(x)$ since Φ is differentiable. Finally, $x \in \mathcal{X}$ by the definition of Bregman projection. □

Before proceding to the proof of the results from Section 6, we need to state on last result about the relation of subgradients and conjugate functions.

Lemma H.4 ((Rockafellar, 1970, Theorem 23.5)). Let $f\colon \mathcal{X} \to \mathbb{R}$, let $x \in \mathcal{X}$ and let $\hat{y} \in \mathbb{R}^n$. Then $\hat{y} \in \partial f(x)$ if and only if x attains $\sup_{x\in\mathbb{R}^n}(\langle \hat{y}, x\rangle - f(x)) = f^*(\hat{y})$.

Proof (of Proposition 6.1). Let $t \geq 1$ and let $F_t\colon \mathcal{D} \to \mathbb{R}$ be the function being minimized on the right-hand side of (6.1). By definition we have $x_{t+1} = \Pi_{\mathcal{X}}^{\Phi}(y_{t+1})$. Using the optimality conditions of the Bregman projection, we have

$$x_{t+1} = \Pi_{\mathcal{X}}^{\Phi}(y_{t+1}) \iff \hat{y}_{t+1} - \hat{x}_{t+1} \in N_{\mathcal{X}}(x_{t+1}), \quad \text{(by Lemma H.3)}$$

By further using the definitions from Algorithm 2 we get

$$\begin{aligned}
\hat{y}_{t+1} - \hat{x}_{t+1} &= \gamma_t(\hat{x}_t - \eta_t\hat{g}_t) + (1-\gamma_t)\hat{x}_1 - \hat{x}_{t+1} \\
&= \gamma_t(\hat{x}_t - \hat{x}_{t+1} - \eta_t\hat{g}_t) + (1-\gamma_t)(\hat{x}_1 - \hat{x}_{t+1}) \\
&= -\gamma_t\big(\nabla(D_\Phi(\cdot, x_t))(x_{t+1}) + \eta_t\hat{g}_t\big) - (1-\gamma_t)\nabla(D_\Phi(\cdot, x_1))(x_{t+1}) \\
&= -\nabla F_t(x_{t+1})
\end{aligned}$$

Thus, we have $-\nabla F_t(x_{t+1}) \in N_{\mathcal{X}}(x_{t+1})$. By the optimality conditions from Lemma H.2 we conclude that $x_{t+1} \in \arg\min_{x\in\mathcal{X}} F_t(x)$, as desired. □

Theorem 6.2 is an easy consequence of the following proposition.

Proposition H.5. Let $\{f_t\}_{t\geq 1}$ with $f_t : \mathcal{X} \to \mathbb{R}$ be a sequence of convex functions and let $\eta\colon \mathbb{N} \to \mathbb{R}_{>0}$ be non-increasing. Let $\{x_t\}_{t\geq 1}$ and $\{\hat{g}_t\}_{t\geq 1}$ be as in Algorithm 2. Define $\gamma^{[i,t]} := \prod_{j=i}^{t} \gamma_j$ for every $i, t \in \mathbb{N}$. Then, there are $\{p_t\}_{t\geq 1}$ with $p_t \in N_X(x_t)$ for each $t \geq 1$ such that

$$\{x_{t+1}\} = \arg\min_{x\in\mathcal{X}}\Big(\sum_{i=1}^{t} \gamma^{[i,t]}\langle \eta_i\hat{g}_i + p_i, x\rangle - \Big(\gamma^{[1,t]} + \sum_{i=1}^{t}\gamma^{[i+1,t]}(1-\gamma_i)\Big)\langle \hat{x}_1, x\rangle + \Phi(x)\Big), \qquad \forall t \geq 0. \tag{H.1}$$

Proof. First of all, in order to prove (H.1) we claim it suffices to prove that there are $\{p_t\}_{t\geq 1}$ with $p_t \in N_X(x_t)$ for each $t \geq 1$ such that

$$\hat{y}_{t+1} = -\sum_{i=1}^{t} \gamma^{[i,t]}(\eta_i\hat{g}_i + p_i) + \Big(\gamma^{[1,t]} + \sum_{i=1}^{t}\gamma^{[i+1,t]}(1-\gamma_i)\Big)\hat{x}_1, \qquad \forall t \geq 0. \tag{H.2}$$

To see the sufficiency of this claim, note that

$$\begin{aligned}
x_{t+1} = \Pi_{\mathcal{X}}^{\Phi}(y_{t+1})) &\iff \hat{y}_{t+1} - \hat{x}_{t+1} \in N_{\mathcal{X}}(x_{t+1}) && \text{(Lemma H.3)} \\
&\iff \hat{y}_{t+1} \in \partial(\Phi + \delta(\cdot \,|\, \mathcal{X}))(x_{t+1}) && (\partial(\delta(\cdot \,|\, \mathcal{X}))(x) = N_{\mathcal{X}}(x)) \\
&\iff x_{t+1} \in \arg\max_{x\in\mathbb{R}^n}\big(\langle \hat{y}_{t+1}, x\rangle - \Phi(x) - \delta(x \,|\, \mathcal{X})\big) && \text{(Lemma H.4)} \\
&\iff x_{t+1} \in \arg\min_{x\in\mathcal{X}}\big(-\langle \hat{y}_{t+1}, x\rangle + \Phi(x)\big).
\end{aligned}$$

The above together with (H.2) yields (H.1). Let us now prove (H.2) by induction on $t \geq 0$.

For $t = 0$ (H.2) holds trivially. Let $t > 0$. By definition, we have $\hat{y}_{t+1} = (1-\gamma_t)(\hat{x}_t - \eta_t \hat{g}_t) + \gamma_t \hat{x}_1$. At this point, to use the induction hypothesis, we need to write $\hat{x}_t$ in function of $\hat{y}_t$. From the definition of Algorithm 2, we have $x_t = \Pi_{\mathcal{X}}^{\Phi}(y_t)$. By Lemma H.3, the latter holds if and only if $\hat{y}_t - \hat{x}_t \in N_{\mathcal{X}}(x_t)$. That is, there is $p_t \in N_{\mathcal{X}}(x_t)$ such that $\hat{x}_t = \hat{y}_t - p_t$. Plugging these facts together and using our induction hypothesis we have

$$\begin{aligned}
\hat{y}_{t+1} &= \gamma_t(\hat{x}_t - \eta_t \hat{g}_t) + (1-\gamma_t)\hat{x}_1 = \gamma_t(\hat{y}_t - \eta_t \hat{g}_t - p_t) + (1-\gamma_t)\hat{x}_1 \\
&\overset{\text{I.H.}}{=} \gamma_t\Big(-\sum_{i=1}^{t-1} \gamma^{[i,t-1]}(\eta_i \hat{g}_i + p_i) - \eta_t \hat{g}_t - p_t + \Big(\gamma^{[1,t-1]} + \sum_{i=1}^{t-1} \gamma^{[i+1,t-1]}(1-\gamma_i)\Big)\hat{x}_1\Big) + (1-\gamma_t)\hat{x}_1 \\
&= -\sum_{i=1}^{t} \gamma^{[i,t]}(\eta_i \hat{g}_i + p_i) + \Big(\gamma^{[1,t]} + \sum_{i=1}^{t} \gamma^{[i+1,t]}(1-\gamma_i)\Big)\hat{x}_1,
\end{aligned}$$

and this finishes the proof of (H.2). □

Proof (of Theorem 6.2). Define $\gamma^{[i,t]}$ for every $i, t \in \mathbb{N}$ as in Proposition H.5. If $\gamma_t = 1$ for all $t \geq 1$, then $\gamma^{[i,t]} = 1$ for any $t, i \geq 1$. Moreover, if $\gamma_t = \frac{\eta_{t+1}}{\eta_t}$ for every $t \geq 1$, then for every $t, i \in \mathbb{N}$ with $t \geq i$ we have $\gamma^{[i,t]} = \frac{\eta_{t+1}}{\eta_i}$, which yields $\gamma^{[i,t]}(\eta_i \hat{g}_i + p_i) = \eta_t \hat{g}_i + \frac{1}{\eta_i} p_i$ and

$$\gamma^{[1,t]} + \sum_{i=1}^{t} \gamma^{[i+1,t]}(1-\gamma_i) = \frac{\eta_{t+1}}{\eta_1} + \sum_{i=1}^{t} \frac{\eta_{t+1}}{\eta_{i+1}}\Big(1 - \frac{\eta_{i+1}}{\eta_i}\Big) = \frac{\eta_{t+1}}{\eta_1} + \eta_{t+1} \sum_{i=1}^{t}\Big(\frac{1}{\eta_{i+1}} - \frac{1}{\eta_i}\Big) = 1.$$

□

I. Dual Stabilized OMD with Composite Functions

In this section we extend the Dual-Stabilized OMD to the case where the functions revealed at each round are composite (Xiao, 2010; Duchi et al., 2010). More specifically, at each round $t \geq 1$ we see a function of the form $f_t + \Psi$, where $f_t \colon \mathcal{X} \to \mathbb{R}$ is convex and Lipschitz continuous and Ψ is a fixed convex function which we assume to be "easy", i.e., such that we know how to efficiently compute points in $\arg\min_{x \in \mathcal{X}}(D_\Phi(x, \bar{x}) + \Psi(x))$. In fact, we could simply use the original Dual-Stabilized OMD in this setting, but this approach has some drawbacks. One issue is that subgradients of Ψ would end-up appearing on the regret bound from Theorem 4.1, which is not ideal: we want bounds which are unaffected by the "easy" function Ψ. Another drawback is that we would not be using the knowledge of the structure of the functions, which is in many cases sub-optimal. For example, in an online learning problem one may artificially add the ℓ_1-norm to each function in order to induce sparsity on the iterates of the algorithm. However, using subgradients of the ℓ_1-norm instead of the norm itself does not induce sparsity (McMahan, 2011). Finally, the analysis of Dual-Stabilized OMD adapted to the composite setting that we develop in this section is an easy extension of the original analysis of Section 4. This is interesting since in the literature the algorithms for the composite setting usually require a more intricate analysis, such as in the analysis of Regularized DA from (Xiao, 2010), or the use of powerful results, such as the duality between strong convexity and strong smoothness used in (McMahan, 2017). An important exception is the analysis of the composite objective mirror descent from (Duchi et al., 2010) is already elegant. Sill, it does not directly applies when we use dual-stabilization and uses proof techniques more specially-tailored for the proximal step form of writing OMD.

In the composite setting we assume without loss of generality that $\mathcal{X} = \mathbb{R}^n$ since we may substitute Ψ by $\Psi + \delta(\cdot \,|\, \mathcal{X})$ where $\delta(x \,|\, \mathcal{X}) = 0$ if $x \in \mathcal{X}$ and is $+\infty$ anywhere else. To avoid comparing the loss of the algorithm with points outside of the effective domain of Ψ in the definition of regret, denoted by $\operatorname{dom}\Psi$, we define the Ψ-regret of the sequence of functions $\{f_t\}_{t \geq 1}$ (against a comparison point $z \in \operatorname{dom}\Psi$) by and iterates $\{x_t\}_{t \geq 1}$ by

$$\operatorname{Regret}^{\Psi}(T, z) \coloneqq \sum_{t=1}^{T}\big(f_t(x_t) + \Psi(x_t)\big) - \sum_{t=1}^{T}\big(f_t(z) + \Psi(z)\big), \quad \forall T \geq 0.$$

To adapt the Dual Stabilization method to this setting, we use the same idea as in (Duchi et al., 2010). Namely, we modify the proximal-like formulation of Dual Stabilization from Proposition 6.1 so that we do not linearize (i.e., take the subgradient) of the function Ψ, which yields

$$\{x_{t+1}\} \coloneqq \arg\min_{x \in \mathcal{X}}\Big(\gamma_t\big(\eta_t(\langle \hat{g}_t, x\rangle + \Psi(x)) + D_\Phi(x, x_t)\big) + (1-\gamma_t)D_\Phi(x, x_1)\Big), \quad \forall t \geq 0.$$

Algorithm 2 Dual-stabilized OMD with dynamic learning rate η_t and additional regularization function Ψ.

```
Input: $x_1 \in \arg\min_{x \in \mathbb{R}^n} \Psi(x),\ \eta : \mathbb{N} \to \mathbb{R}_+,\ \gamma : \mathbb{N} \to [0,1]$
$\hat{y}_1 = \nabla\Phi(x_1)$
for $t = 1, 2, \ldots$ do
    Incur cost $f_t(x_t)$ and receive $\hat{g}_t \in \partial f_t(x_t)$
    $\hat{x}_t = \nabla\Phi(x_t)$                                        ▷ map primal iterate to dual space
    $\hat{w}_{t+1} = \hat{x}_t - \eta_t \hat{g}_t$                        ▷ gradient step in dual space        (I.1)
    $\hat{y}_{t+1} = \gamma_t \hat{w}_{t+1} + (1 - \gamma_t)\hat{x}_1$    ▷ stabilization in dual space        (I.2)
    $y_{t+1} = \nabla\Phi^*(\hat{y}_{t+1})$                               ▷ map dual iterate to primal space
    $\alpha_{t+1} := \eta_t \gamma_t$                                      ▷ compute scaling factor for $\Psi$
    $x_{t+1} = \Pi^{\Phi}_{\alpha_{t+1}\Psi}(y_{t+1})$                    ▷ project onto feasible region       (I.3)
end for
```

Although we do not prove the equivalence for the sake of conciseness, on Algorithm 2 we present the above procedure written in a form closer to Algorithm 2. In this new algorithm, we extend the definition of Bregman Projection and define the Ψ-Bregman projection by $\{\Pi^{\Phi}_{\Psi}(y)\} := \arg\min_{x \in \mathbb{R}^n}(D_\Phi(x, y) + \Psi(y))$. The next lemma shows an analogue of the generalized pythagorean theorem for the Ψ-Bregman Projection.

Lemma I.1. Let $\alpha > 0$ and $\bar{y} := \Pi^{\Phi}_{\alpha\Psi}(y)$. Then,

$$D_\Phi(x, \bar{y}) + D_\Phi(\bar{y}, y) \leq D_\Phi(x, y) + \alpha(\Psi(x) - \Psi(\bar{y})), \qquad \forall x \in \mathbb{R}^n.$$

Proof. By the optimality conditions of the projection, we have $\nabla\Phi(y) - \nabla\Phi(\bar{y}) \in \partial(\alpha\Psi)(\bar{y})$. Using the generalized triangle inequality for Bregman Divergences (Lemma A.8) and the subgradient inequality, we get

$$D_\Phi(x, \bar{y}) + D_\Phi(\bar{y}, y) - D_\Phi(x, y) = \langle \nabla\Phi(y) - \nabla\Phi(\bar{y}), x - \bar{y} \rangle \overset{\text{(i)}}{\leq} \alpha(\Psi(x) - \Psi(\bar{y})).$$

where (i) follows from $\nabla\Phi(y) - \nabla\Phi(\bar{y}) \in \partial(\alpha\Psi)(\bar{y})$ and the convexity of $\alpha\Psi(\cdot)$. □

In the next theorem we show that the regret bound we have for the Dual-Stabilized OMD still holds in this setting when using Algorithm 2, and the proof boils down to simple modifications to the original proof of Theorem 4.1.

Theorem I.2. Suppose that Φ is ρ-strongly convex with respect to a norm $\|\cdot\|$ and Ψ be a convex function such that $\operatorname{dom}\Psi \subseteq \mathcal{X}$. Set $\gamma_t = \eta_{t+1}/\eta_t$ for each $t \geq 1$. Let $x_1 \in \arg\min_{x \in \operatorname{dom}\Psi} \Psi(x)$ and $\{x_t\}_{t \geq 1}$ be the sequence of iterates generated by Algorithm 2. Then for any sequence of convex functions $\{f_t\}_{t \geq 1}$ with each $f_t \colon \mathcal{X} \to \mathbb{R}$ and $z \in \mathcal{X}$,

$$\operatorname{Regret}^{\Psi}(T, z) \;\leq\; \sum_{t=1}^{T} \frac{D_\Phi\big(\substack{x_t \\ x_{t+1}}; w_{t+1}\big)}{\eta_t} + \frac{D_\Phi(z, x_1)}{\eta_{T+1}}, \quad \forall T > 0. \tag{I.4}$$

Proof. Let $z \in \operatorname{dom}\Psi$ and $t \in \mathbb{N}$.

By (4.8), we have

$$f_t(x_t) + \Psi(x_t) - f_t(z) - \Psi(z) \leq \frac{1}{\eta_t}\Big(D_\Phi(x_t, w_{t+1}) + D_\Phi(z, x_t) - D_\Phi\big(z, w_{t+1}\big)\Big) + \Psi(x_t) - \Psi(z). \tag{I.5}$$

As in Theorem 4.1, let us prove bound the above expression by something with telescoping terms.

Claim I.3. Assume that $\gamma_t = \eta_{t+1}/\eta_t \in (0, 1]$. Then

$$\text{(I.5)} \;\leq\; \frac{D_\Phi\big(\substack{x_t \\ x_{t+1}}; w_{t+1}\big)}{\eta_t} + \underbrace{\Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)}_{\text{telescopes}} D_\Phi(z, x_1) + \underbrace{\frac{D_\Phi(z, x_t)}{\eta_t} - \frac{D_\Phi(z, x_{t+1})}{\eta_{t+1}}}_{\text{telescopes}} + \underbrace{\Psi(x_t) - \Psi(x_{t+1})}_{\text{telescopes}}.$$

Proof. Fix $z \in \operatorname{dom}\Psi$. First we derive the inequality

$$\begin{aligned}
&\gamma_t\big(D_\Phi(z,w_{t+1}) - D_\Phi(x_{t+1},w_{t+1})\big) \;+\; (1-\gamma_t)D_\Phi(z,x_1) \\
\geq\;& \gamma_t D_\Phi\big(\genfrac{}{}{0pt}{}{z}{x_{t+1}};w_{t+1}\big) \;+\; (1-\gamma_t)D_\Phi\big(\genfrac{}{}{0pt}{}{z}{x_{t+1}};x_1\big) && \text{(since } D_\Phi(x_{t+1},x_1) \geq 0 \text{ and } \gamma_t \leq 1\text{)} \\
=\;& D_\Phi\big(\genfrac{}{}{0pt}{}{z}{x_{t+1}};y_{t+1}\big) && \text{(by Proposition A.13 and (I.2))} \\
\geq\;& D_\Phi(z,x_{t+1}) + \alpha_{t+1}\big(\Psi(x_{t+1}) - \Psi(z)\big) && \text{(by Lemma I.1 and (I.3)).}
\end{aligned}$$

Rearranging and using both $\gamma_t > 0$ and $\alpha_{t+1} = \eta_t\gamma_t$ yields

$$D_\Phi(z,w_{t+1}) \;\geq\; D_\Phi(x_{t+1},w_{t+1}) - \Big(\frac{1}{\gamma_t} - 1\Big)D_\Phi(z,x_1) + \frac{1}{\gamma_t}D_\Phi(z,x_{t+1}) + \eta_t\big(\Psi(x_{t+1}) - \Psi(z)\big). \tag{I.6}$$

Plugging this into (I.5) yields

$$\begin{aligned}
\text{(I.5)} \;=\;& \frac{1}{\eta_t}\Big(D_\Phi(x_t,w_{t+1}) - D_\Phi(z,w_{t+1}) + D_\Phi(z,x_t)\Big) + \Psi(x_t) - \Psi(z) \\
\leq\;& \frac{1}{\eta_t}\left(D_\Phi\big(\genfrac{}{}{0pt}{}{x_t}{x_{t+1}};w_{t+1}\big) + \Big(\frac{1}{\gamma_t} - 1\Big)D_\Phi(z,x_1) - \frac{1}{\gamma_t}D_\Phi(z,x_{t+1}) + D_\Phi(z,x_t)\right) + \Psi(x_t) - \Psi(x_{t+1}).
\end{aligned}$$

The claim follows by the definition of γ_t. □

The final step is very similar to the standard OMD proof. Summing (I.5) over t and using Claim I.3 leads to the desired telescoping sum.

$$\begin{aligned}
\sum_{t=1}^{T}\big(f_t(x_t) - f_t(z)\big) \;\leq\;& \sum_{t=1}^{T}\left(\frac{D_\Phi\big(\genfrac{}{}{0pt}{}{x_t}{x_{t+1}};w_{t+1}\big)}{\eta_t} \;+\; \Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)D_\Phi(z,x_1) + \frac{D_\Phi(z,x_t)}{\eta_t} - \frac{D_\Phi(z,x_{t+1})}{\eta_{t+1}} \right. \\
&\left. \qquad + \; \Psi(x_t) - \Psi(x_{t+1})\right) \\
\leq\;& \sum_{t=1}^{T}\frac{D_\Phi\big(\genfrac{}{}{0pt}{}{x_t}{x_{t+1}};w_{t+1}\big)}{\eta_t} \;+\; \frac{D_\Phi(z,x_1)}{\eta_{T+1}} + \Psi(x_1) - \Psi(x_{T+1}) \\
\leq\;& \sum_{t=1}^{T}\frac{D_\Phi\big(\genfrac{}{}{0pt}{}{x_t}{x_{t+1}};w_{t+1}\big)}{\eta_t} \;+\; \frac{D_\Phi(z,x_1)}{\eta_{T+1}},
\end{aligned}$$

where in the last step inequality we used $x_1 \in \arg\min_{x\in\mathcal{X}}\Psi(x)$.

□

References

Auer, P., Cesa-Bianchi, N., and Gentile, C. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

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

Bubeck, S. Introduction to online optimization, December 2011. unpublished.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Duchi, J. C., Shalev-shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. In *Proceedings of COLT*, pp. 14–26, 2010.

Gerchinovitz, S. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Université Paris-Sud, 2011.

McMahan, H. B. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of AISTATS'11*, pp. 525–533, 2011.

McMahan, H. B. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18: 90:1–90:50, 2017.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, 1970.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11, 2010.

Zhang, X. Bregman divergence and mirror descent lecture notes, n.d. Available at `http://users.cecs.anu.edu.au/~xzhang/teaching/bregman.pdf`. Last accessed May 7, 2019.