A Greedy Framework for First-Order Optimization

Jacob Steinhardt[*]
Department of Computer Science
Stanford University
Stanford, CA 94305
jsteinhardt@cs.stanford.edu

Jonathan Huggins[*]
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139
jhuggins@mit.edu

Introduction. Recent work has shown many connections between conditional gradient and other first-order optimization methods, such as herding [3] and subgradient descent [2]. By considering a type of *proximal conditional method*, which we call boosted mirror descent (BMD), we are able to unify all of these algorithms into a single framework, which can be interpreted as taking successive arg-mins of a sequence of surrogate functions. Using a standard online learning analysis based on Bregman divergences, we are able to demonstrate an $O(1/T)$ convergence rate for all algorithms in this class.

Setup. More concretely, suppose that we are given a function $L : U \times \Theta \to \mathbb{R}$ defined by

$$L(u, \theta) = h(u) + \langle u, \theta \rangle - R(\theta) \tag{1}$$

and wish to find the *saddle point*

$$L_* \overset{\text{def}}{=} \min_u \max_\theta L(u, \theta). \tag{2}$$

We should think of u as the primal variables and θ as the dual variables; we will assume throughout that h and R are both convex. We will also abuse notation and define $L(u) \overset{\text{def}}{=} \max_\theta L(u, \theta)$; we can equivalently write $L(u)$ as

$$L(u) = h(u) + R^*(u), \tag{3}$$

where R^* is the Fenchel conjugate of R. Note that $L(u)$ is a convex function. Moreover, since $R \leftrightarrow R^*$ is a one-to-one mapping, for *any* convex function L and *any* decomposition of L into convex functions h and R^*, we get a corresponding two-argument function $L(u, \theta)$.

Primal algorithm. Given the function $L(u, \theta)$, we define the following optimization procedure, which will generate a sequence of points $(u_1, \theta_1), (u_2, \theta_2), \dots$ converging to a saddle point of L. First, take a sequence of weights $\alpha_1, \alpha_2, \dots$, and for notational convenience define

$$\hat{u}_t = \frac{\sum_{s=1}^t \alpha_s u_s}{\sum_{s=1}^t \alpha_s} \quad \text{and} \quad \hat{\theta}_t = \frac{\sum_{s=1}^t \alpha_s \theta_s}{\sum_{s=1}^t \alpha_s}.$$

Then the *primal boosted mirror descent* (PBMD) algorithm is the following iterative procedure:

1. $u_1 \in \arg\min_u h(u)$
2. $\theta_t \in \arg\max_{\theta \in \Theta} \langle \theta, u_t \rangle - R(\theta) = \partial R^*(u_t)$
3. $u_{t+1} \in \arg\min_u h(u) + \langle \hat{\theta}_t, u \rangle = \partial h^*(-\hat{\theta}_t)$

As long as h is strongly convex, for the proper choice of α_t we obtain the bound (see Corollary 2):

$$\sup_{\theta \in \Theta} L(\hat{u}_T, \theta) \leq L_* + O(1/T). \tag{4}$$

As an example, suppose that we are given a γ-strongly convex function f: that is, $f(x) = \frac{\gamma}{2}\|x\|_2^2 + f_0(x)$, where f_0 is convex. Then we let $h(x) = \frac{\gamma}{2}\|x\|_2^2$, $R^*(x) = f_0(x)$, and obtain the updates:

[*] Both authors contributed equally to this work.

1. $u_1 = 0$
2. $\theta_t = \partial f_0(u_t)$
3. $u_{t+1} = -\frac{1}{\gamma}\hat{\theta}_t = -\frac{\sum_{s=1}^t \alpha_s \partial f_0(u_s)}{\gamma \sum_{s=1}^t \alpha_s}$

We therefore obtain a variant on subgradient descent where u_{t+1} is a weighted average of the first t subgradients (times a step size $1/\gamma$). Note that these are the subgradients of f_0, which are related to the subgradients of f by $\partial f_0(x) = \partial f(x) - \gamma x$.

Dual algorithm. We can also consider the dual form of our mirror descent algorithm (*dual boosted mirror descent*, or DBMD):

1. $\theta_1 \in \arg\min_\theta R(\theta)$
2. $u_t \in \arg\min_u h(u) + \langle \theta_t, u\rangle = \partial h^*(-\theta_t)$
3. $\theta_{t+1} \in \arg\max_{\theta\in\Theta} \langle \theta, \hat{u}_t\rangle - R(\theta) = \partial R^*(\hat{u}_t)$

Convergence now hinges upon strong convexity of R rather than h, and we obtain the same $1/T$ convergence rate (see Corollary 4):

$$\sup_{\theta\in\Theta} L(\hat{u}_T, \theta) \leq L_* + O(1/T). \tag{5}$$

An important special case is $h(u) = \begin{cases} 0 & : & u \in K \\ \infty & : & u \notin K \end{cases}$, where K is some compact set. Also let $R^* = f$, where f is an arbitrary strongly convex function. Then we are minimizing f over the compact set K, and we obtain the following updates, which are equivalent to (standard) conditional gradient or Frank-Wolfe:

1. $\theta_1 = \partial f(0)$
2. $u_t \in \arg\min_{u\in K} \langle \theta_t, u\rangle$
3. $\theta_{t+1} = \partial f(\hat{u}_t)$

Our notation is slightly different from previous presentations in that we use linear weights (α_t) instead of geometric weights (often denoted ρ_t, as in [2]). However, under the mapping $\alpha_t = \rho_t / \prod_{s=1}^t (1-\rho_s)$, we obtain an algorithm equivalent to the usual formulation.

Optimality. The solutions generated by conditional gradient have an attractive sparsity property: the solution after the t-th iteration is a convex combination of t extreme points of the optimized space. In [4] it is shown that conditional gradient is asymptotically optimal amongst algorithms with sparse solutions. This suggests that the $1/T$ convergence rate obtained in this paper cannot be improved without further assumptions. For further discussion of the properties of conditional gradient, see Jaggi [4].

Primal vs. dual algorithms. It is worth taking a moment to contrast the primal and dual algorithms. Note that *both* algorithms have a primal convergence guarantee (i.e. both guarantee that $\hat{u}_T$ is close to optimal). The sense in which the latter algorithm is a "dual" algorithm is that it hinges on strong convexity of the dual, as opposed to strong convexity of the primal. If we care about dual convergence instead of primal convergence, the only change we need to make is to take $\hat{\theta}_T$ at the end of the algorithm instead of $\hat{u}_T$. (This can be seeing by defining the loss function $L_2(\theta, u) \overset{\text{def}}{=} -L(u, \theta)$, which inverts the role of u and θ but yields the same updates as above.)

Discussion. Our framework is intriguing in that it involves a purely greedy minimization of surrogate loss functions (alternating between the primal and dual), and yet is powerful enough to capture both (a variant of) subgradient descent and conditional gradient descent, as well as a host of other first-order methods.

An example of a more complex first-order method captured by our framework is the low-rank SDP solver introduced by Arora, Hazan, and Kale [1]. Briefly, the AHK algorithm seeks to minimize

$\sum_{j=1}^{m} \frac{1}{2}(\mathrm{Tr}(A_j^T X) - b_j)^2$ subject to the constraints $X \succeq 0$ and $\mathrm{Tr}(X) \leq \rho$.[1] We can then define

$$h(X) = \begin{cases} 0 & : \quad X \succeq 0 \text{ and } \mathrm{Tr}(X) \leq \rho \\ \infty & : \quad \text{else} \end{cases} \tag{6}$$

and

$$R^*(X) = \sum_{j=1}^{m} \frac{1}{2}(\mathrm{Tr}(A_j^T X) - b_j)^2. \tag{7}$$

Note that this decomposition is actually a special case of the conditional gradient decomposition above, and so we obtain the updates

$$X_{t+1} \in \mathrm{argmin}_{X \succeq 0, Tr(X) \leq \rho} \sum_{j=1}^{m} \left[\mathrm{Tr}(A_j^T \hat{X}_t) - b_j\right] \mathrm{Tr}(A_j^T X), \tag{8}$$

whose solution turns out to be $\rho v v^T$, where v is the top singular vector of the matrix $-\sum_{j=1}^{m} \left[\mathrm{Tr}(A_j^T \hat{X}_t) - b_j\right] A_j$. This example serves both to illustrate the flexibility of our framework and to highlight the interesting fact that the Arora-Hazan-Kale SDP algorithm is a special case of conditional gradient (to get the original formulation in [1], we need to replace the function $\frac{1}{2}x^2$ with $x_+ \log x_+$, where $x_+ = \max(x, 0)$). It is worth noting that the analysis in [1] appears to be tighter in their special case than the analysis we give here. We plan to investigate this phenomenon in future work.

Related work. A generalized Frank-Wolfe algorithm has also been proposed in [5], which considers first-order convex surrogate functions more generally. Algorithm 3 of [5] turns out to be equivalent to a line-search formulation of the algorithm we propose in this paper. However, the perspective is quite different. While both algorithms share the intuition of trying to construct a "proximal Frank-Wolfe" algorithm, the general framework of [5] studies convex *upper bounds*, rather than the convex *lower bounds* of this paper (although for the "proximal gradient surrogates" that they suggest using, their Frank-Wolfe algorithm does indeed deal with lower bounds). Another difference is that their analysis focuses on the Lipschitz properties of the objective function, whereas we use the more general notion of strong convexity with respect to an arbitrary norm. We believe that this geometric perspective is useful for tailoring the proximal function to specific applications, as demonstrated in our discussion of q-herding below. Finally, the saddle point perspective on Frank-Wolfe is not present in [5], although they do discuss the general idea of saddle-point methods in relation to convex surrogates.

q-herding. In addition to unifying several existing methods, our framework allows us to extend herding to an algorithm we call *q-herding*. Herding is an algorithm for constructing pseudosamples that match a specified collection of moments from a distribution; it was originally introduced by Welling [6] and was shown to be a special case of conditional gradient by Bach et al. [3]. Let $\mathcal{M}_{\mathcal{X}}$ be the probability simplex over $\mathcal{X}$. Herding can be cast as trying to minimize $\|\mathbb{E}_\mu[\phi(x)] - \bar{\phi}\|_2^2$ over $\mu \in \mathcal{M}_{\mathcal{X}}$, for a given $\phi : \mathcal{X} \to \mathbb{R}^d$ and $\bar{\phi} \in \mathbb{R}^d$. As shown in [3], the herding updates are equivalent to DBMD with $h(\mu) \equiv 0$ and $R(\theta) = \theta^T \bar{\phi} + \frac{1}{2}\|\theta\|_2^2$.

The implicit assumption in the herding algorithm is that $\|\phi(x)\|_2$ is bounded. We are able to construct a more general algorithm that only requires $\|\phi(x)\|_p$ to be bounded for some $p \geq 2$. This *q-herding* algorithm works by taking $R(\theta) = \theta^T \bar{\phi} + \frac{1}{q}\|\theta\|_q^q$, where $\frac{1}{p} + \frac{1}{q} = 1$. In this case our convergence results imply that $\|\mathbb{E}_\mu[\phi(x)] - \bar{\phi}\|_p^p$ decays at a rate of $O(1/T)$ (proof to appear in an extended version of this paper).

Convergence results. We end by stating our formal convergence results. Proofs are given in the Appendix. For the primal algorithm (PBMD) we have:

Theorem 1. *Suppose that h is strongly convex with respect to a norm $\|\cdot\|$ and let $r = \sup_\theta \|\theta\|_*$. Then, for any u^*,*

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{2r^2}{A_T} \sum_{t=1}^{T} \frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2}. \tag{9}$$

[1] This is actually a variant of their algorithm, which we present for ease of exposition.

Corollary 2. *Under the hypotheses of Theorem 1, for $\alpha_t = 1$ we have*

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{2r^2(\log(T)+1)}{T}. \tag{10}$$

and for $\alpha_t = t$ we have

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{8r^2}{T}. \tag{11}$$

Similarly, for the dual algorithm (DBMD) we have:

Theorem 3. *Suppose that R is strongly convex with respect to a norm $\|\cdot\|$ and let $r = \sup_u \|u\|_*$. Then, for any u^*,*

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{2r^2}{A_T} \sum_{t=1}^{T} \frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2}. \tag{12}$$

Corollary 4. *Under the hypotheses of Theorem 3, for $\alpha_t = 1$ we have*

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{2r^2(\log(T)+1)}{T} \tag{13}$$

and for $\alpha_t = t$ we have

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{8r^2}{T} \tag{14}$$

Thus, a step size of $\alpha_t = t$ yields the claimed $O(1/T)$ convergence rate.

Acknowledgements. Thanks to Simon Lacoste-Julien, Martin Jaggi, Percy Liang, Arun Chaganty, and Sida Wang for helpful discussions. Thanks also to the anonymous reviewer for suggesting improvements to the paper. This work was supported by the Hertz Foundation and by the NDSEG graduate fellowship.

References

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *FOCS*, pages 339–348, 2005.

[2] F Bach. Duality between subgradient and conditional gradient methods. *arXiv.org*, November 2012.

[3] F Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the Equivalence between Herding and Conditional Gradient Algorithms. In *ICML*. INRIA Paris - Rocquencourt, LIENS, March 2012.

[4] Martin Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *ICML*, pages 427–435, 2013.

[5] Julien Mairal. Optimization with First-Order Surrogate Functions . In *ICML*, 2013.

[6] Max Welling. Herding dynamical weights to learn. In *ICML*, 2009.

Appendix: Convergence Proofs

We now prove the convergence results given above. Throughout this section, assume that $\alpha_1, \ldots, \alpha_T$ is a sequence of real numbers and that $A_t = \sum_{s=1}^t \alpha_s$. We further require that $A_t > 0$ for all t.

Also recall that the Bregman divergence is defined by $D_f(x_2 \| x_1) \stackrel{\text{def}}{=} f(x_2) - \langle \partial f(x_1), x_2 - x_1 \rangle - f(x_1)$.

Our proofs hinge on the following key lemma:

Lemma 5. *Let $z_1, \ldots, z_T$ be vectors and let $f(x)$ be a strictly convex function. Define $\hat{z}_t$ to be $\frac{1}{A_t} \sum_{s=1}^t \alpha_s z_s$. Let $x_1, \ldots, x_T$ be chosen via $x_{t+1} = \arg\min_x f(x) + \langle \hat{z}_t, x \rangle$. Then for any x^* we have*

$$\frac{1}{A_T} \sum_{t=1}^T \{\alpha_t (f(x_t) + \langle z_t, x_t \rangle)\} \leq f(x^*) + \langle \hat{z}_t, x^* \rangle + \frac{1}{A_T} \sum_{t=1}^T A_t D_f(x_t \| x_{t+1}).$$

Proof. First note that, if $x_0 = \arg\min f(x) + \langle z, x \rangle$, then $\partial f(x_0) = -z$.

Now note that

$$\begin{aligned} \alpha_t z_t &= A_t \hat{z}_t - A_{t-1} \hat{z}_{t-1} && (15) \\ &= -A_t \partial f(x_{t+1}) + A_{t-1} \partial f(x_t), && (16) \end{aligned}$$

so we have

$$\begin{aligned} &\sum_{t=1}^T \{\alpha_t (f(x_t) + \langle z_t, x_t \rangle)\} && (17) \\ &= \sum_{t=1}^T \{\alpha_t f(x_t) + \langle A_{t-1} \partial f(x_t) - A_t \partial f(x_{t+1}), x_t \rangle\} && (18) \\ &= \sum_{t=1}^T \{\alpha_t f(x_t) - \langle A_t \partial f(x_{t+1}), x_t - x_{t+1} \rangle\} && (19) \\ &\quad - A_T \langle \partial f(x_{T+1}), x_{T+1} \rangle && (20) \\ &= \sum_{t=1}^T \{A_t f(x_t) - \langle A_t \partial f(x_{t+1}), x_t - x_{t+1} \rangle - A_t f(x_{t+1})\} && (21) \\ &\quad + A_T (f(x_{T+1}) - \langle \partial f(x_{T+1}), x_{T+1} \rangle) \\ &= \sum_{t=1}^T \{A_t D_f(x_t \| x_{t+1})\} && (22) \\ &\quad + A_T (f(x_{T+1}) - \langle \partial f(x_{T+1}), x_{T+1} \rangle) \\ &= \sum_{t=1}^T \{A_t D_f(x_t \| x_{t+1})\} + A_T (f(x_{T+1}) + \langle \hat{z}_T, x_{T+1} \rangle) && (23) \\ &\leq \sum_{t=1}^T \{A_t D_f(x_t \| x_{t+1})\} + A_T (f(x^*) + \langle \hat{z}_T, x^* \rangle). && (24) \end{aligned}$$

Dividing both sides by A_T completes the proof. □

We also note that $D_f(x_t \| x_{t+1}) = D_{f^*}(\hat{z}_{t+1} \| z_t)$, where $f^*(z) = \sup_x \langle z, x \rangle - f(x)$. This form of the bound will often be more useful to us. Another useful property of Bregman divergences is the following lemma which relates strong convexity of f to strong smoothness of f^*:

Lemma 6. *Suppose that $D_f(x' \| x) \geq \frac{1}{2} \|x - x'\|^2$ for some norm $\| \cdot \|$ and for all x, x'. (In this case we say that f is strongly convex with respect to $\| \cdot \|$.) Then $D_{f^*}(y' \| y) \leq \frac{1}{2} \|y - y'\|_*^2$ for all y, y'.*

We require a final supporting proposition before proving Theorem 1.

Proposition 7 (Convergence of PBMD). *Consider the updates $\theta_t \in \arg\max_\theta \langle \theta, u_t \rangle - R(\theta)$ and $u_{t+1} \in \arg\min_u h(u) + \langle \hat{\theta}_s, u \rangle$. Then we have*

$$\sup_\theta L(\hat{u}_T, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{1}{A_T} \sum_{t=1}^T A_t D_h(u_t \| u_{t+1}). \tag{25}$$

Proof. Note that $L(u_t, \theta_t) = \max_\theta L(u_t, \theta)$ by construction. Also note that, if we invoke Lemma 5 with $f = h$ and $z_t = \theta_t$ then we get the inequality

$$\frac{1}{A_T} \sum_{t=1}^T \alpha_t L(u_t, \theta_t) \leq \frac{1}{A_T} \sum_{t=1}^T \alpha_t L(u^*, \theta_t) + \frac{1}{A_T} \sum_{t=1}^T A_t D_h(u_t \| u_{t+1}). \tag{26}$$

Combining these together, we get the string of inequalities

$$\begin{aligned}
L(\hat{u}_T, \theta) &= L\left(\frac{1}{A_T} \sum_{t=1}^T \alpha_t u_t, \theta\right) \\
&\leq \frac{1}{A_T} \sum_{t=1}^T \alpha_t L(u_t, \theta) \\
&\leq \frac{1}{A_T} \sum_{t=1}^T \alpha_t L(u_t, \theta_t) \\
&\leq \frac{1}{A_T} \sum_{t=1}^T \alpha_t L(u^*, \theta_t) + \frac{1}{A_T} \sum_{t=1}^T A_t D_h(u_t \| u_{t+1}) \\
&\leq \sup_\theta L(u^*, \theta) + \frac{1}{A_T} \sum_{t=1}^T A_t D_h(u_t \| u_{t+1}),
\end{aligned}$$

as was to be shown. □

Proof of Theorem 1. By Lemma 6, we have

$$\begin{aligned}
D_h(u_t \| u_{t+1}) &= D_{h^*}(\hat{\theta}_{t+1} \| \hat{\theta}_t) \\
&\leq \frac{1}{2} \|\hat{\theta}_{t+1} - \hat{\theta}_t\|_*^2 \\
&= \frac{1}{2} \| \frac{\sum_{s \leq t} \alpha_s \theta_s}{\sum_{s \leq t} \alpha_s} - \frac{\sum_{s \leq t+1} \alpha_s \theta_s}{\sum_{s \leq t+1} \alpha_s} \|_*^2 \\
&= \frac{1}{2} \| \frac{\alpha_{t+1}}{A_t A_{t+1}} \sum_{s \leq t} \alpha_s \theta_s - \frac{\alpha_{t+1}}{A_{t+1}} \theta_{t+1} \|_*^2 \\
&\leq \frac{1}{2} \left(\frac{\alpha_{t+1}}{A_t A_{t+1}} \sum_{s \leq t} \alpha_s \|\theta_s\|_*^2 + \frac{\alpha_{t+1}}{A_{t+1}} \|\theta_{t+1}\|_*^2 \right)^2 \\
&\leq \frac{2 r^2 \alpha_{t+1}^2}{A_{t+1}^2}.
\end{aligned}$$

It follows that

$$\frac{1}{A_T} \sum_{t=1}^T A_t D_h(u_t \| u_{t+1}) \leq \frac{2r^2}{A_T} \sum_{t=1}^T \frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2}.$$

□

Proof of Corollary 2. If we let $\alpha_t = 1$, then $A_t = t$ and $\frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2} = \frac{t}{(t+1)^2} \leq \frac{1}{t}$. We therefore get

$$\frac{2r^2}{A_T} \sum_{t=1}^{T} \frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2} \leq \frac{2r^2}{T} \sum_{t=1}^{T} \frac{1}{t} \leq \frac{2r^2(\log(T)+1)}{T+1}. \tag{27}$$

If we let $\alpha_t = t$, then $A_t = \frac{t(t+1)}{2}$ and $\frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2} = \frac{2(t+1)^2 t(t+1)}{(t+1)^2(t+2)^2} = \frac{2t(t+1)}{(t+2)^2} \leq 2$. We therefore get

$$\frac{2r^2}{A_T} \sum_{t=1}^{T} \frac{\alpha_{t+1}^2 A_t}{A_{t+1}^2} \leq \frac{4r^2}{T(T+1)} \sum_{t=1}^{T} 2 = \frac{8r^2}{T+1}, \tag{28}$$

which completes the proof. □

The proof of Theorem 3 requires an analagous supporting proposition to that of Theorem 1.

Proposition 8 (Convergence of DBMD)**.** *Consider the updates* $u_t \in \arg\min_u h(u) + \langle \theta_t, u \rangle$ *and* $\theta_{t+1} \in \arg\max_\theta \langle \theta, \hat{u}_t \rangle - R(\theta)$. *Then we have*

$$\sup_\theta L(\hat{u}, \theta) \leq \sup_\theta L(u^*, \theta) + \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}). \tag{29}$$

Proof. If we invoke Lemma 5 with $f = R$ and $z_t = -u_t$, then for all θ we get the inequality

$$\begin{aligned} \frac{1}{A_T} \sum_{t=1}^{T} -\alpha_t L(u_t, \theta_t) \leq{}& \frac{1}{A_T} \sum_{t=1}^{T} -\alpha_t L(u_t, \theta) \\ &- \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}). \end{aligned} \tag{30}$$

Re-arranging yields

$$\frac{1}{A_T} \sum_{t=1}^{T} \alpha_t L(u_t, \theta) \leq \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t L(u_t, \theta_t) \tag{31}$$

$$+ \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}). \tag{32}$$

Now, we have the following string of inequalities:

$$\begin{aligned} L(\hat{u}, \theta) &= L\left(\frac{1}{A_T} \sum_{t=1}^{T} \alpha_t u_t, \theta \right) \\ &\leq \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t L(u_t, \theta) \\ &\leq \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t L(u_t, \theta_t) + \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}) \\ &= \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t \inf_u L(u, \theta_t) + \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}) \\ &\leq \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t L(u^*, \theta_t) + \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}) \\ &\leq \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t \sup_\theta L(u^*, \theta) + \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}) \end{aligned}$$

$$= \sup_{\theta} L(u^*, \theta) + \frac{1}{A_T} \sum_{t=1}^{T} A_t D_R(\theta_t \| \theta_{t+1}),$$

as was to be shown. □

Proofs of Theorem 3 and Corollary 4. The proofs are identical to those for PBMD. □