# Bounds for Regret-Matching Algorithms

Amy Greenwald, Zheng Li and Casey Marks

Department of Computer Science
Brown University
Providence, Rhode Island 02912

**CS-06-10**
June 2006

Brown University Technical Report (2006) CS-06-10

# Bounds for Regret-Matching Algorithms

**Amy Greenwald** AMY@BROWN.EDU
*Department of Computer Science*
*Brown University, Box 1910*
*Providence, RI 02912*

**Zheng Li** ZHENG@DAM.BROWN.EDU
*Division of Applied Mathematics*
*Brown University, Box F*
*Providence, RI 02912*

**Casey Marks** CASEY@CS.BROWN.EDU
*Department of Computer Science*
*Brown University, Box 1910*
*Providence, RI 02912*

## Abstract

We introduce a general class of learning algorithms, which we call regret-matching algorithms, along with a general framework for analyzing their performance in online decision problems. In an online decision problem, an agent repeatedly chooses from a set of actions and receives a reward that depends on its choice of action. Our analytical framework is based on a set $\Phi$ of transformations over the agent's set of actions. We calculate a $\Phi$-regret vector by comparing the average reward obtained by an agent over some finite sequence of rounds to the average reward that could have been obtained had the agent instead played each transformation $\phi \in \Phi$ of its sequence of actions. Regret-matching algorithms select the agent's next action based on the vector of $\Phi$-regrets and a link function $f$. In this paper, we derive bounds on the regret experienced by $(\Phi, f)$-regret matching algorithms for polynomial and exponential link functions. In special cases, we improve upon the bounds reported in past work. Of greater interest, though, is the fact that our analysis is more general than others because it does not rely on Taylor's theorem. Hence, we can analyze algorithms based on a wider variety of link functions, in particular non-differentiable link functions, such as the class of polynomial link functions.

## 1. Introduction

In this paper, we introduce a general class of learning algorithms, which we call regret-matching algorithms, along with a general framework for analyzing their performance in online decision problems (ODPs). In an ODP, an agent repeatedly chooses from a set of actions and receives a reward. The particular reward function that applies at any given round is not revealed until after the agent has made its choice of action for that round. Additionally, the reward function may change *arbitrarily* from round to round; no underlying assumptions govern the reward dynamics.

Online decision problems encompass a wide variety of machine learning settings. Consider an agent learning in an infinitely repeated matrix game, for example. The reward dynamics are jointly determined by the matrix and the behavior of the other players. An ODP—with arbitrary reward dynamics—is sufficiently general to model the nonstationarity of multiagent learning.

In this paper, we study regret-matching algorithms, a class of "regret-minimization" algorithms. The power of regret-matching algorithms is that they minimize regret in any ODP, that is, an agent learning via regret-matching in an arbitrary ODP does not experience too much regret for not having played alternative sequences of actions other than the one it did in fact play.

Following Greenwald et al. (2006), our analytical framework is based on a set $\Phi$ of transformations over the agent's set of actions. We calculate a $\Phi$-regret vector by comparing the average reward obtained by an agent over some finite sequence of rounds to the average reward that could have been obtained had the agent instead played each transformation $\phi \in \Phi$ of its sequence of actions. Regret-matching algorithms select the agent's next action based on the vector of $\Phi$-regrets and a link function $f$. Many well-studied learning algorithms can be viewed as instances of regret matching (e.g., Freund and Schapire (1997) and Foster and Vohra (1999)).

Like Cesa-Bianchi and Lugosi (2003), we derive bounds on the regret experienced by $(\Phi, f)$-regret matching algorithms for polynomial and exponential link functions. In special cases, we improve upon the bounds reported in their work. Of greater interest, though, is the fact that our analysis is more general than theirs because it does not rely on Taylor's theorem. Hence, we can analyze algorithms based on a wider variety of link functions, in particular non-differentiable link functions, such as the class of polynomial link functions.[1] In ongoing work, we are studying regret-matching algorithms with alternative link functions, beyond polynomial and exponential.

This paper is organized as follows. In Section 2, we formally present our analytical framework, building up to the definition of regret-matching algorithms, which is presented in Section 3. In Section 3, we also prove the regret-matching theorem, which states that regret-matching algorithms satisfy a property related to Blackwell's condition for approachability (1956), applied in the regret domain. Next, in Section 4, we present our main analytical tool, which is inspired by the work of Gordon (2005). In Section 5, we derive and optimize bounds for specific regret-matching algorithms, namely those based on polynomial and exponential link functions.

## 2. Formal Framework

In this section, we present formal definitions of the key concepts in our analytical framework: action transformations, the regret vector, and link functions.

### 2.1 Online Decision Problems

An online decision problem is parameterized by a *reward system* $(A, \mathcal{R})$, where $A$ is a set of actions and $\mathcal{R}$ is a set of rewards. A *real-valued* reward system is one in which $\mathcal{R} \subseteq \mathbb{R}$. When we restrict our attention to *bounded* (real-valued) reward systems, we assume WLOG that $\mathcal{R} = [0, 1]$. Given a reward system $(A, \mathcal{R})$, we let $\Pi = \{f : A \to \mathcal{R}\}$ denote the set of *reward functions*.

In this paper, we assume that the agent's action set $A$ is finite. We denote by $\Delta(A)$ the set of probability distributions over the set $A$, and we allow agents to play *mixed strategies*, which means that rather than selecting an action $a \in A$ to play at each round, the agent learns a probability distribution $q \in \Delta(A)$. Hence, round $t$ (for $t = 1, 2, \ldots$) proceeds as follows:

1. the agent chooses a mixed strategy $q_t \in \Delta(A)$,
2. the agent plays an (pure) action $a_t \sim q_t$,
3. the agent receives reward $r_t(a_t) \in \mathcal{R}$,
4. the agent is informed of $r_t \in \Pi$.

The last step, in which the agent is informed of the reward function $r_t$ (i.e., the rewards associated with actions it did not choose to play) characterizes *informed* ODPs, which are the subject of this work. Without this step, informing the agent of only $r_t(a_t)$, instead the setup would be that of a *naïve* ODP. See Auer et al. (1995), for example, for consideration of the naïve setting.

1. There appears to be a problem with the analysis in Cesa-Bianchi and Lugosi (2003). Taylor's theorem is applied to the polynomial (potential) function $G(x) = \|x^+\|_p^2$ for $p \geq 2$, whose gradient is the polynomial link function. In doing so, the authors implicitly assume that $G$ is twice differentiable everywhere. However, $G$ is not twice differentiable on the boundary of the negative orthant (see Appendix B, Lemma 21).

**Definition 1** *Given a reward system $(A, \mathcal{R})$, a particular instance of an* **online decision problem** *can be described by a* reward schedule*, that is, a sequence of reward functions $\{r_t\}_{t=1}^{\infty}$, where $r_t \in \Pi$, so that $r_t(a) \in \mathcal{R}$ corresponds to the reward for playing action $a$ at time $t \geq 1$. A* **bounded** *ODP is an ODP over a bounded reward system.*

Given a reward system $(A, \mathcal{R})$, *the set of histories of length* $t \geq 1$ is given by $H_t \equiv A^t \times \Pi^t$. We define $H_0$ to be a singleton and we let $\mathcal{H} = \cup_{t=0}^{\infty} H_t$ denote the complete set of histories.

Given an ODP, the particular history $(\{a_\tau\}_{\tau=1}^t, \{r_\tau\}_{\tau=1}^t)$ corresponds to the agent playing $a_\tau$ and observing reward function $r_\tau$ at each time $\tau = 1, \ldots, t$. An *online learning algorithm* is a sequence of functions $\{L_t\}_{t=1}^{\infty}$, where $L_t : H_{t-1} \to \Delta(A)$ so that $L_t(h) \in \Delta(A)$ represents the agent's mixed strategy at time $t \geq 1$, having observed history $h \in H_{t-1}$.

### 2.2 Action Transformations

Fix a (finite) set of actions $A$. An *action transformation* is defined as a function $\phi : A \to \Delta(A)$. We let $\Phi_{\text{ALL}} \equiv \Phi_{\text{ALL}}(A)$ denote the set of all action transformations over the set $A$. Following Blum and Mansour (2005), we also let $\Phi_{\text{SWAP}} \equiv \Phi_{\text{SWAP}}(A) \subseteq \Phi_{\text{ALL}}$ denote the set of all action transformations that map actions to distributions with all their weight on a single action: i.e., pure strategies.

There are two well-studied subsets of $\Phi_{\text{SWAP}}$: the set of external transformations and the set of internal transformations. Let $\delta_a \in \Delta(A)$ denote the distribution with all its weight on action $a$. An *external* action transformation is simply a constant transformation, so for $a \in A$,

$$\phi_{\text{EXT}}^{(a)} : x \mapsto \delta_a, \quad \text{for all } x \in A \tag{1}$$

An *internal* action transformation behaves like the identity, except on one particular input, so for $a, b \in A$

$$\phi_{\text{INT}}^{(a,b)} : x \mapsto \begin{cases} \delta_b & \text{if } x = a \\ \delta_x & \text{otherwise} \end{cases} \tag{2}$$

We let $\Phi_{\text{EXT}} \equiv \Phi_{\text{EXT}}(A)$ denote the set of external transformations and $\Phi_{\text{INT}} \equiv \Phi_{\text{INT}}(A)$ denote the set of internal transformations. Observe that $|\Phi_{\text{SWAP}}| = |A|^{|A|}$, $|\Phi_{\text{INT}}| = |A|^2 - |A| + 1$, and $|\Phi_{\text{EXT}}| = |A|$.

An action transformation can be extended so that its domain is the set of mixed strategies. Given an action transformation $\phi : A \to \Delta(A)$, let $[\phi] : \Delta(A) \to \Delta(A)$ be the linear transformation given by

$$([\phi](q))(a) = \sum_{a' : \phi(a') = a} q(a') \tag{3}$$

for all $q \in \Delta(A)$, for all $a \in A$.

### 2.3 The Regret Vector

Given a reward system $(A, \mathcal{R})$, a reward function $r \in \Pi$, an action $a \in A$, and an action transformation $\phi \in \Phi_{\text{ALL}}$, the *$\phi$-regret* is given by $\rho^{\phi}(a, r) = \mathbb{E}_{a' \sim \phi(a)}[r(a') - r(a)]$. This quantity is the difference between the rewards the agent obtains by playing action $a$ and the rewards that the agent could have expected to obtain by playing the transformed action $\phi(a)$. Given a set of action transformations $\Phi \subseteq \Phi_{\text{ALL}}$, the *$\Phi$-regret vector* is given by $\rho^{\Phi}(a, r) = (\rho^{\phi}(a, r))_{\phi \in \Phi}$. In this work, we restrict our attention to finite action transformation sets $\Phi \subseteq \Phi_{\text{ALL}}$ so that, for real-valued reward systems, the $\Phi$-regret vector is an element of finite-dimensional Euclidean space.

Given an ODP $\{r_\tau\}_{\tau=1}^{\infty}$ and an infinite sequence of the agent's actions $\{a_\tau\}_{\tau=1}^{\infty}$, the *cumulative $\Phi$-regret vector* at time $t \geq 1$, where $\Phi \subseteq \Phi_{\text{ALL}}$, is computed as follows:

$$R_t^{\Phi}\left(\{a_\tau\}, \{r_\tau\}\right) = \sum_{\tau=1}^{t} \rho^{\Phi}(a_\tau, r_\tau) \tag{4}$$

The cumulative $\Phi$-regret vector compares the cumulative rewards obtained by the agent with the rewards that the agent could have expected to obtain by consistently transforming each of its actions according to each $\phi \in \Phi$. We define $R_0^\Phi(\{a_\tau\}, \{r_\tau\}) = 0$.

We sometimes treat cumulative regret as a function from histories of length $T$ to regret vectors, in which case we write $R_t^\Phi(h)$, for $h \in H_T$ and $t \leq T$.

When considering a particular ODP $\{r_t\}_{t=1}^\infty$ together with an online learning algorithm $\{L_t\}_{t=1}^\infty$, we treat cumulative regret as a random vector over a probability space whose universe consists of infinite sequences of the agent's actions $\{a_\tau\}_{\tau=1}^\infty$ and whose measure is defined by the learning algorithm. Formally, this measure can be defined inductively as follows: for all $\alpha \in A$,

$$\Pr\left[a_t = \alpha \mid a_\tau = \alpha_\tau,\ \forall \tau = 1, \ldots, t-1\right] = L_t((\alpha_1, \ldots, \alpha_{t-1}), (r_1, \ldots, r_{t-1}))(\alpha) \tag{5}$$

In this case, we write $R_t^\Phi$ and $\rho_t^\Phi$ for the cumulative and instantaneous $\Phi$-regret vectors, respectively.

### 2.4 Link Functions

Given two sets $X$ and $Y$, $X^Y = \{f : Y \to X\}$. If $|Y| = n < \infty$, then $\mathbb{R}^Y$ is isomorphic to $\mathbb{R}^n$. We denote the positive and negative orthants of $\mathbb{R}^n$ by $\mathbb{R}^n_+ \equiv \{x \in \mathbb{R}^n \mid x_i \geq 0,\ \forall i\}$ and $\mathbb{R}^n_- \equiv \{x \in \mathbb{R}^n \mid x_i \leq 0,\ \forall i\}$, respectively. A *link* function $f : \mathbb{R}^n \to \mathbb{R}^n$ is a function that is subgradient to a convex function $F : \mathbb{R}^n \to \mathbb{R}$, sometimes called a *potential* function. By the next lemma, without loss of generality, we can assume that the co-domain of any link function $f$ is the positive orthant $\mathbb{R}^n_+$ if $f$ is subgradient to a convex function $F : \mathbb{R}^n \to \mathbb{R}$ that is bounded on the negative orthant: i.e., $F(\mathbb{R}^n_-) \subseteq (-\infty, k]$, for some $k < \infty$.

**Lemma 2** *The co-domain of a link function $f : \mathbb{R}^n \to \mathbb{R}^n$ that is subgradient to a convex function $F : \mathbb{R}^n \to \mathbb{R}$ which is bounded on the negative orthant is the positive orthant: i.e., $f : \mathbb{R}^n \to \mathbb{R}^n_+$.*

The proof of this lemma appears in Appendix A.

Although our main analytical tool, which bounds the regret accumulated by regret-matching algorithms, can be applied more broadly, we demonstrate its efficacy on two classes of link functions, the polynomial and the exponential, which give rise to two well-studied classes of regret-matching algorithms. The polynomial link function $f_i(x) = \left(x_i^+\right)^{p-1}$ with parameter $p \in \mathbb{R}$, where $a^+ = \max\{a, 0\}$, for all $a \in \mathbb{R}$. The exponential link function $f_i(x) = e^{\eta x_i}$ with parameter $\eta \in \mathbb{R}$.

## 3. Regret-Matching Algorithms

In this section, we define a general class of online learning algorithms, which we call regret-matching algorithms,[2] that are parameterized by a set of action transformations $\Phi$ and a link function $f$. In addition, we prove the regret-matching theorem, which states that the algorithms in this class satisfy the generalized $(\Phi, f)$-Blackwell regret condition.

### 3.1 Generalized Blackwell Regret Condition

Blackwell's seminal approachability theorem (1956) provides a sufficient condition for the players' time-averaged payoffs to "approach" a closed subset of $\mathbb{R}^n$ in a vector-valued repeated game. Given a set $\Phi$ of action transformations, in the repeated game whose payoffs are $\Phi$-regret vectors, we obtain from Blackwell's theorem a sufficient condition for the players' online learning algorithms to exhibit no-$\Phi$-regret (see Definition 17). We call this the "$\Phi$-Blackwell regret condition." Here we present the generalized $(\Phi, f)$-Blackwell regret condition, which, if the link function is taken to be the polynomial link function with $p = 2$, reduces to the $\Phi$-Blackwell regret condition.

2. We co-opt this terminology from Hart and Mas-Colell (2001), whose regret-matching algorithms based on $\Phi_{\text{EXT}}$ and the polynomial link functions are instances of this class.

**Algorithm 1** $(\Phi, f)$-RegretMatchingAlgorithm(ODP $\{r_t\}_{t=1}^{\infty}$, numRounds $T$)

initialize $X_0^\Phi = 0$
**for** $t = 1, \ldots, T$ **do**
  sample pure action $a_t \sim q_t$
  observe reward function $r_t$
  **for all** $\phi \in \Phi$ **do**
    compute instantaneous regret $x_t^\phi = \rho^\phi(a_t, r_t)$
    update cumulative regret vector $X_t^\phi = X_{t-1}^\phi + x_t^\phi$
  **end for**
  let $q_{t+1} = (\Phi, f)$-ComputeMixedStrategy$(X_t^\Phi)$
**end for**

**Definition 3** *Given a reward system $(A, \mathcal{R})$, a finite set $\Phi \subseteq \Phi_{ALL}$ of action transformations, and a link function $f : \mathbb{R}^\Phi \to \mathbb{R}_+^\Phi$, an online learning algorithm $\{L_t\}_{t=1}^{\infty}$ is said to satisfy the* generalized $(\Phi, f)$-Blackwell regret condition *if $f\left(R_{t-1}^\Phi(h)\right) \cdot \mathbb{E}_{a \sim L_t(h)}\left[\rho^\Phi(a, r)\right] \leq 0$, for all times $t \geq 1$, for all histories $h \in H_{t-1}$ of length $t-1$, and for all reward functions $r : A \to \mathcal{R}$.*

### 3.2 Regret-Matching Algorithms

Fix a real-valued reward system $(A, \mathcal{R})$, a finite set $\Phi \subseteq \Phi_{\text{ALL}}$ of action transformations, and a link function $f : \mathbb{R}^\Phi \to \mathbb{R}_+^\Phi$. For any history $h \in H_{t-1}$, recall that the quantity $R_{t-1}^\Phi(h)$ denotes the cumulative regret, through time $t-1$, for not having played according to the action transformations contained in $\Phi$. If $R_{t-1}^\Phi(h) \in \mathbb{R}_-^\Phi$, so that each element of the regret vector is non-positive, then the agent does not regret its past actions. In this case, a $(\Phi, f)$-regret-matching algorithm leaves the agent's next play unspecified. But if $R_{t-1}^\Phi(h) \notin \mathbb{R}_-^\Phi$, so that the agent "feels" regret in at least one dimension, then applying the link function $f$ to this quantity yields a non-negative vector, call it $Y_t^\Phi \in \mathbb{R}_+^\Phi$. Normalizing this vector, we compute the linear transformation $M_t$ as a convex combination of the linear transformations $[\phi]$ as follows:

$$M_t = \frac{\sum_{\phi \in \Phi} Y_t^\phi [\phi]}{\sum_{\phi \in \Phi} Y_t^\phi} \tag{6}$$

The function $M_t$ maps $\Delta(A)$, a nonempty compact convex set, into itself. Moreover, $M$ is continuous, since all $\phi \in \Phi$ are linear functions in finite-dimensional Euclidean space, and hence continuous. Therefore, by Brouwer's fixed point theorem, $M_t$ has a fixed point $q \in \Delta(A)$. A $(\Phi, f)$-regret-matching algorithm plays the mixed strategy $q$ whenever $R_{t-1}^\Phi(h) \notin \mathbb{R}_-^\Phi$.

The pseudocode for the class of $(\Phi, f)$-regret-matching algorithms is shown in Algorithm 1. The cumulative regret vector is initialized to zero. For all times $t = 1, \ldots, T$, the agent samples a pure action $a_t$ according to the distribution $q_t$, after which it observes the reward function $r_t$. Given $a_t$ and $r_t$, the agent computes its instantaneous regret with respect to each $\phi \in \Phi$, and updates the cumulative $\Phi$-regret vector accordingly. A subroutine is then called to compute the mixed strategy that the agent learns to play at time $t+1$. In this subroutine, the link function $f$ is applied to the cumulative $\Phi$-regret vector. (Recall that the co-domain of a link function is the positive orthant.) If this quantity is zero, then the subroutine returns an arbitrary mixed strategy. Otherwise, the subroutine returns a fixed point of the stochastic matrix $M_t$.

Many well-known online learning algorithms arise as instances, or close cousins, of Algorithm 1. The no-external-regret algorithm of Hart and Mas-Colell (2000) (Theorem B) is the special case of $(\Phi, f)$-regret matching in which $\Phi = \Phi_{\text{EXT}}$ and $f$ is the polynomial link function with $p = 2$. The no-internal-regret algorithm of Foster and Vohra (1999) is equivalent to $(\Phi, f)$-regret matching with

**Algorithm 2** $(\Phi, f)$-ComputeMixedStrategy(regret vector $X_t^\Phi$)

let $Y_t^\Phi = f(X_t^\Phi)$
**if** $Y_t^\Phi = 0$ **then**
  set $q \in \Delta(A)$ arbitrarily
**else**
  let $M_t = \sum_{\phi\in\Phi} Y_t^\phi[\phi] / \sum_{\phi\in\Phi} Y_t^\phi$
  solve for a fixed point $q = M_t(q)$
**end if**
**return** $q$

$\Phi = \Phi_{\text{INT}}$ and the polynomial link function with $p = 2$.[3] If $f$ is the exponential link function, then $(\Phi, f)$-regret matching reduces to Freund and Schapire's Hedge algorithm (1997) when $\Phi = \Phi_{\text{EXT}}$ and a variant of an algorithm discussed by Cesa-Bianchi and Lugosi (2003) when $\Phi = \Phi_{\text{INT}}$.

### 3.3 Complexity

Each iteration of Algorithm 1 has time complexity $O(\max\{|\Phi||A|^2, |A|^3\})$, assuming the time complexity of $f$ is linear in $\Phi$ (as it is for the polynomial and exponential link functions). Updating the cumulative regret vector in steps 5–8 of Algorithm 1 takes time $O(|\Phi||A|)$, since computing instantaneous regret for each $\phi \in \Phi$ (step 6) is an $O(|A|)$ operation. In Subroutine 2, step 1 takes time $O(|\Phi|)$, by assumption. If we view the mixed strategy transformations $[\phi]$ as $|A| \times |A|$ stochastic matrices, then we can also view $M_t$ as an $|A| \times |A|$ stochastic matrix, as it is a convex combination of the elements of $\Phi$. Computing $M_t$ in step 5 takes time $O(|\Phi||A|^2)$, since each matrix $[\phi]$ has dimensions $|A| \times |A|$. Finding the fixed point of an $n \times n$ stochastic matrix, which can be accomplished, for example, via Gaussian elimination, is an $O(n^3)$ operation.

If, however, $\Phi \subseteq \Phi_{\text{SWAP}}(A)$, then this time complexity reduces to $O(\max\{|\Phi||A|, |A|^3\})$, since in this special case, (i) computing instantaneous regret for each $\phi \in \Phi$ (step 6 of Algorithm 1) takes constant time so that updating the cumulative regret vector takes time $O(|\Phi|)$; and (ii) computing the stochastic matrix $M_t$ in step 5 of Subroutine 2 is only an $O(|\Phi||A|)$ operation, since there are only $|A|$ nonzero entries in each $\phi \in \Phi$. In particular, if $\Phi = \Phi_{\text{INT}}(A)$, then the time complexity reduces to $O(|A|^3)$, because $|\Phi_{\text{INT}}(A)| = O(|A|^2)$. Moreover, if $\Phi = \Phi_{\text{EXT}}(A)$, then the time complexity reduces even further to $O(|A|)$, because matrix manipulation is not required in the special case of $\Phi_{\text{EXT}}$-regret matching. The rows of $M$ are constant: each is a copy of the (normalized) cumulative regret vector, which is precisely the fixed point of $M$. In particular, for all $q \in \Delta(A)$,

$$M_t(q) : a \quad \mapsto \quad \frac{\sum_{\phi\in\Phi_{\text{EXT}}} Y_t^\phi[\phi]}{\sum_{\phi\in\Phi_{\text{EXT}}} Y_t^\phi}(q) \tag{7}$$

$$= \quad \frac{\sum_{a'\in A} Y_t^{a'} \begin{cases} 1 & \text{if } a = a' \\ 0 & \text{otherwise} \end{cases}}{\sum_{a'\in A} Y_t^{a'}} \tag{8}$$

$$= \quad \frac{Y_t^a}{\sum_{a'\in A} Y_t^{a'}} \tag{9}$$

Observe that $M_t(q)$ is independent of $q$, so that

$$q : a \mapsto \frac{Y_t^a}{\sum_{a'\in A} Y_t^{a'}} \tag{10}$$

3. Foster and Vohra (1999) calculate the fixed points of a stochastic matrix that is derived from the internal regret vector. Their matrix is identical to $M_t$ in its off-diagonal entries up to normalization; consequently, both matrices have the same set of fixed points (Greenwald et al. (2006)).

**Algorithm 3** ($\Phi_{\text{EXT}}, f$)-ComputeMixedStrategy(regret vector $X_t^A$)

let $Y_t^A = f(X_t^A)$
**if** $Y_t^A = 0$ **then**
  set $q \in \Delta(A)$ arbitrarily
**else**
  **for all** $a \in A$ **do**
    set $q_a = Y_t^a / \sum_{a \in A} Y_t^a$
  **end for**
**end if**
**return** $q$

is the unique fixed point of $M_t$. Subroutine 3 computes the mixed strategy for an external regret-matching algorithm: at time $t+1$, play action $a$ with probability proportional to $Y_t^a$.

The space complexity of Algorithm 1 (and Subroutine 2) is $O(|\Phi||A|^2) = O(\max\{|\Phi||A|^2, |A|^2\})$ because it is necessary to store the $|\Phi|$ matrices, each with dimensions $|A| \times |A|$, and computing the fixed point of an $|A| \times |A|$ stochastic matrix (via Gaussian elimination) requires $O(|A|^2)$ space. If, however, $\Phi \subseteq \Phi_{\text{SWAP}}(A)$, then the space complexity reduces to $O(\max\{|\Phi||A|, |A|^2\})$, since, in this case, there are only $|A|$ nonzero entries in each $\phi \in \Phi$. In particular, if $\Phi = \Phi_{\text{INT}}(A)$ then the space complexity reduces to $O(|A|^2)$, since it suffices to store cumulative regrets in a matrix of size $|A| \times |A|$. Similarly, if $\Phi = \Phi_{\text{EXT}}(A)$ (Subroutine 3), then the space complexity reduces to $O(|A|)$, since it suffices to store cumulative regrets in a vector of size $|A|$. Our discussion of the time and space complexity of Algorithm 1 and its subroutines is summarized in Table 1.

Table 1: Complexity of $(\Phi, f)$-Regret Matching

| | Time | Space |
|---|---|---|
| $\Phi \subseteq \Phi_{\text{ALL}}$ | $O(\max\{\|\Phi\|\|A\|^2, \|A\|^3\})$ | $O(\|\Phi\|\|A\|^2)$ |
| $\Phi \subseteq \Phi_{\text{SWAP}}$ | $O(\max\{\|\Phi\|\|A\|, \|A\|^3\})$ | $O(\max\{\|\Phi\|\|A\|, \|A\|^2\})$ |
| $\Phi = \Phi_{\text{INT}}$ | $O(\|A\|^3)$ | $O(\|A\|^2)$ |
| $\Phi = \Phi_{\text{EXT}}$ | $O(\|A\|)$ | $O(\|A\|)$ |

### 3.4 Regret-Matching Theorem

We now prove the regret-matching theorem, which states that each $(\Phi, f)$-regret-matching algorithm satisfies the generalized $(\Phi, f)$-Blackwell regret condition (with equality).

**Theorem 4 (Regret-Matching Theorem)** *Given a real-valued reward system $(A, \mathcal{R})$, a finite set $\Phi \subseteq \Phi_{ALL}$ of action transformations, and a link function $f : \mathbb{R}^\Phi \to \mathbb{R}^\Phi_+$, the $(\Phi, f)$-regret-matching algorithm $\{L_t\}_{t=1}^\infty$ satisfies the $(\Phi, f)$-Blackwell regret condition (with equality).*

**Proof** For all times $t \geq 1$ and for all histories $h \in H_{t-1}$, we abbreviate as follows: $Y_t^\Phi \equiv f\left(R_{t-1}^\Phi(h)\right)$. Since $f$ is a link function, $Y_t^\Phi$ is non-negative. If $Y_t^\Phi$ is the zero vector, the result follows immediately.

Otherwise, for all reward functions $r : A \to \mathcal{R}$,

$$
\begin{aligned}
& Y_t^{\Phi} \cdot \mathbb{E}_{a \sim L_t(h)} \left[ \rho^{\Phi}(a, r) \right] && (11)\\
= & \sum_{\phi \in \Phi} Y_t^{\phi} \left( \mathbb{E}_{a \sim L_t(h)} \left[ \rho^{\phi}(a, r) \right] \right) && (12)\\
= & \sum_{\phi \in \Phi} Y_t^{\phi} \left( r \cdot [\phi](L_t(h)) - r \cdot L_t(h) \right) && (13)\\
= & \; r \cdot \sum_{\phi \in \Phi} Y_t^{\phi} \left( [\phi](L_t(h)) - L_t(h) \right) && (14)\\
= & \; r \cdot \left( \sum_{\phi \in \Phi} Y_t^{\phi} [\phi](L_t(h)) - \sum_{\phi \in \Phi} Y_t^{\phi} L_t(h) \right) && (15)\\
= & \; r \cdot \sum_{\phi \in \Phi} Y_t^{\phi} \left( \frac{\sum_{\phi \in \Phi} Y_t^{\phi}[\phi]}{\sum_{\phi \in \Phi} Y_t^{\phi}} (L_t(h)) - L_t(h) \right) && (16)\\
= & \; r \cdot \sum_{\phi \in \Phi} Y_t^{\phi} \left( M_t(L_t(h)) - L_t(h) \right) && (17)\\
= & \; 0 && (18)
\end{aligned}
$$

Line (18) follows from the definition of a $(\Phi, f)$-regret-matching algorithm, which ensures that $L_t(h)$ is a fixed point of $M_t$: i.e., $M_t(L_t(h)) - L_t(h) = 0$. Line (13) can be derived as follows: for all $q \in \Delta(A)$, for all $\phi \in \Phi$, for all $a \in A$, for all $r : A \to \mathcal{R}$,

$$
\begin{aligned}
& \mathbb{E}_{a \sim q} \left[ \rho^{\phi}(a, r) \right] && (19)\\
= & \; \mathbb{E}_{a \sim q} \left[ r(\phi(a)) - r(a) \right] && (20)\\
= & \; \mathbb{E}_{a \sim q} \left[ r(\phi(a)) \right] - \mathbb{E}_{a \sim q} \left[ r(a) \right] && (21)\\
= & \sum_{a \in A} r(\phi(a))\, q(a) - \sum_{a \in A} r(a)\, q(a) && (22)\\
= & \sum_{a \in A} r(a) \left( \sum_{a' : \phi(a') = a} q(a') \right) - \sum_{a \in A} r(a)\, q(a) && (23)\\
= & \sum_{a \in A} r(a)([\phi](q))(a) - \sum_{a \in A} r(a)\, q(a) && (24)\\
= & \; r \cdot [\phi](q) - r \cdot q && (25)
\end{aligned}
$$

Line (23) follows from the fact that $\sum_{a \in A} r(\phi(a)) q(a) = \sum_{a' \in A} r(a') \sum_{\{a \mid \phi(a) = a'\}} q(a)$.
Line (24) follows from the definition of $[\phi]$. ■

In the next section, as a consequence of the regret-matching theorem, we derive a general bound on regret-matching algorithms. Later, we instantiate this general bound to derive specific bounds for particular choices of (classes of) link functions and particular sets of action transformations.

## 4. A General Bound on Regret Matching

In this section, we derive a general bound on regret-matching algorithms. Our foundational result is a corollary of (what is essentially) Gordon's Gradient Descent Theorem (2005), also presented in this section, which bounds the growth rate of any real-valued function on $\mathbb{R}^n$.

**Definition 5** *A* Gordon triple $\langle G, g, \gamma \rangle$ *consists of three functions* $G : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^n$, *and* $\gamma : \mathbb{R}^n \to \mathbb{R}$ *such that for all* $x, y \in \mathbb{R}^n$, $G(x + y) \leq G(x) + g(x) \cdot y + \gamma(y)$.

Given a Gordon triple $\langle G, g, \gamma \rangle$, if $G$ is smooth and $g$ is equal to the gradient of $G$, then $\gamma(y)$ is a bound on the higher order terms of the Taylor expansion of $G(x + y)$.

**Theorem 6 (Gordon 2005)** *Assume* $\langle G, g, \gamma \rangle$ *is a Gordon triple and* $C : \mathbb{N} \to \mathbb{R}$. *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* $t \geq 1$,

$$g(X_{t-1}) \cdot \mathbb{E}\left[x_t \mid X_{t-1}\right] + \mathbb{E}\left[\gamma(x_t) \mid X_{t-1}\right] \leq C(t) \quad \text{a.s.} \tag{26}$$

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

$$\mathbb{E}\left[G(X_t)\right] \leq G(X_0) + \sum_{\tau=1}^{t} C(\tau) \tag{27}$$

**Proof** The proof is by induction on $t$. At time $t = 0$, $\mathbb{E}\left[G(X_t)\right] = G(X_0)$ and $\sum_{\tau=1}^{t} C(\tau) = 0$.

At time $t \geq 1$, since $X_t = X_{t-1} + x_t$ and $\langle G, g, \gamma \rangle$ is a Gordon triple,

$$\begin{aligned} G(X_t) &= G(X_{t-1} + x_t) && (28)\\ &\leq G(X_{t-1}) + g(X_{t-1}) \cdot x_t + \gamma(x_t) && (29) \end{aligned}$$

By Assumption 26, taking conditional expectations w.r.t. $X_{t-1}$ yields

$$\begin{aligned} \mathbb{E}\left[G(X_t) \mid X_{t-1}\right] &\leq G(X_{t-1}) + g(X_{t-1}) \cdot \mathbb{E}\left[x_t \mid X_{t-1}\right] + \mathbb{E}\left[\gamma(x_t) \mid X_{t-1}\right] && (30)\\ &\leq G(X_{t-1}) + C(t) \quad \text{a.s.} && (31) \end{aligned}$$

Taking expectations and applying the law of iterated expectations yields

$$\begin{aligned} \mathbb{E}\left[G(X_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[G(X_t) \mid X_{t-1}\right]\right] && (32)\\ &\leq \mathbb{E}\left[G(X_{t-1})\right] + C(t) && (33) \end{aligned}$$

Therefore, by the induction hypothesis,

$$\begin{aligned} \mathbb{E}\left[G(X_t)\right] &\leq \mathbb{E}\left[G(X_{t-1})\right] + C(t) && (34)\\ &\leq G(X_0) + \sum_{\tau=1}^{t-1} C(\tau) + C(t) && (35)\\ &= G(X_0) + \sum_{\tau=1}^{t} C(\tau) && (36) \end{aligned}$$

■

We apply Gordon's theorem to derive the following corollary:

**Corollary 7** *Given a real-valued reward system* $(A, \mathcal{R})$ *and a finite set* $\Phi \subseteq \Phi_{ALL}$ *of action transformations. If* $\langle G, g, \gamma \rangle$ *is a Gordon triple, then a* $(\Phi, g)$*-regret-matching algorithm* $\{L_t\}_{t=1}^{\infty}$ *guarantees*

$$\mathbb{E}\left[G(R_t^{\Phi})\right] \leq G(0) + t \sup_{a,r} \gamma\left(\rho^{\Phi}(a, r)\right) \tag{37}$$

*at all times* $t \geq 0$.

**Proof** For all times $t \geq 1$, for all histories $h \in H^{t-1}$, and for all reward functions $r : A \to \mathcal{R}$, $g\left(R^{\Phi}_{t-1}(h)\right) \cdot \mathbb{E}_{a \sim L_t(h)}\left[\rho^{\Phi}(a, r)\right] = 0$, by the regret-matching theorem. The only part of the history through time $t-1$ that impacts the agent's mixed strategy at time $t$ is the cumulative regret vector; hence, we rewrite $\mathbb{E}_{a \sim L_t(h)}\left[\rho^{\Phi}(a, r)\right]$ as $\mathbb{E}\left[\rho^{\Phi}_t \mid R^{\Phi}_{t-1}\right]$. It follows that $g\left(R^{\Phi}_{t-1}\right) \cdot \mathbb{E}\left[\rho^{\Phi}_t \mid R^{\Phi}_{t-1}\right] = 0$. Now apply Theorem 6 with $x_t = \rho^{\Phi}_t$, $X_t = R^{\Phi}_t$, and $C(t) = \sup_{a,r} \gamma\left(\rho^{\Phi}(a, r)\right)$, noting that $R^{\Phi}_0 = 0$. ■

## 5. Bounds on Specific Regret-Matching Algorithms

We now apply Corollary 7 to particular Gordon triples to derive bounds on the regret experienced by agents that employ polynomial and exponential regret-matching algorithms. In doing so, we rely on three Gordon triples. The proofs that our choices of the functions $G : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^n$, and $\gamma : \mathbb{R}^n \to \mathbb{R}$ are indeed Gordon triples are relegated to Appendix B.

We rely on the following lemma in deriving our bounds.

**Lemma 8** *Given a reward system $(A, \mathcal{R})$ and a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, let $f, f'$ be two link functions, both mapping $\mathbb{R}^{\Phi}$ to $\mathbb{R}^{\Phi}_{+}$. If there exists a strictly positive function $\psi : \mathbb{R}^{\Phi} \to \mathbb{R}$ such that $\psi(x) f(x) = f'(x)$ for all $x \in \mathbb{R}^{\Phi}$, then a $(\Phi, f)$-regret-matching algorithm is also a $(\Phi, f')$-regret-matching algorithm.*

**Proof** For arbitrary time $t \geq 1$ and for history $h \in H_{t-1}$, let $M_t$ and $M'_t$ be defined according to Equation 6 for $f$ and $f'$, respectively. Since $\psi$ is strictly positive, $M_t = M'_t$ so that a $(\Phi, f)$-regret-matching algorithm plays a fixed point of $M'_t$, whenever $R^{\Phi}_{t-1}(h) \notin \mathbb{R}^{\Phi}_{-}$. ■

### 5.1 Polynomial Regret-Matching Bounds

The polynomial regret-matching algorithms are based on the polynomial link function $f_i(x) = \left(x_i^+\right)^{p-1}$ with parameter $p \in \mathbb{R}$. We divide our analyses of these algorithms into two cases: $2 < p < \infty$ and $1 \leq p \leq 2$. In both cases, we rely on the following two lemmas.

**Lemma 9** *If $x$ is a random vector that takes values in $\mathbb{R}^n$, then $\left(\mathbb{E}\left[\max_i x_i\right]\right)^q \leq \mathbb{E}\left[\left\|x^+\right\|_p^q\right]$, for all $p, q \geq 1$.*

**Proof**

$$\left(\mathbb{E}\left[\max_i x_i\right]\right)^q \leq \left(\mathbb{E}\left[\left\|x^+\right\|_\infty\right]\right)^q \leq \left(\mathbb{E}\left[\left\|x^+\right\|_p\right]\right)^q \leq \mathbb{E}\left[\left\|x^+\right\|_p^q\right] \tag{38}$$

The first two inequalities follow from the following two facts, respectively: for all $x \in \mathbb{R}^n$,

- $\max_i x_i \leq \max_i x_i^+ = \max_i |x_i^+| = \|x^+\|_\infty$
- $\|x\|_\infty \leq \|x\|_p$, for all $p \geq 1$

The third inequality follows from Jensen's inequality: in particular, $(\mathbb{E}[x])^q \leq \mathbb{E}[x^q]$, for all $q \geq 1$. ■

Given a (finite) set of actions $A$ and a finite set of action transformations $\Phi \subseteq \Phi_{\text{ALL}}$, the *maximal activation*, denoted $\mu(\Phi)$, is computed by maximizing, over all actions $a \in A$, the number of transformations $\phi$ that alter action $a$: i.e.,

$$\mu(\Phi) = \max_A |\{\phi \in \Phi : \phi(a) \neq a\}| \tag{39}$$

Clearly, $\mu(\Phi) \leq |\Phi|$. In addition, observe that $\mu(\Phi_{\text{EXT}}) = \mu(\Phi_{\text{INT}}) = |A| - 1$.

**Lemma 10** *Given a bounded reward system $(A, [0,1])$ and a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, $\|\rho^\Phi(a,r)\|_p \leq (\mu(\Phi))^{1/p}$. for all reward functions $r : A \to [0,1]$.*

**Proof** Since rewards are bounded in $[0,1]$, regrets are bounded in $[-1,1]$, so that

$$\|\rho^\Phi(a,r)\|_p = \sqrt[p]{\sum_{\phi\in\Phi}(\rho^\phi(a,r))^p} \leq \sqrt[p]{\sum_{\phi\in\Phi}\mathbb{1}_{\phi(a)\neq a}} \leq \sqrt[p]{\mu(\Phi)} \tag{40}$$

■

We are now ready to bound the $\Phi$-regret of polynomial regret-matching, that is, regret-matching with the polynomial link function $f_i(x) = \left(x_i^+\right)^{p-1}$. We state two theorems, the first pertaining to $2 < p < \infty$ and the second pertaining to $1 \leq p \leq 2$.

**Lemma 11** *For $2 < p < \infty$, if $G(x) = \|x^+\|_p^2$,*

$$g_i(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \text{otherwise} \end{cases} \tag{41}$$

*and $\gamma(x) = (p-1)\|x\|_p^2$, then $\langle G, g, \gamma\rangle$ is a Gordon triple.*

**Proof** See Appendix B.

**Theorem 12** *Given a bounded ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and a polynomial link function $f$ with $2 < p < \infty$, a $(\Phi, f)$-regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi\in\Phi}\frac{1}{t}R_t^\phi\right] \leq \sqrt{\frac{p-1}{t}}\sqrt[p]{\mu(\Phi)} \tag{42}$$

*at all times $t \geq 1$.*

**Proof** Let $\langle G, g, \gamma\rangle$ be the Gordon triple defined in Lemma 11. By Lemma 8, a $(\Phi, f)$-regret-matching algorithm is also a $(\Phi, g)$-regret-matching algorithm. Now

$$\begin{aligned}
\left(\mathbb{E}\left[\max_{\phi\in\Phi}R_t^\phi\right]\right)^2 &\leq \mathbb{E}\left[\left\|\left(R_t^\Phi\right)^+\right\|_p^2\right] && (43)\\
&= \mathbb{E}\left[G\left(R_t^\Phi\right)\right] && (44)\\
&\leq G(0) + t\sup_{a,r}\gamma\left(\rho^\Phi(a,r)\right) && (45)\\
&= G(0) + t(p-1)\left(\sup_{a,r}\|\rho^\Phi(a,r)\|^2\right) && (46)\\
&\leq t(p-1)(\mu(\Phi))^{2/p} && (47)
\end{aligned}$$

The first inequality follows from Lemma 9, with $x = R_t^\Phi$, $q = 2$, and $2 < p < \infty$. The second inequality is an application of Corollary 7. The third inequality follows from Lemma 10. Hence,

$$\mathbb{E}\left[\max_{\phi\in\Phi}R_t^\phi\right] \leq \sqrt{t(p-1)}\sqrt[p]{\mu(\Phi)} \tag{48}$$

from which it follows that

$$\mathbb{E}\left[\max_{\phi\in\Phi}\frac{1}{t}R_t^\phi\right] \leq \sqrt{\frac{p-1}{t}}\sqrt[p]{\mu(\Phi)} \tag{49}$$

■

**Lemma 13** *For $1 \leq p \leq 2$, if $G(x) = \|x^+\|_p^p$, $g_i(x) = p(x_i^+)^{p-1}$, and $\gamma(x) = \|x\|_p^p$, then $\langle G, g, \gamma \rangle$ is a Gordon triple.*

**Proof** See Appendix B.

**Theorem 14** *Given a bounded ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and a polynomial link function $f$ with $1 \leq p \leq 2$, a $(\Phi, f)$-regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^\phi\right] \leq t^{\left(\frac{1}{p}-1\right)} \sqrt[p]{\mu(\Phi)} \tag{50}$$

*at all times $t \geq 1$.*

**Proof** Let $\langle G, g, \gamma \rangle$ be the Gordon triple defined in Lemma 13. By Lemma 8, a $(\Phi, f)$-regret-matching algorithm is also a $(\Phi, g)$-regret-matching algorithm. Now

$$\begin{aligned}
\left(\mathbb{E}\left[\max_{\phi \in \Phi} R_t^\phi\right]\right)^p &\leq \mathbb{E}\left[\left\|\left(R_t^\Phi\right)^+\right\|_p^p\right] && (51)\\
&= \mathbb{E}\left[G\left(R_t^\Phi\right)\right] && (52)\\
&\leq t \sup_{a,r} \gamma\left(\rho^\Phi(a,r)\right) && (53)\\
&= t \sup_{a,r} \|\rho^\Phi(a,r)\|_p^p && (54)\\
&\leq t\,\mu(\Phi) && (55)
\end{aligned}$$

The first inequality follows from Lemma 9, with $x = R_t^\Phi$, $q = p$, and $1 \leq p \leq 2$. The second inequality is an application of Corollary 7. The third inequality follows from Lemma 10. Hence,

$$\mathbb{E}\left[\max_{\phi \in \Phi} R_t^\phi\right] \leq \sqrt[p]{t\,\mu(\Phi)} \tag{56}$$

from which it follows that

$$\mathbb{E}\left[\max_{\phi \in \Phi} \frac{1}{t} R_t^\phi\right] \leq t^{\left(\frac{1}{p}-1\right)} \sqrt[p]{\mu(\Phi)} \tag{57}$$

■

### 5.2 Exponential Regret-Matching Bounds

Next, we bound the $\Phi$-regret of exponential regret-matching, that is, regret-matching with the exponential link function $f_i(x) = e^{\eta x_i}$ with parameter $\eta \in \mathbb{R}$.

**Lemma 15** *If*

$$G(x) = \frac{1}{\eta} \ln\left(\sum_i e^{\eta x_i}\right) \tag{58}$$

$$g_i(x) = \frac{e^{\eta x_i}}{\sum_j e^{\eta x_j}} \tag{59}$$

*and $\gamma(x) = \frac{\eta}{2}\|x\|_\infty^2$, then $\langle G, g, \gamma \rangle$ is a Gordon triple.*

**Proof** See Appendix B.

**Theorem 16** *Given a bounded ODP, a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, and an exponential link function $f$ with parameter $\eta > 0$, a $(\Phi, f)$-regret-matching algorithm guarantees*

$$\mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^\phi\right] \leq \frac{\ln|\Phi|}{\eta t} + \frac{\eta}{2} \tag{60}$$

*at all times $t \geq 1$.*

**Proof** Let $\langle G, g, \gamma\rangle$ be the Gordon triple defined in Lemma 15. By Lemma 8, a $(\Phi, f)$-regret-matching algorithm is also a $(\Phi, g)$-regret-matching algorithm. Now

$$\begin{aligned}
\mathbb{E}\left[\max_{\phi\in\Phi} \eta R_t^\phi\right] &\leq \mathbb{E}\left[\ln \sum_{\phi\in\Phi} e^{\eta R_t^\phi}\right] && (61)\\
&= \eta\mathbb{E}\left[G\left(R_t^\Phi\right)\right] && (62)\\
&\leq \eta\left(G(0) + t \sup_{a,r} \gamma\left(\rho^\Phi(a,r)\right)\right) && (63)\\
&= \eta\left(\frac{1}{\eta}\ln|\Phi| + \frac{\eta t}{2} \sup_{a,r} \|\rho^\Phi(a,r)\|_\infty^2\right) && (64)\\
&= \ln|\Phi| + \frac{\eta^2 t}{2} && (65)
\end{aligned}$$

The first inequality follows from the following observation: for all $x \in \mathbb{R}^n$,

$$\max_i x_i = \max_i \ln e^{x_i} = \ln \max_i e^{x_i} \leq \ln \sum_i e^{x_i} \tag{66}$$

The second inequality is an application of Corollary 7. Hence,

$$\mathbb{E}\left[\max_{\phi\in\Phi} R_t^\phi\right] \leq \frac{\ln|\Phi|}{\eta} + \frac{\eta t}{2} \tag{67}$$

from which it follows that

$$\mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^\phi\right] \leq \frac{\ln|\Phi|}{\eta t} + \frac{\eta}{2} \tag{68}$$

■

### 5.3 Summary of Bounds

We have considered two classes of algorithms, polynomial and exponential regret matching, each of which has a single parameter, $p$ and $\eta$, respectively. Table 2 summarizes the bounds we derived on $\mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^\phi\right]$. Our two analyses of polynomial $\Phi$-regret matching (Theorems 12 and 14) agree when $p = 2$. In addition, for finite $\Phi \subseteq \Phi_{\text{ALL}}$, our bounds on polynomial $\Phi$-regret matching for $2 \leq p < \infty$ agree with those of Cesa-Bianchi and Lugosi (2003), although their bounds are computed in terms of the number of experts rather than the number of action transformations. For polynomial external and internal regret matching, however, we improve upon the bounds that can be derived immediately from their results. Though the improvement is small for external regret matching (from a bound proportional to $\sqrt[p]{|A|}$ to a bound proportional to $\sqrt[p]{|A|-1}$), it is more significant for internal regret matching (from $|A|^{2/p}$ to $(|A|-1)^{1/p}$).

Table 2: Bounds for polynomial and exponential regret-matching algorithms.

| $f_i(x)$ | Condition | Bound for finite $\Phi \subseteq \Phi_{\text{ALL}}$ | Bound for $\Phi_{\text{EXT}}$,$\Phi_{\text{INT}}$ |
|---|---|---|---|
| $(x_i^+)^{p-1}$ | $2 < p < \infty$ | $\sqrt{\frac{p-1}{t}}\sqrt[p]{\mu(\Phi)}$ | $\sqrt{\frac{p-1}{t}}\sqrt[p]{\lvert A\rvert - 1}$ |
| $(x_i^+)^{p-1}$ | $1 \le p \le 2$ | $t^{\left(\frac{1}{p}-1\right)}\sqrt[p]{\mu(\Phi)}$ | $t^{\left(\frac{1}{p}-1\right)}\sqrt[p]{\lvert A\rvert - 1}$ |
| $e^{\eta x_i}$ | $\eta > 0$ | $\frac{\ln \lvert\Phi\rvert}{\eta t} + \frac{\eta}{2}$ | $\frac{\ln \lvert\Phi\rvert}{\eta t} + \frac{\eta}{2}$ |

Finally, observe that for any $\Phi \subseteq \Phi_{\text{ALL}}$, polynomial regret matching has the property that

$$\lim_{t\to\infty} \mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^\phi\right] = 0 \tag{69}$$

while exponential $\Phi$-regret matching has the property that

$$\lim_{t\to\infty} \mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R_t^\phi\right] = \frac{\eta}{2} \tag{70}$$

In particular, the bound on the time-averaged $\Phi$-regret of any polynomial algorithm is eventually better than that of any exponential algorithm.

### 5.4 Optimal Parameters

We now consider the task of optimizing the parameters in the polynomial and exponential regret-matching algorithms so as to minimize the corresponding bounds. For the polynomial algorithms, we derive the optimal value of the parameter $p$; previously, only approximately optimal values were reported. Also, we find that no matter how we tune the parameters of an exponential algorithm, any polynomial algorithm eventually achieves a lower bound on its regret. However, for sufficiently large action sets, a properly tuned exponential algorithm obtains a lower bound on its regret at a target time $t$ than any polynomial algorithm.

The bound for polynomial regret matching derived in Theorem 14 for $1 \le p \le 2$ is strictly decreasing in $p$; hence $p = 2$ is the optimal setting. By considering the partial derivative with respect to $p$ of the bound derived in Theorem 12 for $2 < p < \infty$, we find that we can minimize this bound by setting $p = p^*(\Phi)$, defined as

$$p^*(\Phi) = \begin{cases} \psi(\mu(\Phi)) & \text{for } \mu(\Phi) \ge e^2 \\ 2, & \text{otherwise} \end{cases} \tag{71}$$

where $\psi(x) = \ln x + \sqrt{(\ln x)^2 - 2\ln x}$. Combining these results, we see that in fact $p = p^*(\Phi)$ is optimal for all $1 \le p < \infty$. This result is similar to the optimal parameter for the $p$-norm regression algorithm calculated by Gentile (2003). Cesa-Bianchi and Lugosi (2003) suggest setting $p = 2\ln|\Phi|$, which nearly optimal when $\mu(\Phi) = |\Phi|$. In particular, if we choose $p = 2\ln\mu(\Phi)$, although this choice is suboptimal, as Gentile (2003) observes, it yields a simpler bound, which differs from the optimal only by lower order terms.

Considering the partial derivative with respect to the parameter $\eta$ of the bound derived in Theorem 16 for exponential regret matching, we find that the optimal setting of $\eta$ depends on the time $t$ at which we want to optimize the bound. The best bound at time $t^*$ is obtained by setting $\eta = \eta^*(\Phi, t^*)$, where

$$\eta^*(\Phi, t) = \sqrt{\frac{2\ln|\Phi|}{t}}. \tag{72}$$

Plugging $\eta^*(\Phi, t^*)$ into Equation 60 we find that an optimized exponential regret matching algorithm guarantees

$$\mathbb{E}\left[\max_{\phi\in\Phi} \frac{1}{t} R^{\phi}_{t^*}\right] \leq \sqrt{\frac{2\ln|\Phi|}{t^*}}. \tag{73}$$

(This result was obtained by Cesa-Bianchi and Lugosi (2003).) For large enough action sets ($|A| \geq 4$ for $\Phi_{\text{EXT}}$, $|A| \geq 13$ for $\Phi_{\text{INT}}$), an $\eta^*(\Phi, t^*)$ exponential algorithm will have a lower bound than any polynomial algorithm at $t = t^*$. For small action sets, however, an optimal polynomial algorithm will have a lower bound than any exponential algorithm for all times $t$.

## 6. Related Work

The literature is rife with analyses of regret-minimization algorithms defined within a variety of frameworks, giving rise to a variety of definitions of regret. Generally speaking, each regret framework is characterized by a set of transformations of the agent's history of actions. The regret vector is then defined as the difference between the rewards the agent obtained and the rewards the agent could have obtained had it played according to each transformation of its history of actions.

Given a precise definition of regret, there are two broad classes of results about online learning algorithms. *Bounding results*, derived primarily by computer scientists, provide functions that bound the expected cumulative (or time-averaged) regret vector at a particular time $t$ when an agent plays according to a particular algorithm. *Convergence results*, derived primarily by game theorists, establish guarantees on the behavior of the time-averaged regret vector as $t \to \infty$.

A commonly studied regret property is the "approachability" (in the sense of Blackwell (1956)) of the negative orthant by the vector of time-averaged regrets. Equivalently (for finite-dimensional regret vectors), we may say that $\mathbb{R}_-$ is approachable by the maximal entry in the time-averaged regret vector. Approachability necessitates uniform convergence: the sequence in question must converge to the set in question at a uniform rate, regardless of the ODP.

Weaker convergence results are also possible. The time-averaged regret vector may converge to a set almost surely (with probability 1) or, weaker still, in probability. Note that an $o(t)$ (i.e., sublinear) bound on the expected time-averaged regret vector yields an in-probability convergence result (see Appendix C, Proposition 23). Note also that approachability is stronger than both almost surely convergence and convergence in probability.

**Regret Varieties** The three most commonly studied forms of regret are external, internal, and swap. The notion of "no regret" (e.g., no external regret, no internal regret) is used by some to indicate approachability of the negative orthant and by others to indicate almost surely convergence or convergence in probability to the negative orthant. *We adopt the first interpretation.*

**Definition 17** *Given a reward system $(A, \mathcal{R})$ and a finite set of action transformations $\Phi \subseteq \Phi_{ALL}$, an online learning algorithm $\{L_t\}_{t=1}^{\infty}$ is said to exhibit* **no-Φ-regret** *if for all $\epsilon > 0$ there exists $t_0 \geq 0$ such that for any ODP,* $\Pr\left[\exists t > t_0 \text{ s.t. } \max_{\phi\in\Phi} \frac{1}{t} R^{\phi}_t \geq \epsilon\right] < \epsilon$.

The notion of external regret is attributed to Hannan (1957). Indeed, the no-external-regret property is often called "Hannan consistency," although it is also sometimes called "universal consistency" (Fudenberg and Levine (1995)). Foster and Vohra (1999) introduced the notion of internal regret, and Blum and Mansour (2005) introduced the terminology "swap regret."

**Convergence Results** Closest to our framework is the Φ-transformation framework introduced in Greenwald and Jafari (2003) and extended in Greenwald et al. (2006). Here, we consider transformations from the set of pure strategies (i.e., actions) $A$ to the set of mixed strategies $\Delta(A)$. There, a transformation $\phi \in \Phi$ is a mapping from $\Delta(A)$ to $\Delta(A)$. Given a sequence of actions $\{a_t\}_{t=1}^{\infty}$, they compare the rewards the agent obtains by playing $\{a_t\}_{t=1}^{\infty}$ to the expected rewards that could

have been obtained by playing $\phi(a_t)$ at each time step $t$. For all finite $\Phi$, they exhibit no-$\Phi$-regret algorithms based on the polynomial link function with $p = 2$.

Hart and Mas-Colell (2001) limit their investigations to external regret and internal regret, the latter of which they call "conditional regret." They exhibit a class no-regret algorithms paramterized by potential functions, which includes the polynomial regret-matching algorithms studied here.

Lehrer's (2003) approach combines "replacing schemes," which are functions from $\mathcal{H} \times A$ to $A$, with "activeness functions" from $\mathcal{H} \times A$ to $\{0, 1\}$. Given a replacing scheme $g$ and an activeness function $I$, Lehrer's framework compares the agent's rewards to the rewards that could have been obtained by playing action $g(h_t, a_t)$, but only if $I(h_t, a_t) = 1$. Lehrer establishes the existence of (no-regret) algorithms whose regret with respect to any countable set of pairs of replacing schemes and activeness functions, averaged over the number of times each pair is "active," approaches the negative orthant. Lehrer's proof, however, is non-constructive.

Fudenberg and Levine (1999) suppose the existence of a countable set of categories $\Psi$ and consider "classification rules," that is functions from $\mathcal{H} \times A$ to $\Psi$. They then compare the agent's rewards to the rewards that could have been obtained under sequences of actions that are measurable with respect to some classification rule. Their framework is a special case of Lehrer's. Fudenberg and Levine derive a variant of fictitious play, "categorical smooth fictitious play," with parameter $\epsilon$ that guarantees that the lim sup of the maximal entry in the time-averaged regret vector converges to $(-\infty, \epsilon]$ almost surely, a property they call "$\epsilon$-universal conditional consistency."

Young (2004) presents "incremental" conditional regret-matching algorithm, a variant of polynomial internal regret matching with $p = 2$. Rather than playing the fixed point of an $|A| \times |A|$ stochastic matrix, the computation of which is an $O(|A|^3)$ operation, this algorithm incrementally updates its mixed strategy based on the cumulative internal regret vector, whose maintenence is only an $O(|A|^2)$ operation. Young argues that via incremental conditional regret-matching, the time-averaged internal regret vector converges to the negative orthant almost surely. Presumably, this result can be generalized to yield a class of no-regret algorithms parameterized by $\Phi$ and $f$.

**Bounding Results** Freund and Schapire (1997) introduce the Hedge algorithm. Inspired by the method of Littlestone and Warmuth (1994), they derive a bound on its external regret.

Herbster and Warmuth (1998) consider a finite set of "experts," which are functions from $\mathcal{E} : \mathbb{N} \to A$. For each $\mathcal{E}$, they compare the agent's rewards to the rewards that could have been obtained had the agent played according to $\mathcal{E}(t)$ at each time $t$. They present an algorithm and prove a bound on its regret, with respect to alternatives constructed by dividing its history into finite-length segments and choosing the best expert for each segment.

Cesa-Bianchi and Lugosi (2003) develop a framework of "generalized regret." They rely on the same notion of experts as Herbster and Warmuth (1998), but they pair experts $f_1, \ldots, f_N$ with activation functions $I_i : A \times \mathbb{N} \to \{0, 1\}$. At time $t$, for each $i$, if $I_i(a_t, t) = 1$, they compare the agent's rewards to the rewards the agent could have obtained by playing $f_i(t)$. This approach is more general than our $\Phi$-transformation framework in that alternatives may depend on time. At the same time, it is more limited in that it cannot represent swap regret.

Blum and Mansour (2005)'s framework is similar to Lehrer's. Their "modification rules" are the same as Lehrer's replacing schemes, but instead of activeness functions, they pair modification rules with "time selection functions," which are functions from $\mathbb{N}$ to the interval $[0, 1]$. The rewards an agent could have obtained under each modification rule are weighted according to how "awake" the rule is, as indicated by the corresponding time selection function. They present a method that, given a collection of algorithms whose external regret is bounded above by $f(t)$ at time $t$, generates an algorithm whose swap regret (and hence, internal regret) is bounded above by $|A|f(t)$.

**Bounding and Convergence Results** Foster and Vohra (1999) define "no internal regret" to mean "the expected internal regret is $o(t)$" (i.e., convergence in probability to the negative orthant). They derive a learning algorithm that satisfies this notion, and they also prove that the internal

| | Approachability | Almost Surely Convergence | Bounding |
|---|---|---|---|
| SWAP | Lehrer (2003)<br>Greenwald et al. (2006) | Fudenberg and Levine (1999) | Blum and Mansour (2005)<br>*the present study* |
| INT | Hart and Mas-Colell (2001) | Foster and Vohra (1999)<br>Young (2004) | Foster and Vohra (1999)<br>Cesa-Bianchi and Lugosi (2003) |
| EXT | | | Hannan (1957)<br>Freund and Schapire (1997)<br>Herbster and Warmuth (1998) |

Table 3: Related Work. Note that approachability results subsume almost surely convergence results. Also, SWAP results subsume INT results, which in turn subsume EXT results. However, many of these results are more general than their entry in this table suggests. For example, the framework of Herbster and Warmuth (1998) deals with time-varying experts as well as external regret (EXT).

regret obtained by their algorithm almost surely converges to the negative orthant. In fact, the internal regret of their algorithm approaches the negative orthant (Greenwald et al. (2006)).

**Game-Theoretic Results** Finally, some authors consider the relationship between game-theoretic equilibria and the regret experienced by the players of infinitely repeated matrix games.

Foster and Vohra (1999) show that if all players play according to algorithms whose internal regrets converge almost surely to the negative orthant, then the empirical distribution of their play converges almost surely to the set of correlated equilibria. Stoltz and Lugosi (To Appear) generalize this result to the case where the set of actions is a compact and convex subset of a normed space and the reward function is continuous.

Hart and Mas-Colell (2000) exhibit a simple adaptive procedure, the empirical distribution of play of which converges to the set of correlated equilibria. They also prove that the empirical distribution of play converges to the set of correlated equilibria if and only if the internal regret experienced by all players converges to zero. Their simple adaptive procedure, however, does not exhibit no-internal-regret; it is not regret-minimizing regardless of the opponents' behavior.

Hart and Mas-Colell (2001) define a class of learning algorithms (parameterized by potential functions) called "better play" which exhibit no external regret. They show that in repeated two-player zero-sum games, if both players play according to better play algorithms, then each player's empirical distribution of play converges to the set of minimax strategies and the players' average rewards converge to the minimax value of the game, almost surely. Moreover, if all players play using an algorithm in their class of no-conditional-regret learing algorithms, then the players' joint empirical distribution of play converges to the set of correlated equilibria, almost surely.

Greenwald et al. (2006) define a general class of game-theoretic equilibria based on the $\Phi$-transformation framework and show that if each agent plays according to a no-$\Phi$-regret algorithm, then the empirical distribution of play approaches the corresponding set of $\Phi$-equilibria, almost surely. These results generalize the results in Hart and Mas-Colell (2001) for polynomial regret-matching algorithms along the dimension $p = 2$. However, Greenwald et al. (2006) also show that the equilibrium set that corresponds to internal regret is that of correlated equilibrium, and that this notion of equilibrium is the tightest solution concept that can be expressed within their framework.

**Regret Frameworks** The frameworks of Lehrer (2003), Blum and Mansour (2005), and Fudenberg and Levine (1999), and the $\Phi$-transformations frameworks, can all represent external, internal, and swap regret. The frameworks of Cesa-Bianchi and Lugosi (2003), Hart and Mas-Colell (2001), and Foster and Vohra (1999) can represent external and internal regret naturally.

Lehrer's (2003) framework is very general. In fact, it subsumes the frameworks presented in Cesa-Bianchi and Lugosi (2003), Herbster and Warmuth (1998), and Fudenberg and Levine (1999). However, it does not allow for mixed strategies; thus, it does not subsume the $\Phi$-transformation frameworks. It also does not allow for partially awake experts as in Blum and Mansour (2005).

## 7. Conclusion

In this paper, we developed a general framework for analyzing the performance of a general class of online learning algorithms in online decision problems. One advantage of this framework is that it may be used to analyze non-smooth link (or potential) functions.

Using our framework, we derived bounds on the regret experienced by $(\Phi, f)$-regret matching algorithms for $\Phi_{\text{INT}}$ and $\Phi_{\text{EXT}}$ and for polynomial and exponential link functions. We also calculated the parameter settings that optimize these bounds.

In ongoing work, we are researching algorithms based on alternative link functions, beyond polynomial and exponential. We are also generalizing our analytic framework to accommodate activation functions (Lehrer, 2003) and transformations that vary with time.

## Acknowledgements

We are grateful to Geoff Gordon for ongoing discussions that helped to clarify many of the technical points in this paper. We also thank David Gondek for his insightful discussions. This research was supported by NSF Career Grant #IIS-0133689 and NSF IGERT Grant #9870676.

## References

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. ACM Press, November 1995.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

Avrim Blum and Yishay Mansour. From internal to external regret. In *Proceedings of the 2005 Computational Learning Theory Conferences*, pages 621–636, June 2005.

Nicolò Cesa-Bianchi and Gábor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51(3):239–261, 2003.

Dean Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–35, 1999.

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*, 55:119–139, 1997.

D. Fudenberg and D. K. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dyanmics and Control*, 19:1065–1090, 1995.

D. Fudenberg and D. K. Levine. Conditional universal consistency. *Games and Economic Behavior*, 29:104–130, 1999.

Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

G. Gordon. No-regret algorithms for structured prediction problems. Technical Report 112, Carnegie Mellon University, Center for Automated Learning and Discovery, 2005.

A. Greenwald and A. Jafari. A general class of no-regret algorithms and game-theoretic equilibria. In *Proceedings of the 2003 Computational Learning Theory Conference*, pages 1–11, August 2003.

Amy Greenwald, Amir Jafari, and Casey Marks. A general class of no-regret algorithms and game-theoretic equilibria. Extended Version, Submitted for Publication, 2006.

J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A.W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.

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

Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54, 2001.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Ehud Lehrer. A wide range no-regret theorem. *Games and Economic Behavior*, 42(1):101–115, 2003.

Nick Littlestone and Manfred Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212 – 261, 1994.

Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, To Appear.

Peyton Young. *Strategic Learning and its Limits*. Oxford University Press, Oxford, 2004.

## A. Link Functions

**Lemma 18** *Given $x^{(1)}, x^{(2)} \in \mathbb{R}^n$, define $x^{(1)} > x^{(2)}$ if and only if $x_i^{(1)} > x_i^{(2)}$ for all $i = 1, \ldots, n$. If $F : \mathbb{R}^n \to \mathbb{R}$ is a convex function that is bounded on the negative orthant, then $F(x^{(1)}) \geq F(x^{(2)})$ whenever $x^{(1)} > x^{(2)}$.*

**Proof** Suppose not, i.e, there exist $x^{(1)}, x^{(2)} \in \mathbb{R}^n$ such that $x^{(1)} > x^{(2)}$ but $F(x^{(2)}) \geq F(x^{(1)}) + \epsilon$, for some $\epsilon > 0$. Define the sequence $\{x^{(k)}\}_{k=3}^{\infty}$ recursively as follows: $x^{(k)} = 2x^{(k-1)} - x^{(k-2)}$. Note that $x^{(k)} - x^{(k-1)} = x^{(k-1)} - x^{(k-2)} = \ldots = x^{(2)} - x^{(1)}$, for $k = 3, \ldots$. If we let $\Delta = x^{(1)} - x^{(2)} > 0$ so that $x^{(k)} = x^{(1)} - (k-1)\Delta$, then for large enough $k$, $x^{(k)} \in \mathbb{R}_-^n$. Now, since $F$ is convex,

$$F(x^{(k)}) + F(x^{(k-2)}) \geq 2F(x^{(k-1)}) \tag{74}$$

It follows that

$$\begin{aligned} F(x^{(k)}) - F(x^{(k-1)}) &\geq F(x^{(k-1)}) - F(x^{(k-2)}) && (75)\\ &\geq F(x^{(k-2)}) - F(x^{(k-3)}) && (76)\\ &\vdots && (77)\\ &\geq F(x^{(2)}) - F(x^{(1)}) && (78)\\ &\geq \epsilon && (79) \end{aligned}$$

Therefore, $F(x^{(k)}) = F(x^{(k)}) - F(x^{(k-1)}) + F(x^{(k-1)}) - F(x^{(k-2)}) - \ldots + F(x^{(2)}) - F(x^{(1)}) + F(x^{(1)}) \geq (k-1)\epsilon + F(x^{(1)})$. But then, as $k \to \infty$, the function $F(x^{(k)})$ is unbounded. Since $x^{(k)} \in \mathbb{R}_-^n$, this contradicts the assumption that $F$ is bounded on the negative orthant. ■

**Lemma 2** *The co-domain of a link function $f : \mathbb{R}^n \to \mathbb{R}^n$ that is subgradient to a convex function $F : \mathbb{R}^n \to \mathbb{R}$ which is bounded on the negative orthant is the positive orthant: i.e., $f : \mathbb{R}^n \to \mathbb{R}_+^n$.*

**Proof** Since $F$ is convex and $f$ is subgradient to $F$, $F(x+y) \geq F(x) + f(x) \cdot y$. Similarly $F(x-y) \geq F(x) - f(x) \cdot y$. If $y > 0$, so that $x > x - y$, then Lemma 18 implies that $F(x) \geq F(x-y)$. Therefore, $f(x) \cdot y \geq F(x) - F(x-y) \geq 0$.

If, for the sake of contradiction, $f_i(x) < 0$ for some $i$, then we can choose $y_i > 0$ big enough and $y_j > 0$ small enough, for all $j \neq i$, so that $f(x) \cdot y < 0$. In particular, if $y_i = -2f_i(x) + \sum_{j=1}^{n} f_j(x) > 0$ and $y_j = -f_i(x) > 0$, for all $j \neq i$, then

$$\begin{aligned} f(x) \cdot y &= \sum_{i=1}^{n} f_i(x) y_i && (80) \\ &= f_i(x) \left( -2f_i(x) + \sum_{j=1}^{n} f_j(x) \right) + \sum_{j \neq i} f_j(x)(-f_i(x)) && (81) \\ &= -\left(f_i(x)\right)^2 + \sum_{j \neq i} f_i(x) f_j(x) + \sum_{j \neq i} f_j(x)(-f_i(x)) && (82) \\ &= -\left(f_i(x)\right)^2 && (83) \\ &< 0 && (84) \end{aligned}$$

Therefore, $f_i(x) \geq 0$, for all $i = 1, \ldots, n$ and $x \in \mathbb{R}^n$. ■

## B. Gordon Triples

In this appendix, we prove Lemmas 11, 13, and 15. Of particular interest is the proof of Lemma 11, because in this proof we identify a set of points (namely, the boundary of the negative orthant; see Lemma 21) on which the polynomial link function for $p \geq 2$ is not differentiable.

Cesa-Bianchi and Lugosi (2003) apply Taylor's theorem to the polynomial (potential) function $G(x) = \|x^+\|_p^2$ for $p \geq 2$, whose gradient is the polynomial link function. In this application, the authors implicitly assume that $G$ is twice differentiable everywhere; however, it is not.

### B.1 Proof of Lemma 11

In Lemmas 19, 20, and 21, we let $i$ and $j$ range over the set $\{1, \ldots, n\}$.

In addition, we rely on the following functions:

- for $p \geq 1$, $G : \mathbb{R}^n \to \mathbb{R}$, defined by:
$$G(x) = \|x^+\|_p^2 \tag{85}$$
- for $p \geq 2$, $g : \mathbb{R}^n \to \mathbb{R}^n$, where for all $i$,
$$g_i(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \text{otherwise} \end{cases} \tag{86}$$
- for $p > 2$, $h : \mathbb{R}^n \to \mathbb{R}^{2n}$, where
$$h_{ii}(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ 2(2-p)\left(\frac{x_i}{\|x^+\|_p}\right)^{2p-2} + 2(p-1)\left(\frac{x_i}{\|x^+\|_p}\right)^{p-2} & \text{otherwise} \end{cases} \tag{87}$$

and for $i \neq j$

$$h_{ij}(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \text{ or } x_j \leq 0 \\ 2(2-p)\left(\frac{x_i x_j}{\|x^+\|_p^2}\right)^{p-1} & \text{otherwise} \end{cases} \tag{88}$$

**Lemma 19** *For $p \geq 2$, $G$ is $C^1$ (i.e., continuously differentiable) with gradient $g$.*

**Proof**

- $G$ is differentiable with gradient $g$ on the complement of the negative orthant.
  We (de)construct $G$ as follows:

$$\begin{aligned} a_i(x) &= (x_i^+)^p & (89)\\ b(y) &= \sum_i a_i(y) & (90)\\ c(z) &= z^{\frac{2}{p}} & (91)\\ G &= c \circ b & (92) \end{aligned}$$

  Because the $a_i$ are differentiable everywhere, $b$ is $C^1$. Also, $c$ is differentiable on $(0, \infty)$, so $G$ is $C^1$ on the complement of the negative orthant. By straightforward calculus, for $u$ in the complement of the negative orthant, $\left.\frac{\partial G}{\partial x_i}\right|_u = g_i(u)$.

- $G$ is differentiable with gradient $g$ on the negative orthant.
  Let $u \in \mathbb{R}^n_-$. We show the following:

$$\lim_{\delta \to 0} \frac{G(u + \delta e_i) - G(u)}{\delta} = g_i(u) \tag{93}$$

  Since $u_i \leq 0$, $g_i(u) = 0$ and $G(u) = 0$. If $u_i < 0$, then $G(u + \delta e_i) = 0$ for sufficiently small $\delta$. If $u_i = 0$, then

$$\lim_{\delta \to 0} \frac{G(u + \delta e_i) - G(u)}{\delta} = \lim_{\delta \to 0} \frac{G(u + \delta e_i)}{\delta} = \lim_{\delta \to 0} \frac{(\delta^+)^2}{\delta} = 0 \tag{94}$$

  Hence, for $u$ in the negative orthant, $\left.\frac{\partial G}{\partial x_i}\right|_u = g_i(u)$.

- $g_i$ is continuous, for all $i$.
  Let $u \in \mathbb{R}^n$. If $u_i > 0$, then $g_i(x) = \frac{2(x_i)^{p-1}}{\|x^+\|_p^{p-2}}$ on a neighborhood of $u$, so $g_i$ is continuous at $u$.
  Otherwise, for all $x \in \mathbb{R}^n$ such that $x_i > 0$, since, by assumption, $p > 2$, it follows that

$$0 < \frac{2(x_i)^{p-1}}{\|x^+\|_p^{p-2}} \leq \frac{2(x_i)^{p-1}}{|x_i|^{p-2}} = 2x_i, \tag{95}$$

  Hence, for all $x \in \mathbb{R}^n$, $0 \leq g_i(x) \leq 2x_i^+$. Now, if $\{x^{(\tau)}\}$ is a sequence such that $\lim x^{(\tau)} = u$, then $u_i \leq 0 \Rightarrow u_i^+ = 0 \Rightarrow \lim \left(x_i^{(\tau)}\right)^+ = 0 \Rightarrow \lim g_i\left(x^{(\tau)}\right) = 0 \Rightarrow \lim g_i\left(x^{(\tau)}\right) = g_i(u)$.

■

**Lemma 20** *For $p \geq 1$, $G$ is smooth on the complement of the axes: i.e., on the set $\{x \in \mathbb{R}^n \mid x_i \neq 0,$ for all $i\}$. Further, on the set where $G$ is smooth, $\left.\frac{\partial G}{\partial x_i \partial x_j}\right|_u = h_{ij}(u)$, for all $i, j$.*

**Proof** For a point $u$ not on an axis, we define an (everywhere) smooth function $\widetilde{G}_u$ by replacing the + operator for each component of the argument of $G$ with either the identity, if u is positive in that component, or zero, if it is not. Specifically, given $u \in \mathbb{R}^n$ such that $u_i \neq 0$ for all $i$, let

$$\widetilde{G}_u(x) = \left( \sum_{i:u_i>0} x_i^p \right)^{\frac{2}{p}} \tag{96}$$

Observe that $G = \widetilde{G}_u$ on a neighborhood of $u$, and for $v$ in this neighborhood, $\left.\frac{\partial \widetilde{G}_u}{\partial x_i \partial x_j}\right|_v = h_{ij}(v)$. ■

**Lemma 21** *Assume $p > 2$. All second-order partial derivatives of $G$ exist and are continuous at a point $u$ if and only if $u$ is not on the boundary of the negative orthant. Furthermore, on the set where $G$ is twice-differentiable, $\left.\frac{\partial G}{\partial x_i \partial x_j}\right|_u = h_{ij}(u)$, for all $i, j$.*

**Proof** By Lemma 19, $G$ is differentiable with gradient $g$ on this set. By Lemma 20, the result holds for $u$ not on an axis. For $u$ on an axis, we show that for all $i, j$,

$$\lim_{\delta \to 0} \frac{g_i(u + \delta \mathbf{e}_j) - g_i(u)}{\delta} = h_{ij}(u) \tag{97}$$

if and only if $u$ is not on the boundary of the negative orthant.

- Case 1: $u_i = 0$.
  In this case, for all $j$, $h_{ij}(u) = 0$. For $i \neq j$, the limit in Equation 97 is equal to 0. For $i = j$, the limit from the left:
  $$\lim_{\delta \to 0^-} \frac{g_i(u + \delta \mathbf{e}_i) - g_i(u)}{\delta} = \frac{0}{\delta} = 0 \tag{98}$$
  and from the right:
  $$\lim_{\delta \to 0^+} \frac{g_i(u + \delta \mathbf{e}_i) - g_i(u)}{\delta} \tag{99}$$
  $$= \lim_{\delta \to 0^+} \frac{2\delta^{p-1}}{\|(u + \delta \mathbf{e}_i)^+\|_p^{p-2}} \frac{1}{\delta} \tag{100}$$
  $$= \lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{\|(u + \delta \mathbf{e}_i)^+\|_p^{p-2}} \tag{101}$$
  $$= \lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{(C + \delta^p)^{1-\frac{2}{p}}} \tag{102}$$
  where $C = \sum_{k \neq i} (u_k^+)^p$. If $u$ is not in the negative orthant (and thus not on the boundary of the negative orthant), then $C > 0$ and the limit in Equation 102 is equal to 0.[4]
  *If $u$ is in the negative orthant (and thus on the boundary of the negative orthant), then $C = 0$ and the limit in Equation 102 is* not *equal to 0. On the contrary,*
  $$\lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{(\delta^p)^{1-\frac{2}{p}}} = \lim_{\delta \to 0^+} \frac{2\delta^{p-2}}{\delta^{p-2}} = 2 \tag{103}$$
  *In particular, $g_i$ is not differentiable on the boundary of the negative orthant.*

4. Note that Equation 102 simplifies to 2, if $p = 2$; hence, we assume that $p > 2$.

- Case 2: $u_i \neq 0, u_j = 0$.
  In this case, for all $i$, $h_{ij}(u) = 0$. Because $u_i \neq u_j$, it cannot be the case that $i = j$. If $u_i < 0$, then the limit in Equation 97 is equal to 0. If $u_i > 0$, then $(u + \delta \mathbf{e}_j)^+ = u^+$ for $\delta < 0$ (since $u_j = 0$), so the left hand limit of Equation 97 is also equal to 0. Now, for the right-hand limit:

$$\lim_{\delta \to 0^+} \frac{1}{\delta} \left( \frac{2u_i^{p-1}}{\|(u + \delta \mathbf{e}_j)^+\|_p^{p-2}} - \frac{2u_i^{p-1}}{\|u^+\|_p^{p-2}} \right) \tag{104}$$

$$= \lim_{\delta \to 0^+} \frac{2u_i^{p-1}}{\delta} \left( \frac{1}{(C + \delta^p)^{1-\frac{2}{p}}} - \frac{1}{C^{1-\frac{2}{p}}} \right) \tag{105}$$

$$= \lim_{\delta \to 0^+} \frac{2u_i^{p-1}}{\delta} \frac{C^{1-\frac{2}{p}} - (C + \delta^p)^{1-\frac{2}{p}}}{C^{1-\frac{2}{p}} (C + \delta^p)^{1-\frac{2}{p}}} \tag{106}$$

$$= \frac{2u_i^{p-1}}{C^{1-\frac{2}{p}}} \lim_{\delta \to 0^+} \frac{C^{1-\frac{2}{p}} - (C + \delta^p)^{1-\frac{2}{p}}}{\delta (C + \delta^p)^{1-\frac{2}{p}}} \tag{107}$$

$$= \frac{2u_i^{p-1}}{C^{1-\frac{2}{p}}} \lim_{\delta \to 0^+} \frac{(p-2)\delta^{p-1} (C + \delta^p)^{\frac{2}{p}}}{(C + \delta^p)^{1-\frac{2}{p}} + (p-2)\delta^p (C + \delta^p)^{\frac{2}{p}}} \tag{108}$$

$$= 0 \tag{109}$$

  where $C = \sum_{k \neq j} (u_k^+)^p > 0$ since $u_i > 0$. Line 108 follows from L'Hôpital's rule.

- Case 3: $u_i \neq 0$, $u_j \neq 0$.
  To take partial derivatives in the $i$th and $j$th coordinates, we can project onto $\mathbb{R}^2$, holding all other coordinates constant. We then apply the technique of Lemma 20 to this projection.

Finally, we must show that all $h_{ij}$ are continuous on the complement of the boundary of the negative orthant. First, consider $h_{ii}$, for an arbitrary $i$. Clearly, $h_{ii}$ is continuous on the (open) sets $\{x \in \mathbb{R}^n \mid x_i < 0\}$ and $\{x \in \mathbb{R}^n \mid x_i > 0\}$. It remains to show $h_{ii}$ is continuous on the complement of these sets, namely $\{x \in \mathbb{R}^n \mid x_i = 0\}$, except where it intersects the boundary of the negative orthant. At each point $u$ in the set $\{x \mid x_i = 0 \text{ and } \exists j \; x_j > 0\}$, $x_i = 0$, so $h_{ii}(u) = 0$. Moreover, as $x$ approaches each such point $u$, $\frac{x_i}{\|x^+\|_p}$ approaches 0 so that $h_{ii}(x)$ does too. Thus, all $h_{ii}$ are continuous on the complement of the boundary of the negative orthant.

Now consider $h_{ij}$, for arbitrary $i \neq j$. Clearly, $h_{ij}$ is continuous on the (open) sets $\{x \in \mathbb{R}^n \mid x_i < 0 \text{ or } x_j < 0\}$ and $\{x \in \mathbb{R}^n \mid x_i > 0 \text{ and } x_j > 0\}$. It remains to show $h_{ij}$ is continuous on the complement of these sets, except where it intersects the boundary of the negative orthant. At each point $u$ in the relevant set, either $x_i = 0$ or $x_j = 0$, so $h_{ij}(u) = 0$. Moreover, as $x$ approaches each such point $u$, $\frac{x_i x_j}{\|x^+\|_p^2}$ approaches 0 so that $h_{ij}(x)$ does too. Thus, all $h_{ij}$ are continuous on the complement of the boundary of the negative orthant.

■

**Lemma 22** *If $p > 2$, then the functions*

$$G(x) = \|x^+\|_p^2 \tag{110}$$

$$g_i(x) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ \frac{2x_i^{p-1}}{\|x^+\|_p^{p-2}} & \text{otherwise} \end{cases} \tag{111}$$

$$\gamma(x) = (p-1)\|x\|_p^2 \tag{112}$$

*satisfy the condition $G(x + y) \leq G(x) + g(x) \cdot y + \gamma(y)$, for $x, y \in \mathbb{R}^n$ such that the boundary of the negative orthant does not intersect the line segment between them (inclusive).*

**Proof** Let $U$ be an open convex set containing $x$ and $x+y$ but no points on the boundary of the negative orthant. By Lemma 19, $G$ is $C^1$ and the gradient of $G$ is $g$. By Lemma 21, $G$ is $C^2$ on $U$,

$$\left.\frac{\partial^2 G}{\partial x_i^2}\right|_u = \begin{cases} 0 & \text{if } u_i \leq 0 \\ 2(2-p)\left(\frac{u_i}{\|u^+\|_p}\right)^{2p-2} + 2(p-1)\left(\frac{u_i}{\|u^+\|_p}\right)^{p-2} & \text{otherwise} \end{cases} \tag{113}$$

and for $i \neq j$,

$$\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u = \begin{cases} 0 & \text{if } u_i \leq 0 \text{ or } u_j \leq 0 \\ 2(2-p)\left(\frac{u_i u_j}{\|u^+\|_p^2}\right)^{p-1} & \text{otherwise} \end{cases} \tag{114}$$

By Taylor's theorem, for some $u \in U$,

$$G(x+y) = G(x) + g(x)\cdot y + \frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j \tag{115}$$

If $u \leq 0$ so that $u^+ = 0$, then

$$\frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j = 0 \leq (p-1)\|y\|_p^2 \tag{116}$$

Otherwise,

$$\frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j \tag{117}$$

$$= \sum_{i,j}(2-p)\left(\frac{u_i^+ u_j^+}{\|u^+\|_p^2}\right)^{p-1} y_i y_j + \sum_i (p-1)\left(\frac{u_i^+}{\|u^+\|_p}\right)^{p-2} y_i^2 \tag{118}$$

$$= (2-p)\|u^+\|_p^{2-2p}\left(\sum_i (u_i^+)^{p-1} y_i\right)^2 + (p-1)\|u^+\|_p^{2-p}\sum_i (u_i^+)^{p-2} y_i^2 \tag{119}$$

$$\leq (p-1)\|u^+\|_p^{2-p}\sum_i (u_i^+)^{p-2} y_i^2 \tag{120}$$

$$\leq (p-1)\|u^+\|_p^{2-p}\left(\sum_i \left((u_i^+)^{p-2}\right)^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}}\left(\sum_i |y_i|^p\right)^{\frac{2}{p}} \tag{121}$$

$$= (p-1)\|y\|_p^2 \tag{122}$$

Line (120) follows from the fact that $p \geq 2$ (by assumption, $p > 2$). Line (121) is an application of Hölder's inequality. ■

**Lemma 11** *If $p > 2$, then the functions $G$, $g$, and $\gamma$ defined in Lemma 22 form a Gordon triple.*

**Proof** Define $z = x + y$. If the line segment between $x$ and $z$ (inclusive) does not intersect the boundary of the negative orthant. then apply Lemma 22. Otherwise, three cases arise.

- Case 1: $x \in \mathbb{R}^n_-$
  If $x$ is in the negative orthant, then $G(x) = 0$ and $g(x) = 0$. Hence, it suffices to show that

$$G(x+y) = \|(x+y)^+\|_p^2 \leq \|y^+\|_p^2 \leq \|y\|_p^2 \leq (p-1)\|y\|_p^2 = \gamma(y) \tag{123}$$

- Case 2: $x \notin \mathbb{R}^n_-, z \in \mathbb{R}^n_-$
  If $z$ is in the negative orthant, but $x$ is not, then $G(z) = 0$. Hence, it suffices to show that

$$||x^+||_p^2 + \left(\frac{2(x^+)^{p-1}}{||x^+||_p^{p-2}}\right) \cdot y + (p-1)||y||_p^2 \geq 0 \tag{124}$$

  which reduces to

$$||x^+||_p^p + 2(x^+)^{p-1} \cdot y + (p-1)||y||_p^2||x||_p^{p-2} \geq 0 \tag{125}$$

  By Hölder's inequality,

$$||y||_p^2||x||_p^{p-2} \geq \sum_i |y_i|^2 \cdot |x_i|^{p-2} \tag{126}$$

  So in fact, it suffices to show that for any $a, b \in \mathbb{R}$,

$$(a^+)^p + 2(a^+)^{p-1}b + (p-1)b^2|a|^{p-2} \geq 0 \tag{127}$$

  Indeed,

$$\begin{aligned}
& (a^+)^p + 2(a^+)^{p-1}b + (p-1)b^2|a|^{p-2} && (128)\\
\geq\ & (a^+)^p + 2(a^+)^{p-1}b + (p-1)b^2(a^+)^{p-2} && (129)\\
=\ & (a^+)^{p-2}((a^+)^2 + 2a^+b + (p-1)b^2) && (130)\\
\geq\ & (a^+)^{p-2}((a^+)^2 + 2a^+b + b^2) && (131)\\
=\ & (a^+)^{p-2}(a^+ + b)^2 && (132)\\
\geq\ & 0 && (133)
\end{aligned}$$

  Line (130) follows from the fact that $p \geq 2$.

- Case 3: $x \notin \mathbb{R}^n_-, z \notin \mathbb{R}^n_-$
  Neither $x$ nor $z$ is in the negative orthant, but, by assumption, the line segment between $x$ and $z$ intersects the boundary of the negative orthant. We claim the line segment between $x^+$ and $z$ does not intersect this boundary. We prove this claim by contradition.

  Assume, for the sake of contradiction, that there exists $\lambda \in (0,1)$ such that $\lambda x^+ + (1-\lambda)z \in \mathbb{R}^n_-$. Define $I = \{i \mid z_i > 0\}$. Since $z \notin \mathbb{R}^n_-$, the set $I$ is non-empty. Moreover, for all $i \in I$, $x_i < 0$.

  For all $i \in I$, $z_i > 0$ and $x_i^+ = 0$, so $\lambda = 1$ (because $\lambda x^+ + (1-\lambda)z \in \mathbb{R}^n_-$). Consequently, it must be the case that $x^+ \in \mathbb{R}^n_-$. This is a contradiction, since $x \notin \mathbb{R}^n_-$, by assumption.

  Therefore, we can apply Lemma 22 to $x^+$ and $z - x^+$, which yields

$$G(x+y) = G(z) \leq G(x^+) + g(x^+) \cdot (z - x^+) + \gamma(z - x^+). \tag{134}$$

  First, observe that $G(x^+) = G(x)$. Second, for all $i = 1, \ldots, n$, $g_i(x^+) = g_i(x) \geq 0$ and $y_i = z_i - x_i \geq z_i - x_i^+$. Hence, $g(x^+) \cdot (z - x^+) \leq g(x) \cdot (z - x) = g(x) \cdot y$. Third, for all $i = 1, \ldots, n$, $|z_i - x_i^+| \leq |z_i - x_i|$ Thus, $\|z - x^+\|_p^2 \leq \|z - x\|_p^2 = \|y\|_p^2$. Therefore,

$$G(x^+) + g(x^+) \cdot (z - x^+) + \gamma(z - x^+) \leq G(x) + g(x) \cdot y + \gamma(y) \tag{135}$$

  Together, Equations 134 and 135 imply the desired conclusion.

■

### B.2 Proof of Lemma 13

**Lemma 13** *If $1 \leq p \leq 2$, then the following is a Gordon triple: $G(x) = \|x^+\|_p^p$, $g_i(x) = p(x_i^+)^{p-1}$, and $\gamma(x) = \|x\|_p^p$.*

**Proof** Because $\|x^+\|_p^p = \sum_i \left(x_i^+\right)^p$, it suffices to show that for any $a, b \in \mathbb{R}$,

$$((a+b)^+)^p \leq (a^+)^p + p(a^+)^{p-1}b + |b|^p \tag{136}$$

after which the result follows from a component-wise proof.

If $p = 1$, then Line (136) follows immediately because $(a+b)^+ \leq a^+ + b^+$. Otherwise, we use the basic inequality $x^\alpha + y^\alpha \geq (x+y)^\alpha$, for all $x, y \geq 0$ and $\alpha \in [0,1]$. Two cases arise.

- Case 1: $b \geq 0$. Define the function $h_c(z) = z^p + p(c^+)^{p-1}z - ((c+z)^+)^p$ for $c \in \mathbb{R}$ for $z \geq 0$. Taking its derivative yields $h_c'(z) = p\left(z^{p-1} + (c^+)^{p-1} - ((c+z)^+)^{p-1}\right)$. Applying the basic inequality with $x = z$, $y = c^+$, and $\alpha = p-1$ yields

$$\begin{aligned} z^{p-1} + (c^+)^{p-1} &\geq (z + c^+)^{p-1} && (137)\\ &\geq ((c+z)^+)^{p-1} && (138) \end{aligned}$$

  Thus, $h_c'(z) \geq 0$, i.e., $h_c$ is non-decreasing on $[0, \infty)$. In particular,

$$\begin{aligned} h_a(0) &\leq h_a(b) && (139)\\ -(a^+)^p &\leq b^p + p(a^+)^{p-1}b - ((a+b)^+)^p && (140)\\ ((a+b)^+)^p &\leq (a^+)^p + p(a^+)^{p-1}b + b^p && (141)\\ &\leq (a^+)^p + p(a^+)^{p-1}b + |b|^p && (142) \end{aligned}$$

- Case 2: $b < 0$. Define the function $h_c(z) = z^p - p(c^+)^{p-1}z - ((c-z)^+)^p$ for $c \in \mathbb{R}$ for $z \geq 0$. Taking its derivative yields $h_c'(z) = p\left(z^{p-1} - (c^+)^{p-1} + ((c-z)^+)^{p-1}\right)$. Applying the basic inequality with $x = z$, $y = (c-z)^+$, and $\alpha = p-1$ yields

$$\begin{aligned} z^{p-1} + ((c-z)^+)^{p-1} &\geq (z + (c-z)^+)^{p-1} && (143)\\ &\geq (c^+)^{p-1} && (144) \end{aligned}$$

  Thus, $h_c'(z) \geq 0$, i.e., $h_c$ is non-decreasing on $[0, \infty)$. In particular,

$$\begin{aligned} h_a(0) &\leq h_a(-b) && (145)\\ -(a^+)^p &\leq (-b)^p + p(a^+)^{p-1}b - ((a+b)^+)^p && (146)\\ ((a+b)^+)^p &\leq (a^+)^p + p(a^+)^{p-1}b + (-b)^p && (147)\\ &\leq (a^+)^p + p(a^+)^{p-1}b + |b|^p && (148) \end{aligned}$$

■

### B.3 Proof of Lemma 15

**Lemma 15** *If $\delta > 0$, then the following is a Gordon triple:*

$$G(x) = \frac{1}{\delta} \ln \left( \sum_i e^{\delta x_i} \right) \tag{149}$$

$$g_i(x) = \frac{e^{\delta x_i}}{\sum_j e^{\delta x_j}} \tag{150}$$

$$\gamma(x) = \frac{\delta}{2}\|x\|_\infty^2 \tag{151}$$

**Proof** We use the same technique as in the proof of Lemma 22.

Observe that $G$ is smooth and that the gradient of $G$ is $g$. Moreover, for $i \neq j$,

$$\left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u = -\frac{\delta e^{\delta u_i} e^{\delta u_j}}{\left(\sum_i e^{\delta u_i}\right)^2} \tag{152}$$

and otherwise,

$$\left.\frac{\partial^2 G}{\partial x_i^2}\right|_u = -\frac{\delta e^{\delta u_i} e^{\delta u_i}}{\left(\sum_i e^{\delta u_i}\right)^2} + \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} \tag{153}$$

By Taylor's theorem, for some $u \in U$,

$$G(x+y) = G(x) + g(x)\cdot y + \frac{1}{2}\sum_{ij} \left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j \tag{154}$$

Finally,

$$\begin{aligned}
\frac{1}{2}\sum_{ij} \left.\frac{\partial^2 G}{\partial x_i \partial x_j}\right|_u y_i y_j &= \sum_{ij} -\frac{\delta e^{\delta u_i} e^{\delta u_j}}{\left(\sum_i e^{\delta u_i}\right)^2} y_i y_j + \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} y_i^2 && (155)\\
&= -\delta\left(\frac{\sum_i e^{\delta u_i} y_i}{\sum_i e^{\delta u_i}}\right)^2 + \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} y_i^2 && (156)\\
&\leq \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} y_i^2 && (157)\\
&\leq \sum_i \frac{\delta e^{\delta u_i}}{\sum_i e^{\delta u_i}} \|y\|_\infty^2 && (158)\\
&= \delta\|y\|_\infty^2 && (159)
\end{aligned}$$

■

## C. Convergence in Probability

In this appendix, we note that sub-linear regret implies convergence in probability.[5]

**Proposition 23** *If $\{X_t\}$ be a sequence of random variables such that $\mathbb{E}[X_t] \in o(t)$, then, for all $\epsilon, \delta > 0$, there exists $t_0$ such that for all $t \geq t_0$,*

$$Pr\left(\frac{X_t}{t} \geq \delta\right) < \epsilon \tag{160}$$

*That is, $\frac{X_t}{t}$ converges to the set $(-\infty, 0]$ in probability.*

5. We first made this observation during discussions with Geoff Gordon.

**Proof** First apply Markov's inequality to $\frac{X_t}{t}$:

$$Pr\left(\frac{X_t}{t} \geq \delta\right) \leq \frac{\mathbb{E}\left[\frac{X_t}{t}\right]}{\delta} = \frac{\mathbb{E}[X_t]}{t\delta}. \tag{161}$$

Now apply the definition of $o(t)$: for all $c > 0$ there exists $t_0$ such that for all $t \geq t_0$, $0 \leq E[X_t] < ct$. If $c = \epsilon\delta$, then there exists $t_0$ such that for all $t \geq t_0$,

$$\frac{\mathbb{E}[X_t]}{t\delta} < \frac{c}{\delta} = \epsilon, \tag{162}$$

Combining equations 161 and 162 yields the desired result. ■