CONSTANT REGRET IN FINITE GAMES VIA HIGHER-ORDER OPTIMISM

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce an uncoupled learning algorithm which, when employed by all players of an *arbitrary* N-player normal form game with up to K actions per player, guarantees $\mathcal{O}(N^3 \log^2 K)$ individual regret, uniformly over the horizon of play. The proposed algorithm—which we call *higher-order optimism with discounting* (HOOD)—is a variant of optimistic follow-the-regularized-leader (OPTFTRL) that combines a discounted $(N+1)$-th order predictor with entropic regularization over a suitable "lifting" of the game's strategy space. This combination of ingredients is purposefully designed to dampen large oscillations of the induced sequence of play in a controlled manner, removing in this way a key stumbling block of previous attempts to achieve constant regret in general finite games.

1 INTRODUCTION

The standard model for learning in games typically unfolds as follows: at each instance of a repeated decision process (*i*) every participating player concurrently chooses an action; (*ii*) they receive the reward and/or any feedback generated from their chosen actions; (*iii*) they update their strategies, and the process repeats. In this general setting, the overarching objective of each player is to improve their individual payoff over time, ideally reaching a state where no further improvements are possible or, at the very least—and, perhaps, more pragmatically—not regretting their past choices.

The latter criterion leads to the seminal notion of (external) *regret*, first introduced in statistical decision-making by Wald (1950) and Savage (1951), and subsequently pioneered in the context of games and approachability by Hannan (1957) and Blackwell (1954), respectively. At a high level, the individual regret of a player provides a natural "worst-case" performance benchmark, as it compares the player's cumulative payoff over time to that of the best fixed action in hindsight. Accordingly, these early works also established the existence of *no-regret* learning procedures, i.e., update rules whose regret grows sublinearly with the horizon of play in *any* "game against Nature"—that is, against *any* possible behavior of the learner's opponents, rational or otherwise.

Nonetheless, since the behavior of no-regret learners in a normal form game is *not* arbitrary but shaped by the actions of all other players, an important question which has remained open since these early days of the field is whether the accumulation of regret ultimately tapers off when all players adhere to a common learning rule. Somewhat more formally:

Are there learning rules which, if followed by all players,
guarantee constant individual regret to all players, in all games?

An important caveat in the above is that any such rule should be *uncoupled* in the sense of Hart & Mas-Colell (2003, 2006), i.e., the strategy adjustment of any given player should not depend on the payoff functions of the other players (though it may of course depend on the payoff function of the focal player, and through it, the strategies of other players). An additional consideration is that the feedback available to the players must be sufficiently informative to allow them to leverage the—presumably mild—evolution of their learned strategies over time. On that account, the "constant regret" question is framed exclusively in the deterministic, full-information setting, where each player receives as feedback their mixed payoff vector at each stage of the process (but no further information). Without this assumption, standard information-theoretic lower bounds would exclude the possibility of achieving constant regret, except in very special cases (Bubeck & Cesa-Bianchi, 2012; Cesa-Bianchi & Lugosi, 2006).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions in the context of related work. Our main contribution is that, with these qualifications in mind, *our driving question above can be answered in the affirmative.* In more detail, we propose an uncoupled, full-information learning rule, which we call *higher-order optimism with discounting* (HOOD), and which enjoys the following universal guarantee:

Main Theorem (Informal version)**.** *If all players of a finite N-player game with up to K actions per player follow HOOD, the individual regret of each player $i = 1, \dots, N$, is bounded for all T as*

$$\mathrm{Reg}_i(T) \leq 45(N+1)^3[16 + 4\log K + \log^2 K] = \mathcal{O}(N^3 \log^2 K)\,. \tag{1}$$

Two additional appealing features of HOOD are that (*i*) it is horizon-free, i.e., it does not require prior knowledge of the horizon of play; and (*ii*) the method's tuning uses fixed, horizon-independent parameters. As a result, for any fixed N and K, each player incurs only a finite amount of regret over the entire (possibly infinite) trajectory of play. Thus, by standard results in the field, the empirical distribution of play converges to the game's set of coarse correlated equilibria at a rate of $\mathcal{O}(1/T)$.

To provide the proper context for our results, it is instructive to begin with the so-called adversarial framework, where a single learner plays a "game against Nature"—that is, they are facing an arbitrary sequence of payoff vectors, not necessarily coming from a game. In this setting, the exponential weights algorithm for exploration and exploitation (EXP3) guarantees $\mathcal{O}(\sqrt{TK\log K})$ regret when run with bandit, payoff-based feedback, whereas HEDGE, the full-information variant of EXP3, achieves $\mathcal{O}(\sqrt{T\log K})$ regret (Auer et al., 1995, 2002). In either case however, the $\mathcal{O}(\sqrt{T})$ dependence is unavoidable without imposing further assumptions on how the sequence of payoff vectors is generated (Bubeck & Cesa-Bianchi, 2012; Cesa-Bianchi & Lugosi, 2006).

This is where the reward generation mechanism of a game enters the stage: If every player follows an iterative update rule, the individual strategy adjustments of each player are likely to be relatively small from one epoch to the next, so the induced sequence of payoff vectors is likewise expected to be more predictable. This predictability can be leveraged by "extrapolation-based" algorithms, that is, where players try to anticipate their future payoff landscape in order to take a more informed update step at each iteration—as in the original extra-gradient algorithm of Korpelevich (1976), the single-query variant of Popov (1980), and the like.

In the context of regret minimization, the most widely studied algorithmic scheme of this type is the *optimistic FTRL* (OPTFTRL) method of Rakhlin & Sridharan (2013a,b), which was shown to achieve $\mathcal{O}(1)$ regret in two-player zero-sum games under self-play. These results were subsequently expanded upon by Syrgkanis et al. (2015) and, more recently, Tsuchiya (2026), who obtained tight upper and lower regret bounds for optimistic HEDGE (OPTHEDGE)—the entropically regularized instance of OPTFTRL—in zero-sum games. More broadly, constant regret has since been achieved in several structured classes of games, including variationally stable games (Hsieh et al., 2021), potential games (Anagnostides et al., 2022b), and the strategic counterpart of the latter, harmonic games (Legacci et al., 2024).

These results show that attaining constant regret *is* possible in the context of game-theoretic learning, but they are all contingent on the underlying game possessing an amenable "variational" structure. By contrast, in the absence of such a structure, the movement of one player can amplify the movement of another in an unpredictable manner, so it is considerably harder to control the iterates of the algorithm over time. As such, for a long time, it was not known whether the presence of an underlying game mechanism was sufficient to overcome the adversarial worst-case $\mathcal{O}(\sqrt{T})$ regret bound.

The first breakthrough in this direction was obtained by Syrgkanis et al. (2015), who showed that OPTHEDGE guarantees $\mathcal{O}(\sqrt{N}\,\log K \cdot T^{1/4})$ regret in *any* finite N-player game with at most K actions per player. Chen & Peng (2020) subsequently expanded on the stability analysis of Syrgkanis et al. (2015) and sharpened this bound to $\mathcal{O}(T^{1/6})$ for two-player general-sum games. The next major step for N-player games came from the work of Daskalakis et al. (2021) who showed that OPTHEDGE incurs *polylogarithmic* regret in *any* finite game—the first "near-constant" result of its kind. Then, at around the same time, Piliouras et al. (2022) showed that a proximal variant of the exponential weights (EW) algorithm achieves constant regret on a subsequence of the induced sequence of play.

The proof of Daskalakis et al. (2021) in particular introduced a "higher-order smoothness" framework for the analysis, which was used to show that successive temporal differences of the generated sequence of play can be controlled to sufficiently high order. Subsequent work then simplified and

	Individual regret	Parameter tuning
Syrgkanis et al. (2015)	$\mathcal{O}(\sqrt{N}\,\log K \cdot T^{1/4})$	static; function of N and T
Daskalakis et al. (2021)	$\mathcal{O}(N \log K \cdot \log^4 T)$	static; function of N and T
Farina et al. (2022)	$\mathcal{O}(NK \cdot \log T)$	static; function of N
Soleymani et al. (2025a)	$\mathcal{O}(N \log^2 K \cdot \log T)$	adaptive; function of N and K
This paper	$\mathcal{O}(N^3 \log^2 K)$	static; function of N and K

Table 1. Overview of related work; see also Appendix A for an extended version.

strengthened this picture: Farina et al. (2022) reduced the individual regret bound to $\mathcal{O}(NK \log T)$ in general games, Anagnostides et al. (2022a) obtained the same $\mathcal{O}(\log T)$ horizon dependence for the stronger notion of swap regret in finite games, and, more recently, Soleymani et al. (2025a,b) reduced the action dependence to polylogarithmic while keeping a $\mathcal{O}(\log T)$ dependence on the horizon.

Building on these developments, the pressing question that remains is whether it is possible to conclude this progression and descend from unbounded to constant regret in general games, or if there is a fundamental limitation that goes beyond the algorithms and approaches used so far. Our paper provides an affirmative answer to this question by showing that, under HOOD, the players' regret are bounded uniformly over time. In this sense, it completes the sequence of horizon improvements for uncoupled learning in arbitrary games under full information feedback, where we summarize this progression in Table 1, and also describe the specific tuning required for each algorithm.

What goes on under the HOOD. Our algorithm is based on OPTHEDGE, but with two important differences. The first is that we use a judiciously crafted predictor in the algorithm's extrapolation step, with the aim of leveraging the higher-order analysis framework of Daskalakis et al. (2021). In more detail, the algorithm's predictor is chosen in a way that turns the induced prediction error into a geometrically discounted finite difference of order $N + 1$; by contrast, the classical predictor of Popov (1980) and Rakhlin & Sridharan (2013a,b) only uses payoff information from the previous round, ultimately leading to a first-order difference in time, which is much harder to control.

The second has to do with the algorithm's regularizer. Instead of using entropic regularization off the shelf—as in HEDGE and OPTHEDGE—we also incorporate a "lifting" mechanism in the spirit of Farina et al. (2022). From an operational standpoint, this lifting allows us to control the positive part of the regret rather than the signed regret, so high positive regret accumulated by one player cannot be offset by the negative regret contribution of another. At the same time, the chosen regularizer stabilizes the players' updates, which in turn limit the changes in the sequence of payoff vectors encountered. The higher-order predictor of HOOD is meticulously designed to exploit precisely this interaction, that is, as payoff changes propagate through successive player updates, some player's update must be encountered again when expanding the prediction error. At that point, the existing regularization control allows us to bound it in terms of the variation of the players' strategies, which is itself controlled by the prediction error, thereby closing the loop. These two components then dovetail together to—ultimately—yield the desired constant regret bound.

Contemporaneous works. During the final stages of preparing our paper, and within less than four weeks from the submission deadline, a number of preprints appeared online with a motivation similar to our own. For completeness, we discuss these contemporaneous unpublished works in Appendix A.

2 PRELIMINARIES

Basics from game theory. Throughout the sequel, we focus exclusively on finite N-player games in normal form. For simplicity—and without loss of generality for our purposes—we assume that each player $i \in \mathcal{N} \equiv \{1, \dots, N\}$ has K actions at their disposal, indexed by $\alpha_i \in \mathcal{K} \equiv \{1, \dots, K\}$. We then write $\mathcal{A} := \mathcal{K}^N$ for the set of the players' *action profiles* $\alpha = (\alpha_1, \dots, \alpha_N)$, and $u_i \colon \mathcal{A} \to [-1, 1]$ for the (pure) payoff function of player $i \in \mathcal{N}$.

We further assume that each player may randomize their choice of action by playing a *mixed strategy*, that is, a probability distribution drawn from the unit simplex $\Delta(\mathcal{K}) = \{x \in \mathbb{R}^K_+ : \|x\|_1 = 1\}$ of $\mathbb{R}^K$. We likewise write $\mathcal{X} := \Delta(\mathcal{K})^N$ for the game's *strategy space*, i.e., the set of the players' *mixed*

strategy profiles $x = (x_1, \dots, x_N)$. Finally, in a slight abuse of notation, we denote the *mixed payoff* of player $i \in \mathcal{N}$ as

$$u_i(x) = \mathbb{E}_{\alpha \sim x}[u_i(\alpha)] = \sum_{\alpha \in \mathcal{A}} x_\alpha u_i(\alpha) \tag{2}$$

where $x_\alpha = x_{1,\alpha_1} \cdots x_{N,\alpha_N}$ denotes the joint probability of $\alpha \in \mathcal{A}$ under $x \in \mathcal{X}$.

In a similar vein, we define the *mixed payoff vector* of player $i \in \mathcal{N}$ as

$$v_i(x) = (u_i(\alpha_i; x_{-i}))_{\alpha_i \in \mathcal{K}}\,, \tag{3}$$

where we used the standard shorthand $(\alpha_i; x_{-i})$ to denote the strategy profile where player $i \in \mathcal{N}$ plays α_i against the mixed strategy $x_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ of i's opponents. Clearly, by multilinearity we have $u_i(x) = \langle v_i(x), x_i \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^{\mathcal{K}}$.

Repeated play, feedback, and regret. The repeated play process that we consider unfolds as follows: at each stage $t = 1, 2, \dots$, every player $i \in \mathcal{N}$ chooses a mixed strategy $x_{i,t} \in \Delta(\mathcal{K})$, giving rise to the mixed strategy profile $x_t = (x_{i,t})_{i \in \mathcal{N}} \in \mathcal{X}$. Then, in the full information setting that is the focus of our work, each player is assumed to observe their mixed payoff vector and reward, viz.

$$v_{i,t} \coloneqq v_i(x_t) \quad \text{and} \quad u_{i,t} \coloneqq u_i(x_t) = \langle v_{i,t}, x_{i,t} \rangle\,. \tag{4}$$

In this context, the (*external*) *regret* of player $i \in \mathcal{N}$ relative to the sequence of play x_t, $t = 1, 2, \dots$, is defined as

$$\mathrm{Reg}_i(T) \coloneqq \max_{p_i \in \Delta(\mathcal{K})} \sum_{t=1}^{T} [u_i(p_i; x_{-i,t}) - u_i(x_t)] \tag{5}$$

and we say that the player incurs *no regret* if $\mathrm{Reg}_i(T) = o(T)$. In words, the regret of a player represents the cumulative payoff difference between their mixed payoff and that of the best fixed strategy in hindsight; as such, the regret will be our main figure of merit for the sequel.

3 Regularized learning, extrapolation, and lifting

In this section, we proceed to describe the main modules of our method: regularization, extrapolation, and the lifting mechanism. To ease notation, we adopt a single-agent viewpoint and we suppress all player indices for simplicity; we then reinstate them toward the end of this section.

Regularized learning. Our starting point is the algorithmic template known as *follow-the-regularized-leader* (FTRL)—arguably the most widely used family of policies for attaining no regret (Shalev-Shwartz, 2011; Shalev-Shwartz & Singer, 2006, 2007). At a high level, the main idea of FTRL is that, at each stage of the process, the learner plays the mixed strategy that maximizes their cumulative payoff minus a regularization penalty intended to stabilize the induced sequence of play. Formally, given a sequence of payoff vectors $v_t \in \mathbb{R}^K$, $t = 1, 2, \dots$, this leads to the two-step, "stimulus-response" update rule

$$\begin{aligned} &(a)\ \text{Aggregate payoff information (stimulus):} && y_{t+1} = y_t + \eta v_t \\ &(b)\ \text{Update mixed strategies (response):} && x_{t+1} = \arg\max_{x \in \Delta(\mathcal{K})} \{\langle y_{t+1}, x \rangle - h(x)\} \end{aligned} \tag{FTRL}$$

In the above, the penalty function $h\colon \Delta(\mathcal{K}) \to \mathbb{R}$ is referred to as the *regularizer* of the method, and $\eta > 0$ is a tunable parameter—the method's *learning rate*—which controls the sensitivity of the response phase. Standard choices for h include (i) the quadratic regularizer $h(x) = (1/2)\sum_{\alpha=1}^{K} x_\alpha^2$, which gives rise to the lazy projection dynamics (Friedman, 1991; Zinkevich, 2003); and (ii) the entropic regularizer $h(x) = \sum_{\alpha=1}^{K} x_\alpha \log x_\alpha$, which leads to Hedge and the exponential / multiplicative weights template (Auer et al., 1995, 2002; Littlestone & Warmuth, 1994; Vovk, 1990).

While (FTRL) achieves optimal regret guarantees in the (full-information) adversarial setting, it is also well-known to diverge away from equilibrium in certain zero-sum games (Mertikopoulos et al., 2019), thus incurring significant regret under self-play. We address this issue below.

Extrapolation. An effective way to alleviate this undesirable aspect of (FTRL) is to include an "*extrapolation step*" with the aim of stabilizing the method's iterates over time. The idea for this is as follows: First, *before* playing in round t, the learner constructs a *predictor* $\hat{v}_t$ for v_t, and uses it to move the "base state" y_t to an interim, "leading state" $y_{t+1/2}$; subsequently, the learner chooses a strategy x_t based on $y_{t+1/2}$ instead of y_t (this is the extrapolation step), they observe v_t, and only then do they update y_t. Formally, this leads to the *extrapolated FTRL* (FTRL+) update rule:

$$y_{t+1/2} = y_t + \eta\hat{v}_t \qquad x_t = Q(y_{t+1/2}) \qquad y_{t+1} = y_t + \eta v_t \tag{FTRL+}$$

where, for convenience, we introduced the *regularized choice*—or *best response*—*map*

$$Q(y) = \arg\max_{x\in\Delta(\mathcal{K})}\{\langle y, x\rangle - h(x)\} \quad \text{for all } y \in \mathbb{R}^K \tag{6}$$

to encode more concisely the algorithm's strategy selection rule.

Clearly, different predictors can lead to varying degrees of stability in (FTRL+). As an extreme example, when $\hat{v}_t = 0$, the learner essentially gives up on constructing a meaningful predictor, so (FTRL+) reduces to (FTRL); by contrast, when $\hat{v}_t = v_{t-1}$, the learner optimistically seeks to infer the value of v_t from v_{t-1}, in which case we recover the *optimistic FTRL* (OptFTRL) algorithm of Rakhlin & Sridharan (2013a,b). The precise choice of predictor is a major—as well as intricate—element of our analysis so, to streamline our presentation, we postpone the relevant details to Section 4.

The lifting mechanism. The last ingredient of our algorithmic scheme is a "lifting" device in the spirit of Farina et al. (2022). To understand its role in the analysis to come, we begin by noting that, by linearity, the learner's regret can be expressed more simply as

$$\mathrm{Reg}(T) = \max_{\alpha\in\mathcal{K}} \sum_{t=1}^{T} [v_{\alpha,t} - u_t] = \max_{\alpha\in\mathcal{K}} \sum_{t=1}^{T} g_{\alpha,t} \tag{7}$$

where, for convenience, we write $g_t = (g_{\alpha,t})_{\alpha\in\mathcal{K}}$ for the learner's *payoff excess*—or *centered payoff*—*vector*, with components

$$g_{\alpha,t} = v_{\alpha,t} - u_t\,. \tag{8}$$

The sum $\sum_{t=1}^{T} g_{\alpha,t}$ is precisely the driving variable of (FTRL) / (FTRL+), explaining in this way the deep link between regularized learning and regret minimization. That being said, the real object of interest is the *positive part* $[\mathrm{Reg}(T)]_+$ of the regret, which can itself be expressed as

$$[\mathrm{Reg}(T)]_+ = \max\{0, \mathrm{Reg}(T)\} = \max\left\{0, \max_{\alpha\in\mathcal{K}} \sum_{t=1}^{T} g_{\alpha,t}\right\} = \max_{u\in\mathcal{K}_0} \sum_{t=1}^{T} \langle g_t, u\rangle = \max_{z\in\Delta_0(\mathcal{K})} \sum_{t=1}^{T} \langle g_t, z\rangle \tag{9}$$

with

$$\mathcal{K}_0 = \{\mathbf{0}\} \cup \{e_1, \dots, e_K\} \quad \text{and} \quad \Delta_0(\mathcal{K}) = \mathrm{conv}(\mathcal{K}_0) = \{z \in \mathbb{R}^K_+ : \|z\|_1 \le 1\} \tag{10}$$

respectively denoting the set of "lifted" pure strategies and its convex hull, where, in the above and all that follows, e_α denotes the α-th basis vector of $\mathbb{R}^K$, $\mathbf{0} = (0, \dots, 0)$ denotes the all-zeros vector, and $\mathbf{1} = (1, \dots, 1)$ the all-ones vector.

The terminology "lifting" alludes to the fact that $\mathcal{K}_0$ represents an expanded set of pure strategies, where we have adjoined to $\mathcal{K}$ the "opt-out" comparator "$\mathbf{0}$" with a constant payoff excess of zero. In a similar vein, $\Delta_0(\mathcal{K})$ represents the corner-of-cube "lifting" of the standard probability simplex, and each "lifted" strategy includes the probability of opting out.[1] To make this more precise, note that any nonzero $z \in \Delta_0(\mathcal{K})$ can be represented uniquely as $z = \lambda x$, with $\lambda = \|z\|_1$ denoting the *mass* of the lifted strategy—that is, the probability of "opting in"—whereas $x = z/\|z\|_1 \in \Delta(\mathcal{K})$ is the resulting conditional probability distribution on $\mathcal{K}$. In this way, a lifted strategy encodes both the probability of "opting out", as well as the player's original mixed strategy; for an illustration, cf. Figure 1.

The *raison d'être* of this mechanism is that it allows us to construct a lifted version of (FTRL) by playing over $\Delta_0(\mathcal{K})$ instead of $\Delta(\mathcal{K})$. Specifically, given a continuous and 1-strongly convex regularizer ψ on $\Delta_0(\mathcal{K})$ which is differentiable on the relative interior $\mathrm{ri}\,\Delta_0(\mathcal{K})$, we respectively define the associated *lifted choice*, *mass* and *normalized choice maps* as

$$Z(y) = \arg\max_{z\in\Delta_0(\mathcal{K})}\{\langle y, z\rangle - \psi(z)\}, \quad \lambda(y) = \|Z(y)\|_1 \quad \text{and} \quad Q(y) = \frac{Z(y)}{\|Z(y)\|_1} \tag{11}$$

[1] Sensu stricto, the pure payoff of the opt-out strategy would depend on the player's mixed strategy, but we set this detail aside because the above interpretation is intended only as an informal illustration of the lifting trick.

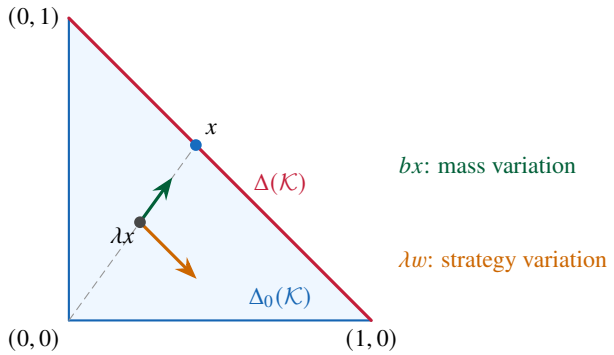


Figure 1. Mass and strategy variations on the lifted simplex for two actions. The red edge is the probability simplex $\mathcal{X}_2$, and the blue shaded triangle is its lift $\Delta_0(\mathcal{K})$. At $z = \lambda x$, a variation ξ can be written uniquely as $\xi = bx + \lambda w$, where the component bx changes the mass λ at rate $b = \langle \mathbf{1}, \xi \rangle$, and the component λw changes the strategy x at rate w, which satisfies $\langle \mathbf{1}, w \rangle = 0$.

for all $y \in \mathbb{R}^K$. In this way, the lifted choice map Z computes the regularized best response to y in the lifted simplex $\Delta_0(\mathcal{K})$, while the normalized choice map Q proceeds to extract the mixed strategy used to select a pure strategy. For this reason, we will also assume that the regularizer is *steep*, meaning that $\|\nabla\psi(z_t)\|_1 \to \infty$ whenever $z_t \in \operatorname{ri}\Delta_0(\mathcal{K})$ approaches the boundary of $\Delta_0(\mathcal{K})$, so that Z takes values in $\operatorname{ri}\Delta_0(\mathcal{K})$ and normalization is always well defined. As we shall see below, a further advantage of this lifting is that it provides additional control over how much the resulting strategy can vary as the score changes.

Putting everything together. We are now in a position to combine the various elements discussed above—regularization, extrapolation, and the lifting mechanism—into a general algorithmic blueprint, which we call *lift-extrapolated FTRL* (LIFTFTRL+). Reinstating player indices for completeness, the LIFTFTRL+ method updates in three interleaved phases as follows:

$$\begin{aligned}
&(a)\ \text{Extrapolate:} && y_{i,t+1/2} = y_{i,t} + \eta\hat{g}_{i,t} && z_{i,t+1/2} = Z(y_{i,t+1/2}) \\
&(b)\ \text{Play / Observe:} && x_{i,t} = Q(y_{i,t+1/2}) && g_{i,t} = v_i(x_t) - u_i(x_t)\mathbf{1} \\
&(c)\ \text{Update:} && y_{i,t+1} = y_{i,t} + \eta g_{i,t} && z_{i,t+1} = Z(y_{i,t+1})
\end{aligned} \qquad \text{(LIFTFTRL+)}$$

In the above, $\hat{g}_{i,t} \in \mathbb{R}^K$ denotes a predictor for the t-th stage excess payoff vector $g_{i,t}$, which, as in the case of (FTRL+), is constructed *before* the i-th player chooses a mixed strategy $x_{i,t}$ for round t. The precise form of $\hat{g}_{i,t}$ is a fundamental—as well as intricate—design choice for (LIFTFTRL+), so we postpone the details to the next section. To set the stage for this discussion, we only present below a basic template inequality for (LIFTFTRL+), which will underlie much of the subsequent analysis. To state it, let

$$e_{i,t} = g_{i,t} - \hat{g}_{i,t} \tag{12}$$

denote the *prediction error* incurred by $\hat{g}_{i,t}$, and, for all $i \in \mathcal{N}$ and all $t = 1, 2, \dots$, set

$$P_{i,t} = D_\psi(z_{i,t+1/2}, z_{i,t}) \qquad \text{and} \qquad I_{i,t} = D_\psi(z_{i,t+1/2}, z_{i,t+1}), \tag{13}$$

where

$$D_\psi(z, z') = \psi(z) - \psi(z') - \langle \nabla\psi(z'), z - z' \rangle \tag{14}$$

denotes the *Bregman divergence* of ψ. The two quantities in (13) basically represent the "*prediction*" and "*innovation*" Bregman movement at time t, in the sense that the former measures the Bregman displacement formed by the prediction, while the latter measures the discrepancy between the prediction and the actual realized update. With this in mind, we obtain the following template inequality for (LIFTFTRL+).

Proposition 1. *Let $H_0 = \max\psi - \min\psi$. Then, under* (LIFTFTRL+)*, we have:*

$$[\operatorname{Reg}_i(T)]_+ \leq \frac{H_0}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T}(I_{i,t} - P_{i,t}) \tag{15}$$

$$\leq \frac{H_0}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|e_{i,t}\|_\infty^2 - \frac{1}{\eta}\sum_{t=1}^{T}P_{i,t}, \tag{RVU}$$

for all $T \geq 1$ and all $i \in \mathcal{N}$.

The proof of Proposition 1 is detailed in Appendix E, and relies on an energy argument, telescoping the Bregman divergence from a fixed comparator to the base iterates, and the observation that

$$2I_{i,t} \leq \eta^2 \|e_{i,t}\|_\infty^2, \tag{16}$$

which is a consequence of the strong convexity of ψ. In this regard, it is worth noting that (RVU) may be seen as a *regret bounded by variation in utilities* (RVU) inequality, where the prediction error represents the "*variation in utilities*" in the language of Syrgkanis et al. (2015).

Proposition 1 also indicates what needs to be proved to obtain a constant regret bound. To see this, let

$$P(T) = \sum_{i\in\mathcal{N}} \sum_{t=1}^{T} P_{i,t}, \quad I(T) = \sum_{i\in\mathcal{N}} \sum_{t=1}^{T} I_{i,t}, \quad \text{and} \quad E(T) = \sum_{i\in\mathcal{N}} \sum_{t=1}^{T} \|e_{i,t}\|_\infty^2, \tag{17}$$

so, by the non-negativity of the lifted regret, Proposition 1 gives

$$P(T) + I(T) \leq NH_0 + \eta^2 E(T). \tag{18}$$

This shows that the prediction error upper bounds the Bregman movement; as such, to close the loop, we need to find a corresponding lower bound of the form

$$\sqrt{E(T)} \leq a + b\eta\sqrt{P(T) + I(T)} + c\sqrt{E(T)}, \tag{19}$$

for positive constants a, b and $c < 1$ that do not depend on T. In that case, substituting (18) gives

$$\sqrt{E(T)} \leq a + b\eta\sqrt{NH_0} + (b\eta^2 + c)\sqrt{E(T)}, \tag{20}$$

so, if η is small enough, the last term can be absorbed into the left hand side to yield a uniform, horizon-independent bound on $E(T)$—and hence, by (RVU), on the learner's regret. In view of this, establishing the target inequality (19) will be the main point driving our proof and our design choices.

4 Higher-order optimism with discounting

In this section, we proceed to specialize the two defining ingredients of (LiftFTRL+), the choice of regularizer and the choice of predictor. These two choices are driven respectively by the need to balance the interplay between the players' mixed strategies and their mass, and the need to exploit and dampen the interactions between players in order to establish sufficient control over the iterates of the algorithm. All this leads to a fairly intricate tapestry, which we outline below.

The lifted regularizer. A reasonable starting point for choosing a regularizer for (LiftFTRL+) is the negative entropy $h(x) = \sum_{\alpha\in\mathcal{K}} x_\alpha \log x_\alpha$ on $\Delta(\mathcal{K})$, which is the basis for the exponential weights template. Under this regularization, the mixed strategy choice map (6) is the *logit choice map*

$$\Lambda(y) = \frac{(\exp(y_1), \dots, \exp(y_K))}{\exp(y_1) + \dots + \exp(y_K)} \quad \text{for all } y \in \mathbb{R}^K, \tag{21}$$

and the associated Bregman divergence is simply the *Kullback–Leibler* (KL) divergence

$$D_{\mathrm{KL}}(p\|x) = \sum_{\alpha\in\mathcal{K}} p_\alpha \log \frac{p_\alpha}{x_\alpha} \quad \text{for all } p \in \Delta(\mathcal{K}) \text{ and all } x \in \operatorname{ri}\Delta(\mathcal{K}). \tag{22}$$

A natural way to extend this regularizer to $\Delta_0(\mathcal{K})$ would be to recall the decomposition $z = \lambda x$ of $z \in \Delta_0(\mathcal{K})$ and consider a separable lifting of the form $\psi(z) = h(x) + \theta(\lambda)$ for some 1-dimensional regularizer $\theta\colon (0,1] \to \mathbb{R}$. Unfortunately, as can be seen by a direct calculation (which we omit), this lifting invariably leads to a normalized choice map of the form $Q(y) = \Lambda(\lambda(y)y)$, for any θ. As it turns out, this form of Q is not suitable for our purposes, so we need a more flexible lifting scheme.

In view of the above, we go one step further and consider instead the *affine* lifting scheme

$$\psi(z) = \rho(\lambda)h(x) + \theta(\lambda) \tag{23}$$

where the coefficients $\rho(\lambda)$ and $\theta(\lambda)$ are to be determined. Solving the choice problem (11) now yields the more flexible expression $Q(y) = \Lambda(\lambda(y)/\rho(\lambda(y)) \cdot y)$, again irrespective of θ (we omit the details). The factor $\rho(\lambda)$ provides an extra degree of freedom which can then be tuned further during

the analysis in order to ensure sufficient stability; doing so ultimately leads to the *square-root lifted entropy*

$$\psi(z) = (\lambda + \sqrt{\lambda})h(x) - 3L\sqrt{\lambda(1-\lambda)}\,, \tag{SRE}$$

where, with a fair amount of hindsight, we set $L = 4 + \log K + \frac{1}{4}\log^2 K$.

The following proposition showcases the main useful properties of this choice.

Proposition 2. *Under* (SRE)*, the induced choice maps Z and Q of* (11) *are C^∞-smooth, and we have*

$$Q(y) = \Lambda\left(\frac{\sqrt{\lambda(y)}\,y}{1+\sqrt{\lambda(y)}}\right) \quad \textit{for all } y \in \mathbb{R}^K\,. \tag{24}$$

The range of ψ is bounded as $H_0 \leq 13L/6$, and the derivatives of Q and λ are bounded as

$$\|DQ(y)\|_{\mathrm{op}} \leq \min\{1, 2\sqrt{\lambda(y)}\} \quad \textit{and} \quad \|\nabla\lambda(y)\|_1 \leq \lambda(y)/4\,, \tag{25}$$

where DQ denotes the Jacobian of Q, and $\|\cdot\|_{\mathrm{op}}$ denotes the (L^∞, L^1) operator norm on $\mathbb{R}^{K\times K}$.

Proposition 2 shows that the multiplicative factor $\rho(\lambda)$ in (SRE) induces a mass-dependent temperature in the logit choice map; thus, when the mass is small, the chosen mixed strategy becomes less sensitive to score perturbations. The derivative bounds (25) subsequently quantify the stability we gain from this choice: *a*) the Jacobian bound shows that the sensitivity of the choice map is of order $\sqrt{\lambda}$, precisely compensating the $\mathcal{O}(1/\sqrt{\lambda})$ sensitivity of strategies in ℓ_1 norm; and *b*) the gradient bound essentially translates to multiplicative control over the mass, so bounded score changes imply that the resulting masses remain within a constant factor of each other. This allows us to preserve the desired cancellation between the $\sqrt{\lambda}$ sensitivity of the choice map and the $1/\sqrt{\lambda}$ scale of the strategy variations across nearby iterates, and is the main property that our predictor will exploit.

The higher-order predictor. We now proceed to define the last ingredient of our algorithm—its predictor. We begin by noting that the standard one-step prediction $\hat{g}_{i,t+1} = g_{i,t}$ of OptFTRL leads to an error of $e_{i,t+1} = \mathcal{O}(\|g_{i,t+1} - g_{i,t}\|_\infty)$. To refine this, we seek an analogous identity involving higher-order differences in the spirit of Daskalakis et al. (2021), so that the prediction rule can exploit the temporal regularity created by all players' updates. To that end, reinstating player indices for completeness, we will consider the *higher-order predictor* (HOP) given by

$$\hat{g}_{i,t} = \sum_{k=1}^{N+1}(-1)^{k+1}\binom{N+1}{k}\left[(1-\delta^k)g_{i,t-k} + \delta^k \hat{g}_{i,t-k}\right], \qquad \delta = \frac{N}{N+1}\,, \tag{HOP}$$

where δ is the predictor's *discount factor*, and we set $g_{i,s} = \hat{g}_{i,s} = 0$ for $s = -N, \dots, 0$. [Note that all quantities in (HOP) are available *before* play, so (HOP) defines a bona fide, *causal* predictor.]

The most mysterious aspect of (HOP) is the combinatorial coefficents that appear therein, and which have been reverse-engineered in order to stabilize the induced prediction error. To understand this, let w_t, $t = 1, 2, \dots$, be a sequence in $\mathbb{R}^K$—with negative indices set to zero—and consider the *time difference* and *discounted sum* operators

$$(\Delta w)_t = w_t - w_{t-1} \qquad \text{and} \qquad (\mathrm{S}_\delta\, w)_t = \sum_{k\geq 0}\delta^k w_{t-k}\,. \tag{26}$$

The prediction error sequence $e_{i,t}$, $t = 1, 2, \dots$, is then obtained by applying an $(N+1)$-fold composition of $\mathrm{S}_\delta\,\Delta$ to the payoff excess sequence $g_i = (g_{i,t})_{t=1,\dots,\infty}$, viz.

$$e_i = (\mathrm{S}_\delta\,\Delta)^{N+1} g_i \tag{27}$$

This higher-order form of the prediction error highlights the role of the predictor, which is adapted to the endogenous variation generated by the players' concurrent evolution, rather than the variation of the payoff sequence viewed in isolation. Indeed, when expanding the prediction error via (27), the order $N+1$ ensures that some player's response is encountered twice in the expansion, at different orders, enabling the previously described cancellation between strategy variation and the choice map's Jacobian. The resulting cross terms can then be controlled by the Bregman movement, which is itself bounded by the prediction error through (18), thereby closing the loop. Finally, the discounting plays a complementary stabilizing role, in that it dampens the exponential amplification which may otherwise result from taking repeated differences, leading in particular to a polynomial, rather than exponential, dependence on N in the final regret bound. We formalize and prove these useful properties of (HOP) in Appendix D.

The algorithm. With all this in hand, we are finally in a position to instantiate (LiftFTRL+) to obtain the *higher-order optimism with discounting* (HOOD) algorithm. In pseudocode form, the algorithm unfolds as follows:

Algorithm 1: Higher-order optimism with discounting (HOOD)

Initialize $y_{i,1} \leftarrow \mathbf{0}$; for $-N \leq k \leq 0$, set $g_{i,k}, \hat{g}_{i,k} \leftarrow \mathbf{0}$
for $t = 1, 2, \dots$ **do**
  compute $\hat{g}_{i,t}$ from (HOP), set $y_{i,t+1/2} \leftarrow y_{i,t} + \eta \hat{g}_{i,t}$
  update $x_{i,t} \leftarrow Q(y_{i,t+1/2})$, play $x_{i,t}$
  observe $v_{i,t} = v_i(x_t)$, set $g_{i,t} \leftarrow v_{i,t} - u_i(x_t)\mathbf{1}$ and $y_{i,t+1} \leftarrow y_{i,t} + \eta g_{i,t}$
end for

The resulting learning rule is of course uncoupled, i.e., each player only uses their own mixed strategy and payoff information. Then, as a result of all the design choices made along the way, we obtain the following regret guarantee:

Theorem 1. *Suppose that every player $i \in \mathcal{N}$ runs Algorithm 1 with learning rate*

$$0 < \eta \leq 1/[60(N+1)^3]\,. \tag{28}$$

We then have

$$\mathrm{Reg}_i(T) \leq 3L/\eta \quad \text{for all } T \geq 1. \tag{29}$$

In particular, for $\eta = 1/[60(N+1)^3]$, we get the uniform regret bound

$$\mathrm{Reg}_i(T) \leq 45(N+1)^3[16 + 4\log K + \log^2 K] = \mathcal{O}(N^3 \log^2 K) \quad \text{for all } T \geq 1. \tag{30}$$

The technical scaffolding for the proof of Theorem 1 is in turn provided by the explicit bound below.

Proposition 3. *With assumptions as in Theorem 1, the aggregate prediction error (17) is bounded as*

$$\sqrt{E(T)} \leq 5\sqrt{N}(N+1) + 2000(N+1)^5\eta\sqrt{P(T) + I(T)} + \tfrac{3}{10}\sqrt{E(T)} \quad \text{for all } T \geq 1. \tag{31}$$

By the discussion at the end of Section 3—and Eq. (19) in particular—the proof of Theorem 1 is fairly direct once the estimate (31) is in place. However, the proof of Proposition 3 is quite arduous and takes up most of the appendix; for that reason, we provide a detailed proof overview in Appendix B, highlighting the intricate interplay between the chosen regularizer and predictor of HOOD.

From an operational standpoint, two further remarks are in order: First, the admissible range of learning rates η in Theorem 1 is *independent* of the horizon of play T. Second, regarding equilibrium outcomes, recall that a distribution π on $\mathcal{A}$ is an *ε-coarse correlated equilibrium* (ε-CCE) if

$$\mathbb{E}_{\alpha\sim\pi}[u_i(\beta_i, \alpha_{-i}) - u_i(\alpha)] \leq \varepsilon \quad \text{for all } i \in \mathcal{N} \text{ and all } \beta_i \in \mathcal{A}. \tag{32}$$

Then, by standard results, Theorem 1 readily yields an $\mathcal{O}(1/T)$ rate of convergence to the set of CCE.

Corollary 1. *With assumptions as in Theorem 1, let $\pi_T = T^{-1}\sum_{t=1}^{T}\prod_i x_{i,t}$ denote the empirical distribution of play up to time T. Then π_T is a $3L/(\eta T)$-approximate CCE.*

Finally, we can also exploit the boundedness of the regret to further obtain robustness against adversarial payoff sequences via a very simple switching rule—it suffices for each player to follow HOOD until the regret bound is exceeded, and then switch permanently to any algorithm with adversarial regret guarantees. We provide the details of this guarantee in Appendix G.

5 Concluding Remarks

Theorem 1 bridges an important gap in the literature by showing that uncoupled learning *can* achieve $\mathcal{O}(1)$ regret in general games under full information. The algorithmic innovations required to achieve this involve a fairly intricate regularizer / predictor structure, so the behavior of HOOD relative to other performance metrics remains completely open. In particular, a very fruitful direction would be to see if the sequence of play generated by HOOD converges to Nash equilibrium in those classes of games where other regularized learning algorithms are known to converge—such as potential games, two-player zero-sum games, and/or harmonic games. We defer this question to future work.

AI USE STATEMENT

Some design choices in the algorithm and its analysis, including aspects of the mathematical reasoning and proof development, were assisted by generative AI tools. All AI-assisted material was independently reviewed and verified by the authors, who take responsibility for the final content of the paper.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

The main results of this paper are theoretical. We detail each underlying assumption before stating any result. Complete proofs are given in the Appendix.

REFERENCES

I. Anagnostides, G. Farina, C. Kroer, C.-W. Lee, H. Luo, and T. Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. In *Advances in Neural Information Processing Systems 35*, 2022a.

I. Anagnostides, I. Panageas, G. Farina, and T. Sandholm. On last-iterate convergence beyond zero-sum games. In *Proc. 39th Internat. Conf. Machine Learning*, 2022b.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. 36th Annual Sympos. Foundations of Computer Science*, 1995.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

D. Blackwell. Controlled random walks. In *Proc. Internat. Congr. Math.*, volume 3, pp. 336–338, 1954.

A. Blum and Y. Mansour. From external to internal regret. *J. Mach. Learn. Res.*, 8:1307–1324, 2007.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

X. Chen and B. Peng. Hedging in games: Faster convergence of external and swap regrets. In *Advances in Neural Information Processing Systems 33*, pp. 18990–18999, 2020.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

C. Daskalakis, M. Fishelson, and N. Golowich. Near-optimal no-regret learning in general games. In *Advances in Neural Information Processing Systems 34*, 2021.

G. Farina, I. Anagnostides, H. Luo, C.-W. Lee, C. Kroer, and T. Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Advances in Neural Information Processing Systems 35*, 2022.

D. Friedman. Evolutionary games in economics. *Econometrica*, 59(3):637–666, 1991.

J. Ha. A logarithmic regret bound for Optimistic Hedge in general-sum games. https://arxiv.org/abs/2609.19677, 2026a.

J. Ha. A horizon-independent regret bound for Optimistic Hedge in general-sum games. https://arxiv.org/abs/2609.22839, 2026b.

J. Hannan. Approximation to bayes risk in repeated play. In *Contributions to the Theory of Games, Vol. III*, pp. 97–139. Princeton University Press, 1957.

S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *Amer. Econ. Rev.*, 93(5): 1830–1836, 2003.

S. Hart and A. Mas-Colell. Stochastic uncoupled dynamics and Nash equilibrium. *Games Econom. Behav.*, 57: 286–303, 2006.

Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Proc. 34th Conf. Learning Theory*, 2021.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

D. Legacci, P. Mertikopoulos, C. H. Papadimitriou, G. Piliouras, and B. S. R. Pradelski. No-regret learning in harmonic games: Extrapolation in the face of conflicting interests. In *Advances in Neural Information Processing Systems 37*, 2024.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2): 212–261, 1994.

M. Liu, G. Farina, and A. Ozdaglar. Constant individual regret in general games. https://arxiv.org/abs/2608.31166, 2026.

T. Mai. Constant swap regret in general-sum games via two-scale higher-order optimism. https://arxiv.org/abs/2609.16751, 2026.

P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

G. Piliouras, R. Sim, and S. Skoulakis. Beyond time-average convergence: Near-optimal uncoupled online learning via Clairvoyant Multiplicative Weights Update. In *Advances in Neural Information Processing Systems 35*, 2022.

L. D. Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Math. Notes Acad. Sci. USSR*, 28(5):845–848, 1980.

A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26*, 2013a.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proc. 26th Annual Conf. Learning Theory*, pp. 993–1019, 2013b.

I. Sason and S. Verdú. f-divergence inequalities. *IEEE Trans. Inform. Theory*, 62(11):5973–6006, 2016.

L. J. Savage. The theory of statistical decision. *J. Amer. Statist. Assoc.*, 46(253):55–67, March 1951.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. In *NIPS' 06: Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 1265–1272. MIT Press, 2006.

S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Journal of Machine Learning Research*, 69(2):115–142, 2007.

A. Soleymani and G. Piliouras. Multiplicative optimism for constant regret in games. https://arxiv.org/abs/2609.21976, 2026.

A. Soleymani, G. Piliouras, and G. Farina. Faster rates for no-regret learning in general games via cautious optimism. In *Proc. 57th Annual ACM Sympos. Theory of Computing*, 2025a.

A. Soleymani, G. Piliouras, and G. Farina. Cautious optimism: A meta-algorithm for near-constant regret in general games, 2025b. Extended abstract in Proc. 26th ACM Conf. Economics and Computation; arXiv:2506.05005.

V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28*, 2015.

T. Tsuchiya. Tight regret upper and lower bounds for Optimistic Hedge in two-player zero-sum games. In *Proc. Internat. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2026.

V. G. Vovk. Aggregating strategies. In *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*, pp. 371–383, 1990.

A. Wald. *Statistical Decision Functions*. Wiley, 1950.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML '03: Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003.

A Contemporaneous work

During the final preparation stages of our manuscript, and within less than four weeks from the submission deadline, a number of preprints appeared online with motivation and results comparable to our own. Even though these concurrent unpublished works clearly constitute "contemporaneous papers" under the ICLR reviewing guidelines (https://iclr.cc/Conferences/2027/ReviewerGuidelines), we discuss them below for completeness.

Liu et al. (2026). The first of these papers appeared on arxiv on Aug. 31, 2026, and, completely independently, obtained a constant $\mathcal{O}(N^{21}\log^4 K)$ regret bound—by comparison, the guarantee of HOOD is $\mathcal{O}(N^3\log^2 K)$. The bound of Liu et al. (2026) was achieved via the use of a higher-order predictor similar to (HOP), but of order N instead of $N+1$ (the predictor in question is called an exponential moving average cascade in the paper). Liu et al. (2026) also employ a lifting mechanism with regularizer $\psi(z) = \sqrt{\lambda}[h(x) - A] - \sqrt{1-\lambda}$ for some positive constant A, which has a different boundary behavior than our own—and hence leads to different stabilization estimates. By far the greatest difference with Liu et al. (2026) lies in their proof approach, which relies on filter theory—and, in particular, on the specific filtering induced by these choices. Their proof techniques, motivated by signal processing and transfer function design, leverage a series of time-domain filters and EMA kernels, together with a recursive variation argument whose dependence on the number of players is reminiscent of—but substantially different from—our higher-order expansion analysis.

Mai (2026). This paper appeared on arxiv on Sep. 15, 2026, and used higher-order techniques to prove a constant *swap regret* bound of $\mathcal{O}(\sqrt{N}K\log K\log^{5/2}(NK))$. Higher-order optimism plays a particularly important role in the two-scale arguments of the paper, which make heavy use of the well-known Blum & Mansour (2007) reduction technique.

Ha (2026a,b). These two papers appeared on arxiv on Sep. 17 and Sep. 19, 2026, respectively, and focus on the regret guarantees of OptHedge under self-play. Ha (2026a) proves an $\mathcal{O}(\sqrt{N}\log K\log T)$ regret bound for OptHedge by sharpening the higher-order smoothness analysis of Daskalakis et al. (2021). By contrast, Ha (2026b) proves a non-constructive constant regret bound for OptHedge, using an exotic set of tools from real-analytic geometry. More specifically, using the fact that OptHedge defines a real-analytic recurrence on a compact space, the proof considers the ring of germs of real-analytic functions around this compact space and makes heavy use of the fact that this ring is Noetherian. Since, in a Noetherian ring, there can be no infinite ascending chain of ideals, this yields the existence of a finite and exact higher-order relation that closes the proof. However, owing to its non-constructive character, the paper leaves open and implicit both the tuning of the algorithm's learning rate and the functional dependence on N and K in the derived regret bound; put differently, this is more of an existence result, in that there exists *an instance* of OptHedge which achieves constant regret in T, but with an *unknown* dependence on N and K.

Soleymani & Piliouras (2026). This paper appeared on arxiv on Sep. 18, 2026, a few hours before the ICLR registration deadline. This paper considers a multiplicative approximant of OptHedge, and uses the stability provided by this multiplicative structure to tackle directly the evolution of a *first-order* predictor, and they obtain an anytime $\mathcal{O}(\sqrt{N}\log K)$ regret guarantee. In more detail, the update structure of Soleymani & Piliouras (2026) is closely related to that of OptHedge, but they replace the log-sum-exp regularization potential of OptHedge with a power-law function that is reminiscent of—but not equivalent to—Tsallis-type regularization. In addition, Soleymani & Piliouras (2026) also change the way the optimistic prediction is incorporated into the update, from exponential to multiplicative. In more detail, the aggregate payoffs are still incorporated via an exponential, but this is then multiplied directly by the optimistic correction instead of taking the exponential of the latter as well.

B Proof overview

In this appendix, we give an overview of the proof of Theorem 1, explaining how the regularizer, the higher-order predictor, and the finite expansion combine to control the prediction error and close

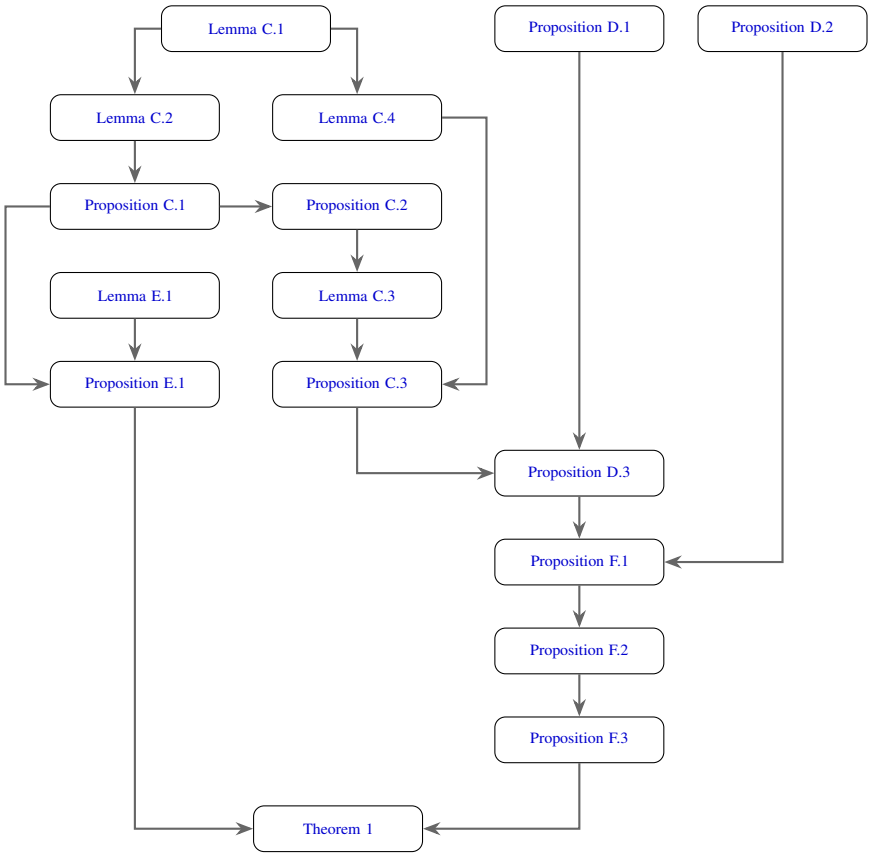


Figure 2. Dependency graph of the proof of Theorem 1.

the regret bound, all the precise proofs and statements are in the next appendices. We start from the template inequality (RVU) in the main text, proved in Proposition E.1:

$$[\mathrm{Reg}_i(T)]_+ \le \frac{H_0}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|e_{i,t}\|_\infty^2 - \frac{1}{\eta}\sum_{t=1}^{T} P_{i,t}, \tag{B.1}$$

where the positive term is the cumulative prediction error, whereas the negative term is the prediction Bregman movement. Moreover, the innovation bound (16), proved in Proposition E.1, controls the innovation movement by the same prediction error, so the nonnegativity of lifted regret gives, after summing over the players,

$$P(T) + I(T) \le NH_0 + \eta^2 E(T), \tag{B.2}$$

which is the aggregate movement bound (18) from the main text. Thus prediction error already controls total Bregman movement, and constant regret will follow once the converse bound controls prediction error by that movement, up to a sufficiently small residual copy of prediction error. Obtaining this converse bound requires local control of the strategy movement, enough successive payoff differences to repeat the resulting argument, and a summable expansion of the terms that remain after each difference. These requirements determine, in turn, the roles of the regularizer, discounting, and the higher-order predictor.

B.1 Bregman movement and the regularizer

The Bregman movement above concerns the lifted iterate λx, whereas payoff variation will depend on changes of the normalized strategy x. This becomes problematic when the lifted mass is small, indeed, for a naive lift of the form $\psi(\lambda x) = \lambda h(x) + \theta(\lambda)$ and equal masses,

$$D_\psi(\lambda x', \lambda x) = \lambda D_{\mathrm{KL}}(x'\|x) \ge \frac{\lambda}{2}\|x' - x\|_1^2, \tag{B.3}$$

so estimating the strategy variation $x' - x$ from lifted movement loses a factor $\lambda^{-1/2}$. Since a change of strategy is produced by a perturbation of the score through $x = Q(y)$, the choice map's sensitivity to scores must compensate for this loss. For this purpose, the square-root entropic regularizer (SRE)—whose relevant global and local properties are established in Propositions C.1 and C.2—was chosen so that the identities (24)–(25) give

$$Q(y) = \Lambda\left(\frac{\sqrt{\lambda(y)}}{1+\sqrt{\lambda(y)}}y\right) \quad \text{and} \quad \|\mathcal{J}(y)\|_{\mathrm{op}} \le \min\{1, 2\sqrt{\lambda(y)}\}, \tag{B.4}$$

where $\mathcal{J} = DQ$. Thus, writing

$$\lambda_{i,t} = \lambda(y_{i,t}), \qquad J_{i,t} = \eta\mathcal{J}(y_{i,t}), \tag{B.5}$$

the local lifted geometry—quantified in Proposition D.3—gives

$$\left\|\Delta x_{i,t}\right\|_1^2 \le \frac{5}{\lambda_{i,t}}(P_{i,t} + I_{i,t-1}). \tag{B.6}$$

Multiplying (B.6) by the Jacobian bound in (B.4) yields the crucial cancellation

$$\begin{aligned}\|J_{i,t}\|_{\mathrm{op}}^2\left\|\Delta x_{i,t}\right\|_1^2 &\le (4\eta^2\lambda_{i,t})\frac{5(P_{i,t} + I_{i,t-1})}{\lambda_{i,t}} \\ &= 20\eta^2(P_{i,t} + I_{i,t-1}),\end{aligned} \tag{B.7}$$

so the factor λ^{-1} lost when normalizing the lifted movement is exactly compensated by the factor λ supplied by the squared Jacobian. This cancellation will later be needed when the two factors are evaluated at nearby rounds rather than at the same score. Note also that, since $\lambda(y) > 0$, the second bound in (25) may be written as

$$|D\log\lambda(y)[q]| \le \frac{1}{4}\|q\|_\infty, \tag{B.8}$$

since, by ℓ_1–ℓ_∞ duality,

$$\sup_{\|q\|_\infty \le 1} |D\log\lambda(y)[q]| = \|\nabla\log\lambda(y)\|_1 = \frac{\|\nabla\lambda(y)\|_1}{\lambda(y)}. \tag{B.9}$$

In this regard, as an immediate consequence of Proposition C.2, the above bound gives, after integration,

$$\exp\left(-\frac{1}{4}\|y' - y\|_\infty\right) \le \frac{\lambda(y')}{\lambda(y)} \le \exp\left(\frac{1}{4}\|y' - y\|_\infty\right), \tag{B.10}$$

so nearby score arguments have comparable masses and the same $\lambda \times \lambda^{-1}$ cancellation survives up to a harmless constant. Moreover, local control of Jacobian variation by the Bregman divergence gives

$$\|J_{i,t+1} - J_{i,t}\|_{\mathrm{op}}^2 \le 4100\eta^2(P_{i,t} + I_{i,t}), \tag{B.11}$$

as proved in Proposition D.3, where we will call (B.11) the *Jacobian variation bound*.

B.2 Higher-order prediction and discounting

Our first observation is that a payoff time difference can be converted into strategy differences, and rewriting further those strategy differences through the score update may leave another payoff contribution, effectively creating a feedback loop which we could hope to make self-stabilizing. For this purpose, it is thus handier to possess a stable supply of successive time differences rather than only one such as in classical optimism. The recurrence (HOP) is thus designed precisely so that its prediction error already has this higher-order form (proved in Appendix D)

$$e_i = (\mathrm{S}_\delta\,\Delta)^{N+1} g_i, \tag{B.12}$$

hence this identity supplies the successive differences needed to repeat the payoff expansion, but before using them, we must ensure that their size does not grow exponentially with the order. Indeed, ordinary differences satisfy

$$\Delta^q w_t = \sum_{r=0}^{q}(-1)^r\binom{q}{r}w_{t-r}, \qquad \|\Delta^q w_t\|_\infty \le \sum_{r=0}^{q}\binom{q}{r}\|w_{t-r}\|_\infty, \tag{B.13}$$

and the absolute coefficient sum is therefore $\sum_{r=0}^{q}\binom{q}{r} = 2^q$. With $\delta = N/(N+1)$, the discounted sum operator offsets this growth, in that, for every square summable scalar sequence a and every $0 \le q \le N+1$,

$$\left(\sum_{t\in\mathbb{Z}} |((\mathrm{S}_\delta\,\Delta)^q a)_t|^2\right)^{1/2} \le 2\left(\sum_{t\in\mathbb{Z}} |a_t|^2\right)^{1/2}, \tag{B.14}$$

as established in the proof of Proposition D.1, which allows us to avoid the exponential cost in N. Thus, the discounted summation acts as a smoothing filter that counterbalances the amplification introduced by repeated differences. Next, to follow through with the expansion and apply the time differences one at a time, we set, for $0 \le p \le N+1$,

$$\mathrm{D}_p = \mathrm{S}_\delta{}^{N+1}\,\Delta^{N+1-p}, \quad \text{so that} \quad \mathrm{D}_p = \mathrm{D}_{p+1}\Delta, \tag{B.15}$$

where the second identity holds for $p \le N$. In particular, (B.12) becomes $e_i = \mathrm{D}_1\Delta g_i$, and passing from order p to order $p+1$ reveals exactly one further ordinary difference. In this sense, the family D_p interpolates between smoothing and differencing while exposing the latter one order at a time. Generally, for a sequence w vanishing at nonpositive indices, we have

$$\left(\sum_{t=1}^{T}\left\|(\mathrm{D}_p w)_t\right\|_\infty^2\right)^{1/2} \le 4(N+1)^{p+1}\left(\sum_{t=1}^{T}\|w_t\|_\infty^2\right)^{1/2}, \tag{B.16}$$

as proved in Proposition D.1, so, for the expansion still, exposing one more difference will cost us only one additional factor $N+1$, noting also that the zero extension for negative indices contributes only a finite initialization term, and the learning rate restriction (28) keeps all score displacements generated over the required $N+1$ rounds inside the local regime described by (B.10) and (B.11) (Propositions C.3 and D.1 and Lemma C.3).

B.3 Payoff differences and the finite expansion

We now apply the first exposed difference in $e_i = \mathrm{D}_1\Delta g_i$. The basic idea is already visible for a bilinear payoff (two player game) with payoff matrix A:

$$\langle a_{t+1}, Ab_{t+1}\rangle - \langle a_t, Ab_t\rangle = \langle \Delta a_{t+1}, Ab_{t+1}\rangle + \langle a_t, A\Delta b_{t+1}\rangle, \tag{B.17}$$

so one time difference lands on one multilinear argument at a time. Applying the same product rule to the payoff excess and accounting for the zero initialization gives an expansion of the form

$$\Delta g_{i,t} = \mathbb{1}_{\{t=1\}}\gamma_{i,0} + \sum_{j\neq i} f^v_{ij,t}\Delta x_{j,t} + \sum_{j=1}^{N} f^u_{ij,t}\Delta x_{j,t}, \tag{B.18}$$

as proved in Proposition D.2, where the linear "coefficient maps" f^v and f^u have norm at most one when their remaining arguments are strategies. Apart from the finite initialization term, a payoff difference has therefore been reduced to strategy differences. These strategy differences are nonlinear functions of the scores, so we expand the choice map to first order around the corresponding base score. With the Jacobian $J_{i,t}$ from (B.5), define the discrete *Taylor remainder* $r_{i,t}$ by

$$\Delta x_{i,t} = J_{i,t}\frac{\Delta y_{i,t+1/2}}{\eta} + r_{i,t}. \tag{B.19}$$

Using the score update in Algorithm 1, this becomes

$$\begin{aligned}\Delta y_{i,t+1/2} &= \eta(g_{i,t} - \Delta e_{i,t}),\\ \Delta x_{i,t} &= r_{i,t} + J_{i,t}g_{i,t} - J_{i,t}\Delta e_{i,t}.\end{aligned} \tag{B.20}$$

with the corresponding identities given in (D.43)–(D.45). The first and third contributions are already controlled, and the same local choice map bounds that gave (B.6) yield

$$\|r_{i,t}\|_1^2 \le 300000(N+1)^2\eta^2(P_{i,t} + I_{i,t-1}), \tag{B.21}$$

as proved in Proposition D.3, while, since $\|J_{i,t}\|_{\mathrm{op}} \le \eta$ by Proposition C.2,

$$\|J_{i,t}\Delta e_{i,t}\|_1^2 \le 2\eta^2\left(\|e_{i,t}\|_\infty^2 + \|e_{i,t-1}\|_\infty^2\right), \tag{B.22}$$

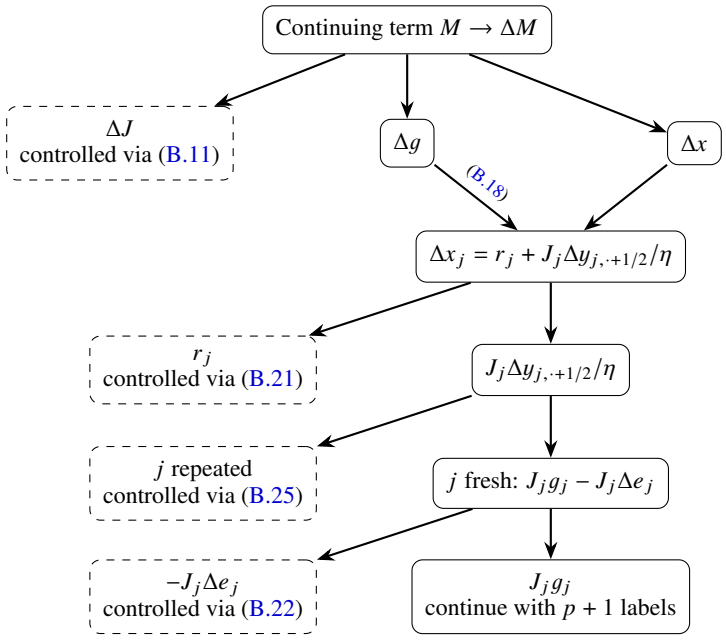


Figure 3. One step of the order p expansion.

as follows from (D.46), so it is already controlled by the prediction error and, consequently, only the payoff contribution $J_{i,t} g_{i,t}$ may require another difference. To proceed, we call a term *continuing* when, after the contributions already controlled by (B.21) and (B.22) are removed, it still contains such a payoff factor and must therefore use the next time difference in the expansion. To see this more concretely, at first order, every continuing term, as formalized in Proposition F.1, has the form

$$\mathrm{D}_1\big(f^v_{ij} J_j g_j\big) \quad \text{or} \quad \mathrm{D}_1\big(f^u_{ij} J_j g_j\big), \tag{B.23}$$

then, writing either coefficient as f_t and applying the next difference gives

$$\begin{aligned} \Delta(f J_j g_j)_t = {} & (f_t - f_{t-1}) J_{j,t} g_{j,t} \\ & + f_{t-1}(J_{j,t} - J_{j,t-1}) g_{j,t} \\ & + f_{t-1} J_{j,t-1} \Delta g_{j,t}. \end{aligned} \tag{B.24}$$

We call the contributions created when the difference lands on an existing strategy factor, Jacobian factor, or payoff factor, respectively, a *strategy term*, *Jacobian term*, or *payoff term*. Thus the second line of (B.24) is a Jacobian term and stops by the Jacobian variation bound (B.11); the coefficient difference in the first line is a sum of strategy terms by multilinearity, and the payoff term in the third line becomes strategy terms by (B.18). Substituting (B.20) into each resulting strategy difference again removes the remainder and prediction error contributions, leaving only another factor $J_j g_j$ as a possible continuing contribution, as formalized in Proposition F.1.

At this point several player indices may occur in the same expression, hence we shall call the index attached to a factor J_j or Δx_j its *player label*. Suppose now that, at order p, a continuing term contains Jacobians with p distinct player labels and a strategy difference carries label j, then, if j is already among those labels, a Jacobian which appeared at a prior order can be paired with the linear part of that strategy difference, and we can use (B.7) to bound this pairing. Indeed, the strategy difference has the inverse mass factor from (B.6) and the existing Jacobian cancels it through (B.4), using (B.10) to keep the two masses comparable across the at most $N+1$ relevant round shifts in the expansion. After accounting for the already controlled remainder in (B.21), this gives, for $|t - t'| \le N+1$, the *repeated player bound*

$$\|J_{j,t'}\|^2_{\mathrm{op}} \left\| J_{j,t} \frac{\Delta y_{j,t+1/2}}{\eta} \right\|_1^2 \le 90\eta^2 (P_{j,t} + I_{j,t-1}). \tag{B.25}$$

as proved in Proposition D.3. Therefore, a repeated label is stopped by Bregman movement. Then, if j is not already among the Jacobian labels, we call it *fresh*, and (B.20) shows that only $J_j g_j$ can remain,

so this continuing term has one additional Jacobian label, hence once $p = N$ no further continuation is possible, as proved in Proposition F.1; this explains in particular why the $N + 1$ differences in (B.12) are sufficient. The main repeated step in the expansion is summarized in Figure 3.

With all this done, it now remains to sum the finite expansion without losing control of its size. At order p, a continuing term contains $s \le p$ payoff terms, and any fixed player label can occur among its existing strategy terms at most p times. For a fresh player label j, since the payoff excess in (B.18) contains one payoff vector contribution and one expected payoff contribution, j can appear at most twice when differencing a payoff factor, as shown in Proposition D.2, thus giving at most $2s$ possibilities, meanwhile, differencing the strategy factors already present in the expansion gives at most p further possibilities. For these reasons, a fresh label can generate at most

$$2s + p \le 3p \tag{B.26}$$

continuing branches in the expansion and, since p of the N labels are already present, only $N - p$ labels may remain fresh and there are therefore at most $3p(N - p)$ continuing terms at the next order, as proved in Proposition F.1. Then, passing from one such term to the next order costs one factor $N + 1$ through (B.16), while the newly introduced factor $J_j g_j$ has size at most 2η because $\|J_j\|_{\mathrm{op}} \le \eta$ by Proposition C.2 and $\|g_j\|_\infty \le 2$. Hence, by (28), the total weighted continuation factor satisfies

$$\big[3p(N - p)\big](N + 1)(2\eta) \le \frac{1}{40}. \tag{B.27}$$

which is the contraction used in Proposition F.2. Thus successive orders contract geometrically, apart from the single explicit factor p contributed by the controlled terms at order p. Summing over all branches and using $\sum_{p\ge1} p(1/40)^{p-1} < 11/10$ as in Proposition F.2, then adding the finite initialization contribution and the remainder and prediction error contributions already separated in (B.20), gives the self-bound of Proposition F.3

$$\sqrt{E(T)} \le 5\sqrt{N}(N + 1) + 2000(N + 1)^5\eta\sqrt{P(T) + I(T)} + \frac{3}{10}\sqrt{E(T)}, \tag{B.28}$$

where the three terms on the right are respectively the initialization cost, the contributions controlled by Bregman movement, and the contributions controlled by prediction error.

Next, folding the movement bound (B.2) into (B.28) and using $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}$, we obtain

$$\begin{aligned}\sqrt{E(T)} \le{}& 5\sqrt{N}(N + 1) + 2000(N + 1)^5\eta\sqrt{NH_0} \\ &+ \left(2000(N + 1)^5\eta^2 + \frac{3}{10}\right)\sqrt{E(T)}.\end{aligned} \tag{B.29}$$

By the range bound in Proposition 2 (proved in Proposition C.1) and the learning rate choice (28), the first two terms on the right are together smaller than $26(N + 1)^3\sqrt{L}$, while the coefficient of $\sqrt{E(T)}$ is at most $26/45 < 1$. Hence

$$\sqrt{E(T)} < 26(N + 1)^3\sqrt{L} + \frac{26}{45}\sqrt{E(T)}, \tag{B.30}$$

and rearranging, using $26 \cdot 45/19 < 62$, gives the horizon-independent bound

$$E(T) < 62^2(N + 1)^6 L. \tag{B.31}$$

as established in Appendix F. Returning to the template inequality (B.1) and discarding its nonpositive movement term now yields

$$[\mathrm{Reg}_i(T)]_+ \le \frac{H_0}{\eta} + \frac{\eta}{2}E(T) < \left(\frac{13}{6} + \frac{5}{9}\right)\frac{L}{\eta} < \frac{3L}{\eta}, \tag{B.32}$$

where we used Proposition 2, (28), and (B.31), with the complete calculation given in Appendix F. Finally, we have obtained (29), and the choices in Theorem 1 then give (30).

C Lifted regularizer and choice map

In this appendix, we establish the regularity properties of the regularizer that will provide the control needed for the expansion underlying the proof of Theorem 1.

C.1 Global properties of the regularizer

To prepare these bounds, for $x \in \Delta_K$ and values a_α defined on the support of x, write

$$\mathrm{Var}_x(a) = \sum_{\alpha: x_\alpha > 0} x_\alpha \left(a_\alpha - \sum_{\beta: x_\beta > 0} x_\beta a_\beta \right)^2. \tag{C.1}$$

For $p \in \Delta_K$ and $q \in \mathrm{ri}\,\Delta_K$, we will use the following KL bounds (Cover & Thomas, 2006; Sason & Verdú, 2016):

$$D_{\mathrm{KL}}(p\|q) \geq \frac{1}{2}\|p - q\|_1^2, \qquad D_{\mathrm{KL}}(p\|q) \geq \left\|\sqrt{p} - \sqrt{q}\right\|_2^2. \tag{C.2}$$

Likewise, for $x \in \Delta_K$, Gibbs' inequality gives (Cover & Thomas, 2006)

$$-\log K \leq h(x) \leq 0, \qquad \log K + h(x) = D_{\mathrm{KL}}(x\|K^{-1}\mathbf{1}) \geq 0. \tag{C.3}$$

Finally, for $z = \lambda x \in \mathrm{ri}\,\Delta_0(\mathcal{K})$, we use the decomposition illustrated in Figure 1:

$$\xi = bx + \lambda w, \qquad b = \langle \mathbf{1}, \xi \rangle, \qquad \langle \mathbf{1}, w \rangle = 0. \tag{C.4}$$

The following lemma supplies the logarithmic moment bound used to control the derivatives of the regularizer.

Lemma C.1. *For every $x \in \Delta_K$,*

$$\mathrm{Var}_x(\log x) \leq \frac{1}{4}\log^2 K + \log K + 2 = L - 2. \tag{C.5}$$

Proof. We begin with two moment bounds, used below to bound the variance and later in Lemma C.4 to control its variation. For $p \in \mathbb{R}_+^K$ with $\sum_\alpha p_\alpha \leq 1$, we claim that

$$\sum_{\alpha: p_\alpha > 0} p_\alpha |\log p_\alpha|^2 \leq 4L, \tag{C.6}$$

$$\sum_{\alpha: p_\alpha > 0} p_\alpha |\log p_\alpha|^4 \leq 26L^2. \tag{C.7}$$

Indeed, for $s \geq 0$, summing only over actions with $p_\alpha > 0$ gives

$$\sum_{\alpha: -\log p_\alpha \geq s} p_\alpha \leq \min\{1, Ke^{-s}\}. \tag{C.8}$$

Integrating (C.8) against the identity $[a]_+^q = q\int_0^\infty u^{q-1}\mathbb{1}_{\{a \geq u\}}\,du$, for every integer $q \geq 1$ we obtain

$$\begin{aligned}\sum_{\alpha: p_\alpha > 0} p_\alpha [-\log p_\alpha - \log K]_+^q &= q\int_0^\infty u^{q-1} \sum_{\alpha: -\log p_\alpha \geq \log K + u} p_\alpha\,du \\ &\leq q\int_0^\infty u^{q-1}e^{-u}\,du = q!.\end{aligned} \tag{C.9}$$

Minkowski's inequality in the weighted sum, applied with $-\log p_\alpha \leq \log K + [-\log p_\alpha - \log K]_+$, yields

$$\left(\sum_{\alpha: p_\alpha > 0} p_\alpha |\log p_\alpha|^q\right)^{1/q} \leq \log K + (q!)^{1/q}. \tag{C.10}$$

Specializing (C.10) to $q = 2$ and recalling that $2\sqrt{L} \geq \log K + 2$ proves (C.6). For $q = 4$, the inequalities $24^{1/4} < 5/2$ and $\sqrt{L} \geq 2$ imply

$$\log K + 24^{1/4} < \log K + \frac{5}{2} \leq 2\sqrt{L} + \frac{1}{2} \leq \frac{9}{4}\sqrt{L}. \tag{C.11}$$

Consequently,

$$\sum_{\alpha: p_\alpha > 0} p_\alpha |\log p_\alpha|^4 \leq \left(\frac{9}{4}\right)^4 L^2 < 26L^2, \tag{C.12}$$

which proves (C.7). Taking now $p = x \in \Delta_K$ and using $\mathrm{Var}_x(\log x) = \mathrm{Var}_x(-\log x)$, we may bound the latter by its second moment about $\frac{1}{2}\log K$ to obtain

$$\mathrm{Var}_x(\log x) \le \sum_{\alpha: x_\alpha > 0} x_\alpha \left(-\log x_\alpha - \frac{1}{2}\log K\right)^2. \tag{C.13}$$

For actions satisfying $-\log x_\alpha \le \log K$, the squared term is at most $(\log K)^2/4$. On the complementary set,

$$\left(-\log x_\alpha - \frac{1}{2}\log K\right)^2 = \frac{(\log K)^2}{4} + \int_{\log K}^{-\log x_\alpha} 2\left(s - \frac{1}{2}\log K\right)\, ds. \tag{C.14}$$

Summing over the actions and applying (C.8) inside the integral, we obtain

$$\begin{aligned}\sum_{\alpha: x_\alpha > 0} x_\alpha \left(-\log x_\alpha - \frac{1}{2}\log K\right)^2 &\le \frac{(\log K)^2}{4} + \int_{\log K}^{\infty} 2\left(s - \frac{1}{2}\log K\right) K e^{-s}\, ds \\ &= \frac{1}{4}\log^2 K + \log K + 2 = L - 2.\end{aligned} \tag{C.15}$$

Substituting this bound into the variance bound proves (C.5). ∎

To apply the logarithmic moment bound to the regularizer, separate its square root terms: for nonzero $z = \lambda x \in \Delta_0(\mathcal{K})$, define

$$\varphi(z) = \sqrt{\lambda}\, h(x) - 3L\sqrt{\lambda(1-\lambda)}, \qquad \varphi(\mathbf{0}) = 0. \tag{C.16}$$

The next lemma applies the logarithmic moment bound to obtain bounds on the Hessian with respect to a variation $\xi = bx + \lambda w$ in lifted space.

Lemma C.2. *Let $z = \lambda x \in \mathrm{ri}\,\Delta_0(\mathcal{K})$ and let $\xi = bx + \lambda w$ satisfy (C.4). Then*

$$D^2\varphi(z)[\xi,\xi] \ge \frac{Lb^2}{2\lambda^{3/2}} \ge 0. \tag{C.17}$$

Moreover,

$$D^2\varphi(z)[\xi,\xi] \ge \frac{1}{2}\sqrt{\lambda}\sum_\alpha \frac{w_\alpha^2}{x_\alpha} + 3b^2. \tag{C.18}$$

Proof. Along the line $z(q) = z + q\xi = \lambda(q)x(q)$, the decomposition (C.4) gives

$$\lambda(q) = \lambda + qb, \qquad x(q) = x + \frac{q\lambda}{\lambda + qb}w, \tag{C.19}$$

and hence $x'(0) = w$ and $x''(0) = -2bw/\lambda$. Since $\langle \mathbf{1}, w\rangle = 0$, we also have

$$Dh(x)[w] = \langle w, \log x\rangle, \qquad D^2h(x)[w,w] = \sum_\alpha \frac{w_\alpha^2}{x_\alpha}. \tag{C.20}$$

Two differentiations of $\varphi(z(q))$ at $q = 0$ now give

$$D^2\varphi(z)[\xi,\xi] = \lambda^{-3/2}\left[\lambda^2\sum_\alpha \frac{w_\alpha^2}{x_\alpha} - \lambda b\langle w, \log x\rangle + \frac{b^2}{4}\left(\frac{3L}{(1-\lambda)^{3/2}} - h(x)\right)\right]. \tag{C.21}$$

To control the mixed term in (C.21), use $\langle \mathbf{1}, w\rangle = 0$ and Cauchy–Schwarz to write

$$|\langle w, \log x\rangle| = |\langle w, \log x - h(x)\mathbf{1}\rangle| \le \sqrt{\mathrm{Var}_x(\log x)}\left(\sum_\alpha \frac{w_\alpha^2}{x_\alpha}\right)^{1/2}. \tag{C.22}$$

Writing $S = \sum_\alpha w_\alpha^2/x_\alpha$ and $V = \mathrm{Var}_x(\log x)$, we complete the square to obtain

$$\lambda^2 S - \lambda b\langle w, \log x\rangle \ge -\frac{Vb^2}{4}. \tag{C.23}$$

The bounds (C.3) and (C.5), together with $(1-\lambda)^{-3/2} \geq 1$, show that the bracket in (C.21) is at least

$$\frac{b^2}{4}(3L - V - h(x)) \geq \frac{2L+2}{4}b^2 \geq \frac{L}{2}b^2, \tag{C.24}$$

which proves (C.17). To obtain the stronger bound, we retain half of the positive strategy term when bounding the cross term, which gives

$$\lambda^2 S - \lambda b\langle w, \log x\rangle \geq \frac{1}{2}\lambda^2 S - \frac{1}{2}Vb^2. \tag{C.25}$$

Upon writing $s = \sqrt{\lambda}$, substitution of (C.5) into (C.21), with $L \geq 4$, yields

$$\begin{aligned} D^2\varphi(z)[\xi,\xi] &\geq \frac{1}{2}\sqrt{\lambda}S + \lambda^{-3/2}\frac{b^2}{4}\left(\frac{3L}{(1-\lambda)^{3/2}} - 2V - h(x)\right) \\ &\geq \frac{1}{2}\sqrt{\lambda}S + s^{-3}b^2\left(\frac{3}{(1-s^2)^{3/2}} - 1\right). \end{aligned} \tag{C.26}$$

To finish, the function $3(1-s^2)^{-3/2} - 1 - 3s^3$ has value 2 at $s = 0$ and derivative $9s[(1-s^2)^{-5/2} - s] \geq 0$ on $[0, 1)$. Hence the coefficient of b^2 in the preceding display is at least 3, proving (C.18). ∎

The preceding Hessian bounds now give the global properties of the regularizer and its choice map.

Proposition C.1. *The regularizer* (SRE) *is continuous on* $\Delta_0(\mathcal{K})$*, steep and* 1*-strongly convex with respect to the* ℓ_1 *norm. For every* $y \in \mathbb{R}^K$*, the maximizer* $Z(y)$ *is unique, belongs to* $\operatorname{ri}\Delta_0(\mathcal{K})$*, and satisfies*

$$\nabla\psi(Z(y)) = y, \qquad Q(y) = \Lambda\left(\frac{\sqrt{\lambda(y)}}{1+\sqrt{\lambda(y)}}y\right). \tag{C.27}$$

Moreover, Z *and* Q *are* C^∞*,* $\max_{\Delta_0(\mathcal{K})}\psi = 0$*, and*

$$\max_{\Delta_0(\mathcal{K})}\psi - \min_{\Delta_0(\mathcal{K})}\psi \leq \frac{13}{6}L. \tag{C.28}$$

Proof. Equations (SRE) and (C.3), together with $r\log r \to 0$ as $r \downarrow 0$, show that both ψ and φ are continuous on $\Delta_0(\mathcal{K})$. The positive semidefiniteness in (C.17) makes φ convex on $\operatorname{ri}\Delta_0(\mathcal{K})$, and continuity extends this property to $\Delta_0(\mathcal{K})$. Adding the perspective entropy to the resulting convex term, for $z = \lambda x \in \operatorname{ri}\Delta_0(\mathcal{K})$ we have

$$\psi(z) = \lambda h(x) + \varphi(z). \tag{C.29}$$

For $\xi = bx + \lambda w$ as in (C.4), direct differentiation of the perspective term gives

$$D^2[\lambda h(x)][\xi,\xi] = \lambda\sum_\alpha \frac{w_\alpha^2}{x_\alpha}. \tag{C.30}$$

Adding this identity to (C.18), we obtain

$$D^2\psi(z)[\xi,\xi] \geq \left(\lambda + \frac{1}{2}\sqrt{\lambda}\right)\sum_\alpha \frac{w_\alpha^2}{x_\alpha} + 3b^2 \geq \|\xi\|_1^2. \tag{C.31}$$

For the last inequality, Cauchy–Schwarz gives $\|w\|_1^2 \leq \sum_\alpha w_\alpha^2/x_\alpha$. Setting $a = |b|$ and $c = \lambda\|w\|_1$, and using $\lambda \leq 1$, bounds the middle expression in (C.31) below by $3a^2 + \frac{3}{2}c^2 \geq (a+c)^2 \geq \|\xi\|_1^2$. Integration along line segments in $\operatorname{ri}\Delta_0(\mathcal{K})$, followed by approximation of boundary points, then shows that ψ is 1-strongly convex with respect to the ℓ_1 norm on $\Delta_0(\mathcal{K})$, and in particular that $D^2\psi(z)$ is positive definite on $\operatorname{ri}\Delta_0(\mathcal{K})$.

For every y, continuity and strong convexity now give a unique maximizer in (11). To locate it, fix $0 < \lambda \leq 1$, for which the x-dependent part of the objective is

$$\lambda\langle y, x\rangle - (\lambda + \sqrt{\lambda})h(x). \tag{C.32}$$

Division by $\lambda + \sqrt{\lambda}$ reduces this optimization to the entropic choice formula (21), so at any positive optimizing mass, $x = \Lambda(\sqrt{\lambda}\, y/(1+\sqrt{\lambda}))$, which is (24). The maximized value at mass $\lambda \in (0,1)$ is therefore

$$V_y(\lambda) = (\lambda + \sqrt{\lambda}) \log\left(\sum_{\alpha=1}^{K} \exp\left(\frac{\sqrt{\lambda}}{1+\sqrt{\lambda}} y_\alpha\right)\right) + 3L\sqrt{\lambda(1-\lambda)}. \tag{C.33}$$

This function extends continuously to $[0,1]$ with $V_y(0) = 0$, and

$$\lim_{\lambda \downarrow 0} \frac{V_y(\lambda)}{\sqrt{\lambda}} = \log K + 3L > 0. \tag{C.34}$$

The positive limit excludes $\lambda = 0$, and to rule out the other endpoint we differentiate the objective at the unique smooth strategy $x(\lambda)$ determined by the logit formula for $0 < \lambda < 1$. Its contribution to the derivative vanishes because the x-gradient is a scalar multiple of $\mathbf{1}$ and $\langle \mathbf{1}, x'(\lambda) \rangle = 0$, leaving

$$V_y'(\lambda) = \langle y, x \rangle - \left(1 + \frac{1}{2\sqrt{\lambda}}\right) h(x) + \frac{3L}{2} \frac{1-2\lambda}{\sqrt{\lambda(1-\lambda)}}. \tag{C.35}$$

As $\lambda \uparrow 1$, the last term tends to $-\infty$ and all other terms remain bounded, so $V_y' < 0$ on some interval $(1-\varepsilon, 1)$. Integration over that interval gives $V_y(1) < V_y(1-\varepsilon)$, excluding $\lambda = 1$ and placing the unique maximizer in $\operatorname{ri} \Delta_0(\mathcal{K})$.

At this maximizer,

$$\nabla \psi(Z(y)) = y, \tag{C.36}$$

and since the regularizer is C^∞ on $\operatorname{ri} \Delta_0(\mathcal{K})$ by (SRE) with positive definite Hessian by (C.31), the implicit function theorem applied to (C.36) shows that $Z \in C^\infty(\mathbb{R}^K)$. Normalization then gives $Q \in C^\infty(\mathbb{R}^K)$. Finally, ψ is steep: indeed, if $z_t \in \operatorname{ri} \Delta_0(\mathcal{K})$ approaches $\partial\, \Delta_0(\mathcal{K})$ and $\|\nabla \psi(z_t)\|_1$ does not tend to infinity, then, along a subsequence, $\nabla \psi(z_t) \to q$ for some $q \in \mathbb{R}^K$ and, since $z_t = Z(\nabla \psi(z_t))$, continuity of Z gives $z_t \to Z(q) \in \operatorname{ri} \Delta_0(\mathcal{K})$, a contradiction.

For the range bound, the nonpositivity of both terms in (SRE), together with $\psi(\mathbf{0}) = 0$, gives $\max_{\Delta_0(\mathcal{K})} \psi = 0$. Moreover, the identity

$$L - 3 \log K = \frac{1}{4} (\log K - 4)^2 \geq 0, \tag{C.37}$$

implies $\log K / L \leq 1/3$, and hence

$$\frac{-\psi(\lambda x)}{L} \leq \frac{\log K}{L} (\lambda + \sqrt{\lambda}) + 3\sqrt{\lambda(1-\lambda)} \leq \frac{2}{3} + \frac{3}{2} = \frac{13}{6}. \tag{C.38}$$

Since this bound holds throughout $\Delta_0(\mathcal{K})$, it proves the stated range bound. ∎

C.2 Local stability of the choice map

We next differentiate the choice formula of Proposition C.1, retaining its dependence on both mass and strategy. To this end, denote $\odot$ for coordinate-by-coordinate multiplication, write $Z(y) = \lambda x$ and, for $x \in \operatorname{ri} \Delta_K$, set

$$\mu_x = x \odot (\log x - h(x)\mathbf{1}), \qquad \Gamma_x = \operatorname{Diag}(x) - x x^\top. \tag{C.39}$$

Then

$$\Gamma_x v = x \odot (v - \langle x, v \rangle \mathbf{1}), \qquad \Gamma_x \log x = \mu_x, \qquad \|\mu_x\|_1 \leq \sqrt{\operatorname{Var}_x(\log x)} \leq \sqrt{L-2}. \tag{C.40}$$

For $\lambda \in (0,1)$ and $x \in \operatorname{ri} \Delta_K$, also define

$$\mathcal{D}(\lambda, x) = \frac{3L}{4(1-\lambda)^{3/2}} - \frac{1}{4}\left[h(x) + \frac{\operatorname{Var}_x(\log x)}{1+\sqrt{\lambda}}\right]. \tag{C.41}$$

The entropy and variance bounds (C.3)–(C.5) imply

$$\mathcal{D}(\lambda, x) \geq \frac{L}{2(1-\lambda)^{3/2}} + \frac{1}{2} \geq \frac{L+1}{2}. \tag{C.42}$$

Indeed, the first inequality uses $\mathrm{Var}_x(\log x) \le L-2$ and $(1-\lambda)^{-3/2} \ge 1$. Finally, for $\lambda \in (0,1)$ and $x \in \operatorname{ri}\Delta_K$, set

$$\nu_{\lambda,x} = x + \frac{\mu_x}{2(1+\sqrt{\lambda})}. \tag{C.43}$$

The next proposition bounds the resulting Jacobian and the logarithmic mass derivative, which will allow the local bounds to be compared at nearby scores.

Proposition C.2. *For every $y, q \in \mathbb{R}^K$,*

$$\|\mathcal{J}(y)\|_{\mathrm{op}} \le \min\{1, 2\sqrt{\lambda(y)}\}, \qquad |D\log\lambda(y)[q]| \le \frac{1}{4}\|q\|_\infty. \tag{C.44}$$

Proof. The first order condition (C.35) for the optimizing mass reads

$$\langle y, x\rangle = \left(1 + \frac{1}{2\sqrt{\lambda}}\right) h(x) - \frac{3L}{2}\frac{1-2\lambda}{\sqrt{\lambda(1-\lambda)}}. \tag{C.45}$$

Differentiating (24) with respect to the mass, on the other hand, yields

$$\partial_\lambda x = \frac{1}{2\sqrt{\lambda}(1+\sqrt{\lambda})^2}\Gamma_x y = \frac{\mu_x}{2\lambda(1+\sqrt{\lambda})}, \tag{C.46}$$

where the second identity follows from $\Gamma_x y = (1+\sqrt{\lambda})\mu_x/\sqrt{\lambda}$. Consequently,

$$\partial_\lambda h(x) = \frac{\mathrm{Var}_x(\log x)}{2\lambda(1+\sqrt{\lambda})}, \qquad \partial_\lambda\langle y, x\rangle = \frac{\mathrm{Var}_x(\log x)}{2\lambda^{3/2}}. \tag{C.47}$$

Differentiation of (C.35), followed by collection of terms, gives

$$\begin{aligned} V_y''(\lambda) &= \lambda^{-3/2}\left[\frac{\mathrm{Var}_x(\log x)}{4(1+\sqrt{\lambda})} + \frac{h(x)}{4} - \frac{3L}{4(1-\lambda)^{3/2}}\right] \\ &= -\lambda^{-3/2}\mathcal{D}(\lambda, x). \end{aligned} \tag{C.48}$$

At fixed λ, differentiating (24) gives $D_y x[q] = \sqrt{\lambda}\Gamma_x q/(1+\sqrt{\lambda})$. Since $\Gamma_x y = (1+\sqrt{\lambda})\mu_x/\sqrt{\lambda}$ and $\langle \mathbf{1}, D_y x[q]\rangle = 0$, we obtain

$$\begin{aligned} D_y V_y'(\lambda)[q] &= \langle x, q\rangle + \langle y, D_y x[q]\rangle - \left(1 + \frac{1}{2\sqrt{\lambda}}\right)\langle \log x, D_y x[q]\rangle \\ &= \langle x, q\rangle + \frac{1}{2(1+\sqrt{\lambda})}\langle \mu_x, q\rangle = \langle \nu_{\lambda,x}, q\rangle. \end{aligned} \tag{C.49}$$

Differentiating $V_y'(\lambda(y)) = 0$ and substituting (C.48), we obtain

$$D\lambda(y)[q] = \frac{\lambda^{3/2}}{\mathcal{D}(\lambda, x)}\langle \nu_{\lambda,x}, q\rangle. \tag{C.50}$$

The chain rule applied to (24) now yields

$$DQ(y)[q] = \frac{\sqrt{\lambda}}{1+\sqrt{\lambda}}\Gamma_x q + \frac{D\lambda(y)[q]}{2\lambda(1+\sqrt{\lambda})}\mu_x. \tag{C.51}$$

Substituting (C.50) into this identity yields

$$\mathcal{J}(y) = \frac{\sqrt{\lambda}}{1+\sqrt{\lambda}}\left[\Gamma_x + \frac{1}{2\mathcal{D}(\lambda, x)}\mu_x \nu_{\lambda,x}^\top\right]. \tag{C.52}$$

For $\|q\|_\infty \le 1$, (C.40) further gives

$$\|\Gamma_x q\|_1 \le \sqrt{\mathrm{Var}_x(q)} \le 1, \qquad \|\mu_x\|_1 \le \sqrt{L}, \qquad \|\nu_{\lambda,x}\|_1 \le 1 + \frac{\sqrt{L}}{2} \le \sqrt{L}. \tag{C.53}$$

Inserting these bounds in (C.42), we obtain

$$\frac{\|\mu_x\|_1 \|\nu_{\lambda,x}\|_1}{2\mathcal{D}(\lambda,x)} \le \frac{L}{L+1} < 1. \tag{C.54}$$

Equation (C.52) now implies

$$\|DQ(y)[q]\|_1 \le 2\frac{\sqrt{\lambda}}{1+\sqrt{\lambda}}\|q\|_\infty. \tag{C.55}$$

The inequality $\sqrt{\lambda}/(1+\sqrt{\lambda}) \le \min\{1/2, \sqrt{\lambda}\}$ proves the Jacobian bounds. For the mass derivative, (C.42), $\|\nu_{\lambda,x}\|_1 \le \sqrt{L}$, and (C.50) give

$$|D\lambda(y)[q]| \le \frac{2\sqrt{L}}{L+1}\lambda^{3/2}\|q\|_\infty \le \lambda^{3/2}\|q\|_\infty. \tag{C.56}$$

To sharpen this to the stated logarithmic derivative bound, put $s = \sqrt{\lambda}$ and retain the stronger variance bound in (C.40). Since $\sqrt{L-2} \le \sqrt{L} - 1/\sqrt{L}$ and $3\sqrt{L}/8 + 1/(2\sqrt{L}) \ge 1$ for $\sqrt{L} \ge 2$, we obtain

$$\|\nu_{\lambda,x}\|_1 \le 1 + \frac{1}{2}\sqrt{L-2} \le \frac{7}{8}\sqrt{L}. \tag{C.57}$$

Moreover, (C.3) and (C.5) give

$$4\mathcal{D}(\lambda,x) \ge L\left[\frac{3}{(1-s^2)^{3/2}} - \frac{1}{1+s}\right]. \tag{C.58}$$

For $0 \le s < 1$, the bracket satisfies

$$\frac{3}{(1-s^2)^{3/2}} - \frac{1}{1+s} \ge 7s. \tag{C.59}$$

Indeed, multiplying the desired inequality by $1-s^2 > 0$ reduces it to

$$\frac{3}{\sqrt{1-s^2}} - (1-s) - 7s(1-s^2) \ge 0. \tag{C.60}$$

The bound $(1-s^2)^{-1/2} \ge 1 + s^2/2$ reduces the claim to $2 - 6s + \frac{3}{2}s^2 + 7s^3 \ge 0$. If $s \le 1/6$, this follows from $2 - 6s \ge 1$ and the nonnegativity of the remaining terms. For $s \ge 1/6$, the same polynomial may instead be written as

$$(s - \tfrac{1}{2})^2(7s + \tfrac{17}{2}) + \frac{3s}{4} - \frac{1}{8} \ge 0. \tag{C.61}$$

For $s > 0$, the preceding inequality implies $\mathcal{D}(\lambda,x) \ge 7Ls/4$. Substituting this bound and the bound on $\nu_{\lambda,x}$ into (C.50), with $\sqrt{L} \ge 2$, we obtain

$$|D\log\lambda(y)[q]| \le \frac{s\|\nu_{\lambda,x}\|_1}{\mathcal{D}(\lambda,x)}\|q\|_\infty \le \frac{1}{2\sqrt{L}}\|q\|_\infty \le \frac{1}{4}\|q\|_\infty. \tag{C.62}$$

■

As an immediate consequence, for $y, y' \in \mathbb{R}^K$ with $Z(y) = \lambda x$ and $Z(y') = \lambda' x'$, integrating (C.62) gives

$$\exp\left(-\frac{1}{4}\|y'-y\|_\infty\right) \le \frac{\lambda'}{\lambda} \le \exp\left(\frac{1}{4}\|y'-y\|_\infty\right). \tag{C.63}$$

The decomposition (C.29) further gives

$$D_\psi(\lambda' x', \lambda x) = \lambda' D_{\mathrm{KL}}(x'\|x) + D_\varphi(\lambda' x', \lambda x). \tag{C.64}$$

The mass comparison (C.63) and the decomposition (C.64) now reduce local coercivity to a lower bound on the mass contribution of φ.

Lemma C.3. *Let $y, y' \in \mathbb{R}^K$. Suppose $Z(y) = \lambda x$, $Z(y') = \lambda' x'$, and $\|y' - y\|_\infty \le M \le 1$. Then*

$$D_\psi(Z(y'), Z(y)) \ge e^{-M/4}\left[\lambda D_{\mathrm{KL}}(x'\|x) + \frac{(\sqrt{\lambda'} - \sqrt{\lambda})^2}{\sqrt{\lambda}}\right], \tag{C.65}$$

and

$$\|x' - x\|_1^2 \le \frac{2e^{M/4}}{\lambda} D_\psi(Z(y'), Z(y)). \tag{C.66}$$

Proof. We begin by specializing (C.63) to the assumed score displacement, obtaining

$$e^{-M/4}\lambda \le \lambda' \le e^{M/4}\lambda. \tag{C.67}$$

Consequently,

$$\lambda' D_{\mathrm{KL}}(x'\|x) \ge e^{-M/4}\lambda D_{\mathrm{KL}}(x'\|x). \tag{C.68}$$

Along the line $z(\theta) = (1-\theta)Z(y) + \theta Z(y')$, (C.67) bounds the mass $\lambda(\theta) = \langle \mathbf{1}, z(\theta)\rangle$ by $e^{M/4}\lambda$. Applying (C.17) under the integral representation of the Bregman divergence therefore yields

$$\begin{aligned} D_\varphi(Z(y'), Z(y)) &= \int_0^1 (1-\theta) D^2\varphi(z(\theta))[Z(y') - Z(y), Z(y') - Z(y)]\ d\theta \\ &\ge \frac{L}{4} e^{-3M/8} \frac{(\lambda' - \lambda)^2}{\lambda^{3/2}} \\ &\ge e^{-M/4} \frac{(\sqrt{\lambda'} - \sqrt{\lambda})^2}{\sqrt{\lambda}}. \end{aligned} \tag{C.69}$$

For the last inequality, (C.67) yields $\sqrt{\lambda'/\lambda} \ge e^{-M/8}$, and $L \ge 4$ together with $e^{-M/8} \ge 1 - M/8 \ge 7/8$ implies

$$\frac{L}{4} e^{-M/8}(1 + e^{-M/8})^2 \ge \frac{7}{8}\left(\frac{15}{8}\right)^2 > 1. \tag{C.70}$$

Folding (C.68) and the bound on the φ divergence back into (C.64) proves (C.65), and applying (C.2) to its KL term yields

$$D_\psi(Z(y'), Z(y)) \ge \frac{1}{2} e^{-M/4}\lambda \|x' - x\|_1^2, \tag{C.71}$$

which is (C.66). ∎

To use (C.65) in the Jacobian factorization (C.52), we next bound the changes in entropy, logarithmic moments, and the covariance matrix in terms of $\left\|\sqrt{x'} - \sqrt{x}\right\|_2$.

Lemma C.4. *Let $x, x' \in \operatorname{ri}\Delta_K$ and $\lambda, \lambda' \in (0,1)$. With*

$$\varepsilon = \left\|\sqrt{x'} - \sqrt{x}\right\|_2, \tag{C.72}$$

one has

$$|h(x') - h(x)| \le 5\sqrt{L}\,\varepsilon, \tag{C.73}$$

$$\|\mu_{x'} - \mu_x\|_1 \le 14\sqrt{L}\,\varepsilon, \tag{C.74}$$

$$|\operatorname{Var}_{x'}(\log x') - \operatorname{Var}_x(\log x)| \le 31 L\,\varepsilon, \tag{C.75}$$

$$\|\Gamma_{x'} - \Gamma_x\|_{\mathrm{op}} \le 3\varepsilon. \tag{C.76}$$

Moreover,

$$\left\|\nu_{\lambda,x}\right\|_1 \le \sqrt{L}, \tag{C.77}$$

$$\left\|\nu_{\lambda',x'} - \nu_{\lambda,x}\right\|_1 \le 8\sqrt{L}\,\varepsilon + \frac{1}{2}\sqrt{L}\,|\sqrt{\lambda'} - \sqrt{\lambda}|. \tag{C.78}$$

Proof. For $0 \le \theta \le 1$, consider the segment

$$a(\theta) = (1-\theta)\sqrt{x} + \theta\sqrt{x'}, \qquad p_\alpha(\theta) = a_\alpha(\theta)^2. \tag{C.79}$$

The bound $\|a(\theta)\|_2 \le (1-\theta)\|\sqrt{x}\|_2 + \theta\|\sqrt{x'}\|_2 = 1$ shows that the weights $p_\alpha(\theta)$ have total mass at most one, so the moment bounds (C.6)–(C.7) apply along the segment. Setting $f_1(s) = s^2 \log s^2$ and $f_2(s) = s^2(\log s^2)^2$, with values and derivatives at zero defined by continuity, we have

$$|f_1'(s)|^2 = 4s^2(1+\log s^2)^2, \qquad |f_2'(s)|^2 = 4s^2(\log s^2)^2(2+\log s^2)^2. \tag{C.80}$$

The cross terms involving odd powers of $\log p_\alpha(\theta)$ are nonpositive. Consequently, (C.6)–(C.7) give

$$\begin{aligned}
\sum_\alpha |f_1'(a_\alpha(\theta))|^2 &\le 4 + 4\sum_\alpha p_\alpha(\theta)|\log p_\alpha(\theta)|^2 \le 16L + 4, \\
\sum_\alpha |f_2'(a_\alpha(\theta))|^2 &\le 4\sum_\alpha p_\alpha(\theta)|\log p_\alpha(\theta)|^4 + 16\sum_\alpha p_\alpha(\theta)|\log p_\alpha(\theta)|^2 \\
&\le 104L^2 + 64L.
\end{aligned} \tag{C.81}$$

For $L \ge 4$, the two right-hand sides are smaller than $25L$ and $121L^2$, respectively. The fundamental theorem of calculus and Cauchy–Schwarz then yield, for $j = 1, 2$,

$$\sum_\alpha |f_j(a_\alpha(1)) - f_j(a_\alpha(0))| \le \varepsilon \int_0^1 \left(\sum_\alpha |f_j'(a_\alpha(\theta))|^2\right)^{1/2} d\theta. \tag{C.82}$$

Substituting the two derivative bounds into this bound yields

$$\begin{aligned}
\sum_\alpha |x'_\alpha \log x'_\alpha - x_\alpha \log x_\alpha| &\le 5\sqrt{L}\,\varepsilon, \\
\sum_\alpha |x'_\alpha (\log x'_\alpha)^2 - x_\alpha (\log x_\alpha)^2| &\le 11L\,\varepsilon.
\end{aligned} \tag{C.83}$$

We also have

$$\|x' - x\|_1 \le \left\|\sqrt{x'} - \sqrt{x}\right\|_2 \left\|\sqrt{x'} + \sqrt{x}\right\|_2 \le 2\varepsilon. \tag{C.84}$$

Equation (C.73) follows by summing the coordinate differences in the first line of (C.83). Substituting this bound and (C.84) into $\mu_x = x \odot \log x - h(x)x$, and recalling $|h(x)| \le \log K \le 2\sqrt{L}$, we obtain

$$\begin{aligned}
\|\mu_{x'} - \mu_x\|_1 &\le 5\sqrt{L}\,\varepsilon + |h(x') - h(x)| + |h(x)|\|x' - x\|_1 \\
&\le 14\sqrt{L}\,\varepsilon.
\end{aligned} \tag{C.85}$$

The preceding display proves (C.74), and subtracting the second moments and the squared means yields the corresponding variance bound:

$$\begin{aligned}
|\operatorname{Var}_{x'}(\log x') - \operatorname{Var}_x(\log x)| &\le 11L\,\varepsilon + |h(x') - h(x)|\big(|h(x')| + |h(x)|\big) \\
&\le 31L\,\varepsilon,
\end{aligned} \tag{C.86}$$

which proves (C.75). For the covariance matrices, we change one factor at a time in the factorization

$$\Gamma_x = \operatorname{Diag}(\sqrt{x})\big(\mathrm{I} - \sqrt{x}(\sqrt{x})^\top\big)\operatorname{Diag}(\sqrt{x}) \tag{C.87}$$

to obtain

$$\begin{aligned}
\Gamma_{x'} - \Gamma_x = {} & \operatorname{Diag}(\sqrt{x'} - \sqrt{x})\big(\mathrm{I} - \sqrt{x'}(\sqrt{x'})^\top\big)\operatorname{Diag}(\sqrt{x'}) \\
& + \operatorname{Diag}(\sqrt{x})\big(\sqrt{x}(\sqrt{x})^\top - \sqrt{x'}(\sqrt{x'})^\top\big)\operatorname{Diag}(\sqrt{x'}) \\
& + \operatorname{Diag}(\sqrt{x})\big(\mathrm{I} - \sqrt{x}(\sqrt{x})^\top\big)\operatorname{Diag}(\sqrt{x'} - \sqrt{x}).
\end{aligned} \tag{C.88}$$

To bound the three terms in the preceding display, for vectors c, d and a matrix A we use

$$\|\operatorname{Diag}(c)A\operatorname{Diag}(d)\|_{\mathrm{op}} \le \|c\|_2\|A\|_2\|d\|_2. \tag{C.89}$$

where $\|A\|_2$ is for the L_2 operator norm. Indeed, for $\|q\|_\infty \le 1$, Cauchy–Schwarz gives

$$\|\operatorname{Diag}(c)A\operatorname{Diag}(d)q\|_1 \le \|c\|_2\|A\operatorname{Diag}(d)q\|_2 \le \|c\|_2\|A\|_2\|d\|_2. \tag{C.90}$$

Applied to the first and third terms, this inequality bounds each by ε, because $\left\|\sqrt{x}\right\|_2 = \left\|\sqrt{x'}\right\|_2 = 1$. For the middle term, put $u = \sqrt{x}$, $v = \sqrt{x'}$, and $c = \langle u, v\rangle \in [0,1]$. If $c = 1$, the identity $\|u - v\|_2^2 = 2 - 2c = 0$ makes the two projections equal. If $c < 1$, define instead

$$a = \frac{u+v}{\sqrt{2+2c}}, \qquad b = \frac{u-v}{\sqrt{2-2c}}. \tag{C.91}$$

The identities $\|u+v\|_2^2 = 2+2c$, $\|u-v\|_2^2 = 2-2c$, and $\langle u+v, u-v\rangle = 0$ show that a, b are orthonormal, and expanding the two outer products gives

$$uu^\top - vv^\top = \sqrt{1-c^2}(ab^\top + ba^\top). \tag{C.92}$$

Consequently, for every vector q,

$$\left\|(uu^\top - vv^\top)q\right\|_2^2 = (1-c^2)\left(\langle a, q\rangle^2 + \langle b, q\rangle^2\right) \le (1-c^2)\|q\|_2^2, \tag{C.93}$$

with equality at $q = a$. Both cases consequently satisfy

$$\begin{aligned} \|\sqrt{x}(\sqrt{x})^\top - \sqrt{x'}(\sqrt{x'})^\top\|_{\mathrm{op}}^2 &= 1 - \left\langle \sqrt{x}, \sqrt{x'}\right\rangle^2 \\ &\le 2\left(1 - \left\langle \sqrt{x}, \sqrt{x'}\right\rangle\right) \\ &= \left\|\sqrt{x'} - \sqrt{x}\right\|_2^2 = \varepsilon^2. \end{aligned} \tag{C.94}$$

The middle term is therefore also at most ε, and the corresponding bounds for the other two terms prove (C.76). Finally, (C.40) and $L \ge 4$ imply

$$\left\|\nu_{\lambda,x}\right\|_1 \le 1 + \frac{1}{2}\|\mu_x\|_1 \le 1 + \frac{1}{2}\sqrt{L-2} \le \sqrt{L}, \tag{C.95}$$

which proves (C.77). Subtracting the two expressions in (C.43) and invoking (C.84), (C.74), and $|(1+\sqrt{\lambda'})^{-1} - (1+\sqrt{\lambda})^{-1}| \le |\sqrt{\lambda'} - \sqrt{\lambda}|$, we further obtain

$$\begin{aligned} \left\|\nu_{\lambda',x'} - \nu_{\lambda,x}\right\|_1 &\le 2\varepsilon + 7\sqrt{L}\,\varepsilon + \frac{1}{2}\sqrt{L}\,|\sqrt{\lambda'} - \sqrt{\lambda}| \\ &\le 8\sqrt{L}\,\varepsilon + \frac{1}{2}\sqrt{L}\,|\sqrt{\lambda'} - \sqrt{\lambda}|, \end{aligned} \tag{C.96}$$

where the last inequality uses $\sqrt{L} \ge 2$, and this proves (C.78). ∎

The factorization (C.52), together with Lemma C.4, expresses the Jacobian variation in the two quantities controlled by Lemma C.3.

Proposition C.3. *Let $y, y' \in \mathbb{R}^K$ and suppose $\|y' - y\|_\infty \le M \le 1$. Then*

$$\|\mathcal{J}(y') - \mathcal{J}(y)\|_{\mathrm{op}}^2 \le 2000 e^{M/4} D_\psi(Z(y'), Z(y)). \tag{C.97}$$

Moreover, for $y, v \in \mathbb{R}^K$ with $\|v\|_\infty \le M \le 1$,

$$\|Q(y+v) - Q(y) - \mathcal{J}(y)v\|_1^2 \le 2000 e^{M/4}\|v\|_\infty^2 D_\psi(Z(y+v), Z(y)). \tag{C.98}$$

Proof. We begin from the factorization (C.52), writing $Z(y) = \lambda x$, $Z(y') = \lambda' x'$, and $\varepsilon = \left\|\sqrt{x'} - \sqrt{x}\right\|_2$. To separate its scalar coefficient from the remaining matrix, set

$$B(\lambda, x) = \Gamma_x + \frac{\mu_x \nu_{\lambda,x}^\top}{2\mathcal{D}(\lambda, x)}. \tag{C.99}$$

The bounds (C.40), (C.77), and (C.42) give

$$\|\Gamma_x\|_{\mathrm{op}} \le 1, \qquad \frac{\|\mu_x\|_1 \left\|\nu_{\lambda,x}\right\|_1}{2\mathcal{D}(\lambda, x)} \le \frac{L}{L+1} < 1. \tag{C.100}$$

Hence $\|B(\lambda,x)\|_{\mathrm{op}} \le 2$. Changing the scalar coefficient and the matrix one at a time, we use

$$\left|\frac{\sqrt{\lambda'}}{1+\sqrt{\lambda'}} - \frac{\sqrt{\lambda}}{1+\sqrt{\lambda}}\right| \le |\sqrt{\lambda'} - \sqrt{\lambda}|, \tag{C.101}$$

and (C.52) then yields

$$\|\mathcal{J}(y') - \mathcal{J}(y)\|_{\mathrm{op}} \le \sqrt{\lambda}\,\|B(\lambda',x') - B(\lambda,x)\|_{\mathrm{op}} + 2|\sqrt{\lambda'} - \sqrt{\lambda}|. \tag{C.102}$$

Indeed, this follows by writing the difference as $\frac{\sqrt{\lambda}}{1+\sqrt{\lambda}}(B(\lambda',x') - B(\lambda,x)) + (\frac{\sqrt{\lambda'}}{1+\sqrt{\lambda'}} - \frac{\sqrt{\lambda}}{1+\sqrt{\lambda}})B(\lambda',x')$ and using $\sqrt{\lambda}/(1+\sqrt{\lambda}) \le \sqrt{\lambda}$. To bound the matrix difference in (C.102), first hold λ fixed and apply (C.73) and (C.75) to obtain

$$|\mathcal{D}(\lambda,x') - \mathcal{D}(\lambda,x)| \le \frac{1}{4}(5\sqrt{L} + 31L)\varepsilon. \tag{C.103}$$

Dividing by $\mathcal{D}(\lambda,x')\mathcal{D}(\lambda,x)$ and invoking (C.42) at both strategies yields

$$|\mathcal{D}(\lambda,x')^{-1} - \mathcal{D}(\lambda,x)^{-1}| \le \frac{35}{L}\varepsilon. \tag{C.104}$$

Indeed, the stated constant follows because $\sqrt{L} \le L/2$ and (C.42) give

$$\frac{5\sqrt{L} + 31L}{(L+1)^2} < \frac{35}{L}. \tag{C.105}$$

For fixed x, on the other hand, the function $s \mapsto \mathcal{D}(s^2,x)$ satisfies

$$\frac{d}{ds}\mathcal{D}(s^2,x) = \frac{9Ls}{4(1-s^2)^{5/2}} + \frac{\mathrm{Var}_x(\log x)}{4(1+s)^2}. \tag{C.106}$$

For $0 < s < 1$, (C.42) and $\mathrm{Var}_x(\log x) \le L$ imply

$$\begin{aligned}\left|\frac{d}{ds}\mathcal{D}(s^2,x)^{-1}\right| &\le \frac{9}{L}s\sqrt{1-s^2} + \frac{1}{L}\frac{(1-s^2)^3}{(1+s)^2} \\ &\le \frac{11}{2L} < \frac{6}{L}.\end{aligned} \tag{C.107}$$

The mean value theorem, together with $s\sqrt{1-s^2} \le 1/2$, gives

$$|\mathcal{D}(\lambda',x)^{-1} - \mathcal{D}(\lambda,x)^{-1}| \le \frac{6}{L}|\sqrt{\lambda'} - \sqrt{\lambda}|. \tag{C.108}$$

Adding (C.104) and (C.108), we find

$$|\mathcal{D}(\lambda',x')^{-1} - \mathcal{D}(\lambda,x)^{-1}| \le \frac{35}{L}\varepsilon + \frac{6}{L}|\sqrt{\lambda'} - \sqrt{\lambda}|. \tag{C.109}$$

To separate the three changes in the outer product, abbreviate $\mu = \mu_x$, $\nu = \nu_{\lambda,x}$, and $d = \mathcal{D}(\lambda,x)$, with primed symbols denoting evaluation at (λ',x'). We then have

$$\frac{\mu'\nu'^{\top}}{2d'} - \frac{\mu\nu^{\top}}{2d} = \frac{(\mu'-\mu)\nu'^{\top}}{2d'} + \frac{\mu(\nu'-\nu)^{\top}}{2d'} + \frac{\mu\nu^{\top}}{2}\left(\frac{1}{d'} - \frac{1}{d}\right). \tag{C.110}$$

Applying $\|ab^{\top}\|_{\mathrm{op}} = \|a\|_1\|b\|_1$ to these three terms and substituting the preceding reciprocal bound, (C.76), (C.74), (C.77), (C.78), and $2\mathcal{D} \ge L+1$, we obtain

$$\begin{aligned}\|B(\lambda',x') - B(\lambda,x)\|_{\mathrm{op}} &\le 3\varepsilon + 14\varepsilon + (8\varepsilon + \tfrac{1}{2}|\sqrt{\lambda'} - \sqrt{\lambda}|) + (\tfrac{35}{2}\varepsilon + 3|\sqrt{\lambda'} - \sqrt{\lambda}|) \\ &\le 43\varepsilon + 4|\sqrt{\lambda'} - \sqrt{\lambda}|.\end{aligned} \tag{C.111}$$

Substituting this bound into (C.102) and using $\sqrt{\lambda} \le 1$ yields

$$\|\mathcal{J}(y') - \mathcal{J}(y)\|_{\mathrm{op}} \le 43\sqrt{\lambda}\varepsilon + 6|\sqrt{\lambda'} - \sqrt{\lambda}|. \tag{C.112}$$

Cauchy–Schwarz, (C.2), and $0 < \sqrt{\lambda} \le 1$ now imply

$$\|\mathcal{J}(y') - \mathcal{J}(y)\|_{\mathrm{op}}^2 \le (43^2 + 6^2)(\lambda\varepsilon^2 + |\sqrt{\lambda'} - \sqrt{\lambda}|^2)$$
$$\le 1900 \left[\lambda D_{\mathrm{KL}}(x'\|x) + \frac{|\sqrt{\lambda'} - \sqrt{\lambda}|^2}{\sqrt{\lambda}}\right]. \tag{C.113}$$

After enlarging 1900 to 2000, (C.65) turns the last display into (C.97). To obtain the Taylor bound, integrate this Jacobian difference starting from

$$Q(y+v) - Q(y) - \mathcal{J}(y)v = \int_0^1 [\mathcal{J}(y+\theta v) - \mathcal{J}(y)]v \; d\theta. \tag{C.114}$$

Along the same segment,

$$\frac{d}{d\theta} D_\psi(Z(y+\theta v), Z(y)) = \langle \theta v, DZ(y+\theta v)[v] \rangle \ge 0, \tag{C.115}$$

because $DZ(y+\theta v) = [D^2\psi(Z(y+\theta v))]^{-1}$ is positive definite by (C.31), so the divergence along the segment is at most its value at $\theta = 1$. The triangle inequality, Cauchy–Schwarz, and (C.97) now give

$$\|Q(y+v) - Q(y) - \mathcal{J}(y)v\|_1^2 \tag{C.116}$$
$$\le \|v\|_\infty^2 \int_0^1 \|\mathcal{J}(y+\theta v) - \mathcal{J}(y)\|_{\mathrm{op}}^2 \; d\theta \tag{C.117}$$
$$\le 2000 e^{M/4} \|v\|_\infty^2 D_\psi(Z(y+v), Z(y)), \tag{C.118}$$

which proves (C.98).

■

D Higher-order predictor and one step expansions

In this appendix, we establish the finite difference and payoff difference bounds that will allow us to relate the higher-order prediction error to changes in the players' strategies.

D.1 Discounted finite differences

We begin with the discount factor and operators of Section 4, including the zero extension specified there. These operators act on the round index of the optimistic score and lifted iterate sequences, as in $\Delta w_{t+3/2} = w_{t+3/2} - w_{t+1/2}$, whereas, when time indices are suppressed, x_i denotes the sequence $(x_{i,t})_{t\in\mathbb{Z}}$. The series defining $\mathrm{S}_\delta w$ converges absolutely for bounded w and is finite at each fixed time whenever w vanishes at nonpositive indices, and directly from its definition,

$$(\mathrm{S}_\delta w)_t - \delta(\mathrm{S}_\delta w)_{t-1} = w_t. \tag{D.1}$$

Conversely, if $w_t = \mathbf{0}$ for $t \le 0$, the following sum is finite at every fixed t and telescopes:

$$\sum_{k=0}^{\infty} \delta^k (w_{t-k} - \delta w_{t-k-1}) = w_t. \tag{D.2}$$

Hence S_δ inverts $w \mapsto (w_t - \delta w_{t-1})_t$ on sequences vanishing at nonpositive indices, and absolute convergence extends the same telescoping calculation to bounded sequences. We also have

$$\mathrm{S}_\delta \Delta = \Delta \mathrm{S}_\delta, \tag{D.3}$$

because both sides at time t equal $\sum_{k\ge 0} \delta^k (w_{t-k} - w_{t-k-1})$. With these identities in place, the prediction error formula follows directly from the recurrence.

Proof of (27). For $t \ge 1$, Algorithm 1 and (12) give

$$e_{i,t} = \sum_{k=0}^{N+1} (-1)^k \binom{N+1}{k} g_{i,t-k} - \sum_{k=1}^{N+1} (-1)^k \binom{N+1}{k} \delta^k e_{i,t-k}. \tag{D.4}$$

Since both sides vanish for $t \le 0$, rearranging the identity gives, for every $t \in \mathbb{Z}$,

$$\sum_{k=0}^{N+1} (-1)^k \binom{N+1}{k} \delta^k e_{i,t-k} = \Delta^{N+1} g_{i,t}. \tag{D.5}$$

The left-hand side of (D.5) applies $w \mapsto (w_t - \delta w_{t-1})_t$ successively $N+1$ times to e_i. Because S_δ is the inverse, applying it the same number of times proves (27). ∎

Motivated by this identity, define, for $0 \le p \le N+1$,

$$\mathrm{D}_p = {\mathrm{S}_\delta}^{N+1} \Delta^{N+1-p}. \tag{D.6}$$

Commutation of S_δ and Δ gives, for $0 \le p \le N$,

$$\mathrm{D}_p = \mathrm{D}_{p+1}\Delta. \tag{D.7}$$

To bound discounted differences of payoffs, strategies, and Jacobians, we use, respectively, the ℓ_∞, ℓ_1, and ℓ_∞-to-ℓ_1 operator norms. In this subsection, $\|w_t\|$ denotes the applicable norm, with absolute value used for scalars, and

$$\|w\|_{\ell_2} = \left(\sum_{t \in \mathbb{Z}} \|w_t\|^2 \right)^{1/2}. \tag{D.8}$$

The following proposition collects the discounted finite difference bounds needed later to control the initialization terms and the sequences produced by the expansion.

Proposition D.1. *For every $1 \le p \le N+1$ and every sequence of these vectors or matrices with $\|w\|_{\ell_2} < \infty$,*

$$\left\|\mathrm{D}_p w\right\|_{\ell_2} \le 4(N+1)^{p+1} \|w\|_{\ell_2}. \tag{D.9}$$

For any integer t_0 and vector or matrix v, if $w_t = \mathbb{1}_{\{t=t_0\}} v$, then

$$\left\|\mathrm{D}_p w\right\|_{\ell_2} \le 2(N+1)^p \|v\|. \tag{D.10}$$

For every bounded sequence w,

$$\sup_t \left\|((\mathrm{S}_\delta \Delta)^{N+1} w)_t\right\| \le 4(N+1) \sup_t \|w_t\|. \tag{D.11}$$

For sequences vanishing at nonpositive indices, the same squared sum bounds hold with both sums restricted to $1 \le t \le T$.

Proof. For every finitely supported scalar sequence a, expanding the squares gives

$$\begin{aligned} &4 \sum_t (a_t - \delta a_{t-1})^2 - (1+\delta)^2 \|\Delta a\|_{\ell_2}^2 \\ &\qquad = 2(1-\delta)^2 \sum_t (a_t^2 + a_t a_{t-1}) \ge 0, \end{aligned} \tag{D.12}$$

where the last inequality follows by summing $2a_t a_{t-1} \ge -a_t^2 - a_{t-1}^2$. Since the finite difference operators in (D.12) are bounded on scalar ℓ_2, this inequality extends by density to every scalar ℓ_2 sequence. Minkowski's inequality also yields

$$\|\mathrm{S}_\delta w\|_{\ell_2} \le \sum_{k \ge 0} \delta^k \|(w_{t-k})_t\|_{\ell_2} = (N+1)\|w\|_{\ell_2}, \tag{D.13}$$

and, since

$$(\mathrm{S}_\delta w)_t - \delta(\mathrm{S}_\delta w)_{t-1} = w_t, \qquad \Delta \mathrm{S}_\delta w = \mathrm{S}_\delta \Delta w, \tag{D.14}$$

applying (D.12) to $\mathrm{S}_\delta w$, for a scalar sequence w, yields

$$\|(\mathrm{S}_\delta \Delta) w\|_{\ell_2} \le \frac{2}{1+\delta} \|w\|_{\ell_2}. \tag{D.15}$$

To iterate this scalar bound over the required orders, set

$$\beta = \frac{2}{1+\delta} = \left(1 - \frac{1}{2(N+1)}\right)^{-1}. \tag{D.16}$$

Iterating (D.15), for $0 \le q \le N+1$ we obtain

$$\|(\mathrm{S}_\delta \Delta)^q w\|_{\ell_2} \le \beta^q \|w\|_{\ell_2}. \tag{D.17}$$

The strict form of Bernoulli's inequality, applied with $N+1 \ge 2$, gives

$$\beta^{-(N+1)} = \left(1 - \frac{1}{2(N+1)}\right)^{N+1} > 1 - \frac{N+1}{2(N+1)} = \frac{1}{2}. \tag{D.18}$$

It follows that $\beta^q \le \beta^{N+1} < 2$ and, for $0 \le q \le N+1$,

$$\|(\mathrm{S}_\delta \Delta)^q w\|_{\ell_2} \le 2\|w\|_{\ell_2}. \tag{D.19}$$

To pass from the scalar bound to the payoff, strategy, and Jacobian norms, we need the absolute sum of the convolution coefficients. By the definition of S_δ,

$$((\mathrm{S}_\delta \Delta)v)_t = \sum_{k \ge 0} b_k v_{t-k}, \tag{D.20}$$

where $b_0 = 1$ and, for $k \ge 1$,

$$b_k = -(1-\delta)\delta^{k-1}. \tag{D.21}$$

For $k < 0$, set $b_k = 0$. Starting from $a_0^{(0)} = 1$ and $a_k^{(0)} = 0$ for $k \ne 0$, define inductively, for $q \ge 0$ and $k \ge 0$,

$$a_k^{(q+1)} = \sum_{\ell=0}^{k} b_\ell a_{k-\ell}^{(q)}, \tag{D.22}$$

and set $a_k^{(q)} = 0$ for $k < 0$. Since $\sum_{k \ge 0} |b_k| = 2$, induction shows that $a^{(q)}$ is absolutely summable and that

$$((\mathrm{S}_\delta \Delta)^q v)_t = \sum_{k \ge 0} a_k^{(q)} v_{t-k}. \tag{D.23}$$

We can now bound this kernel by applying (D.15) repeatedly to the scalar sequence supported at time zero with value one, obtaining

$$\left(\sum_{k \ge 0} |a_k^{(q)}|^2\right)^{1/2} \le \beta^q. \tag{D.24}$$

The squared sum alone does not bound the absolute sum of the coefficients, so we also bound their weighted squared sum. To this end, set

$$s_q = \left(\sum_{k \ge 0} k^2 |a_k^{(q)}|^2\right)^{1/2}. \tag{D.25}$$

Multiplying (D.22) by k and writing $k = (k-\ell) + \ell$ gives

$$k a_k^{(q+1)} = \sum_{\ell=0}^{k} b_\ell (k-\ell) a_{k-\ell}^{(q)} + \sum_{\ell=0}^{k} \ell b_\ell a_{k-\ell}^{(q)}, \tag{D.26}$$

so (D.15) applied to the first sum and Minkowski's inequality applied to the second yield

$$\begin{aligned} s_{q+1} &\le \beta s_q + \left(\sum_{\ell \ge 0} \ell |b_\ell|\right)\left(\sum_{k \ge 0} |a_k^{(q)}|^2\right)^{1/2} \\ &\le \beta s_q + (N+1)\beta^q, \end{aligned} \tag{D.27}$$

where we used

$$\sum_{\ell \ge 0} \ell |b_\ell| = (1-\delta) \sum_{\ell \ge 1} \ell \delta^{\ell-1} = \frac{1}{1-\delta} = N+1. \tag{D.28}$$

Starting from $s_0 = 0$, the preceding recurrence proves finiteness by induction and yields, for $1 \le q \le N+1$,

$$s_q \le q(N+1)\beta^{q-1} < 2q(N+1). \tag{D.29}$$

Equations (D.24) and (D.29), together with $\beta^q < 2$, imply

$$\sum_{k \ge 0} |a_k^{(q)}|^2 < 4, \qquad \sum_{k \ge 0} k^2 |a_k^{(q)}|^2 < 4q^2(N+1)^2. \tag{D.30}$$

For $1 \le q \le N+1$, Cauchy–Schwarz then gives

$$\left(\sum_{k\ge 0} |a_k^{(q)}|\right)^2 \le \left(\sum_{k\ge 0} \frac{1}{1+(k/[q(N+1)])^2}\right) \times \left(\sum_{k\ge 0} |a_k^{(q)}|^2 \left(1+\frac{k^2}{q^2(N+1)^2}\right)\right). \tag{D.31}$$

For the first factor, comparison with an integral gives

$$\sum_{k\ge 0} \frac{1}{1+(k/[q(N+1)])^2} \le 1 + \int_0^\infty \frac{dx}{1+(x/[q(N+1)])^2} = 1 + \frac{\pi q(N+1)}{2} \le 2(N+1)^2, \tag{D.32}$$

where the last inequality uses $\pi < 7/2$, $q \le N+1$, and $N+1 \ge 2$. Equations (D.24) and (D.29) bound the remaining factor by 8. Hence, for $1 \le q \le N+1$,

$$\sum_{k\ge 0} |a_k^{(q)}| < 4(N+1). \tag{D.33}$$

The definition of $a^{(0)}$ yields the same bound when $q = 0$. Consequently, for each of the three norms specified in (D.8), Minkowski's inequality and (D.33) imply

$$\|(\mathrm{S}_\delta \Delta)^q w\|_{\ell_2} \le \sum_{k\ge 0} |a_k^{(q)}| \|(w_{t-k})_t\|_{\ell_2} \le 4(N+1)\|w\|_{\ell_2}. \tag{D.34}$$

Likewise, repeated application of the corresponding bound for S_δ gives

$$\|{\mathrm{S}_\delta}^p w\|_{\ell_2} \le (N+1)^p \|w\|_{\ell_2}. \tag{D.35}$$

The factorization

$$\mathrm{D}_p = {\mathrm{S}_\delta}^p (\mathrm{S}_\delta \Delta)^{N+1-p}, \tag{D.36}$$

and the preceding two bounds imply

$$\left\|\mathrm{D}_p w\right\|_{\ell_2} \le 4(N+1)^{p+1} \|w\|_{\ell_2}, \tag{D.37}$$

which proves (D.9). When w is supported at time t_0 with value v, the more precise formula

$$((\mathrm{S}_\delta \Delta)^{N+1-p} w)_t = a_{t-t_0}^{(N+1-p)} v, \tag{D.38}$$

together with (D.24) and $\beta^{N+1-p} < 2$, gives

$$\left\|(\mathrm{S}_\delta \Delta)^{N+1-p} w\right\|_{\ell_2} \le 2\|v\|. \tag{D.39}$$

The bound for ${\mathrm{S}_\delta}^p$ now proves (D.10), and the triangle inequality in (D.23), followed by (D.33) with $q = N+1$, gives (D.11). For the finite horizon bounds, replace w_t by zero after T and apply the preceding bounds to this truncated input. Since every output depends only on present and earlier inputs, its values up to T remain unchanged, which proves the final assertion. ∎

Specializing (D.11) to (27) and using $\left\|g_j\right\|_\infty \le 2$, we obtain

$$\left\|e_{i,t}\right\|_\infty \le 8(N+1). \tag{D.40}$$

Recalling that $\hat{g}_{i,t} = g_{i,t} - e_{i,t}$ and $N+1 \ge 2$, this implies

$$\left\|\hat{g}_{i,t}\right\|_\infty \le 9(N+1). \tag{D.41}$$

These bounds keep the score displacements considered below within the local regime of Lemma C.3 and Proposition C.3, and we now turn to the payoff and strategy differences that enter the expansion.

D.2 Payoff and strategy differences

We begin by extending the base and optimistic scores by zero before the first round, and extend their lifted iterates and strategies by $Z(\mathbf{0})$ and $Q(\mathbf{0})$, respectively. For $t \le 0$, we also set $P_{i,t} = I_{i,t} = 0$, whereas the payoff excess, predictor, and prediction error retain the zero extension of Section 4. For every player i and integer t, write

$$\lambda_{i,t} = \lambda(y_{i,t}), \qquad J_{i,t} = \eta \mathcal{J}(y_{i,t}). \tag{D.42}$$

With this notation, the update (LiftFTRL+) and the definition (12) give

$$\Delta y_{i,t+1/2} = \eta(g_{i,t} - \Delta e_{i,t}). \tag{D.43}$$

We therefore define $r_{i,t}$ by

$$\Delta x_{i,t} = J_{i,t} \frac{\Delta y_{i,t+1/2}}{\eta} + r_{i,t}. \tag{D.44}$$

For $t \le 0$, both differences vanish, so $r_{i,t} = \mathbf{0}$, and substituting (D.43) into (D.44) yields

$$\Delta x_{i,t} = r_{i,t} + J_{i,t} g_{i,t} - J_{i,t} \Delta e_{i,t}. \tag{D.45}$$

We will also repeatedly use the elementary bound

$$\left\|\Delta e_{i,t}\right\|_\infty^2 \le 2\left\|e_{i,t}\right\|_\infty^2 + 2\left\|e_{i,t-1}\right\|_\infty^2. \tag{D.46}$$

To expand payoff differences without repeatedly separating the initialization round, we introduce an extension of the payoff excess sequence. Namely, for every player i and integer $t \in \mathbb{Z}$, define

$$\gamma_{i,t} = v_i(x_t) - u_i(x_t)\mathbf{1}, \tag{D.47}$$

Thus, on the actual rounds $t \ge 1$, this is exactly the payoff excess, $\gamma_{i,t} = g_{i,t}$, while for $t \le 0$ the two sequences differ only because $g_{i,t}$ is set to zero by convention whereas $\gamma_{i,t}$ retains the payoff excess induced by the extended strategy profile, hence we have, for every $t \in \mathbb{Z}$,

$$\Delta g_{i,t} = \mathbb{1}_{\{t=1\}} \gamma_{i,0} + \Delta \gamma_{i,t}. \tag{D.48}$$

The payoff range $[-1, 1]$ then gives also

$$\left\|\gamma_{i,t}\right\|_\infty \le 2. \tag{D.49}$$

Next, we state the basic product rule that we shall use in all that follows: if F is q-linear, then

$$\begin{aligned} &F(z_{1,t}, \dots, z_{q,t}) - F(z_{1,t-1}, \dots, z_{q,t-1}) \\ &\quad = \sum_{k=1}^{q} F(z_{1,t-1}, \dots, z_{k-1,t-1}, \Delta z_{k,t}, z_{k+1,t}, \dots, z_{q,t}). \end{aligned} \tag{D.50}$$

The next proposition applies this product rule to the payoff excess to obtain a decomposition in term of strategy differences.

Proposition D.2. *For every player i and every $t \in \mathbb{Z}$, there are linear maps $f^v_{ij,t}$ ($j \ne i$) and $f^u_{ij,t}$ ($j \in \mathcal{N}$) such that*

$$\Delta \gamma_{i,t} = \sum_{j \ne i} f^v_{ij,t} \Delta x_{j,t} + \sum_{j=1}^{N} f^u_{ij,t} \Delta x_{j,t}, \tag{D.51}$$

with

$$\left\|f^v_{ij,t} w\right\|_\infty \le \|w\|_1, \qquad \left\|f^u_{ij,t} w\right\|_\infty \le \|w\|_1. \tag{D.52}$$

Each coefficient contains at most $N-1$ strategy terms, evaluated at round t or $t-1$, with each player label appearing at most once. Each fixed player labels a differenced strategy at most twice in (D.51).

Proof. Applying (D.50) to the maps $x_{-i} \mapsto u_i(\cdot, x_{-i})$ and $x \mapsto -u_i(x)\mathbf{1}$ proves (D.51). More precisely, $f^v_{ij,t}$ leaves player j's argument free in the first map and fixes the remaining arguments at the strategies prescribed by the product rule, and $f^u_{ij,t}$ is defined analogously for the second map. Hence each coefficient contains at most $N-1$ strategy terms, with every label occurring at most once, and each differenced player can occur once in each of the two sums. For the norm bounds, fix an

integer $q \geq 1$, an array $(a_\beta)_{\beta \in \mathcal{A}^q}$ with $|a_\beta| \leq 1$, and vectors $z_1, \ldots, z_q \in \mathbb{R}^K$. The triangle inequality yields

$$\left| \sum_{\beta \in \mathcal{A}^q} a_\beta \prod_{k=1}^{q} z_{k,\beta_k} \right| \leq \prod_{k=1}^{q} \|z_k\|_1. \tag{D.53}$$

Taking one argument equal to w and the others in Δ_K proves (D.52). More generally, if $f[z_1, \ldots, z_q]$ denotes either coefficient map with its q strategy arguments replaced by arbitrary vectors, the same calculation with w in the free argument gives

$$\left\| f[z_1, \ldots, z_q]w \right\|_\infty \leq \|w\|_1 \prod_{a=1}^{q} \|z_a\|_1. \tag{D.54}$$

Here the empty product is one, and the same coordinatewise calculation covers the minus sign and factor $\mathbf{1}$ in the second coefficient map. Thus, (D.54) also applies when subsequent differences replace some of the fixed strategy arguments by arbitrary vectors. ∎

The payoff decomposition identifies the strategy differences that enter the expansion, and we next establish the pointwise bounds that control them together with the accompanying Jacobian and remainder terms.

Proposition D.3. *Assume (28). For every player i and round $t \geq 1$,*

$$\left\| \Delta x_{i,t} \right\|_1^2 \leq \frac{5}{\lambda_{i,t}} (P_{i,t} + I_{i,t-1}), \tag{D.55}$$

$$\left\| r_{i,t} \right\|_1^2 \leq 300000 (N+1)^2 \eta^2 (P_{i,t} + I_{i,t-1}), \tag{D.56}$$

$$\|J_{i,t+1} - J_{i,t}\|_{\mathrm{op}}^2 \leq 4100 \eta^2 (P_{i,t} + I_{i,t}). \tag{D.57}$$

Moreover, for every integer t' with $|t - t'| \leq N + 1$,

$$\|J_{i,t'}\|_{\mathrm{op}}^2 \left\| J_{i,t} \frac{\Delta y_{i,t+1/2}}{\eta} \right\|_1^2 \leq 90 \eta^2 (P_{i,t} + I_{i,t-1}). \tag{D.58}$$

Proof. For every integer t, (D.42) and Proposition C.2 yield the two bounds $\|J_{i,t}\|_{\mathrm{op}} \leq \eta$ and $\|J_{i,t}\|_{\mathrm{op}} \leq 2\eta\sqrt{\lambda_{i,t}}$, and (LiftFTRL+) together with (12) implies

$$y_{i,t+1/2} = y_{i,t} + \eta \hat{g}_{i,t} = y_{i,t+1} - \eta e_{i,t}. \tag{D.59}$$

Taking differences recovers (D.43), and, more specifically, for a fixed $t \geq 0$, (D.59) places the two consecutive strategies about the same base score:

$$x_{i,t+1} = Q\left(y_{i,t+1} + \eta \hat{g}_{i,t+1}\right), \qquad x_{i,t} = Q\left(y_{i,t+1} - \eta e_{i,t}\right), \tag{D.60}$$

and $\Delta y_{i,t+3/2} = \eta(\hat{g}_{i,t+1} + e_{i,t})$. By (D.40)–(D.41) and (28), the two displacements from $y_{i,t+1}$ satisfy

$$\eta \left\| \hat{g}_{i,t+1} \right\|_\infty < \frac{1}{16}, \qquad \eta \left\| e_{i,t} \right\|_\infty < \frac{1}{16}. \tag{D.61}$$

Indeed, their respective upper bounds are $9/[60(N+1)^2] \leq 3/80$ and $8/[60(N+1)^2] \leq 1/30$. Applying (C.66) to these displacements now yields

$$\left\| x_{i,t+1} - Q(y_{i,t+1}) \right\|_1^2 \leq \frac{2e^{1/64}}{\lambda_{i,t+1}} P_{i,t+1}, \qquad \left\| x_{i,t} - Q(y_{i,t+1}) \right\|_1^2 \leq \frac{2e^{1/64}}{\lambda_{i,t+1}} I_{i,t}. \tag{D.62}$$

The triangle inequality therefore gives

$$\left\| \Delta x_{i,t+1} \right\|_1^2 \leq \frac{4e^{1/64}}{\lambda_{i,t+1}} (P_{i,t+1} + I_{i,t}) \leq \frac{5}{\lambda_{i,t+1}} (P_{i,t+1} + I_{i,t}). \tag{D.63}$$

We have obtained (D.55) at round $t+1$, and expanding both strategies about that same base score in (D.44) expresses the remainder as

$$r_{i,t+1} = \left(x_{i,t+1} - Q(y_{i,t+1}) - J_{i,t+1} \hat{g}_{i,t+1}\right) - \left(x_{i,t} - Q(y_{i,t+1}) + J_{i,t+1} e_{i,t}\right). \tag{D.64}$$

For each bracket, (C.98) applies with $M = 1/16$, and the inequality $e^a \le (1-a)^{-1}$ for $0 \le a < 1$ implies $2000e^{1/64} < 2050$. Since

$$D_\psi(z_{i,t+3/2}, z_{i,t+1}) = P_{i,t+1}, \qquad D_\psi(z_{i,t+1/2}, z_{i,t+1}) = I_{i,t}, \tag{D.65}$$

the triangle inequality and (D.40)–(D.41) consequently give

$$\|r_{i,t+1}\|_1^2 \le 2050(N+1)^2\eta^2 \left(9\sqrt{P_{i,t+1}} + 8\sqrt{I_{i,t}}\right)^2. \tag{D.66}$$

Invoking $(9a+8b)^2 \le 145(a^2+b^2)$ reduces this to

$$\|r_{i,t+1}\|_1^2 \le 300000(N+1)^2\eta^2(P_{i,t+1} + I_{i,t}). \tag{D.67}$$

The preceding bound proves (D.56) at round $t+1$, after which we control the Jacobian difference by inserting the optimistic score $y_{i,t+1/2}$ between the two base scores. Each resulting displacement has ℓ_∞ norm smaller than $1/16$, and two applications of (C.97) yield

$$\begin{aligned}
\|J_{i,t+1} - J_{i,t}\|_{\mathrm{op}}^2 &\le 2\eta^2\|\mathcal{J}(y_{i,t+1}) - \mathcal{J}(y_{i,t+1/2})\|_{\mathrm{op}}^2 \\
&\quad + 2\eta^2\|\mathcal{J}(y_{i,t+1/2}) - \mathcal{J}(y_{i,t})\|_{\mathrm{op}}^2 \\
&\le 4000e^{1/64}\eta^2(P_{i,t} + I_{i,t}) \\
&\le 4100\eta^2(P_{i,t} + I_{i,t}),
\end{aligned} \tag{D.68}$$

where the base point in each application is chosen to match the orientation of the divergence in (13). Thus (D.57) holds, including at $t = 0$, where both sides vanish by initialization. To prove the remaining repeated player bound, let t' satisfy $|t+1-t'| \le N+1$. The score update and $\|g_{i,t}\|_\infty \le 2$ then imply

$$\|y_{i,t'} - y_{i,t+1}\|_\infty \le 2\eta(N+1) \le \frac{1}{120}. \tag{D.69}$$

Our extension ensures that $J_{i,t'} = \eta\mathcal{J}(y_{i,t'})$ also for $t' \le 0$. Equations (C.63) and (25) then give

$$\|J_{i,t'}\|_{\mathrm{op}}^2 \le 4e^{1/480}\eta^2\lambda_{i,t+1}. \tag{D.70}$$

Likewise, by (D.44),

$$\left\|J_{i,t+1}\frac{\Delta y_{i,t+3/2}}{\eta}\right\|_1^2 \le 2\|\Delta x_{i,t+1}\|_1^2 + 2\|r_{i,t+1}\|_1^2. \tag{D.71}$$

Multiplying the preceding two bounds and substituting (D.55) and (D.56) into the resulting right-hand side, we obtain

$$\begin{aligned}
&\|J_{i,t'}\|_{\mathrm{op}}^2\left\|J_{i,t+1}\frac{\Delta y_{i,t+3/2}}{\eta}\right\|_1^2 \\
&\quad\le 4e^{1/480}\eta^2\left(10 + 600000(N+1)^2\eta^2\right)(P_{i,t+1} + I_{i,t}) \\
&\quad\le 90\eta^2(P_{i,t+1} + I_{i,t}),
\end{aligned} \tag{D.72}$$

where (28) implies $600000(N+1)^2\eta^2 < 11$, and $e^{1/480} < 480/479$. This proves (D.58), with the mass factor from the second Jacobian cancelling the inverse mass in (D.55). ∎

E THE TEMPLATE INEQUALITY

In this appendix, we derive the template inequality for (LIFTFTRL+) that will reduce the proof of Theorem 1 to controlling the cumulative prediction error.

Under the assumptions of Section 3, interiority of the lifted mirror map gives, for every score y,

$$\nabla\psi(Z(y)) = y. \tag{E.1}$$

For $p \in \Delta_0(\mathcal{K})$ and $z \in \operatorname{ri}\Delta_0(\mathcal{K})$, strong convexity likewise gives

$$D_\psi(p, z) \ge \frac{1}{2}\|p - z\|_1^2. \tag{E.2}$$

Continuity in the first argument of (14) extends this bound to boundary comparators without requiring a derivative there. The following lemma combines the extended bound with the optimality identity (E.1).

Lemma E.1. *For $p \in \Delta_0(\mathcal{K})$ and $y, w_1, w_2 \in \mathbb{R}^K$, set $z = Z(y)$ and, for $k = 1, 2$, $z_k^+ = Z(y + w_k)$. Then*

$$D_\psi(p, z_2^+) - D_\psi(p, z) = \langle w_2, z_1^+ - p \rangle + \langle w_2 - w_1, z_2^+ - z_1^+ \rangle - D_\psi(z_2^+, z_1^+) - D_\psi(z_1^+, z) \tag{E.3}$$

$$= \langle w_2, z_1^+ - p \rangle + D_\psi(z_1^+, z_2^+) - D_\psi(z_1^+, z), \tag{E.4}$$

and

$$D_\psi(z_1^+, z_2^+) \le \frac{1}{2}\|w_2 - w_1\|_\infty^2. \tag{E.5}$$

Proof. For $a, c \in \operatorname{ri} \Delta_0(\mathcal{K})$ and $p \in \Delta_0(\mathcal{K})$, expanding (14) yields

$$D_\psi(p, c) - D_\psi(p, a) = \langle \nabla\psi(c) - \nabla\psi(a), c - p \rangle - D_\psi(c, a). \tag{E.6}$$

Apply this identity with $a = z$ and $c = z_2^+$, first to p and then to z_1^+. Subtracting the two instances and invoking (E.1), we obtain

$$D_\psi(p, z_2^+) - D_\psi(p, z) - D_\psi(z_1^+, z_2^+) + D_\psi(z_1^+, z) \tag{E.7}$$

$$= \langle \nabla\psi(z_2^+) - \nabla\psi(z), z_1^+ - p \rangle = \langle w_2, z_1^+ - p \rangle, \tag{E.8}$$

which proves (E.4). On the other hand, adding the opposite divergences gives

$$D_\psi(z_1^+, z_2^+) + D_\psi(z_2^+, z_1^+) = \langle w_2 - w_1, z_2^+ - z_1^+ \rangle. \tag{E.9}$$

Substituting (E.9) into (E.4) proves (E.3). Finally, writing $a = \|w_2 - w_1\|_\infty$ and $d = \|z_2^+ - z_1^+\|_1$, we obtain from (E.2), (E.9), and Hölder's inequality that

$$D_\psi(z_1^+, z_2^+) \le ad - \frac{d^2}{2} = \frac{a^2}{2} - \frac{(d-a)^2}{2} \le \frac{a^2}{2}, \tag{E.10}$$

which proves (E.5). ∎

With Lemma E.1 in hand, we can now derive the template inequality for the lifted OptFTRL updates.

Proposition E.1. *Under (LiftFTRL+), for every player i and horizon $T \ge 1$,*

$$[\mathrm{Reg}_i(T)]_+ \le \frac{H_0}{\eta} + \frac{1}{\eta}\sum_{t=1}^T (I_{i,t} - P_{i,t}) \tag{E.11}$$

$$\le \frac{H_0}{\eta} + \frac{\eta}{2}\sum_{t=1}^T \|e_{i,t}\|_\infty^2 - \frac{1}{\eta}\sum_{t=1}^T P_{i,t}, \tag{E.12}$$

and, for every round $t \ge 1$,

$$I_{i,t} \le \frac{\eta^2}{2}\|e_{i,t}\|_\infty^2. \tag{E.13}$$

Proof. For a fixed player i, comparator $p \in \Delta_0(\mathcal{K})$, and round $t \ge 1$, we begin by applying Lemma E.1 with the substitutions

$$\begin{aligned} y &= y_{i,t}, \quad & w_1 &= \eta\hat{g}_{i,t}, \quad & w_2 &= \eta g_{i,t}, \\ z &= z_{i,t}, \quad & z_1^+ &= z_{i,t+1/2}, \quad & z_2^+ &= z_{i,t+1}. \end{aligned} \tag{E.14}$$

Under these substitutions, (LiftFTRL+) and (13) turn (E.4) into

$$D_\psi(p, z_{i,t+1}) - D_\psi(p, z_{i,t}) = \eta\langle g_{i,t}, z_{i,t+1/2} - p \rangle + I_{i,t} - P_{i,t}. \tag{E.15}$$

For these same substitutions, (E.5) and (12) establish the innovation bound (16). We may now sum (E.15) from $t = 1$ to T and rearrange the telescoping divergences to obtain

$$\sum_{t=1}^T \langle g_{i,t}, p - z_{i,t+1/2} \rangle = \frac{D_\psi(p, z_{i,1}) - D_\psi(p, z_{i,T+1})}{\eta} + \frac{1}{\eta}\sum_{t=1}^T (I_{i,t} - P_{i,t}). \tag{E.16}$$

The zero initialization, (11), and (E.1) show that $z_{i,1} = Z(\mathbf{0})$ minimizes ψ and has zero gradient. Consequently, for every $p \in \Delta_0(\mathcal{K})$, the endpoint divergences satisfy

$$D_\psi(p, z_{i,1}) = \psi(p) - \min_{\Delta_0(\mathcal{K})} \psi \le H_0, \qquad D_\psi(p, z_{i,T+1}) \ge 0. \tag{E.17}$$

Substituting these endpoint bounds into (E.16), maximizing over p, and invoking (9) for the optimistic lifted iterates proves (15), from which (RVU) follows by (16). ∎

For HOOD, Proposition C.1 verifies the assumptions on (SRE), and the predictor (HOP) is causal, so the template applies to Algorithm 1 for every $\eta > 0$. Summing (RVU) and (16) over the players, we obtain

$$\sum_{i\in\mathcal{N}} [\mathrm{Reg}_i(T)]_+ \le \frac{NH_0}{\eta} + \frac{\eta}{2}E(T) - \frac{P(T)}{\eta}, \qquad I(T) \le \frac{\eta^2}{2}E(T). \tag{E.18}$$

Since the left-hand side of the first inequality is nonnegative, $P(T) \le NH_0 + \eta^2 E(T)/2$, and adding the second inequality in (E.18) proves (18).

F Proof of the main theorem

With the local bounds of Appendices C and D in place, we now obtain the prediction error bound of Proposition F.3 by expanding one payoff difference and summing the resulting contributions. The proof of Theorem 1 will then follow by inserting this bound into (18) and invoking Proposition 1.

Fix $T \ge 1$ and use the norms of (D.8) with sums restricted to $1 \le t \le T$. Since every sequence to which D_p is applied vanishes for $t \le 0$, Proposition D.1 gives (D.9) and (D.10) directly for these finite horizon norms. Nonnegative round shifts do not increase any of the norms used below: if $w_t = \mathbf{0}$ for $t \le 0$ and $q \in \mathbb{Z}_{\ge 0}$, then

$$\sum_{t=1}^{T} \|w_{t-q}\|^2 \le \sum_{t=1}^{T} \|w_t\|^2. \tag{F.1}$$

In particular, (F.1) applies to remainders, error differences, and Jacobian differences, whose prehistories vanish. Summing the pointwise bounds (D.56), (D.57), and (D.46) over rounds and players, and using $I_{i,0} = 0$ and $J_{i,1} = J_{i,0}$, we obtain

$$\sum_{j=1}^{N} \|r_j\|_{\ell_2}^2 \le 300000(N+1)^2\eta^2(P(T) + I(T)), \tag{F.2}$$

$$\sum_{j=1}^{N} \|\Delta J_j\|_{\ell_2}^2 \le 4100\eta^2(P(T) + I(T)), \tag{F.3}$$

$$\sum_{j=1}^{N} \|\Delta e_j\|_{\ell_2}^2 \le 4E(T). \tag{F.4}$$

In each sum, shifting a nonnegative movement by one round only removes its final term because the initial term is zero. For integers $0 \le \ell, u \le N+1$, the corresponding shifted bound follows directly from (D.58):

$$\sum_{t=1}^{T} \|J_{j,t-\ell}\|_{\mathrm{op}}^2 \left\| J_{j,t-u} \frac{\Delta y_{j,t-u+1/2}}{\eta} \right\|_1^2 \le 90\eta^2 \sum_{k=1}^{T} (P_{j,k} + I_{j,k}). \tag{F.5}$$

Indeed, whenever $t - u > 0$, (D.58) applies at time $t - u$ because $|u - \ell| \le N + 1$, and the summand vanishes at all remaining indices. To sum contributions carrying the same player label, we will also use the following consequence of Cauchy–Schwarz. Suppose that $c, u_j \ge 0$, $\|w_\nu\|_{\ell_2} \le c\, u_{j_\nu}$, the index set $\mathcal{I}$ has size at most n_0, and each label occurs at most n_1 times. Then

$$\left\| \sum_{\nu\in\mathcal{I}} w_\nu \right\|_{\ell_2} \le c\sqrt{n_0 n_1} \left(\sum_{j=1}^{N} u_j^2 \right)^{1/2}. \tag{F.6}$$

Indeed, the triangle inequality followed by Cauchy–Schwarz gives

$$\left\| \sum_{\nu\in\mathcal{I}} w_\nu \right\|_{\ell_2} \le c \sum_{\nu\in\mathcal{I}} u_{j_\nu} \le c\sqrt{|\mathcal{I}| \sum_{\nu\in\mathcal{I}} u_{j_\nu}^2}, \tag{F.7}$$

and, since each label occurs at most n_1 times, $\sum_{\nu\in\mathcal{I}} u_{j_\nu}^2 \le n_1 \sum_{j=1}^{N} u_j^2$, which proves (F.6).

We may now state the one step expansion. A *strategy term* is a strategy vector, a *payoff term* is a payoff excess vector, a *Jacobian term* is a factor J_j, and a *remainder term* is a factor r_j from (D.44).

Each expression below is a fixed multilinear map of these factors evaluated at rounds $t-q$, with only its arguments depending on t. After replacing its Jacobian, payoff, and strategy factors by arbitrary linear maps $B_a : \mathbb{R}^K \to \mathbb{R}^K$ and vectors v_a, z_a, denote the resulting value by $\mathcal{T}[(B_a);(v_a);(z_a)]$. The maps generated by the expansion satisfy

$$\|\mathcal{T}[(B_a);(v_a);(z_a)]\|_\infty \le \prod_a \|B_a\|_{\mathrm{op}} \prod_a \|v_a\|_\infty \prod_a \|z_a\|_1. \tag{F.8}$$

Proposition F.1. *Assume (28) and fix $T \ge 1$. Fix $1 \le p \le N$, and let M be a sequence obtained from a fixed multilinear map as above, containing p Jacobian terms with distinct player labels $j_1, \dots, j_p$, together with s payoff terms and h strategy terms, where $1 \le s \le p$ and $0 \le h \le Np$. Assume that all round shifts lie in $\{0, \dots, p\}$, that each player occurs at most p times among the strategy terms, and that its underlying multilinear map satisfies (F.8) for arbitrary arguments. Then*

$$\Delta M = R_M + \sum_{M' \in \mathcal{C}(M)} M', \tag{F.9}$$

where $\mathcal{C}(M)$ is an indexed family, counted with multiplicity, of continuing terms, and R_M collects the controlled terms and satisfies

$$\begin{aligned} \left\|\mathrm{D}_{p+1} R_M\right\|_{\ell_2} \le\ & 2p(N+1)^{p+1}(2\eta)^p \\ & + 815p\sqrt{N}\,(N+1)^{p+2}(2\eta)^{p-1}\eta\sqrt{P(T)+I(T)} \\ & + 12p\sqrt{N}\,(N+1)^{p+2}(2\eta)^{p+1}\sqrt{E(T)}. \end{aligned} \tag{F.10}$$

Moreover, every $M' \in \mathcal{C}(M)$ satisfies the same conditions as M with p replaced by $p+1$, and

$$|\mathcal{C}(M)| \le 3p(N-p). \tag{F.11}$$

In particular, $\mathcal{C}(M) = \varnothing$ when $p = N$.

Proof. We begin by applying (D.50) to M, so exactly one factor is differenced in each summand and every round shift increases by at most one. For each payoff difference, substitution of (D.48) and (D.51) separates the initialization contribution and leaves strategy differences. Thus, the s payoff terms and h strategy terms produce at most

$$(2N-1)s + h \le 3Np \tag{F.12}$$

strategy differences. Each player occurs at most twice per payoff term by Proposition D.2 and, by hypothesis, at most p times among the strategy terms. Its total multiplicity is therefore at most

$$2s + p \le 3p \tag{F.13}$$

times. Applying (D.44) to any of these differences produces a remainder controlled by (F.2). If $j \in \{j_1, \dots, j_p\}$, its linear part can be paired with the existing Jacobian for j. Both round shifts lie in $\{0, \dots, p+1\}$, so their difference is at most $N+1$ and (F.5) applies. For a new label, (D.45) leaves only $J_j g_j$ to be expanded further. There are $N-p$ possible new labels and each appears at most $3p$ times, proving (F.11). In particular, no label remains when $p = N$.

Substitution of Bv for a differenced strategy term preserves (F.8) because $\|Bv\|_1 \le \|B\|_{\mathrm{op}}\|v\|_\infty$. When a payoff term is differenced, Proposition D.2 introduces at most $N-1$ strategy arguments, and (D.54) yields

$$\left\|f[z_1, \dots, z_q]Bv\right\|_\infty \le \|B\|_{\mathrm{op}}\|v\|_\infty \prod_{a=1}^{q} \|z_a\|_1. \tag{F.14}$$

Substituting this bound into (F.8) preserves that bound for arbitrary arguments. If the difference falls on a payoff, at most $N-1$ strategy terms are added and each player occurs at most once among them. A difference falling on a strategy replaces that strategy term by one payoff term. Hence a continuing expression contains at most $N(p+1)$ strategy terms, at most $p+1$ occurrences of any player among them, and between 1 and $p+1$ payoff terms. Its $p+1$ distinct Jacobian labels and shifts in $\{0, \dots, p+1\}$ complete the hypotheses required at the next order.

To bound the remaining contributions, use (F.1) for every round shift and decompose

$$R_M = R_M^J + R_M^0 + R_M^r + R_M^e + R_M^y, \tag{F.15}$$

where the five pieces contain, respectively, a Jacobian variation, an initialization term, a remainder r_j, a term $-J_j\Delta e_j$, and a repeated player term $J_j\frac{\Delta y_{j,\cdot+1/2}}{\eta}$. By $\|g_j\|_\infty \le 2$, (25), and (D.42), the product of the bounds for the p Jacobian terms, s payoff terms, and all strategy terms is at most

$$\eta^p 2^s \le (2\eta)^p. \tag{F.16}$$

We will use this product bound after removing whichever factor is being bounded. First, applying the triangle inequality to (F.15) separates the required bounds as

$$\begin{aligned}\left\|\mathrm{D}_{p+1}R_M\right\|_{\ell_2} &\le \left\|\mathrm{D}_{p+1}R^0_M\right\|_{\ell_2} && \text{(F.17a)}\\ &+ \left\|\mathrm{D}_{p+1}R^J_M\right\|_{\ell_2} + \left\|\mathrm{D}_{p+1}R^r_M\right\|_{\ell_2} + \left\|\mathrm{D}_{p+1}R^y_M\right\|_{\ell_2} && \text{(F.17b)}\\ &+ \left\|\mathrm{D}_{p+1}R^e_M\right\|_{\ell_2}. && \text{(F.17c)}\end{aligned}$$

Bounding (F.17a)**.** Replacing a payoff from round $t-q$ by its initialization contribution produces a term supported only at $t = q+1$. Equation (D.49) bounds the substituted vector by 2, with the product of the other factors bounded by $\eta^p 2^{s-1}$. Thus the entire term has norm at most $(2\eta)^p$. There are at most $s \le p$ such terms, and (D.10) yields

$$\left\|\mathrm{D}_{p+1}R^0_M\right\|_{\ell_2} \le 2p(N+1)^{p+1}(2\eta)^p. \tag{F.18}$$

Bounding (F.17b)**.** When one of the p Jacobian factors is replaced by ΔJ_j, the remaining factors contribute at most $2(2\eta)^{p-1}$. Applying (D.9), followed by Cauchy–Schwarz over the p distinct Jacobian labels and (F.3), yields

$$\begin{aligned}\left\|\mathrm{D}_{p+1}R^J_M\right\|_{\ell_2} &\le 4(N+1)^{p+2}\,2(2\eta)^{p-1}\sqrt{p}\left(\sum_{j=1}^N\|\Delta J_j\|^2_{\ell_2}\right)^{1/2}\\ &\le 8\sqrt{4100}\sqrt{p}\,(N+1)^{p+2}(2\eta)^{p-1}\eta\sqrt{P(T)+I(T)}\\ &\le 520\sqrt{p}\,(N+1)^{p+2}(2\eta)^{p-1}\eta\sqrt{P(T)+I(T)},\end{aligned} \tag{F.19}$$

where $8\sqrt{4100} < 520$.

For the remainder contribution, each of the at most $3Np$ strategy differences leaves factors bounded by $(2\eta)^p$ after r_j is removed. Equation (F.13) bounds each label's multiplicity by $3p$, so (F.6) applies with

$$n_0 = 3Np, \qquad n_1 = 3p, \tag{F.20}$$

and, after invoking (F.2) and (D.9), we obtain

$$\begin{aligned}\left\|\mathrm{D}_{p+1}R^r_M\right\|_{\ell_2} &\le 4(N+1)^{p+2}(2\eta)^p\sqrt{(3Np)(3p)}\left(\sum_{j=1}^N\|r_j\|^2_{\ell_2}\right)^{1/2}\\ &\le 12\sqrt{300000}\,p\sqrt{N}\,(N+1)^{p+3}(2\eta)^p\eta\sqrt{P(T)+I(T)}\\ &\le 6600p\sqrt{N}\,(N+1)^{p+3}(2\eta)^p\eta\sqrt{P(T)+I(T)},\end{aligned} \tag{F.21}$$

where $12\sqrt{300000} < 6600$.

For R^y_M, pair each repeated player term $J_j\frac{\Delta y_{j,\cdot+1/2}}{\eta}$ with the existing Jacobian carrying the same player label. Invoking (F.5) leaves a product bounded by $2(2\eta)^{p-1}$. We may therefore apply (F.6) and (D.9) to obtain

$$\begin{aligned}\left\|\mathrm{D}_{p+1}R^y_M\right\|_{\ell_2} &\le 4(N+1)^{p+2}\,2(2\eta)^{p-1}\sqrt{(3Np)(3p)}\sqrt{90}\,\eta\sqrt{P(T)+I(T)}\\ &\le 240p\sqrt{N}\,(N+1)^{p+2}(2\eta)^{p-1}\eta\sqrt{P(T)+I(T)},\end{aligned} \tag{F.22}$$

where $24\sqrt{90} < 240$.

Bounding (F.17c)**.** For a term $-J_j\Delta e_j$, removing Δe_j leaves $p+1$ Jacobian factors and at most p payoff factors, whose product is bounded by

$$\eta^{p+1}2^p = \frac{1}{2}(2\eta)^{p+1}. \tag{F.23}$$

Taking again $n_0 = 3Np$ and $n_1 = 3p$, and invoking (F.6), (F.4), and (D.9), we obtain

$$\begin{aligned}\left\|\mathrm{D}_{p+1}R^{e}_{M}\right\|_{\ell_2} &\le 4(N+1)^{p+2}\frac{1}{2}(2\eta)^{p+1}\sqrt{(3Np)(3p)}\left(\sum_{j=1}^{N}\left\|\Delta e_j\right\|_{\ell_2}^{2}\right)^{1/2}\\ &\le 12p\sqrt{N}\,(N+1)^{p+2}(2\eta)^{p+1}\sqrt{E(T)}.\end{aligned} \tag{F.24}$$

Folding (F.18), (F.19), (F.21), (F.22), and (F.24) back into (F.17), we obtain

$$\begin{aligned}\left\|\mathrm{D}_{p+1}R_{M}\right\|_{\ell_2} &\le 2p(N+1)^{p+1}(2\eta)^{p}\\ &\quad+\left[520\sqrt{p}+240p\sqrt{N}+13200p\sqrt{N}(N+1)\eta\right]\\ &\quad\times(N+1)^{p+2}(2\eta)^{p-1}\eta\sqrt{P(T)+I(T)}\\ &\quad+12p\sqrt{N}(N+1)^{p+2}(2\eta)^{p+1}\sqrt{E(T)}.\end{aligned} \tag{F.25}$$

Our choice (28) ensures that $2(N+1)\eta \le 1/120$, reducing the last summand in the square bracket to at most $55p\sqrt{N}$, and the inequality $\sqrt{p} \le p\sqrt{N}$ then bounds the entire bracket by $(520+240+55)p\sqrt{N} = 815p\sqrt{N}$. This proves (F.10). ∎

The count (F.11) now allows the bound (F.10) to be summed over all expansion orders.

Proposition F.2. *Assume* (28)*, and let M satisfy the hypotheses of Proposition F.1 with $p = 1$. Then, for every $T \ge 1$,*

$$\begin{aligned}\left\|\mathrm{D}_1 M\right\|_{\ell_2} &\le 5(N+1)^2\eta\\ &\quad+900\sqrt{N}\,(N+1)^3\eta\sqrt{P(T)+I(T)}\\ &\quad+53\sqrt{N}\,(N+1)^3\eta^2\sqrt{E(T)}.\end{aligned} \tag{F.26}$$

Proof. Every expression requiring further expansion contains at least one payoff term by Proposition F.1, and hence vanishes for $t \le 0$. Starting from $\mathcal{F}_1 = \{M\}$, define recursively, for $1 \le p \le N$,

$$\mathcal{F}_{p+1} = \biguplus_{M' \in \mathcal{F}_p} \mathcal{C}(M'), \tag{F.27}$$

where each occurrence is retained separately, even if two expressions coincide. Thus $\mathcal{F}_p$ counts, with multiplicity, all terms requiring expansion at order p. For $1 \le p \le N$, substitute (F.9) into $\mathrm{D}_p = \mathrm{D}_{p+1}\Delta$ at each such term to obtain

$$\sum_{M' \in \mathcal{F}_p} \mathrm{D}_p M' = \sum_{M' \in \mathcal{F}_p} \mathrm{D}_{p+1} R_{M'} + \sum_{M' \in \mathcal{F}_{p+1}} \mathrm{D}_{p+1} M'. \tag{F.28}$$

Upon summing over p, the continuing terms cancel at consecutive orders, and the endpoint identities $\mathcal{F}_1 = \{M\}$ and $\mathcal{F}_{N+1} = \varnothing$ leave

$$\mathrm{D}_1 M = \sum_{p=1}^{N} \sum_{M' \in \mathcal{F}_p} \mathrm{D}_{p+1} R_{M'}. \tag{F.29}$$

At successive orders, (F.11) gives

$$|\mathcal{F}_{p+1}| \le 3p(N-p)|\mathcal{F}_p|. \tag{F.30}$$

Apart from the explicit factor p, the powers of $N+1$ and 2η in each term of (F.10) gain a factor $2(N+1)\eta$ when the order increases from p to $p+1$. Together with the preceding count, this yields

$$6p(N-p)(N+1)\eta \le \frac{3}{2}N^2(N+1)\eta \le \frac{1}{40}, \tag{F.31}$$

where the last inequality follows from (28). Starting from $|\mathcal{F}_1| = 1$, induction gives

$$|\mathcal{F}_p|(N+1)^{p+2}(2\eta)^{p-1} \le (N+1)^3\left(\frac{1}{40}\right)^{p-1},$$

$$|\mathcal{F}_p|(N+1)^{p+1}(2\eta)^p \le (N+1)^2(2\eta)\left(\frac{1}{40}\right)^{p-1}. \tag{F.32}$$

Finally, the identity

$$\sum_{p\ge 1} p\left(\frac{1}{40}\right)^{p-1} = \left(\frac{40}{39}\right)^2 < \frac{11}{10}, \tag{F.33}$$

allows us to sum (F.10) over all $M' \in \mathcal{F}_p$ and all p, yielding (F.26) with the three numerical constants following from

$$2\cdot 2\cdot\frac{11}{10} < 5, \qquad 815\cdot\frac{11}{10} < 900, \qquad 12\cdot 4\cdot\frac{11}{10} < 53. \tag{F.34}$$

■

The preceding summation bound reduces the remaining argument to making the first payoff difference explicit and separating the initialization, Taylor remainder, and error terms from those to which Proposition F.2 applies.

Proposition F.3. *Assume* (28). *Then, for every horizon* $T \ge 1$,

$$\sqrt{E(T)} \le 5\sqrt{N}(N+1) + 2000(N+1)^5\eta\sqrt{P(T)+I(T)} + \frac{3}{10}\sqrt{E(T)}. \tag{F.35}$$

Proof. We begin by substituting the payoff decomposition (D.51) and the strategy expansion (D.45) into $e_i = \mathrm{D}_1\Delta g_i$. To write the four resulting contributions, for every player i and integer t set

$$b_{i,t} = \mathbb{1}_{\{t=1\}}\gamma_{i,0}, \qquad A_i^r = \sum_{j\ne i} f_{ij}^v r_j + \sum_{j=1}^N f_{ij}^u r_j, \tag{F.36}$$

and

$$A_i^e = \sum_{j\ne i} f_{ij}^v J_j\Delta e_j + \sum_{j=1}^N f_{ij}^u J_j\Delta e_j, \qquad A_i^g = \sum_{j\ne i} f_{ij}^v J_j g_j + \sum_{j=1}^N f_{ij}^u J_j g_j. \tag{F.37}$$

All products in these definitions are evaluated pointwise in time. For example, at round t,

$$(f_{ij}^v J_j g_j)_t = f_{ij,t}^v J_{j,t} g_{j,t}. \tag{F.38}$$

Substituting (D.48), (D.51), (D.44), and (D.43) into the recursion (D.7), we obtain

$$e_i = \mathrm{D}_1 b_i + \mathrm{D}_1(A_i^r - A_i^e + A_i^g). \tag{F.39}$$

The triangle inequality now yields the bound

$$\sqrt{E(T)} \le \left(\sum_i \|\mathrm{D}_1 b_i\|_{\ell_2}^2\right)^{1/2} \tag{F.40a}$$

$$+ \left(\sum_i \|\mathrm{D}_1 A_i^r\|_{\ell_2}^2\right)^{1/2} \tag{F.40b}$$

$$+ \left(\sum_i \|\mathrm{D}_1 A_i^e\|_{\ell_2}^2\right)^{1/2} \tag{F.40c}$$

$$+ \left(\sum_i \|\mathrm{D}_1 A_i^g\|_{\ell_2}^2\right)^{1/2}. \tag{F.40d}$$

Bounding (F.40a)**.** A direct application of (D.10) and (D.49) gives

$$\left(\sum_i \|\mathrm{D}_1 b_i\|_{\ell_2}^2\right)^{1/2} \le 4\sqrt{N}\,(N+1). \tag{F.41}$$

Bounding (F.40b)**.** By Proposition D.2, A_i^r contains $2N-1$ terms, every coefficient has norm at most one, and a fixed player label occurs no more than twice. Cauchy–Schwarz therefore yields, for each i and t,

$$\left\|A_{i,t}^r\right\|_\infty^2 \le 2(2N-1)\sum_j \left\|r_{j,t}\right\|_1^2. \tag{F.42}$$

Summing this bound over t and i, and then applying (D.9) and (F.2), we obtain

$$\begin{aligned}\left(\sum_i \left\|\mathrm{D}_1 A_i^r\right\|_{\ell_2}^2\right)^{1/2} &\le 4\sqrt{300000}\,\sqrt{2N(2N-1)}(N+1)^3\eta\sqrt{P(T)+I(T)} \\ &\le 2200\sqrt{2N(2N-1)}(N+1)^3\eta\sqrt{P(T)+I(T)}.\end{aligned} \tag{F.43}$$

Here the last inequality uses $4\sqrt{300000} < 2200$.

Bounding (F.40c)**.** For the prediction error contribution, use the same multiplicity bound together with $\|J_{j,t}\|_{\mathrm{op}} \le \eta$ from (D.42) and Proposition C.2. The resulting pointwise bound is, for each i and t,

$$\left\|A_{i,t}^e\right\|_\infty^2 \le 2(2N-1)\eta^2 \sum_j \left\|\Delta e_{j,t}\right\|_\infty^2. \tag{F.44}$$

Summing this bound over t and i, and applying (D.9) and (F.4), we obtain

$$\begin{aligned}\left(\sum_i \left\|\mathrm{D}_1 A_i^e\right\|_{\ell_2}^2\right)^{1/2} &\le 4\cdot 2\,\sqrt{2N(2N-1)}(N+1)^2\eta\sqrt{E(T)} \\ &= 8\sqrt{2N(2N-1)}(N+1)^2\eta\sqrt{E(T)}.\end{aligned} \tag{F.45}$$

Bounding (F.40d)**.** To apply Proposition F.2, observe that each term in A_i^g contains one Jacobian and one payoff factor at round t, and that, by Proposition D.2, its coefficient contains at most $N-1$ strategy factors with distinct player labels evaluated at t or $t-1$. Equation (D.54) supplies the bound (F.8) for arbitrary arguments, so each of the $2N-1$ terms meets the hypotheses of Proposition F.2. Applying (F.26) to these terms, summing the resulting bounds, and taking the square root of the sum of squares over i, we obtain

$$\begin{aligned}\left(\sum_i \left\|\mathrm{D}_1 A_i^g\right\|_{\ell_2}^2\right)^{1/2} &\le 5\sqrt{N}(2N-1)(N+1)^2\eta \\ &\quad + 900N(2N-1)(N+1)^3\eta\sqrt{P(T)+I(T)} \\ &\quad + 53N(2N-1)(N+1)^3\eta^2\sqrt{E(T)}.\end{aligned} \tag{F.46}$$

Folding (F.41), (F.43), (F.45), and (F.46) back into (F.40), we obtain

$$\begin{aligned}\sqrt{E(T)} &\le 4\sqrt{N}(N+1) + 5\sqrt{N}(2N-1)(N+1)^2\eta \\ &\quad + \left[2200\sqrt{2N(2N-1)} + 900N(2N-1)\right] \\ &\quad\quad \times (N+1)^3\eta\sqrt{P(T)+I(T)} \\ &\quad + \left[8\sqrt{2N(2N-1)}(N+1)^2\eta\right. \\ &\quad + \left.53N(2N-1)(N+1)^3\eta^2\right]\sqrt{E(T)}.\end{aligned} \tag{F.47}$$

For the initialization contribution, (28) and $2N-1 \le (N+1)^2$ imply

$$\begin{aligned}&4\sqrt{N}\,(N+1) + 5\sqrt{N}(2N-1)(N+1)^2\eta \\ &\le \left(4+\frac{1}{12}\right)\sqrt{N}\,(N+1) < 5\sqrt{N}\,(N+1).\end{aligned} \tag{F.48}$$

For the coefficient of $\sqrt{P(T)+I(T)}$, we use

$$\sqrt{2N(2N-1)} \le 2N, \qquad 2N-1 \le 2N. \tag{F.49}$$

It follows that

$$\begin{aligned} &2200\sqrt{2N(2N-1)} + 900N(2N-1) \\ &\quad \le 4400N + 1800N^2 \\ &\quad \le 2000(N+1)^2, \end{aligned} \tag{F.50}$$

where the last inequality follows because

$$2000(N+1)^2 - (1800N^2 + 4400N) = 200(N^2 - 2N + 10) > 0. \tag{F.51}$$

After multiplication by the common factor $(N+1)^3\eta\sqrt{P(T)+I(T)}$, this contribution is bounded by

$$2000(N+1)^5\eta\sqrt{P(T)+I(T)}. \tag{F.52}$$

Finally, for the coefficient of $\sqrt{E(T)}$, (28) gives

$$\begin{aligned} 8\sqrt{2N(2N-1)}(N+1)^2\eta &\le \frac{4N}{15(N+1)} < \frac{4}{15}, \\ 53N(2N-1)(N+1)^3\eta^2 &\le \frac{106N^2}{60^2(N+1)^3} \le \frac{53}{3600} < \frac{1}{60}, \end{aligned} \tag{F.53}$$

where we used

$$\frac{N^2}{(N+1)^3} \le \frac{1}{N+1} \le \frac{1}{2}. \tag{F.54}$$

The numerical bound

$$\frac{4}{15} + \frac{1}{60} = \frac{17}{60} < \frac{3}{10}, \tag{F.55}$$

shows that the total $\sqrt{E(T)}$ contribution is at most $\frac{3}{10}\sqrt{E(T)}$, and substituting the three bounds into the initial bound proves (31). ∎

The bound of Proposition F.3 supplies the remaining ingredient for (18) and the template inequality, so we can now complete the proof of Theorem 1.

Proof of Theorem 1. Substituting (18) into (31) and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain

$$\sqrt{E(T)} \le 5\sqrt{N}\,(N+1) + 2000(N+1)^5\eta\sqrt{NH_0} + 2000(N+1)^5\eta^2\sqrt{E(T)} + \frac{3}{10}\sqrt{E(T)}. \tag{F.56}$$

Specializing Proposition C.1 and (28), we have $L \ge 4$ and $(N+1)^3\eta \le 1/60$. The three coefficients consequently satisfy

$$\begin{aligned} 5\sqrt{N}\,(N+1) &\le \frac{5}{8}(N+1)^3\sqrt{L}, \\ 2000(N+1)^5\eta\sqrt{NH_0} &\le \frac{100}{3}(N+1)^2\sqrt{\frac{13NL}{6}} < 25(N+1)^3\sqrt{L}, \\ 2000(N+1)^5\eta^2 &\le \frac{5}{18}. \end{aligned} \tag{F.57}$$

Here the first inequality follows from $\sqrt{L} \ge 2$ and $(N+1)^2 \ge 4\sqrt{N}$, the second uses $N+1 \ge 2\sqrt{N}$ and $\sqrt{13/6} < 3/2$, and the third follows from $2000(N+1)^5\eta^2 \le 5/[9(N+1)] \le 5/18$. Consequently,

$$\sqrt{E(T)} < 26(N+1)^3\sqrt{L} + \frac{26}{45}\sqrt{E(T)}. \tag{F.58}$$

Rearranging the preceding inequality and using $26 \cdot 45/19 < 62$ gives

$$E(T) < 62^2(N+1)^6L. \tag{F.59}$$

Returning to the template inequality (RVU) and using the nonnegativity of $P_{i,t}$, we obtain

$$[\mathrm{Reg}_i(T)]_+ \le \frac{H_0}{\eta} + \frac{\eta}{2}E(T). \tag{F.60}$$

We now invoke Proposition C.1, (F.59), and $\eta^2(N+1)^6 \le 1/60^2$. Since $62^2\eta^2(N+1)^6/2 \le 961/1800 < 5/9$, the preceding template bound becomes

$$[\mathrm{Reg}_i(T)]_+ < \left(\frac{13}{6} + \frac{5}{9}\right)\frac{L}{\eta} = \frac{49}{18}\frac{L}{\eta} < \frac{3L}{\eta}. \tag{F.61}$$

Since the bound is independent of T, it proves (29), and substituting $\eta = 1/(60(N+1)^3)$ then gives (30). ∎

G Adversarial robustness

Assume (28). For the switching rule stated after Theorem 1, set

$$B = \frac{3L}{\eta}. \tag{G.1}$$

With this threshold fixed, consider a full information adversarial rule whose guarantee is valid at every horizon. More precisely, for each $s \ge 1$ and payoff sequence $w_1, \dots, w_s \in [-1,1]^K$, its strategies q_k, formed from $w_1, \dots, w_{k-1}$ after a fresh start, satisfy

$$\max_{\alpha \in \mathcal{A}} \sum_{k=1}^{s} \langle \alpha - q_k, w_k \rangle \le H(s), \qquad H(s) = o(s). \tag{G.2}$$

Standard examples include Hedge / exponential weights with well-chosen learning rates. With $\mathrm{Reg}_i(0) = 0$, define

$$\tau_i = \inf\{t \ge 1 : \mathrm{Reg}_i(t) > B\}, \qquad \inf \varnothing = \infty. \tag{G.3}$$

The player follows Algorithm 1 through round τ_i and, if $\tau_i < \infty$, restarts the adversarial rule at round $\tau_i + 1$ and uses it thereafter.

The following proposition records the two guarantees of this rule, namely that no player switches when every player follows it and that each player retains vanishing average regret against arbitrary deviations by the others.

Proposition G.1. *If every player uses the above rule, then $\tau_i = \infty$ for every player. Against arbitrary deviations of the other players, the same rule satisfies*

$$\limsup_{T \to \infty} \frac{\mathrm{Reg}_i(T)}{T} \le 0. \tag{G.4}$$

Proof. The definition of τ_i bounds regret by B at every horizon before the switch, and over the switching round, $\|g_j\|_\infty \le 2$ allows each action's cumulative payoff excess, and hence their maximum in (5), to increase by at most 2. Consequently, if $\tau_i < \infty$,

$$\mathrm{Reg}_i(\tau_i) \le B + 2. \tag{G.5}$$

For $T > \tau_i$, bound the prefix through τ_i by $B + 2$ and apply (G.2) to the restarted rule with payoff vectors $v_{i,\tau_i+1}, \dots, v_{i,T}$. Indeed, for each fixed action,

$$\langle \alpha_i - x_{i,t}, v_{i,t} \rangle = u_i(\alpha_i, x_{-i,t}) - u_i(x_t), \tag{G.6}$$

and splitting cumulative regret at τ_i, with the maximum of the sum bounded by the sum of the maxima, therefore gives

$$\mathrm{Reg}_i(T) \le \mathrm{Reg}_i(\tau_i) + H(T - \tau_i) \le B + 2 + H(T - \tau_i), \tag{G.7}$$

where the suffix is controlled by (G.2), for any realized play of the other players.

Suppose now that every player uses the switching rule. If $\tau = \min_j \tau_j < \infty$, play through round τ agrees with a trajectory of Algorithm 1, because the first switch takes effect only at round $\tau + 1$. The bound (29) in Theorem 1 would imply $\mathrm{Reg}_j(\tau) \le B$ for every j, contradicting $\mathrm{Reg}_j(\tau) > B$ for a player with $\tau_j = \tau$. Therefore $\tau_i = \infty$ for every player.

Against arbitrary opponents, the case of no switch has $\mathrm{Reg}_i(T) \le B$ for every T and already implies (G.4). If instead $\tau_i < \infty$, divide the prefix and suffix bound by T to obtain, for $T > \tau_i$,

$$\frac{\mathrm{Reg}_i(T)}{T} \le \frac{B+2}{T} + \frac{T - \tau_i}{T}\frac{H(T - \tau_i)}{T - \tau_i}. \tag{G.8}$$

Letting $T \to \infty$ and invoking $H(s) = o(s)$ makes both terms tend to zero, which proves (G.4) also in this case. ∎