Bounds for Approximate Regret-Matching Algorithms

Ryan D'Orazio, Dustin Morrill, James R. Wright
Department of Computing Science
University of Alberta
{rdorazio, morrill, james.wright}@ualberta.ca

Abstract

A dominant approach to solving large imperfect-information games is *Counterfactual Regret Minimization (CFR)*. In CFR, many regret minimization problems are combined to solve the game. For very large games, abstraction is typically needed to render CFR tractable. Abstractions are often manually tuned, possibly removing important strategic differences in the full game and harming performance. Function approximation provides a natural solution to finding good abstractions to approximate the full game. A common approach to incorporating function approximation is to learn the inputs needed for a regret minimizing algorithm, allowing for generalization across many regret minimization problems. This paper gives regret bounds when a regret minimizing algorithm uses estimates instead of true values. This form of analysis is the first to generalize to a larger class of (Φ, f)-regret matching algorithms, and includes different forms of regret such as swap, internal, and external regret. We demonstrate how these results give a slightly tighter bound for Regression Regret-Matching (RRM), and present a novel bound for combining regression with Hedge.

1   Introduction

The dominant framework for approximating Nash equilibria in extensive-form games with imperfect information is Counterfactual Regret Minimization (CFR), and it has successfully been used to solve and expertly play human-scale poker games [1, 2, 3, 4]. This framework is built on the idea of decomposing a game into a network of simple regret minimizers [5, 6]. For very large games, abstraction is typically used to reduce the size and yield a strategically similar game that is feasible to solve with CFR [5, 7, 8, 9].

Function approximation is a natural generalization of abstraction. In CFR, this amounts to estimating the regrets for each regret minimizer instead of storing them all in a table [10, 11, 12, 13, 14, 15]. Function approximation can be competitive with domain specific state abstraction [10, 11, 12], and in some cases is able to outperform tabular CFR without abstraction if the players are optimizing against their best responses [14].

Combining regression and regret-minimization with applications to CFR was initially studied by Waugh et. al. [10], introducing the RRM theorem—giving a sufficient condition for function approximator error to still achieve no external regret. In this paper we generalize the RRM theorem to a larger class of regret-minimizers and Φ-regret—a set of regret metrics that include external regret, internal regret, and swap regret. Extending to a larger class of regret-minimizers provides insight into the effectiveness of combining function approximation and regret minimization—the effect of function approximation error on the bounds will vary between algorithms. Furthermore, extending to other algorithms can give theory for existing or future methods. For example, there

has been interest in a functional version of hedge, an algorithm within the studied class, for general multiagent and non-stationary settings that can outperform softmax policy gradient methods [13]. Extending to a more general class of regret metrics such as internal regret allows for potentially-novel applications of regret minimization and function approximation including finding an approximate correlated equilibrium [16].

2   Preliminaries

We adopt the notation from Greenwald et al. [17] to describe an online decision problem (ODP). An ODP consists of a set of possible actions A and set of possible rewards $\mathcal{R}$. In this paper we assume a finite set of actions and bounded[1] $\mathcal{R} \subset \mathbb{R}_+$ where $\sup \mathcal{R} = U$. The tuple $(A, \mathcal{R})$ fully characterizes the problem and is referred to as a reward system. Furthermore, let Π denote the set of reward functions $r : A \mapsto \mathcal{R}$.

At each round t an agent selects a distribution over actions $q_t \in \Delta(A)$[2], samples an action $a_t \sim q_t$ and then receives the reward function $r_t \in \Pi$. The agent is able to compute the rewards for actions that were not taken at time t in contrast to the bandit setting where the agent only observes $r_t(a_t)$. Crucially, each r_t is allowed to be selected arbitrarily from Π. As a consequence, this ODP model is flexible enough to encompass multi-agent, adversarial interactions, and game theoretic equilibrium concepts even though it is described from the perspective of a single agent's decisions.

A learning algorithm in an ODP selects q_t using information from the history of observations and actions previously taken. We denote this information at time t as history $h \in H_t := A^t \times \Pi^t$, where $H_0 := \{\emptyset\}$. Formally, an online learning algorithm is a sequence of functions $\{L_t\}_{t=1}^\infty$, where $L_t : H_{t-1} \mapsto \Delta(A)$.

2.1   Action Transformations

To generalize the analysis to different forms of regret (e.g. swap, internal, and external regret), it is useful to define action transformations. Action transformations are functions of the form $\phi : A \mapsto \Delta(A)$, giving a distribution over actions for each action input. Let $\Phi_{ALL} := \Phi_{ALL}(A)$ denote the set of all action transformations for the set of actions A and $\Phi_{SWAP} := \Phi_{SWAP}(A)$ the set of all action transformations with codomain as the set of all pure strategies for action set A.

Two important subsets of Φ_{SWAP} are Φ_{EXT} and Φ_{INT}. Φ_{EXT} denotes the set of all external transformations—the set of constant action transformations in Φ_{SWAP}. More formally, if $\delta_a \in \Delta(A)$ is the distribution with full weight on action a, then $\Phi_{EXT} = \{\phi : \exists y \in A \, \forall x \in A \quad \phi(x) = \delta_y\}$.

Φ_{INT} consists of the set of all possible internal transformations for action set A, where an internal transformation from action a to action b is defined as $\phi_{INT}^{(a,b)}(x) = \delta_b$ if $x = a$, $\phi_{INT}^{(a,b)}(x) = \delta_x$ otherwise.

We have that $|\Phi_{SWAP}| = |A|^{|A|}, \Phi_{EXT} = |A|, |\Phi_{INT}| = |A|^2 - |A| + 1$ [17].

We will also make use of the linear transformation $[\phi] : \Delta(A) \mapsto \Delta(A)$ defined as $[\phi](q) = \sum_{a \in A} q(a)\phi(a)$.

2.2   Regret

For a given action transformation ϕ we can compute the difference in expected reward for a particular action and reward function. This expected difference, known as ϕ-regret, is denoted by $\rho^\phi(a, r) = \mathbb{E}_{s \sim \phi(a)}[r(s)] - r(a)$. For a set of action transformations Φ, the Φ-regret vector is $\rho^\Phi(a, r) = (\rho^\phi(a, r))_{\phi \in \Phi}$. Note the expected value of ϕ-regret if the agent chooses $q \in \Delta(A)$ is $\mathbb{E}_{a \sim q}[\rho^\phi(a, r)]$.

For an ODP with observed history h at time t, with reward functions $\{r_s\}_{s=1}^t$ and actions $\{a_s\}_{s=1}^t$, the cumulative Φ-regret for time t and action transformations Φ is $R_t^\Phi(h) = \sum_{s=1}^t \rho^\Phi(a_s, r_s)$. For brevity we will omit the h argument, and for convenience we set $R_0^\Phi = 0$. Note that R_t^Φ is a random

---

[1]The restriction of positive rewards is without loss of generality and is only used for convenience.
[2]$\Delta(A)$ is the set of all probability distributions over actions in A.

vector, and we seek to bound

$$\mathbb{E}\left[\frac{1}{t}\max_{\phi\in\Phi}R_t^\phi\right].\tag{1}$$

Choosing $\Phi_{EXT},\Phi_{INT},\Phi_{SWAP}$ for (1) amounts to minimizing external regret, internal regret, and swap regret respectively. One can also change (1) by interchanging the max and the expectation. In RRM, $\max_{\phi\in\Phi}\mathbb{E}\left[\frac{1}{t}R_t^\phi\right]$ is bounded [10, 11], however, bounds for (1) still apply [17, Corollary 18].

3   Approximate Regret-Matching

Given a set of action transformations Φ and a link function $f:\mathbb{R}^{|\Phi|}\mapsto\mathbb{R}_+^{|\Phi|}$ that is subgradient to a convex potential function $G:\mathbb{R}^{|\Phi|}\mapsto\mathbb{R}$, where $\mathbb{R}_+^N$ denotes the N-dimensional positive orthant[3], we can define a general class of online learning algorithms known as (Φ,f)-*regret-matching* algorithms [17]. A (Φ,f)-*regret-matching* algorithm at time t chooses $q\in\Delta(A)$ that is a fixed point[4] of $M_t=\sum_{\phi\in\Phi}Y_t^\phi[\phi]/\sum_{\phi\in\Phi}Y_t^\phi$ where $Y_t^\Phi=(Y_t^\phi)_{\phi\in\Phi}=f(R_{t-1}^\Phi)$ when $R_{t-1}^\Phi\in\mathbb{R}_+^{|\Phi|}\setminus\{0\}$ and arbitrarily otherwise. If $\Phi=\Phi_{EXT}$ then the fixed point of M_t is a distribution $q\propto Y_t^\Phi$ [18]. Examples of (Φ,f)-regret-matching algorithms include Hart's algorithm [19]–typically called "regret-matching" or the polynomial weighted average forecaster [16]–and Hedge [20]–the exponentially weighted average forecaster [16], with link functions $f(x)_i=(x_i^+)^{p-1}$ for $p\geq1$, and $f(x)_i=e^{\eta x_i}$ with parameter $\eta>0$, respectively.

A useful technique to bounding regret when estimates are used in place of true values is to define an $\epsilon-$Blackwell condition, as was used in the RRM theorem [10]. The analysis in RRM was specific to $\Phi=\Phi_{EXT}$ and the polynomial link f with $p=2$. To generalize across different link functions and $\Phi\subseteq\Phi_{ALL}$ we define the (Φ,f,ϵ)-Blackwell condition.

Definition 1 ((Φ,f,ϵ)-Blackwell Condition). *For a given reward system $(A,\mathcal{R})$, finite set of action transformations $\Phi\subseteq\Phi_{ALL}$, and link function $f:\mathbb{R}^{|\Phi|}\mapsto\mathbb{R}_+^{|\Phi|}$, a learning algorithm satisfies the (Φ,f,ϵ)-Blackwell condition with value ϵ if $f(R_{t-1}^\Phi(h))\cdot\mathbb{E}_{a\sim L_t(h)}[\rho^\Phi(a,r)]\leq\epsilon$.*

The Regret Matching Theorem [17] shows that the (Φ,f)-Blackwell condition ($\epsilon=0$) holds with equality for (Φ,f)-*regret-matching* algorithms for any finite set of action transformations Φ and link function f.

We seek to bound objective (1) when an algorithm at time t chooses the fixed point of $\tilde{M}_t=\sum_{\phi\in\Phi}\tilde{Y}_t^\phi[\phi]/\sum_{\phi\in\Phi}\tilde{Y}_t^\phi$, when $\tilde{R}_{t-1}^\Phi\in\mathbb{R}_+^{|\Phi|}\setminus\{0\}$ and arbitrarily otherwise, where $\tilde{Y}_t^\Phi=f(\tilde{R}_{t-1}^\Phi)$ and $\tilde{R}_{t-1}^\Phi$ is an estimate of R_{t-1}^Φ, possibly from a function approximator. Such an algorithm is referred to as approximate (Φ,f)-regret-matching.

Similarly to the RRM theorem [10, 11], we show that the ϵ parameter of the (Φ,f,ϵ)-Blackwell condition depends on the error in approximating the exact link outputs, $\|Y_t^\Phi-\tilde{Y}_t^\Phi\|_1$.

Theorem 1. *Given reward system $(A,\mathcal{R})$, a finite set of action transformations $\Phi\subseteq\Phi_{ALL}$, and link function $f:\mathbb{R}^{|\Phi|}\mapsto\mathbb{R}_+^{|\Phi|}$, then an approximate (Φ,f)-regret-matching algorithm, $\{L_t\}_{t=1}^\infty$, satisfies the (Φ,f,ϵ)-Blackwell Condition with $\epsilon\leq2U\|Y_t^\Phi-\tilde{Y}_t^\Phi\|_1$, where $Y_t^\Phi=f(R_{t-1}^\Phi)$, and $\tilde{Y}_t^\Phi=f(\tilde{R}_{t-1}^\Phi)$.*

All proofs are deferred to the appendix.

For a (Φ,f)-regret-matching algorithm, an approach to bounding (1) is to use the (Φ,f)-Blackwell condition and provide a bound on $\mathbb{E}[G(R_t^\Phi)]$ for an appropriate potential function G [17, 16]. Bounding the regret (1) for an approximate (Φ,f)-regret-matching algorithm will be done similarly, except the bound on ϵ from Theorem 1 will be used. Proceeding in this fashion yields the following theorem:

---

[3] Note that as long as G is bounded from above on the negative orthant then the codomain of f is the positive orthant.

[4] Note that since M_t is a linear operator over the simplex $\Delta(A)$, the fixed point always exists by the Brouwer fixed point theorem.

Theorem 2. *Given a real-valued reward system $(A, \mathcal{R})$ a finite set $\Phi \subseteq \Phi_{ALL}$ of action transformations. If $\langle G, g, \gamma \rangle$ is a Gordon triple[5], then an approximate (Φ, g)-regret-matching algorithm $\{L_t\}_{t=1}^{\infty}$ guarantees at all times $t \geq 0$*

$$\mathbb{E}[G(R_t^{\Phi})] \leq G(0) + t \sup_{a \in A, r \in \Pi} \gamma(\rho^{\Phi}(a, r)) + 2U \sum_{s=1}^{t} \|g(R_{s-1}^{\Phi}) - g(\tilde{R}_{s-1}^{\Phi})\|_1.$$

4 Bounds

4.1 Polynomial Link

Given the polynomial link function $f(x)_i = (x_i^+)^{p-1}$ we consider two cases $2 < p < \infty$ and $1 < p \leq 2$. For the following results it is useful to denote the maximal activation $\mu(\Phi) = \max_{a \in A} |\{\phi \in \Phi : \phi(a) \neq \delta_a\}|$ [17].

For the case $p > 2$ we have the following bound on (1).

Theorem 3. *Given an ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and the polynomial link function f with $p > 2$, then an approximate (Φ, f)- regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^{\phi}\right] \leq \frac{1}{t} \sqrt{t(p-1)U^2(\mu(\Phi))^{2/p} + 2U \sum_{k=1}^{t} \|g(R_{k-1}^{\Phi}) - g(\tilde{R}_{k-1}^{\Phi})\|_1}$$

where $g : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$ and $g(x)_i = 0$ if $x_i \leq 0$ otherwise $g(x)_i = \frac{2(x_i)^{p-1}}{\|x^+\|_p^{p-2}}$.

Similarly for the case $1 < p \leq 2$ we have the following.

Theorem 4. *Given an ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and the polynomial link function f with $1 < p \leq 2$, then an approximate (Φ, f)- regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^{\phi}\right] \leq \frac{1}{t} \left(tU^p \mu(\Phi) + 2U \sum_{k=1}^{t} \|g(R_{k-1}^{\Phi}) - g(\tilde{R}_{k-1}^{\Phi})\|_1\right)^{1/p}$$

where $g : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$ and $g(x)_i = p(x_i^+)^{p-1}$.

In comparison to the RRM theorem [11], the above bound is tighter as there is no $\sqrt{|A|}$ term in front of the errors and the $|A|$ term has been replaced by[6] $|A| - 1$. These improvements are due to the tighter bound in Theorem 1 and the original Φ-regret analysis [17], respectively. Aside from these differences, the bounds coincide.

4.2 Exponential Link

Theorem 5. *Given an ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and an exponential link function $f(x)_i = e^{\eta x_i}$ with $\eta > 0$, then an approximate (Φ, f)- regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^{\phi}\right] \leq \frac{1}{t} \left(\frac{\ln|\Phi|}{\eta} + 2U \sum_{k=1}^{t} \|g(R_{k-1}^{\Phi}) - g(\tilde{R}_{k-1}^{\Phi})\|_1\right) + \frac{\eta U^2}{2}$$

where $g : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$ and $g(x)_i = e^{\eta x_i} / \sum_j e^{\eta x_j}$.

The Hedge algorithm corresponds to the exponential link function $f(x)_i = e^{\eta x_i}$ when $\Phi = \Phi_{EXT}$, so Theorem 5 provides a bound on a regression Hedge algorithm. Note that in this case, the approximation error term is not inside a root function as it is under the polynomial link function. This seems to imply that at the level of link outputs, polynomial link functions have a better dependence on the approximation errors. However, g in the exponential link function bound is normalized to the simplex while the polynomial link functions can take on larger values. So which link function has a better dependence on the approximation errors depends on the magnitude of the cumulative regrets, which depends on the environment and the algorithm's empirical performance.

---

[5]See definition 2 in appendix.
[6]For $\Phi = \Phi_{EXT}, \mu(\Phi) = |A| - 1$.

Acknowledgments

References

[1] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

[2] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[3] Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[4] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[5] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, pages 1729–1736, 2008.

[6] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Regret circuits: Composability of regret minimizers. In *International Conference on Machine Learning*, pages 1863–1872, 2019.

[7] Kevin Waugh, David Schnizlein, Michael Bowling, and Duane Szafron. Abstraction pathologies in extensive games. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 781–788. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[8] Michael Johanson, Neil Burch, Richard Valenzano, and Michael Bowling. Evaluating state-space abstractions in extensive-form games. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 271–278. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[9] Sam Ganzfried and Tuomas Sandholm. Action translation in extensive-form games with large action spaces: Axioms, paradoxes, and the pseudo-harmonic mapping. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[10] Kevin Waugh, Dustin Morrill, James Andrew Bagnell, and Michael Bowling. Solving games with functional regret estimation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[11] Dustin Morrill. *Using Regret Estimation to Solve Games Compactly*. Master's thesis, University of Alberta, 2016.

[12] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML-19)*, pages 793–802, 2019.

[13] Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, and Karl Tuyls. Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*, 2019.

[14] Edward Lockhart, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr Timbers, and Karl Tuyls. Computing approximate equilibria in sequential adversarial games by exploitability descent. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 464–470. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[15] Eric Steinberger. Single deep counterfactual regret minimization. *arXiv preprint arXiv:1901.07621*, 2019.

[16] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[17] Amy Greenwald, Zheng Li, and Casey Marks. Bounds for regret-matching algorithms. In *ISAIM*, 2006.

[18] Amy Greenwald, Zheng Li, and Casey Marks. Bounds for regret-matching algorithms. Technical Report CS-06-10, Brown University, Department of Computer Science, 2006.

[19] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[20] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[21] Geoffrey J Gordon. No-regret algorithms for structured prediction problems. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2005.

[22] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[23] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Appendices

Below we recall results from Greenwald et. al.[17] and include the detailed proofs omitted in the main body of the paper.

Many of the following results make use of a Gordon triple. We restate the definition from Greenwald et. al. below.

Definition 2. *A Gordon triple $\langle G, g, \gamma \rangle$ consists of three functions $G : \mathbb{R}^n \mapsto \mathbb{R}$, $g : \mathbb{R}^n \mapsto \mathbb{R}^n$, and $\gamma : \mathbb{R}^n \mapsto \mathbb{R}$ such that for all $x, y \in \mathbb{R}^n$, $G(x + y) \leq G(x) + g(x) \cdot y + \gamma(y)$.*

A Existing Results

Lemma 1. *If x is a random vector that takes values in $\mathbb{R}^n$, then $(\mathbb{E}[\max_i x])^q \leq \mathbb{E}[\|x^+\|_p^q]$ for $p, q \geq 1$.*

See [Lemma 21][17].

Lemma 2. *Given a reward system $(A, \mathcal{R})$ and a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, then $\|\rho^\Phi(a, r)\|_p \leq U(\mu(\Phi))^{1/p}$ for any reward function $r \in \Pi$.*

The proof is indentical to [Lemma 22][17] except we have that regrets are bounded in $[-U, U]$ instead of $[-1, 1]$. Also note that by assumption $\mathcal{R}$ is bounded.

Theorem 6 (Gordon 2005). *Assume $\langle G, g, \gamma \rangle$ is a Gordon triple and $C : \mathcal{N} \mapsto \mathbb{R}$. Let $X_0 \in \mathbb{R}^n$, let $x_1, x_2, \ldots$ be a sequence of random vectors over $\mathbb{R}^n$, and define $X_t = X_{t-1} + x_t$ for all times $t \geq 1$. If for all times $t \geq 1$,*

$$g(X_{t-1}) \cdot \mathbb{E}[x_t | X_{t-1}] + \mathbb{E}[\gamma(x_t) | X_{t-1}] \leq C(t) \quad a.s.$$

then, for all times $t \geq 0$,

$$\mathbb{E}[G(X_t)] \leq G(X_0) + \sum_{\tau=1}^{t} C(\tau).$$

It should be noted that the above theorem was originally proved by Gordon [21].

B Proofs

Theorem 1. *Given reward system (A,$\mathcal{R}$), a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and link function $f : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$, then an approximate (Φ, f)-regret-matching algorithm, $\{L_t\}_{t=1}^{\infty}$, satisfies the (Φ, f, ϵ)-Blackwell Condition with $\epsilon \leq 2U\|Y_t^{\Phi} - \tilde{Y}_t^{\Phi}\|_1$, where $Y_t^{\Phi} = f(R_{t-1}^{\Phi})$, and $\tilde{Y}_t^{\Phi} = f(\tilde{R}_{t-1}^{\Phi})$.*

Proof. We denote $r = (r'(a))_{a \in A}$ as the reward vector for an arbitrary reward function $r' : A \mapsto \mathbb{R}$. Since by construction this algorithm chooses L_t at each timestep t to be the fixed point of $\tilde{M}_t$, all that remains to be shown is that this algorithm satisfies the (Φ, f, ϵ)-Blackwell condition with $\epsilon \leq 2U\|Y_t^{\Phi} - \tilde{Y}_t^{\Phi}\|_1, t > 0$.

By expanding the value of interest in the (Φ, f)-Blackwell condition and applying elementary upper bounds, we arrive at the desired bound. For simplicity, we omit timestep indices and set $L := L_t(h)$. First, suppose $\sum_{\phi \in \Phi} \tilde{Y}_t^{\Phi} \neq 0$:

$$Y^{\Phi} \cdot \mathbb{E}_{a \sim L}[\rho^{\Phi}(a, r)] = \sum_{\phi \in \Phi} Y^{\phi}(r \cdot [\phi](L) - r \cdot L)$$

$$= r \cdot \left(\sum_{\phi \in \Phi} Y^{\phi}[\phi]L - L \right) \text{. By adding and subtracting } \tilde{Y}^{\Phi},$$

$$= r \cdot \left(\sum_{\phi \in \Phi} (\tilde{Y}^{\phi} - \tilde{Y}^{\phi} + Y^{\phi})([\phi](L) - L) \right) \text{. By expanding, as well as multiplying and dividing by } \left(\sum_{\phi \in \Phi} \tilde{Y}^{\phi} \right),$$

$$= r \cdot \left((\sum_{\phi \in \Phi} \tilde{Y}^{\phi}) \sum_{\phi \in \Phi} \tilde{M}L - L + \sum_{\phi \in \Phi} (Y^{\phi} - \tilde{Y}^{\phi})([\phi](L) - L) \right) \text{. Since } L \text{ is a fixed point of } \tilde{M},$$

$$= r \cdot \left(\sum_{\phi \in \Phi} (Y^{\phi} - \tilde{Y}^{\phi})([\phi](L) - L) \right) \text{. By the generalized Cauchy-Schwarz inequality [22, 23],}$$

$$\leq \|r\|_{\infty} \| \sum_{\phi \in \Phi} (Y^{\phi} - \tilde{Y}^{\phi})([\phi](L) - L)\|_1$$

$$\leq \|r\|_{\infty} \sum_{\phi \in \Phi} |Y^{\phi} - \tilde{Y}^{\phi}|(\|[\phi](L)\|_1 + \|L\|_1) \text{. Since } [\phi](L) \text{ and } L \text{ are both distributions,}$$

$$\leq \|r\|_{\infty} \sum_{\phi \in \Phi} |Y^{\phi} - \tilde{Y}^{\phi}|(1 + 1)$$

$$\leq 2U\|Y^{\Phi} - \tilde{Y}^{\Phi}\|_1.$$

If $\sum_{\phi \in \Phi} \tilde{Y}^{\Phi} = 0$ it is easy to see the inequality still holds.

Therefore, $\{L_t\}_{t=1}^{\infty}$ satisfies the (Φ, f, ϵ)-Blackwell condition with $\epsilon \leq 2U\|Y^{\Phi} - \tilde{Y}^{\Phi}\|_1$, as required to complete the argument.

$\square$

An important observation of Theorem 1 is the following corollary:

Corollary 1. *For a reward system $(A, \mathcal{R})$, finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and two link functions f and f', if there exists a strictly positive function $\psi : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}$ such that $f'(x) = \psi(x)f(x)$ then for any $\epsilon \in \mathbb{R}$, then an approximate (Φ, f)-regret-matching algorithm satisfies*

$$f'(R_{t-1}^{\Phi}(h)) \cdot \mathbb{E}_{a \sim L_t(h)}[\rho^{\Phi}(a, r)] \leq 2U\|f'(R_{t-1}^{\Phi}) - f'(\tilde{R}_{t-1}^{\Phi})\|_1.$$

Proof. The reasoning is similar to [Lemma 20][17]. The played fixed point is the same under both link functions, thus following the same steps to Theorem 1 provides the above bound. $\square$

Theorem 2. *Given a real-valued reward system $(A, \mathcal{R})$ a finite set $\Phi \subseteq \Phi_{ALL}$ of action transformations. If $\langle G, g, \gamma \rangle$ is a Gordon triple[7], then an approximate (Φ, g)-regret-matching algorithm $\{L_t\}_{t=1}^{\infty}$ guarantees at all times $t \geq 0$*

$$\mathbb{E}[G(R_t^{\Phi})] \leq G(0) + t \sup_{a \in A, r \in \Pi} \gamma(\rho^{\Phi}(a, r)) + 2U \sum_{s=1}^{t} \|g(R_{s-1}^{\Phi}) - g(\tilde{R}_{s-1}^{\Phi})\|_1.$$

Proof. The proof is similar to [Corollary 7][17] except that the learning algorithm is playing the approximate fixed point with respect to the link function g. From Theorem 1 we have $g(R_{t-1}^{\Phi}(h)) \cdot \mathbb{E}_{a \sim L_t(h)}[\rho^{\Phi}(a, r)] \leq 2U\|g(R_{t-1}^{\Phi}) - g(\tilde{R}_{t-1}^{\Phi})\|_1$. Noticing that $\mathbb{E}_{a \sim L_t(h)}[\rho^{\Phi}(a, r)] = \mathbb{E}[\rho^{\Phi}(a, r)|R_{t-1}^{\Phi}]$ and taking $x_t = \rho^{\Phi}(a, r), X_t = R_t^{\Phi}$ we have

$$g(X_{t-1}) \cdot \mathbb{E}[x_t | X_{t-1}] + \mathbb{E}[\gamma(x_t) | X_{t-1}] \leq 2U\|g(R_{t-1}^{\Phi}) - g(\tilde{R}_{t-1}^{\Phi})\|_1 + \sup_{a \in A, r \in \Pi} \gamma(\rho^{\Phi}(a, r)).$$

The result directly follows from Theorem 6 by taking $C(\tau) = 2U\|g(R_{\tau-1}^{\Phi}) - g(\tilde{R}_{\tau-1}^{\Phi})\|_1 + \sup_{a \in A, r \in \Pi} \gamma(\rho^{\Phi}(a, r))$. $\qquad\square$

Theorem 3. *Given an ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and the polynomial link function f with $p > 2$, then an approximate (Φ, f)- regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^{\phi}\right] \leq \frac{1}{t}\sqrt{t(p-1)U^2(\mu(\Phi))^{2/p} + 2U \sum_{k=1}^{t} \|g(R_{k-1}^{\Phi}) - g(\tilde{R}_{k-1}^{\Phi})\|_1}$$

where $g : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$ and $g(x)_i = 0$ if $x_i \leq 0$ otherwise $g(x)_i = \frac{2(x_i)^{p-1}}{\|x^+\|_p^{p-2}}$.

Proof. The proof follows closely to [Theorem 9][17]. Taking $G(x) = \|x^+\|_p^2$ and $\gamma(x) = (p-1)\|x\|_p^2$ then $\langle G, g, \gamma \rangle$ is a Gordon triple [17]. Given the above gordon triple we have

$$\left(\mathbb{E}\left[\max_{\phi \in \Phi} R_t^{\phi}\right]\right)^2 \leq \mathbb{E}[\|(R_t^{\Phi})^+\|_p^2] \tag{2}$$

$$= \mathbb{E}[G(R_t^{\Phi})] \tag{3}$$

$$\leq G(0) + t \sup_{a \in A, r \in \Pi} \gamma(\rho^{\Phi}(a, r)) + 2U \sum_{s=1}^{t} \|g(R_{s-1}^{\Phi}) - g(\tilde{R}_{s-1}^{\Phi})\|_1 \tag{4}$$

$$\leq G(0) + t(p-1)U^2(\mu(\Phi))^{2/p} + 2U \sum_{k=1}^{t} \|g(R_{k-1}^{\Phi}) - g(\tilde{R}_{k-1}^{\Phi})\|_1 \tag{5}$$

The first inequality is from Lemma 1. The second inequality follows from Corollary 1 and theorem 2. The third inequality is an application of Lemma 2. The result then immediately follows. $\qquad\square$

Theorem 4. *Given an ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and the polynomial link function f with $1 < p \leq 2$, then an approximate (Φ, f)- regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^{\phi}\right] \leq \frac{1}{t}\left(tU^p\mu(\Phi) + 2U \sum_{k=1}^{t} \|g(R_{k-1}^{\Phi}) - g(\tilde{R}_{k-1}^{\Phi})\|_1\right)^{1/p}$$

where $g : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$ and $g(x)_i = p(x_i^+)^{p-1}$.

---

[7]See definition 2 in appendix.

8

Proof. The proof follows closely to [Theorem 11][17]. Taking $G(x) = \|x^+\|_p^p$ and $\gamma(x) = (p-1)\|x\|_p^p$ then $\langle G, g, \gamma \rangle$ is a Gordon triple [17]. Given the above Gordon triple we have

$$\left(\mathbb{E} \left[\max_{\phi \in \Phi} R_t^\phi \right] \right)^p \leq \mathbb{E}[\|(R_t^\Phi)^+\|]_p^p \tag{6}$$

$$= \mathbb{E}[G(R_t^\Phi)] \tag{7}$$

$$\leq G(0) + t \sup_{a \in A, r \in \Pi} \gamma(\rho^\Phi(a,r)) + 2U \sum_{s=1}^{t} \|g(R_{s-1}^\Phi) - g(\tilde{R}_{s-1}^\Phi)\|_1 \tag{8}$$

$$\leq G(0) + tU^p(\mu(\Phi)) + 2U \sum_{k=1}^{t} \|g(R_{k-1}^\Phi) - g(\tilde{R}_{k-1}^\Phi)\|_1 \tag{9}$$

The first inequality is from Lemma 1. The second inequality follows from Corollary 1 and theorem 2. The third inequality is an application of Lemma 2. The result then immediately follows. $\square$

Theorem 5. *Given an ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and an exponential link function $f(x)_i = e^{\eta x_i}$ with $\eta > 0$, then an approximate (Φ, f)- regret-matching algorithm guarantees*

$$\mathbb{E} \left[\max_{\phi \in \Phi} \frac{1}{t} R_t^\phi \right] \leq \frac{1}{t} \left(\frac{\ln |\Phi|}{\eta} + 2U \sum_{k=1}^{t} \|g(R_{k-1}^\Phi) - g(\tilde{R}_{k-1}^\Phi)\|_1 \right) + \frac{\eta U^2}{2}$$

where $g : \mathbb{R}^{|\Phi|} \mapsto \mathbb{R}_+^{|\Phi|}$ and $g(x)_i = e^{\eta x_i} / \sum_j e^{\eta x_j}$.

Proof. The proof follows closely to [Theorem 13][17]. Taking $G(x) = \frac{1}{\eta}\ln\left(\sum_i e^{\eta x_i}\right)$ and $\gamma(x) = \frac{\eta}{2}\|x\|_\infty^2$ then $\langle G, g, \gamma \rangle$ is a Gordon triple [17]. Given the above Gordon triple we have

$$\mathbb{E} \left[\max_{\phi \in \Phi} \eta R_t^\phi \right] = \mathbb{E} \left[\ln e^{\max_{\phi \in \Phi} \eta R_t^\phi} \right] \tag{10}$$

$$= \mathbb{E} \left[\ln \max_{\phi \in \Phi} e^{\eta R_t^\phi} \right] \tag{11}$$

$$\leq \mathbb{E} \left[\ln \sum_{\phi \in \Phi} e^{\eta R_t^\phi} \right] \tag{12}$$

$$= \eta \mathbb{E}[G(R_t^\Phi)] \tag{13}$$

$$\leq \eta \left(G(0) + t \sup_{a \in A, r \in \Pi} \gamma(\rho^\Phi(a,r)) + 2U \sum_{s=1}^{t} \|g(R_{s-1}^\Phi) - g(\tilde{R}_{s-1}^\Phi)\|_1 \right) \tag{14}$$

$$\leq \eta \left(G(0) + t\frac{\eta}{2}U^2 + 2U \sum_{s=1}^{t} \|g(R_{s-1}^\Phi) - g(\tilde{R}_{s-1}^\Phi)\|_1 \right) \tag{15}$$

The second inequality follows from Corollary 1 and theorem 2. The result then immediately follows. $\square$