Regret Analysis for Continuous Dueling Bandit

Wataru Kumagai
Center for Advanced Intelligence Project
RIKEN
1-4-1, Nihonbashi, Chuo, Tokyo 103-0027, Japan
wataru.kumagai@riken.jp

Abstract

The dueling bandit is a learning framework wherein the feedback information in the learning process is restricted to a noisy comparison between a pair of actions. In this research, we address a dueling bandit problem based on a cost function over a continuous space. We propose a stochastic mirror descent algorithm and show that the algorithm achieves an $O(\sqrt{T \log T})$-regret bound under strong convexity and smoothness assumptions for the cost function. Subsequently, we clarify the equivalence between regret minimization in dueling bandit and convex optimization for the cost function. Moreover, when considering a lower bound in convex optimization, our algorithm is shown to achieve the optimal convergence rate in convex optimization and the optimal regret in dueling bandit except for a logarithmic factor.

1 Introduction

Information systems and computer algorithms often have many parameters which should be tuned. When cost or utility are explicitly given as numerical values or concrete functions, the system parameters can be appropriately determined depending on the the values or the functions. However, in a human-computer interaction system, it is difficult or impossible for users of the system to provide user preference as numerical values or concrete functions. *Dueling bandit* is introduced to model such situations in Yue and Joachims (2009) and enables us to appropriately tune the parameters based only on comparison results on two parameters by the users. In the learning process of a dueling bandit algorithm, the algorithm chooses a pair of parameters called actions (or arms) and receives only the corresponding comparison result. Since dueling bandit algorithms do not require an individual evaluation value for each action, they can be applied for wider areas that cannot be formulated using the conventional bandit approach.

When action cost (or user utility) implicitly exists, the comparison between two actions is modeled via a cost (or utility) function, which represents the degree of the cost (or utility), and a link function, which determines the noise in the comparison results. We refer to such a modeling method as cost-based (or utility-based) approach and employ it in this research. Yue and Joachims (2009) first introduced the utility-based approach as a model for a dueling bandit problem.

The cost-based dueling bandit relates to function optimization with noisy comparisons (Jamieson et al., 2012; Matsui et al., 2016) because in both frameworks an oracle compare two actions and the feedback from the oracle is represented by binary information. In particular, the same algorithm can be applied to both frameworks. However, as different performance measures are applied to the algorithms in function optimization and dueling bandit, it has not been demonstrated that an algorithm that works efficiently in one framework will also perform well in the other framework. This study clarifies relation between function optimization and dueling bandit thorough their regret analysis.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

1.1 Problem Setup

In the learning process of the dueling bandit problem, a learner presents two points, called actions in a space $\mathcal{A}$, to an oracle and the oracle returns one-bit feedback to the learner based on which action wins (i.e., which action is more preferable for the oracle). Here, we denote by $a \succ a'$ the event that a wins a' and by $P(a \succ a')$ the probability that $a \succ a'$ happens. In other words, we assume that the feedback from the oracle follows the following two-valued random variable:

$$F(a, a') := \begin{cases} 1 & w.p. \quad P(a \succ a') \\ 0 & w.p. \quad 1 - P(a \succ a'), \end{cases} \tag{1}$$

where the probability $P(a \succ a')$ is determined by the oracle. We refer to this type of feedback as *noisy comparison feedback*. Unlike conventional bandit problems, the leaner has to make a decision that is based only on the noisy comparison feedback and cannot access the individual values of the cost (or utility) function. We further assume that each comparison between a pair of actions is independent of other comparisons.

The learner makes a sequence of decisions based on the noisy comparisons provided by the oracle. After receiving $F(a_t, a'_t)$ at time t, the learner chooses the next two action (a_{t+1}, a'_{t+1}). As a performance measure for an action a, we introduce the minimum win probability:

$$P^*(a) = \inf_{a' \in \mathcal{A}} P(a \succ a').$$

We next quantify the performance of the algorithm using the expected regret as follows:[1)]

$$Reg_T^{DB} = \sup_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \{(P^*(a) - P^*(a_t)) + (P^*(a) - P^*(a'_t))\}\right]. \tag{2}$$

1.2 Modeling Assumption

In this section, we clarify some of the notations and assumptions. Let an action space $\mathcal{A} \subset \mathbb{R}^d$ be compact convex set with non-empty interior. We denote the Euclidean norm by $\|\cdot\|$.

Assumption 1. *There exist functions $f : \mathcal{A} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to [0, 1]$ such that the probability in noisy comparison feedback can be represented as follows:*

$$P(a \succ a') = \sigma(f(a') - f(a)). \tag{3}$$

In the following, we call f in Assumption 1 a cost function and and σ a link function. Here, the cost function and the link function are fixed for each query to the oracle. In this sense, our setting is different from online optimization where the objective function changes.

Definition 1. *(Strong Convexity) A function $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex over the set $\mathcal{A} \subset \mathbb{R}^d$ if for all $x, y \in \mathcal{A}$ it holds that*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2.$$

Definition 2. *(Smoothness) A function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth over the set $\mathcal{A} \subset \mathbb{R}^d$ if for all $x, y \in \mathcal{A}$ it holds that*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|^2.$$

Assumption 2. *The cost function $f : \mathcal{A} \to \mathbb{R}$ is twice continuously differentiable, L-Lipschitz, α-strongly convex and β-smooth with respect to the Euclidean norm.*

From Assumption 2, there exists a unique minimizer a^* of the cost function f since f is strictly convex. We set $B := \sup_{a,a' \in \mathcal{A}} f(a') - f(a)$.

Assumption 3. *The link function $\sigma : \mathbb{R} \to [0, 1]$ is three times differentiable and rotation-symmetric (i.e., $\sigma(-x) = 1 - \sigma(x)$). Its first derivative is positive and monotonically non-increasing on $[0, B]$.*

[1)] Although the regret in (2) appears superficially different from that in Yue and Joachims (2009), two regrets can be shown to coincide with each other under Assumptions 1-3 in Subsection 1.2.

For examples, the standard logistic distribution function, the cumulative standard Gaussian distribution function and the linear function $\sigma(x) = (1 + x)/2$ can be taken to be link functions that satisfy Assumption 3. We note that link functions often behave like cumulative probability distribution functions. This is because the sign of the difference between two noisy function values can be regarded as the feedback (1) which satisfies Assumption 1, and then, the link function σ coincides with the cumulative probability distribution function of the noise (see Section 2 of Jamieson et al. (2012) for more details). We will discuss the relation of noisy comparison feedback to noisy function values in Section 5.

1.3 Related Work and Our Contributions

Dueling bandit on the continuous action space relates with various optimization methods. We summarize related studies in the following.

Dueling bandit problem: Yue and Joachims (2009) formulated information retrieval systems as a dueling bandit problem. They reduced this to a problem of optimizing an "almost"-concave function and presented a stochastic gradient ascent algorithm based on one-point bandit feedback. Subsequently, they showed that their algorithm achieves an $O(T^{3/4})$-regret bound under the differentiability and the strict concavity for a utility function. Ailon et al. (2014) presented reduction methods from dueling bandit to the conventional bandit under the strong restriction that the link function is linear and showed that their algorithm achieves an $O(\sqrt{T \log^3 T})$-regret bound. We note that dueling bandit has a number of other formulations (Yue and Joachims, 2011; Yue et al., 2012; Busa-Fekete et al., 2013, 2014; Urvoy et al., 2013; Zoghi et al., 2014; Jamieson et al., 2015).

Optimization with one-point bandit feedback: In conventional bandit settings, various convex optimization methods have been studied. Flaxman et al. (2005) showed that the gradient of smoothed version of a convex function can be estimated from a one-point bandit feedback and proposed a stochastic gradient descent algorithm which achieves an $O(T^{3/4})$ regret bound under the Lipschitzness condition. Moreover, assuming the strong convexity and the smoothness for the convex function such as (2), Hazan and Levy (2014) proposed a stochastic mirror descent algorithm which achieves an $O(\sqrt{T \log T})$ regret bound and showed that the algorithm is near optimal because the upper bound matched the lower bound of $\Omega(\sqrt{T})$ derived by Shamir (2013) up to a logarithmic factor in bandit convex optimization.

Optimization with two-point bandit feedback: Dueling bandit algorithms require two actions at each round in common with two-point bandit optimization. In the context of online optimization, Agarwal et al. (2010) first considered convex optimization with two-point feedback. They proposed a gradient descent-based algorithm and showed that the algorithm achieves the regret bounds of under the Lipschitzness condition and $O(\log T)$ under the strong convexity condition. In stochastic convex optimization, Duchi et al. (2015) showed that a stochastic mirror descent algorithm achieves an $O(\sqrt{T})$ regret bound under the Lipschitzness (or the smoothness) condition and proved the upper bound to be optimal deriving a matching lower bound $\Omega(\sqrt{T})$. Moreover, in both of online and stochastic convex optimization, Shamir (2017) showed that a gradient descent-based algorithm achieves an $O(\sqrt{T})$ regret bound with optimal dependence on the dimension under the Lipschitzness condition. However, those two-point bandit algorithms strongly depend on the availability of the difference of function values and cannot be directly applied to the case of dueling bandit where the difference of function values is compressed to one bit in noisy comparison feedback.

Optimization with noisy comparison feedback: The cost-based dueling bandit relates to function optimization with noisy comparisons (Jamieson et al., 2012; Matsui et al., 2016) because in both frameworks, the feedback is represented by preference information. Jamieson et al. (2012) proposed a coordinate descent algorithm and proved that the convergence rate of the algorithm achieved an optimal order.[2)] Matsui et al. (2016) proposed a Newton method-based algorithm and proved that its convergence rate was almost equivalent to that of Jamieson et al. (2012). They further showed that their algorithm could easily be parallelized and performed better numerically than the dueling bandit algorithm in Yue and Joachims (2009). However, since they considered only the unconstrained case in which $\mathcal{A} = \mathbb{R}^d$, it is not possible to apply their algorithm to the setting considered here, in which the action space is compact.

[2)] The optimal order changes depending on the model parameter $\kappa \geq 1$ of the pairwise comparison oracle in Jamieson et al. (2012).

Appendix for **Regret Analysis for Continuous Dueling Bandit**

A Appendix: Technical Proofs

[**Proof of Lemma 5**] From direct calculation,

$$
\begin{aligned}
Reg_T^{DB} &= \mathbb{E}\left[\sum_{t=1}^{T}\{2\sigma(f(a_t)-f(a_t))-\sigma(f(a^*)-f(a_t))-\sigma(f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t))\}\right] \\
&= 2\mathbb{E}\left[\sum_{t=1}^{T}\{\sigma(f(a_t)-f(a_t))-\sigma(f(a^*)-f(a_t))\}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\{\sigma(f(a^*)-f(a_t))-\sigma(f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t))\}\right]. \\
&= 2\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t)-P_t(a^*))\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\{\sigma(f(a^*)-f(a_t))-\sigma(f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t))\}\right].
\end{aligned}
$$

Here, we note that $f(a^*)-f(a)\leq 0$ for any $a\in\mathcal{A}$ due to the definition of a^* and that σ is convex on $(-\infty,0)$ because the link function is rotation symmetric and its derivative is monotonically non-increasing on positive real numbers from Assumption 3. Thus, Jensen's inequality derives

$$
\begin{aligned}
& \mathbb{E}\left[\sigma(f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t))|a_t\right] \\
\geq\ & \sigma(\mathbb{E}[f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)|a_t]) \\
=\ & \sigma(f(a^*)-\hat{f}(a_t)).
\end{aligned}
$$

In addition, $f(a_t)\leq\hat{f}(a_t)$ holds due to the convexity of f. As σ is monotonically non-decreasing from Assumption 3, we have

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{t=1}^{T}\{\sigma(f(a^*)-f(a_t))-\sigma(f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t))\}\right] \\
=\ & \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^{T}\{\sigma(f(a^*)-f(a_t))-\sigma(f(a^*)-f(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t))\}|a_t\right]\right] \\
\leq\ & \mathbb{E}\left[\sum_{t=1}^{T}\{\sigma(f(a^*)-f(a_t))-\sigma(f(a^*)-\hat{f}(a_t))\}\right] \\
\leq\ & LL_0\mathbb{E}\left[\sum_{t=1}^{T}(\hat{f}(a_t)-f(a_t))\right] \\
\leq\ & \frac{LL_0\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{x\in\mathbb{B}}[\|\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x\|^2]\right] \\
\leq\ & \frac{LL_0\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\lambda\eta t}\right] \\
\leq\ & \frac{LL_0\beta}{\lambda\eta}\log T,
\end{aligned}
$$

where we used the property of the β-smoothness of f in the third inequality. Thus, we obtain (5)

$$
Reg_T^{DB} \leq 2\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t)-P_t(a^*))\right]+\frac{LL_0\beta}{\lambda\eta}\log T.
$$

Here, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t)-P_t(a^*))\right] = \mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t)-P_t(a_T^*))\right]+\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_T^*)-P_t(a^*))\right].$$

From the definition of a_T^*, we have

$$P_t(a_T^*)-P_t(a^*)\leq LL_0\|a_T^*-a^*\|\leq\frac{LL_0R}{T},$$

where R is the diameter of $\mathcal{A}$. Thus (5) is obtained. ■

[**Proof of Lemma 6**] From direct calculation, we obtain that

$$\nabla P_b(a) = \sigma'(f(a)-f(b))\nabla f(a), \tag{15}$$

$$\nabla^2 P_b(a) = \sigma'(f(a)-f(b))\nabla^2 f(a)+\sigma''(f(a)-f(b))\nabla f(a)\nabla f(a)^\top. \tag{16}$$

Then, it is sufficient to give upper bounds on the first and second terms in (16) in the sense of matrix inequalities as

$$\sigma'(f(a)-f(b))\nabla^2 f(a) \leq L_0\beta I, \tag{17}$$

$$\sigma''(f(a)-f(b))\nabla f(a)\nabla f(a)^\top \leq B_2L^2I, \tag{18}$$

where I is the $d\times d$ identity matrix. The inequality (17) follows from the L_0-Lipschitzness of σ and the β-smoothness of f. The inequality (18) follows from the B_2-boundedness of σ'' and the L-Lipschitzness of f. ■

[**Proof of Lemma 7**] We first show that P_b is $l_0\alpha$-strongly convex on $\mathcal{L}(a^*,b)$. Since (16) holds for any $a\in\mathcal{L}(a^*,b)$, it is sufficient to give lower bounds on the first and second terms in (16) for a in $\mathcal{A}_t$ in the sense of matrix inequalities. In the following, we show

$$\sigma'(f(a)-f(b))\nabla^2 f(a) \geq l_0\alpha I, \tag{19}$$

$$\sigma''(f(a)-f(b))\nabla f(a)\nabla f(a)^\top \geq 0. \tag{20}$$

Since $l_0\leq\sigma'$ and f is α-strongly convex, we obtian (19). Next, we show (20). Since σ' is monotonically non-decreasing on $[-B,0]$, $\sigma''(y)$ is negative only if y is positive. Note that $f(a)-f(b)\leq 0$ for any $a\in\mathcal{L}(a^*,b)$ since $-f(b)+f(a^*)\leq 0$ and f is convex. Thus, we have (20).

Next, we show that P_b is $\frac{1}{2}l_0\alpha$-strongly convex on $\mathcal{A}_\delta(a^*,b)$ when $\delta\leq\frac{l_0\alpha}{4L_0^3L_2}$. For an arbitrary $\tilde{a}\in\mathcal{A}_\delta(a^*,b)$, there exists $a\in\mathcal{L}(a^*,b)$ and $y\in\mathbb{B}(0,\delta)$ such that $\tilde{a}=a+y$ by the definition of $\mathcal{A}_\delta(a^*,b)$. Since (16) holds, it is sufficient to give lower bounds on the first and second terms in (16) for a in $\mathcal{A}_t$ in the sense of matrix inequalities. Since (19) holds all we have to do is to show

$$\sigma''(f(a+y)-f(b))\nabla f(a+y)\nabla f(a+y)^\top \geq -\frac{1}{2}l_0\alpha I. \tag{21}$$

Since σ' is monotonically non-decreasing on $[-B,0]$, $\sigma''(z)$ is negative only if z is positive. Note that $f(a)-f(b)\leq 0$ for any $a\in\mathcal{L}(a^*,b)$ since $f(a^*)-f(b)\leq 0$ and f is convex. Thus, we have

$$\begin{aligned}&\sigma''(f(a+y)-f(b))\nabla f(a+y)\nabla f(a+y)^\top\\ \geq&\begin{cases}\sigma''(f(a+y)-f(b))L^2I & if\quad f(a+y)-f(b)>0\\ 0 & if\quad f(a+y)-f(b)\leq 0.\end{cases}\end{aligned} \tag{22}$$

When $f(a+y)-f(b)>0$ and $a\in\mathcal{L}(b,a^*)$, the following holds by the L-Lipshitzness of f:

$$\begin{aligned}f(a+y)-f(b) &= f(a)-f(b)-f(a)+f(a+y)\\ &\leq -f(a)+f(a+y)\\ &\leq L\delta.\end{aligned}$$

where we used again $f(a)-f(b)\leq 0$ for any $a\in\mathcal{L}(a^*,b)$. Thus

$$\begin{aligned}\sigma''(f(a+y)-f(b)) &= \sigma''(f(a+y)-f(b))-\sigma''(0)\\ &\geq -L_2(f(a+y)-f(b))\\ &\geq -LL_2\delta.\end{aligned} \tag{23}$$

Combining (22) and (23), we have

$$\sigma''(f(a+y)-f(b))\nabla f(a+y)\nabla f(a+y)^\top \geq -L^3L_2\delta I. \tag{24}$$

Thus, when $\delta\leq\frac{l_0\alpha}{2L_0^3L_2}$, we obtain (21). ■

[**Proof of Lemma 8**] We have

$$\begin{aligned}
\mathbb{E}[\hat{g}_t|a_t] &= \mathbb{E}_{u_t}[\mathbb{E}[\hat{g}_t|a_t, u_t]] \\
&= \mathbb{E}_{u_t}[d\mathbb{E}[P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t, u_t]] \\
&= d\mathbb{E}[P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t] \\
&= \nabla\mathbb{E}_{x\in\mathbb{B}}[P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x)|a_t] \qquad (25) \\
&= \nabla\hat{P}_t(a_t),
\end{aligned}$$

where we used Stokes' theorem in (25). ■

[**Proof of Lemma 10**] We can divide the left hand side of (8) into three parts:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t) - P_t(a))\right] \\
= \ &\mathbb{E}\left[\sum_{t=1}^{T}(\hat{P}_t(a_t) - \hat{P}_t(a))\right] + \mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t) - \hat{P}_t(a_t))\right] + \mathbb{E}\left[\sum_{t=1}^{T}(\hat{P}_t(a) - P_t(a))\right].
\end{aligned}$$

Here, we bound the above three terms, respectively. First, let us recall that P_t is strongly convex on $\mathbb{B}(a_t, \frac{l_0\alpha}{2L_0^3L_2}) \cap \mathcal{A}$ due to Lemma 7. By the definition of $\mathcal{R}_t$ and the conditions for λ and μ, it holds that $a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x \in \mathbb{B}(a_t, \frac{l_0\alpha}{2L_0^3L_2}) \cap \mathcal{A}$ for any $x \in \mathbb{B}$. Thus, from the local convexity of P_t in Lemma 7 and Jensen's inequality, $P_t(a_t) - \hat{P}_t(a_t) \leq 0$ holds. Next, from the smoothness of P_t in Lemma 6, we have

$$\hat{P}_t(a) - P_t(a) \leq \frac{L_0\beta + B_2L^2}{2}\|\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u\|^2 \leq \frac{L_0\beta + B_2L^2}{2\lambda\eta t},$$

where the first inequality follows from, for example, Lemma 7 of Hazan and Levy (2014) and the second inequality follows from the definition of $\mathcal{R}_t$. Hence, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}(\hat{P}_t(a) - P_t(a))\right] \leq \frac{L_0\beta + B_2L^2}{\lambda\eta}\log T.$$

Finally, we bound the first term of the upper bound of the regret. We have the following inequalities:

$$\begin{aligned}
&\mathbb{E}[\hat{P}_t(a_t) - \hat{P}_t(a)] \\
\leq \ &\mathbb{E}\left[\nabla\hat{P}_t(a_t)^\top(a_t - a) - \frac{l_0\alpha}{4}\|a_t - a\|^2\right] \\
= \ &\mathbb{E}\left[\hat{g}_t^\top(a_t - a) - \frac{l_0\alpha}{4}\|a_t - a\|^2\right] \\
= \ &\eta^{-1}\mathbb{E}\left[(\nabla\mathcal{R}_t(a_{t+1}) - \nabla\mathcal{R}_t(a_t))^\top(a - a_t) - \frac{l_0\alpha\eta}{4}\|a_t - a\|^2\right] \\
= \ &\eta^{-1}\mathbb{E}\left[D_{\mathcal{R}_t}(a, a_t) + D_{\mathcal{R}_t}(a_t, a_{t+1}) - D_{\mathcal{R}_t}(a, a_{t+1}) - \frac{l_0\alpha\eta}{4}\|a_t - a\|^2\right],
\end{aligned}$$

where the first inequality follows from the local convexity of $\hat{P}_t$, the first equality is derived by Lemma 8 and the second equality holds due to the definition of a_{t+1}. Summing up both sides,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}(\hat{P}_t(a_t)-\hat{P}_t(a)\right] \\
\leq\ & \eta^{-1}\mathbb{E}\left[D_{\mathcal{R}_1}(a,a_1)-D_{\mathcal{R}_T}(a,a_{T+1})+\sum_{t=1}^{T}D_{\mathcal{R}_t}(a_t,a_{t+1})\right] \\
&+\eta^{-1}\mathbb{E}\left[\sum_{t=2}^{T}\left(D_{\mathcal{R}_t}(a,a_t)-D_{\mathcal{R}_{t-1}}(a,a_t)-\frac{l_0\alpha\eta}{4}\|a_t-a\|^2\right)\right] \\
\leq\ & \eta^{-1}\mathbb{E}\left[D_{\mathcal{R}_1}(a,a_1)-D_{\mathcal{R}_T}(a,a_T+1)+\sum_{t=1}^{T}D_{\mathcal{R}_t}(a_t,a_{t+1})\right] \\
&+\eta^{-1}\mathbb{E}\left[\sum_{t=2}^{T}\left(D_{\mathcal{R}_t}(a,a_t)-D_{\mathcal{R}_{t-1}}(a,a_t)-\frac{\lambda\eta}{2}\|a_t-a\|^2\right)\right] \\
=\ & \eta^{-1}\mathbb{E}\left[D_{\mathcal{R}_1}(a,a_1)-D_{\mathcal{R}_T}(a,a_T+1)+\sum_{t=1}^{T}D_{\mathcal{R}_t}(a_t,a_{t+1})\right] \\
\leq\ & \eta^{-1}\mathbb{E}\left[D_{\mathcal{R}_1}(a,a_1)+\sum_{t=1}^{T}D_{\mathcal{R}_t}(a_t,a_{t+1})\right] \\
=\ & \eta^{-1}\left(\mathcal{R}(a)-\mathcal{R}(a_1)+\mathbb{E}\left[\sum_{t=1}^{T}D_{\mathcal{R}_t^*}(\nabla\mathcal{R}_t(a_t)-\eta\hat{g}_t,\nabla\mathcal{R}_t(a_t))\right]\right)
\end{aligned}$$

where we used the positivity of the Bregman divergence in the third inequality and $\nabla\mathcal{R}(a_1)=0$ because $a_1=\mathrm{argmin}\mathcal{R}$ in the last equation. Combining the above discussion, we obtain (8). ■

[**Proof of Lemma 11**] Taylar's theorem guarantees the existence of $\delta_t\in(0,1)$ such that

$$D_{\mathcal{R}_t{}^*}(\nabla\mathcal{R}_t(a_t)-\eta\hat{g}_t,\nabla\mathcal{R}_t(a_t)) \quad=\quad \eta^2\hat{g}_t{}^\top\nabla^2\mathcal{R}_t^*(\nabla\mathcal{R}_t(a_t)-\delta_t\eta\hat{g}_t)\hat{g}_t. \tag{26}$$

Then using the self-concordant property of $\mathcal{R}_t^*$ (see e.g. (F.2) of Griva et al. (2009)),

$$\hat{g}_t{}^\top\nabla^2\mathcal{R}_t^*(\nabla\mathcal{R}_t(a_t)-\delta_t\eta\hat{g}_t)\hat{g}_t \quad\leq\quad \left(\frac{\|\hat{g}_t\|^*_{\nabla\mathcal{R}_t(a_t)}}{1-\delta_t\eta\|\hat{g}_t\|^*_{\nabla\mathcal{R}_t(a_t)}}\right)^2. \tag{27}$$

where $\|x\|_y^*=\sqrt{x^\top\nabla^2\mathcal{R}_t^*(y)x}$. Here, we note that

$$\begin{aligned}
\|\hat{g}_t\|^*_{\nabla\mathcal{R}_t(x_t)} &= \sqrt{\hat{g}_t{}^\top\nabla^2\mathcal{R}_t^*(\nabla\mathcal{R}_t(x_t))\hat{g}_t} \\
&= \sqrt{\hat{g}_t{}^\top\nabla^2\mathcal{R}_t(x_t)^{-1}\hat{g}_t} \\
&\leq d\sqrt{u_t^\top\nabla^2\mathcal{R}_t(x_t)^{\frac{1}{2}}\nabla^2\mathcal{R}_t(x_t)^{-1}\nabla^2\mathcal{R}_t(x_t)^{\frac{1}{2}}u_t} \\
&= d
\end{aligned} \tag{28}$$

and thus, $\delta_t\eta\|\hat{g}_t\|^*_{\nabla\mathcal{R}(a_t)}<\frac{1}{2}$ when $\eta\leq\frac{1}{2d}$. Consequently, (26), (27) and (28) derives (9). ■

[**Proof of Lemma 12**] We show the second inequality of (11). From the definition, Since $l_0\leq\alpha_t'$ and $\sigma(0)=\frac{1}{2}$ from Assumption 3, we have

$$\begin{aligned}
Reg_T^{DB} &= \sup_{a\in\mathcal{A}}\mathbb{E}\left[\sum_{t=1}^{T}\left(\{\sigma(f(a_t)-f(a))-\sigma(0)\}+\{\sigma(f(a_t')-f(a))-\sigma(0)\}\right)\right] \\
&\geq l_0\sup_{a\in\mathcal{A}}\mathbb{E}\left[\sum_{t=1}^{T}\left(\{f(a_t)-f(a)\}+\{f(a_t')-f(a)\}\right)\right] \\
&= l_0 Reg_T^{FO}.
\end{aligned}$$

The first inequality of (11) can be proven in a similar manner. ■