On the Optimality of Dilated Entropy and Lower Bounds for Online Learning in Extensive-Form Games

Zhiyuan Fan
MIT
fanzy@mit.edu

Christian Kroer
Columbia University
christian.kroer@columbia.edu

Gabriele Farina
MIT
gfarina@mit.edu

Abstract

First-order methods (FOMs) are arguably the most scalable algorithms for equilibrium computation in large extensive-form games. To operationalize these methods, a distance-generating function, acting as a regularizer for the strategy space, must be chosen. The ratio between the strong convexity modulus and the diameter of the regularizer is a key parameter in the analysis of FOMs. A natural question is then: what is the *optimal* distance-generating function for extensive-form decision spaces? In this paper, we make a number of contributions, ultimately establishing that the weight-one dilated entropy (DilEnt) distance-generating function is optimal up to logarithmic factors. The DilEnt regularizer is notable due to its iterate-equivalence with Kernelized OMWU (KOMWU)—the algorithm with state-of-the-art dependence on the game tree size in extensive-form games—when used in conjunction with the online mirror descent (OMD) algorithm. However, the standard analysis for OMD is unable to establish such a result; the only current analysis is by appealing to the iterate equivalence to KOMWU. We close this gap by introducing a pair of primal-dual *treeplex* norms, which we contend form the natural analytic viewpoint for studying the strong convexity of DilEnt. Using these norm pairs, we recover the diameter-to-strong-convexity ratio that predicts the same performance as KOMWU. Along with a new regret lower bound for online learning in sequence-form strategy spaces, we show that this ratio is nearly optimal. Finally, we showcase our analytic techniques by refining the analysis of Clairvoyant OMD when paired with DilEnt, establishing an $\mathcal{O}(n \log |\mathcal{V}| \log T/T)$ approximation rate to coarse correlated equilibrium in n-player games, where $|\mathcal{V}|$ is the number of reduced normal-form strategies of the players, establishing the new state of the art.

1 Introduction

Extensive-form games (EFG) are a popular framework for modeling sequential games with imperfect information. The framework has been widely used to build superhuman AI agents in real-world imperfect information games [5, 31, 6, 8]. Several notions of equilibrium, including Nash equilibrium [32] in two-player zero-sum and coarse correlated equilibrium in general multiplayer EFGs, can be computed in polynomial time in the size of the game tree under the standard hypothesis of perfect recall [41, 35, 23, 24]. These polynomial-time algorithms, however, require running the ellipsoid method or polynomial algorithm for linear programming, both of which are impractical for large-scale games, due to the high memory usage and large per-iteration computational costs [38].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2804

Instead, fast iterative methods based on convex first-order optimization methods (FOMs) [43, 12, 40, 7, 15, 27, 16, 28–30] are commonly used to find an approximate equilibrium. These iterative methods define strategy update rules that each player can apply iteratively while training in self-play with other players, and that guarantee ergodic convergence to the set of equilibria in the long run. Three popular classes of such FOMs are employed in EFGs: methods based on online mirror descent (OMD) [4, 12], methods based on the counterfactual regret minimization framework [43], and (in the context of two-player zero-sum games specifically) accelerated offline methods such as mirror prox [33] and the excessive gap technique [34] algorithm. In general, these methods are all proximal methods—that is, they perform a generalized notion of projected gradient descent step at each iteration. Some do this explicitly, including OMD and mirror prox, while others do it implicitly, including the counterfactual regret minimization algorithm, which runs proximal steps locally at each decision point [17].

In all methods mentioned above, except CFR,[1] the constraint set for the proximal step (*i.e.*, the set on which gradient steps must be projected onto) is the strategy polytope of the EFG. The proximal steps are parameterized by a choice of *distance-generating function (DGF)* for the strategy polytope, which acts as a regularizer. The performance of FOMs is sensitive to the properties of the DGF. In particular, two qualities are often desired: (1) the ratio between the diameter of the feasible domain (as measured with the DGF) and the strong convexity modulus, with respect to a given norm, of the DGF must be as small as possible; and (2) projections with respect to the DGF onto the feasible set should take linear time in the dimension of the set.

In EFGs, the only DGF family that satisfies the second requirement is based on the framework of *dilated* regularization introduced by Hoda et al. [22]. Within this framework, Kroer et al. [27] gave the first explicit strong convexity bounds based on the dilation framework, specifically for the *dilated entropy DGF*. By combining optimistic regret minimizers for general convex sets with this DGF, one gets an algorithm that achieves a T^{-1} convergence rate for two-player zero-sum EFGs. Subsequent work by Farina et al. [16] introduced the *dilated global entropy* DGF with an improved diameter-to-strong-convexity ratio. By plugging their DGF results into the genereic OMD regret bound, one immediately achieves a regret bound of $\mathcal{O}(\|\mathcal{Q}\|_1\sqrt{\log|\mathcal{A}|}\sqrt{T})$. This was the state-of-the-art regret bound when introduced, in terms of dependence on game constants. Moreover, until now, it was the best bound known to be achievable through the direct application of OMD regret bounds combined with DGF results. However, Farina et al. [20] developed a, seemingly, different approach based on *kernelization*, which is a way to simulate, in linear time in the EFG size, the results of applying optimistic multiplicative weights (OMWU) on the normal-form reduction of an EFG. They call their algorithm *KOMWU*. KOMWU achieves a better, and now state-of-the-art, regret bound $\mathcal{O}(\sqrt{\log|\mathcal{V}|}\sqrt{T})$ in online learning with full-information feedback. Based on their result, two open questions emerged: 1) is this the best possible regret bound that one can achieve? 2) is it possible to achieve such a bound directly using the standard OMD machinery, without resorting to this kernelization trick? Bai et al. [2] made highly interesting progress on the second question: they show that, in fact, the KOMWU algorithm is iterate equivalent to OMD, with the specific version of dilated entropy that uses weight one everywhere. However, their result only shows a state-of-the-art rate by equivalence to KOMWU, and it is still unknown whether this state-of-the-art rate is achieveable directly through results on DGF properties and the standard OMD regret bound. In this paper, we answer these two open questions, by answering the following question:

What is the optimal DGF for FOMs in solving EFGs?

We show that weight-one dilated entropy (DilEnt) is indeed the optimal DGF for FOMs in solving EFGs with full-information feedback, in terms of the diameter-to-strong-convexity ratio ($|\mathcal{D}|/\mu$), up to logarithmic factors. We note that the diameter-to-strong-convexity ratio of the regularizer is a key factor in the performance of FOMs. Intuitively, performance degrades as the diameter increases (since there is more "space" to search), and improves as the regularizer becomes more bowl-shaped (i.e., strongly convex). Consequently, a smaller diameter-to-strong-convexity ratio leads to better performance of the corresponding FOM. Our contributions can be summarized as follows:

- We introduce a pair of primal-dual treeplex norms for the extensive-form decision space. These norms establish an improved framework for analyzing FOMs in EFGs, leading to results with better dependence on the size of the game. Based on this framework, we derive a new state-of-the-art

[1] In CFR, the gradient steps are projected onto the nonnegative cone locally at each decision point of the game, and then renormalized to be a valid probability distribution over the actions.

Regularizer	Norm pair	$\lvert\mathcal{D}\rvert/\mu$ ratio	Max gradient norm
Dilated Entropy [27]	ℓ_1 and ℓ_∞ norms	$\mathcal{O}(2^D \lVert\mathcal{Q}\rVert_1^2 \log \lvert\mathcal{A}\rvert)$	≤ 1
Dilated Gl. Entropy [16]	ℓ_1 and ℓ_∞ norms	$\mathcal{O}(\lVert\mathcal{Q}\rVert_1^2 \log \lvert\mathcal{A}\rvert)$	≤ 1
DilEnt (**this paper**)	treeplex norms	$\ln \lvert\mathcal{V}\rvert$	≤ 1

Table 1: Comparison of the diameter-to-strong-convexity ($\lvert\mathcal{D}\rvert/\mu$) ratio with prior results in DGFs for EFGs, where the "Norm pair" indicates the primal norm used in establishing the strong convexity, and its dual. "Max gradient norm" indicates the maximum norm—measured in the dual of the norm with respect to which each DGF is strongly convex—of any reward vector, or the gradient of utility function, that can be encountered during optimization, assuming that all payoffs at the terminal nodes of the EFG are in the range $[0, 1]$. D denotes the depth of the tree, $\lVert\mathcal{Q}\rVert_1$ the tree size (see Section 3), $\lvert\mathcal{A}\rvert$ the maximum number of actions, and $\lvert\mathcal{V}\rvert$ the number of reduced normal-form strategies. We remark that $\ln \lvert\mathcal{V}\rvert \leq \mathcal{O}(\lVert\mathcal{Q}\rVert_1 \log \lvert\mathcal{A}\rvert)$.

diameter-to-strong-convexity ratio among all known DGFs for the EFG strategy spaces (see Table 1 for comparison). By combining this new results with the standard OMD regret bound, we establish a regret upper bound that aligns with the results achieved by KOMWU.

- By establishing a matching regret lower bound, we identify a minimum diameter-to-strong-convexity ratio for any regularizer. We find that the DilEnt regularizer achieves the optimal ratio up to logarithmic factors, making it a natural candidate for FOMs on EFGs.
- An advantage of our new results, as compared to showing a regret bound through the KOMWU equivalence, is that our DGF result can also be combined with other algorithmic setups. As an example of our results, we show that by equipping Clairvoyant OMD [19] with the DilEnt DGF, we enable convergence to a coarse correlated equilibrium at a rate of $\mathcal{O}(n \log \lvert\mathcal{V}\rvert \log T/T)$ in n-player EFGs. This improves upon the previous results of $\mathcal{O}(n \log \lvert\mathcal{V}\rvert \log^4 T/T)$ by Farina et al. [20] and $\mathcal{O}(n \cdot \lvert\mathcal{J} \times \mathcal{A}\rvert \lVert\mathcal{Q}\rVert_1^2 \log T/T)$ by Farina et al. [18], establishing a new state-of-the-art rate.

2 Related Works

Equilibrium Computation in General-Sum Extensive-Form Games The first line work in the equilibrium computation on general-sum EFGs used linear program (LP) which can be solved efficiently [35, 24]. However, due to the large exponent of LP solvers, it is impractical to run such algorithms on large-scale games. The modern equilibrium computation using fast-iterative methods: Syrgkanis et al. [39] introduced the RVU property on the regret bound for a broad class of optimistic no-regret learning algorithms. With that property, they demonstrated that the individual regret of each player grows as $T^{1/4}$ in general games, thus leading to a $T^{-3/4}$ converge rate to the coarse correlated equilibrium (CCE). A near-optimal bound of order $\log^4(T)$ was established by Daskalakis et al. [13], which implies a fast convergence rate of order $\widetilde{\mathcal{O}}(1/T)$. Subsequent work by Farina et al. [20] generalized the result to a class of polyhedral games that includes EFG. Concurrently, Piliouras et al. [36] introduced the Clairvoyant MWU. Although the algorithm is not non-regret learning, a subset of the steps converge to CCE with a rate of $\log T/T$. Farina et al. [19] showed that the algorithm is an instantiation of the conceptual proximal methods, which has been studied in the literature of FOMs [10, 33]. Using another technique, Farina et al. [18] also achieved this rate with worse game size dependence. Another class of fast iterative methods follows from counterfactual regret minimization (CFR) [43], which guarantees a regret bound of order $\sqrt{T}$. Farina et al. [14] showed that running OMWU at each decision point achieves a $T^{1/4}$ external regret, thus leading to a $T^{-3/4}$ approximation rate to the CCE. Although CFR has a weaker guarantee on convergence rate, variants of the algorithm are widely used in practice due to their superior practical performance [26].

Regret Lower Bounds in Extensive-Form Games Several works have studied lower bounds in EFGs across various settings. Koolen et al. [25] established a lower bound dependent on the number of orthogonal strategy profiles in the decision set for structured games, including EFGs, resulting in a bound of $\Omega(\sqrt{T \log \lvert\mathcal{A}\rvert})$. Syrgkanis et al. [39] demonstrated that in two-player zero-sum games, if one player uses MWU while the other best responds, the former must endure a regret of at least

https://doi.org/10.52202/079017-2804

$\Omega(\sqrt{T})$. Similarly, Chen and Peng [11] gave the same lower bound when both players use MWU. For equilibrium computation, Anagnostides et al. [1] analyzed the sparse-CCE in EFGs, showing that under certain assumptions, no polynomial-time algorithm can learn an ε-CCE with less than $2^{\log_2^{1/2-o(1)}|\mathcal{T}|}$ oracle accesses to the game for even constantly large $\varepsilon > 0$, where $|\mathcal{T}|$ is the number of nodes of the EFG. In the context of stochastic bandit, Bai et al. [3], Fiegel et al. [21] investigated online learning in EFGs with bandit feedback, establishing matched lower and upper bounds.

3 Preliminaries

General Notation We use lowercase boldface letters, such as $\mathbf{x}$, to denote vectors. Let $\mathbf{x} \odot \mathbf{y}$ represent the element-wise product of two vectors, and $|\mathbf{x}|$ the element-wise absolute value. For an index set $\mathcal{C}$, denote by $\mathbf{x}[\mathcal{C}] \in \mathbb{R}^{\mathcal{C}}$ the entries of $\mathbf{x}$ at indices in $\mathcal{C}$, and by $|\mathcal{C}|$ the set cardinality. Let $[\![k]\!]$ be the set $\{1, 2, \ldots, k\}$ and $\varnothing$ the empty set. Denote the simplex over the set $\mathcal{C}$ by $\Delta^{\mathcal{C}}$. The logarithm of x to base 2 is denoted as $\log x$. For non-negative sequences $\{a_n\}$ and $\{b_n\}$, $a_n \leq \mathcal{O}(b_n)$ or $b_n \geq \Omega(a_n)$ indicates the existence of a global constant $C > 0$ such that $a_n \leq C b_n$ for all $n > 0$.

Extensive-Form Games An extensive-form game (EFG) is an n-player game with a sequential structure that can be represented using a tree. A detailed definition of EFG is available in Appendix A. Each node represents a game state where an agent (a.k.a. player) $i \in [\![n]\!]$ or the environment takes action. We use superscript (i) to denote properties of player i, but also omitting the superscript when context allows. Internal nodes branch into feasible actions. At these nodes, the designated player selects an action, advancing the game to the subsequent state according to the tree. The game concludes at a terminal node $z \in \mathcal{Z}$, where players receive a reward $\mathbf{u}[z]$. The goal of each player is to maximize their expected reward. We assume the reward for each player is bounded by 1 as follows:

Assumption 3.1. The reward received by player i at any terminal node $z \in \mathcal{Z}$ satisfies $\mathbf{u}^{(i)}[z] \in [0, 1]$.

Tree-Form Sequential Decision Process In an EFG, an individual player i's decision problem can be modeled by a tree-form sequential decision process (TFSDP). Let $\mathcal{J}$ denote the set of decision points, where each point $j \in \mathcal{J}$ corresponds to an information set in the EFG. At each decision point, the player is provided with a set of available actions $\mathcal{A}_j$ and must select an action $a \in \mathcal{A}_j$. After an action a is taken at decision point j, the game either concludes before the player acts again or continues to a set of possible next decision points determined by actions of other players or by stochastic events. We denote the set of potential subsequent decision points as $\mathcal{C}_{ja} \subseteq \mathcal{J}$, which are reached immediately after action a at decision point j. The tree structure guarantees non-overlapping successors, meaning $\mathcal{C}_{ja} \cap \mathcal{C}_{j'a'} = \varnothing$ for any distinct pairs ja and $j'a'$, where $j \neq j'$ or $a \neq a'$. This encapsulation of all past actions and outcomes at each decision point is known as *perfect recall*.

The point-action pair ja, where action a is taken at decision point j, is referred to as an *observation point*. This leads to a new state influenced by other agents and the environment. We denote the set of all point-action pairs as $\Sigma^+ := \{ja \mid j \in \mathcal{J}, a \in \mathcal{A}_j\}$. Each decision point $j \in \mathcal{J}$ has a parent p_j, the last observation point on the path from the root of the decision process to j. If no action precedes j, p_j defaults to the special observation point $\varnothing$. We define $\Sigma := \Sigma^+ \cup \{\varnothing\}$ as the set of observation points, each also called a *sequence*. The total set of points in the TFSDP, $\mathcal{H} := \mathcal{J} \cup \Sigma$, includes both decision points and sequences. We use $h \in \mathcal{H}$ for unspecified point types. The TFSDP concludes at terminal observation points $\mathcal{E} = \{\sigma \in \Sigma : \mathcal{C}_\sigma = \varnothing\}$, where reward $\mathbf{r}[\sigma]$ is observed. Under Assumption 3.1, it holds that $\mathbf{r}[\sigma] \in [0, 1]$.

Strategies and Transition Kernels A strategy profile for a player in a TFSDP is an assignment of probability distributions over actions $\mathcal{A}_j$ at each decision point $j \in \mathcal{J}$. As customary when using convex optimization techniques in EFGs, we represent strategies in the *sequence-form representation* [41]. This representation stores a player's strategy as a vector whose entries represent the probability of *all* of the player's actions on the *path* from the root to the points. Since products of probabilities on paths are stored directly as variables, expected utilities are *multilinear* in the sequence-form representation of the players' strategies. For symmetry reasons which will become apparent later, we slightly depart from the typical definition of the sequence form, by storing the product of a player's action probabilities on paths from the root to *all* points in the tree—not only those that belong to the player. We call this representation the *extended sequence-form representation*. For an extended

sequence-form strategy to be valid, probability conservation constraints must be satisfied at every point of the tree. Specifically, the set of all valid extended sequence-form strategies is given by

$$\widehat{\mathcal{Q}} := \left\{ \mathbf{x} \in [0,1]^{\mathcal{H}} : \begin{array}{ll} \mathbf{x}[\varnothing] = 1 & \\ \mathbf{x}[\sigma] = \mathbf{x}[j] & \forall \sigma \in \Sigma \setminus \mathcal{E}, j \in \mathcal{C}_\sigma \\ \mathbf{x}[j] = \sum_{a \in \mathcal{A}_j} \mathbf{x}[ja] & \forall j \in \mathcal{J} \end{array} \right\},$$

The distribution of observation outcomes at each observation point, or the transition kernel, is determined by the strategy played by other agents as well as the environment. It can viewed as an opponent who only acts at the observation points. This allows us to encode the transition kernel using a vector $\mathbf{y} \in [0,1]^{\mathcal{H}}$ similar to the sequence-form strategy. The entry corresponding to each point $h \in \mathcal{H}$ represents the product of transition probabilities on the path from the root to the point. Formally, the transition kernel space is given by

$$\widehat{\mathcal{Y}} := \left\{ \mathbf{y} \in [0,1]^{\mathcal{H}} : \begin{array}{ll} \mathbf{y}[\varnothing] = 1 & \\ \mathbf{y}[\sigma] = \sum_{j \in \mathcal{C}_\sigma} \mathbf{y}[j] & \forall \sigma \in \Sigma \setminus \mathcal{E} \\ \mathbf{y}[j] = \mathbf{y}[ja] & \forall j \in \mathcal{J}, a \in \mathcal{A}_j \end{array} \right\},$$

We parameterize the vector spaces primarily over the terminals, since the reach of internal points can be uniquely determined by terminal reaches. We define the compressed extensive-form decision space $\mathcal{Q} := \{\mathbf{x}[\mathcal{E}] \mid \mathbf{x} \in \widehat{\mathcal{Q}}\}$ and the compressed transition kernel $\mathcal{Y} := \{\mathbf{y}[\mathcal{E}] \mid \mathbf{y} \in \widehat{\mathcal{Y}}\}$, representing the projection of the corresponding spaces onto the vector space generated by terminal observation points. The existing of one-to-one mapping guarantees that $\mathcal{Q}$ and $\mathcal{Y}$ are homogeneous to $\widehat{\mathcal{Q}}$ and $\widehat{\mathcal{Y}}$. For each non-terminal point $h \in \mathcal{H} \setminus \mathcal{E}$, we denote by $\mathbf{x}[h]$ the value of $\widehat{\mathbf{x}}[h]$ in $\widehat{\mathcal{Q}}$, corresponding to the compressed strategy profile $\mathbf{x}$. When the agent adopts strategy profile $\mathbf{x} \in \mathcal{Q}$ while the transition kernel aligns with $\mathbf{y} \in \mathcal{Y}$, the reach probability of terminal point $\sigma \in \mathcal{E}$ is given by $\mathbf{x}[\sigma]\mathbf{y}[\sigma]$. The expected reward of the player can be computed from $u(\mathbf{x}; \mathbf{w}) = \langle \mathbf{x}, \mathbf{w} \rangle$, where $\mathbf{w} := \mathbf{r} \odot \mathbf{y}$ is the reward vector, or the gradient of utility.

To assess the complexity of the game, we use several complexity measures for the extensive-form decision space. We define the *tree size* and *leaf count*, denoted as $\|\mathcal{Q}\|_1$ and $\|\mathcal{Q}\|_\perp$, as the maximum number of observation points and terminal observation points that can be reached among all pure strategy profiles, respectively. Formally, we write

$$\|\mathcal{Q}\|_1 := \sup_{\mathbf{x} \in \mathcal{Q}} \|\mathbf{x}[\Sigma]\|_1 = \sup_{\mathbf{x} \in \mathcal{Q}} \sum_{\sigma \in \Sigma} \mathbf{x}[\sigma], \qquad \|\mathcal{Q}\|_\perp := \sup_{\mathbf{x} \in \mathcal{Q}} \|\mathbf{x}\|_1 = \sup_{\mathbf{x} \in \mathcal{Q}} \sum_{\sigma \in \mathcal{E}} \mathbf{x}[\sigma],$$

where we implicitly extend the domain of $\mathbf{x}$ to $\widehat{\mathcal{Q}}$ when writing $\mathbf{x}[\Sigma]$. We further define $\mathcal{V} := \mathcal{Q} \cap \{0,1\}^{\mathcal{E}}$ as the vertices in the extensive-form decision space. Each vertex refers to a pure strategy profile of the player, which reduced to a norm-form strategy. The number of reduced normal-form strategy is given by $|\mathcal{V}|$. We remark that both $\|\mathcal{Q}\|_1$ and $|\mathcal{V}|$ have been used in the literature [e.g., 20].

Subtree As we will often incorporate the recursive structure in TFSDP, we define a *subtree* as the subgame starting from some internal point $h \in \mathcal{H}$. For two points $h, h' \in \mathcal{H}$, we write $h' \succeq h$ if h' is reachable from h in TFSDP. Let $\mathcal{H}_h = \{h' \in \mathcal{H} : h' \succeq h\}$ and $\mathcal{E}_h = \{\sigma \in \mathcal{E} : \sigma \succeq h\}$ be the sets of points and terminals reachable from h. We denote by $\mathcal{Q}_h$ and $\mathcal{Y}_h$ the projected spaces of $\mathcal{Q}$ and $\mathcal{Y}$ over $[0,1]^{\mathcal{E}_h}$, with restrictions $\mathbf{x}[h] = 1$ and $\mathbf{y}[h] = 1$, respectively. Formally, for any point $h \in \mathcal{H}$, we define the compressed projected decision space as $\mathcal{Q}_h := \{\mathbf{x}[\mathcal{E}_h] \mid \mathbf{x} \in \mathcal{Q}, \mathbf{x}[h] = 1\}$. From the definition of the compressed extensive-form decision space, this space can be seen as a projection of

$$\widehat{\mathcal{Q}}_h := \left\{ \mathbf{x} \in [0,1]^{\mathcal{H}_h} : \begin{array}{ll} \mathbf{x}[h] = 1 & \\ \mathbf{x}[\sigma] = \mathbf{x}[j] & \forall \sigma \in \Sigma \setminus \mathcal{E}, j \in \mathcal{C}_\sigma \\ \mathbf{x}[j] = \sum_{a \in \mathcal{A}_j} \mathbf{x}[ja] & \forall j \in \mathcal{J} \end{array} \right\}.$$

Similarly, we define the compressed projected transition kernel space as $\mathcal{Y}_h := \{\mathbf{y}[\mathcal{E}_h] \mid \mathbf{y} \in \mathcal{Y}, \mathbf{y}[h] = 1\}$. It is important to note that this space exhibits a similar closed form to that of the compressed extensive-form decision space.

https://doi.org/10.52202/079017-2804

Proximal Methods We review the standard objects and notations that relate to proximal methods. For a given decision set $\mathcal{Q}$, the proximal method requires a distance generating function (DGF) $\varphi : \mathcal{Q} \to \mathbb{R}$ defined on the decision set. The algorithm is valid when the DGF is μ-strongly convex with respect to some norm $\|\cdot\|$. The DGF induces a generalized notion of distance $\mathcal{D}_\varphi : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}_{\geq 0}$, referred to as the *Bregman Divergence*, which is defined by

$$\mathcal{D}_\varphi(\widehat{\mathbf{x}} \| \mathbf{x}) := \varphi(\widehat{\mathbf{x}}) - \varphi(\mathbf{x}) - \langle \nabla \varphi(\mathbf{x}), \widehat{\mathbf{x}} - \mathbf{x} \rangle.$$

We define the *proximal operator* with respect to the feasible space $\mathcal{Q}$ and the DGF φ. Given a pivot point $\mathbf{x}$ and a gradient vector $\mathbf{g} \in \mathbb{R}^{\mathcal{E}}$, the proximal operator $\Pi_\varphi(\mathbf{g}, \mathbf{x})$ generalizes the notion of a gradient ascent step, and is defined as

$$\Pi_\varphi(\mathbf{g}, \mathbf{x}) := \operatorname*{argmax}_{\widehat{\mathbf{x}} \in \mathcal{Q}} \{ \langle \mathbf{g}, \widehat{\mathbf{x}} \rangle - \mathcal{D}_\varphi(\widehat{\mathbf{x}} \| \mathbf{x}) \}.$$

For the extensive-form decision space $\mathcal{Q}$, the DGF is usually restricted to the dilated DGF [22] so that the proximal operator can be efficiently computed. Moreover, it is well known that the proximal operator is Lipschitz continuous: [e.g. 33, Lemma 2.1]

Lemma 3.2. For any $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^{\mathcal{E}}$, it satisfies that $\|\Pi_\varphi(\mathbf{g}, \mathbf{x}) - \Pi_\varphi(\mathbf{g}', \mathbf{x})\| \leq \mu^{-1} \|\mathbf{g} - \mathbf{g}'\|_*$.

4 Primal-Dual Treeplex Norms

We first introduce the treeplex ℓ_1 norm $\|\cdot\|_{\mathcal{H},1}$ and the treeplex ℓ_∞ norm $\|\cdot\|_{\mathcal{H},\infty}$, which are a primal-dual norm pair defined over the vector space $\mathbb{R}^{\mathcal{E}}$, with respect to a given TFSDP with the point set $\mathcal{H}$. As we will show later, these norms enable a better framework for analyzing FOMs in EFGs. Specifically, in the analysis of OMD, we use the fact that the ℓ_∞ norm for any feasible reward vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_\infty \leq 1$. Although treeplex ℓ_∞ norm is a relaxation of the ℓ_∞ norm, it can still preserve the same guarantee such that $\|\mathbf{w}\|_{\mathcal{H},\infty} \leq 1$. With the relaxation, we have that the treeplex ℓ_1 norm generates a smaller distance compared to the ℓ_1 norm, which allows us to provide a better strong convexity modulus for the regularizer, finally improving the induced regret upper bound.

Both treeplex norms are defined as the support functions with respect to the vector of element-wise absolute values. Specifically, the support function of treeplex ℓ_1 norm is defined using the transition kernel space $\mathcal{Y}$, while the support function of treeplex ℓ_∞ norm is defined using the extensive-form decision space $\mathcal{Q}$. Formally, for some vector $\mathbf{u} \in \mathbb{R}^{\mathcal{E}}$, we denote

$$\|\mathbf{u}\|_{\mathcal{H},1} := \sup_{\mathbf{y} \in \mathcal{Y}} \langle |\mathbf{u}|, \mathbf{y} \rangle = \sup_{\mathbf{y} \in \mathcal{Y}} \sum_{\sigma \in \mathcal{E}} |\mathbf{u}[\sigma]| \cdot \mathbf{y}[\sigma],$$

$$\|\mathbf{u}\|_{\mathcal{H},\infty} := \sup_{\mathbf{x} \in \mathcal{Q}} \langle |\mathbf{u}|, \mathbf{x} \rangle = \sup_{\mathbf{x} \in \mathcal{Q}} \sum_{\sigma \in \mathcal{E}} |\mathbf{u}[\sigma]| \cdot \mathbf{x}[\sigma].$$

We remark that the treeplex ℓ_∞ norm has been used by Zhang et al. [42] for analyzing low-degree swap regret minimization. When the EFG degenerates to an NFG (normal-form game), i.e., $|\mathcal{J}| = 1$, the extensive-form decision space $\mathcal{Q}$ in the TFSDP becomes a simplex $\Delta^{\mathcal{E}}$, and the transition kernel $\mathcal{Y} = \{\mathbf{1}\}$ only contains the all-one vector. It follows that the treeplex ℓ_1 norm $\|\cdot\|_{\mathcal{H},1}$ and the treeplex ℓ_∞ norm $\|\cdot\|_{\mathcal{H},\infty}$ degenerate to the conventional ℓ_1 norm and ℓ_∞ norm for the vector space, respectively. The following lemma verifies that both treeplex ℓ_1 norm and treeplex ℓ_∞ norm are norms in the technical sense. The missing proofs in this section are provided in Appendix C.

Lemma 4.1. The functions $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ are norms defined on the space $\mathbb{R}^{\mathcal{E}}$.

Thanks to the recursive structure of TFSDP, the maximization among $\mathcal{Y}$ or $\mathcal{Q}$ in the treeplex norms can be decomposed at each point $h \in \mathcal{H}$. This decomposition allows us to compute both treeplex norms in a recursive manner.

Lemma 4.2. Let $\mathbf{u} \in \mathbb{R}^{\mathcal{E}_h}$ be a vector with respect to some point $h \in \mathcal{H}$. The treeplex ℓ_1 norm and the treeplex ℓ_∞ norm of vector $\mathbf{u}$ over $\mathcal{H}_h$ can be computed recursively as follows.

- If $h = \sigma \in \mathcal{E}$ is a terminal observation point, then:

$$\|\mathbf{u}\|_{\mathcal{H}_\sigma,1} := |\mathbf{u}[\sigma]|, \qquad \|\mathbf{u}\|_{\mathcal{H}_\sigma,\infty} := |\mathbf{u}[\sigma]|.$$

- If $h = j \in \mathcal{J}$ is a decision point, then:
$$\|\mathbf{u}\|_{\mathcal{H}_j,1} := \sum_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},1}, \qquad \|\mathbf{u}\|_{\mathcal{H}_j,\infty} := \max_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},\infty}.$$
- If $h = \sigma \in \Sigma \setminus \mathcal{E}$ is a non-terminal observation point, then:
$$\|\mathbf{u}\|_{\mathcal{H}_j,1} := \max_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},1}, \qquad \|\mathbf{u}\|_{\mathcal{H}_j,\infty} := \sum_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},\infty}.$$

Equipped with the recursive formula, we are able to show that the two treeplex norms with respect to the same TFSDP are a pair of primal-dual norms.

Theorem 4.3. We have $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ is a pair of primal-dual norms for a given TFSDP with point set $\mathcal{H}$. Specifically, for any vector $\mathbf{u} \in \mathbb{R}^{\mathcal{E}}$,
$$\|\mathbf{u}\|^*_{\mathcal{H},1} := \sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}}} \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\|\mathbf{v}\|_{\mathcal{H},1}} = \|\mathbf{u}\|_{\mathcal{H},\infty}.$$

Furthermore, the recursive formula also enables us to bound the treeplex norms for specific vectors.

Lemma 4.4. We have $\|\mathbf{x}\|_{\mathcal{H},1} = 1$ for any strategy profile $\mathbf{x} \in \mathcal{Q}$, $\|\mathbf{y}\|_{\mathcal{H},\infty} = 1$ for any transition kernel $\mathbf{y} \in \mathcal{Y}$, and $\|\mathbf{w}\|_{\mathcal{H},\infty} \le 1$ for any feasible reward vector $\mathbf{w}$ under Assumption 3.1.

5 Metric Properties of the DilEnt Regularizer and Improved Regret Bounds

In this section, we study the strong convexity modulus of the weight-one dilated entropy (DilEnt) function with respect to the treeplex norms defined above. The DilEnt regularizer is an instantiation of the more general dilated DGFs framework [22]. Specifically, a dilated DGF for an extensive-form decision space is constructed by taking a weighted sum over suitable *local* regularizers φ_j for each $j \in \mathcal{J}$, and is of the form
$$\varphi : \mathcal{Q} \ni \mathbf{x} \mapsto \sum_{j\in\mathcal{J}} \alpha_j \varphi_j^{\square}(\mathbf{x}[p_j], \{\mathbf{x}[ja]\}_{a\in\mathcal{A}_j}),$$
where
$$\varphi_j^{\square}(\mathbf{x}[p_j], \{\mathbf{x}[ja]\}_{a\in\mathcal{A}_j}) := \begin{cases} 0 & \text{if } \mathbf{x}[p_j] = 0 \\ \mathbf{x}[p_j]\varphi_j\left(\frac{\{\mathbf{x}[ja]\}_{a\in\mathcal{A}_j}}{\mathbf{x}[p_j]}\right) & \text{otherwise} \end{cases}$$
and $\alpha_j > 0$ are flexible weight terms that can be chosen to ensure good properties. Note that we have implicitly extended the domain of $\mathbf{x}$ to $\widehat{\mathcal{Q}}$. Each local function $\varphi_j : \Delta^{\mathcal{A}_j} \to \mathbb{R}$ is required to be continuously differentiable and strongly convex on the relative interior of the local probability simplex $\Delta^{\mathcal{A}_j}$. They show that the proximal steps on the dilated DGF can be efficiently computed, provided that the proximal steps for each individual φ_j can be efficiently computed.

The DilEnt regularizer $\varphi_{\mathbf{1}} : \mathcal{Q} \to \mathbb{R}$ is a specific instantiation of the dilated DGF with $\alpha_j = 1$ and each local regularizer d_j being the negative entropy function. It has been used as a specific instantiation for practical implementations [e.g. 28]. The function has the following closed form.
$$\varphi_{\mathbf{1}} : \mathbf{x} \mapsto \sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{x}[ja] \ln\left(\frac{\mathbf{x}[ja]}{\mathbf{x}[p_j]}\right).$$
Prior to our work, weighted variants of the dilated entropy had been the only variants known to have concrete strong convexity bounds, all with weights that grew with the size of the decision space beneath a given decision point [27, 16]. These results used the standard ℓ_1 norm as the corresponding norm for showing strong convexity. With the help of our new primal-dual treeplex norms, we can show that the DilEnt regularizer enjoys very strong properties on the extensive-form decision space. We inspect the following $\mathbf{y}$-weighted dilated entropy for $\mathbf{y} \in \mathcal{Y}$:
$$\varphi_{\mathbf{y}} : \mathbf{x} \mapsto \sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{y}[ja]\mathbf{x}[ja] \ln\left(\frac{\mathbf{x}[ja]}{\mathbf{x}[p_j]}\right).$$
By showing that the function is equivalent to the $\mathbf{y}$-weighted negative entropy on the terminal reach, we are able to prove $\varphi_{\mathbf{y}}$ is 1-strongly convex with respect to the $\mathbf{y}$-weighted ℓ_1 norm. Since the difference $\varphi_{\mathbf{1}} - \varphi_{\mathbf{y}}$ is a summation of convex functions, it can be finally demonstrated that the DilEnt regularizer is 1-strongly convex with respect to the treeplex ℓ_1 norm.

https://doi.org/10.52202/079017-2804

Lemma 5.1. The weight-one dilated entropy (DilEnt) is 1-strongly convex within the extensive-form decision space $\mathcal{Q}$ with respect to the treeplex ℓ_1 norm $\|\cdot\|_{\mathcal{H},1}$. Specifically, for any vector $\mathbf{z} \in \mathbb{R}^{\mathcal{E}}$ and strategy profile $\mathbf{x} \in \mathcal{Q}$, we have that $\|\mathbf{z}\|^2_{\nabla^2 \varphi_{\mathbf{1}}(\mathbf{x})} \geq \|\mathbf{z}\|^2_{\mathcal{H},1}$.

The complete proof is provided in Appendix D. Next, we inspect the diameter of the decision space measured by DilEnt. Using an induction statement, we show that $\ln|\mathcal{V}| \leq \varphi_{\mathbf{1}}(\mathbf{x}) \leq 0$ holds for any strategy profile $\mathbf{x} \in \mathcal{Q}$. By choosing the initial strategy that minimizes the DilEnt regularizer, we upper bound the diameter of the sequence-form decision space with respect to the DilEnt regularizer.

Lemma 5.2. Let $\mathbf{x}_1 := \operatorname{argmin}_{\mathbf{x}\in\mathcal{Q}} \varphi_{\mathbf{1}}(\mathbf{x})$ be the strategy profile minimize the DilEnt regularizer. The Bregman divergence generated by the DilEnt regularizer between $\mathbf{x}_1$ and any $\mathbf{x}_* \in \mathcal{Q}$ can be upper bounded by $\mathcal{D}_{\varphi_{\mathbf{1}}}(\mathbf{x}_*, \mathbf{x}_1) \leq \ln|\mathcal{V}|$.

By combining the above two lemmas, we have that the DilEnt regularizer achieves $|\mathcal{D}|/\mu \leq \ln|\mathcal{V}|$. Using this result, we can establish performance guarantees for FOMs with the DilEnt regularizer. We list these results in the following sections.

5.1 Results on Online Mirror Descent

We first inspect online learning in TFSDP with full-information feedback. Consider the use of (Predictive) OMD [12]. The pseudocode of the algorithm can be found in Algorithm 1 in Appendix B. The algorithm starts from $\widetilde{\mathbf{x}}_1 \leftarrow \operatorname{argmin}_{\mathbf{x}\in\mathcal{Q}} \varphi(\mathbf{x})$ and follows a straightforward structure in each episode t: Take a proximal gradient step from $\widetilde{\mathbf{x}}_t$ according to the prediction $\mathbf{m}_t$ to get the policy $\mathbf{x}_t$; Execute policy $\mathbf{x}_t$; Take another proximal gradient step from $\widetilde{\mathbf{x}}_t$ according to the observed reward vector $\mathbf{w}_t$ to get $\widetilde{\mathbf{x}}_{t+1}$. The algorithm takes some DGF φ to execute proximal steps:

$$\mathbf{x}_t \leftarrow \Pi_{\varphi}(\eta \mathbf{m}_t, \widetilde{\mathbf{x}}_t), \quad \widetilde{\mathbf{x}}_{t+1} \leftarrow \Pi_{\varphi}(\eta \mathbf{w}_t, \widetilde{\mathbf{x}}_t).$$

The value of prediction $\mathbf{m}_t$ depends on the specific variant used (e.g. $\mathbf{m}_t \leftarrow \mathbf{w}_{t-1}$ in Optimistic OMD). For the non-predictive variant, we set $\mathbf{m}_t \leftarrow \mathbf{0}$, and thus $\mathbf{x}_t = \widetilde{\mathbf{x}}_t$. It is known that the algorithm has the following regret bound with respect to a given pair of primal-dual norms.

Theorem 5.3 (Regret Bound for (Predictive) OMD, Rakhlin and Sridharan [37], Syrgkanis et al. [39])**.** Let $\|\cdot\|$ and $\|\cdot\|_*$ be a pair of primal-dual norm defined on $\mathbb{R}^{\mathcal{E}}$. Let φ be a DGF that is μ-strongly convex on $\|\cdot\|$. Denote $\mathbf{w}_t$ as the reward gradient received in episode t. The cumulative regret of running (Predictive) OMD with DGF φ and learning rate η can be upper bounded by

$$\text{Regret}(T) := \max_{\mathbf{x}_*\in\mathcal{Q}} \sum_{t=1}^{T} \langle \mathbf{x}_* - \mathbf{x}_t, \mathbf{w}_t \rangle \leq \frac{1}{\eta} \mathcal{D}_{\varphi}(\mathbf{x}_*, \mathbf{x}_1) + \frac{\eta}{2\mu} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{m}_t\|^2_*.$$

Consider using non-predictive OMD with the DilEnt regularizer $\varphi_{\mathbf{1}}$. The performance of the algorithm can be analyzed by selecting the treeplex ℓ_1 norm and the treeplex ℓ_∞ norm as the desired pair of primal-dual norms. Using the diameter-to-strong-convexity ratio of DilEnt, we can immediately get a regret upper bound that recovers the state-of-the-art result given by KOMWU [20].

Theorem 5.4. Let φ be a regularizer for extensive-form decision space $\mathcal{Q}$ which is μ-strongly convex on $\|\cdot\|_{\mathcal{H},1}$ and has a diameter $|\mathcal{D}| := \sup_{x_*\in\mathcal{Q}} \mathcal{D}_{\varphi}(x_*, x_1)$. Under Assumption 3.1, the cumulative regret of running OMD with regularizer φ and learning rate $\eta := \sqrt{2|\mathcal{D}|/(\mu T)}$ is upper bounded by

$$\text{Regret}(T) \leq \sqrt{2|\mathcal{D}|/\mu}\sqrt{T}.$$

Moreover, if we use the DilEnt regularizer $\varphi_{\mathbf{1}}$ in proximal steps, the result can be specified as

$$\text{Regret}(T) \leq \sqrt{2\ln|\mathcal{V}|}\sqrt{T}.$$

5.2 Results on Clairvoyant Online Mirror Descent

Consider equilibrium computation in n-player EFGs. In this scenario, a group of agents aim to jointly learn the coarse correlated equilibrium (CCE) given only oracle access to the game (See Appendix A for detailed definition). We adopt Clairvoyant OMD to compute CCE, introduced by Piliouras et al.

[36], Farina et al. [19]. The pseudocode of the algorithm can be found in Algorithm 2 in Appendix B. The algorithm can be viewed as a specialized form of predictive OMD, in each episode $t \in [\![K]\!]$, an additional routines is introduced to compute the prediction vector $\mathbf{m}_t$ for each player i (we omit the superscript). The prediction in the episode is calculated through L steps of fixed-point iteration starting from $\mathbf{x}_{t,1} \leftarrow \widetilde{\mathbf{x}}_t$. In each step $l \in [\![L]\!]$, each player i computes proximal step

$$\mathbf{x}_{t,l+1} \leftarrow \Pi_{\varphi}(\eta \mathbf{w}_{t,l}, \widetilde{\mathbf{x}}_t).$$

where we denote by $\mathbf{w}_{t,l}$ the reward vector observed by the player joint policy is corresponding to $\mathbf{x}_{t,l}$. Clairvoyant OMD finally sets the prediction $\mathbf{m}_t$ to the iteration result $\mathbf{w}_{t,L}$. In this case, the committed policy $\mathbf{x}_t \leftarrow \Pi_{\varphi}(\eta \mathbf{m}_t, \widetilde{\mathbf{x}}_t)$ in the OMD framework is equal to $\mathbf{x}_{t,L}$ and the reward vector $\mathbf{w}_t$ is $\mathbf{w}_{t,L+1}$. We show the fixed-point iteration achieves linear convergence. Thus, the difference $\|\mathbf{w}_{t,L+1} - \mathbf{w}_{t,L}\|_{\mathcal{H},\infty}$ can be made as arbitrarily small. The proof starts from the following inequality, establishing that the reward vector is Lipschitz continuous with respect to the joint strategy:

$$\|\mathbf{w}_1^{(i)} - \mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty} \le \sum_{j=1}^{n} \|\mathbf{x}_1^{(j)} - \mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(j)},1}.$$

We denote by $\mathbf{w}_1^{(i)}$ the reward vector of player i when all the players align with joint policy $\{\mathbf{x}_1^{(j)}\}_{j=1}^n$. Together with the fact that the proximal operator is Lipschitz (Lemma 3.2), we can show that the fixed-point iteration achieves a linear convergence rate when the learning rate η is sufficiently small.

Lemma 5.5. Under Assumption 3.1, when running COMD with DilEnt, the reward vector $\mathbf{w}_{t,l}^{(i)}$ received by player i in any $(t,l) \in [\![K]\!] \times [\![L]\!]$ satisfies $\|\mathbf{w}_{t,l+1}^{(i)} - \mathbf{w}_{t,l}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \le 2(n\eta)^{l-1}$.

Therefore, with a logarithmic number of iterations, the discrepancy between the reward vector and the prediction in the OMD framework can be made as small as $\|\mathbf{w}_t - \mathbf{m}_t\|_{\mathcal{H},\infty} = \|\mathbf{w}_{t,L+1} - \mathbf{w}_{t,L}\|_{\mathcal{H},\infty} \le 1/K$. Substituting this result into Theorem 5.3 implies the average joint policy given by all $\mathbf{x}_t$ among $t \in [\![K]\!]$ episodes in Clairvoyant OMD only causes a constant regret. Using the standard online-to-batch conversion [9], we can demonstrate that Clairvoyant OMD finds an ϵ-CCE with only a near-linear number of oracle accesses to the game, establishing the new state of the art.

Theorem 5.6. Under Assumption 3.1, if every player runs Clairvoyant OMD with DilEnt regularizer and learning rate $\eta = 1/(2n)$ for K episodes. With $L = \lceil \log K \rceil$ steps of inner iterations, the average joint policy $\bar{\pi}_K$ is an ε-CCE for $\varepsilon \le \mathcal{O}(n \ln |\mathcal{V}|/K)$. This implies the algorithm converge to a CCE at rate $\mathcal{O}(n \log |\mathcal{V}| \log T/T)$ where $T = KL$ is the number of oracle access to the game.

6 Lower Bounds for Regret Minimization in EFGs and Optimality of DilEnt

In this section, we show that the DilEnt regularizer has a nearly optimal diameter-to-strong-convexity ratio within the extensive-form decision space. To establish this, we prove a lower bound for online learning in TFSDP with full-information feedback. We show that every algorithm must suffer a regret lower bound that matches our regret upper bound in Theorem 5.4. The optimality of the DilEnt regularizer is demonstrated by contradiction: If there were a regularizer with a much better diameter-to-strong-convexity ratio, then the regret of running OMD with that regularizer would violate the established regret lower bound. We prove the lower bound by constructing a hard instance that is completely random. In this scenario, no online learning algorithm can benefit from historical data, while the cumulative reward of the optimal policy benefits from the anti-concentration properties of the maximum among random distributions.

Theorem 6.1. Given a TFSDP with decision space $\mathcal{Q}$, there is an EFG satisfying Assumption 3.1 such that: when the other players are controlled by the adversary, any algorithm `Alg` incurs an expected regret of at least $\Omega(\sqrt{\|\mathcal{Q}\|_{\perp} \log |\mathcal{A}_0| T})$ for a given of episode number $T \ge \|\mathcal{Q}\|_{\perp}$, where $|\mathcal{A}_0| := \min_{j \in \mathcal{J}} |\mathcal{A}_j|$ is the size of the minimum action set.

We provide missing proofs in Appendix E. Comparing Theorem 6.1 with Theorem 5.4, we establish a lower bound for the diameter-to-strong-convexity ratio, $|\mathcal{D}|/\mu \ge \Omega(\|\mathcal{Q}\|_{\perp} \log |\mathcal{A}_0|)$ for any regularizer on the extensive-form decision space with respect to our new treeplex norms. Recall the diameter-to-strong-convexity ratio of the DilEnt regularizer is at most $|\mathcal{D}|/\mu \le \ln |\mathcal{V}|$, derived from combining Lemma 5.1 and Lemma 5.2. We establish connections between these two quantities using the following lemma, implying that the ratio achieved by the DilEnt regularizer is nearly optimal.

https://doi.org/10.52202/079017-2804

Lemma 6.2. Consider a TFSDP with a given point set $\mathcal{H}$. Define $|\mathcal{A}| := \max_{j \in \mathcal{J}} |\mathcal{A}_j|$ as the size of the largest action set. If there is no non-root observation point yields exactly one observation outcome, that is, $|\mathcal{C}_\sigma| \geq 2$ for any $\sigma \in \Sigma^+ \setminus \mathcal{E}$, then it follows that $\ln |\mathcal{V}| \leq \mathcal{O}(\|\mathcal{Q}\|_\perp \log |\mathcal{A}|)$. Without this structural condition, we have $\ln |\mathcal{V}| \leq \mathcal{O}(\|\mathcal{Q}\|_\perp \log |\mathcal{J} \times \mathcal{A}|)$ in general.

According to Lemma 6.2, if every each action set has the same number of actions $|\mathcal{A}_0| = |\mathcal{A}|$, and no non-root observation point yields only one outcome, we have $|\mathcal{D}|/\mu \leq \ln |\mathcal{V}| \leq \mathcal{O}(\|\mathcal{Q}\|_\perp \log |\mathcal{A}_0|)$, implying the DilEnt regularizer achieves the optimal diameter-to-strong-convexity ratio up to constant factors in this scenario. If the action sets vary in size, it creates a gap logarithmic to the size of the maximal action set. If there is some observation point that yields only one observation, the gap inflates with another factor of logarithmic to the number of decision points. All in all, the DilEnt regularizer achieves the optimal diameter-to-strong-convexity ratio up to only logarithmic factors.

7 Conclusion, Limitations, and Open Questions

In this paper, we introduce a new primal-dual norm pair for studying the strong convexity properties of distance-generating functions for sequence-form strategy polytopes arising in extensive-form games. Quantifying these properties is a key component in the construction of efficient first-order optimization methods for equilibrium computation. Our techniques enable us to explain the strong theoretical performance of the DilEnt regularizer, for which no meaningful strong convexity bounds were previously known. In fact, we find that among all convex regularizers for extensive-form games, DilEnt is optimal up to logarithmic factors. To establish this result, we introduced a new regret lower bound for learning in extensive-form games, which is likely of independent relevance.

We remark that our lower bound only applies to extensive-form games with full-information feedback, a setting common in self-playing algorithms. Thus, the optimality of the DilEnt regularizer may not extend to scenarios with stochastic feedback. It would be interesting to study tight lower bounds for learning under other types of feedback. While Fiegel et al. [21] gave matching lower and upper bounds under trajectory bandit feedback, results for external sampling remain open to our knowledge.

Furthermore, we can only prove tight upper and lower bounds up to constant factors for the diameter-to-strong-convexity ratio in a specific family of TFSDPs. There still remains a logarithmic gap related to the total number of sequences in general. In principle, it is possible that this gap could be further reduced, yielding a different regularizer that offers logarithmic advantages over the DilEnt regularizer. Overcoming this technical hurdle and showing that the DilEnt regularizer indeed achieves the optimal rate remains an interesting direction of research. Notably, Fiegel et al. [21] were also only able to prove lower bounds in specific games under bandit feedback, alluding to the intrinsic hardness of proving lower bounds without slight restrictions to the game class.

Acknowledgments and Disclosure of Funding

Christian Kroer was supported by the Office of Naval Research awards N00014-22-1-2530 and N00014-23-1-2374, and the National Science Foundation awards IIS-2147361 and IIS-2238960.

References

[1] Ioannis Anagnostides, Alkis Kalavasis, Tuomas Sandholm, and Manolis Zampetakis. On the complexity of computing sparse equilibria and lower bounds for no-regret learning in games. *arXiv preprint arXiv:2311.14869*, 2023.

[2] Yu Bai, Chi Jin, Song Mei, Ziang Song, and Tiancheng Yu. Efficient Φ-regret minimization in extensive-form games via online mirror descent. *Advances in Neural Information Processing Systems*, 35:22313–22325, 2022.

[3] Yu Bai, Chi Jin, Song Mei, and Tiancheng Yu. Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning*, pages 1337–1382. PMLR, 2022.

[4] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

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

[7] Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1829–1836, 2019.

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

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

[10] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[11] Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33:18990–18999, 2020.

[12] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pages 6–1. JMLR Workshop and Conference Proceedings, 2012.

[13] Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34:27604–27616, 2021.

[14] Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-predictive optimistic counterfactual regret minimization. In *International conference on machine learning*, pages 1853–1862. PMLR, 2019.

[15] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Online convex optimization for sequential decision processes and extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1917–1925, 2019.

[16] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Better regularization for sequential decision spaces: Fast convergence rates for nash, correlated, and team equilibria. In *Proceedings of the 2021 ACM Conference on Economics and Computation*, 2021.

[17] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5363–5371, 2021.

 https://doi.org/10.52202/079017-2804

[18] Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. *Advances in Neural Information Processing Systems*, 35:39076–39089, 2022.

[19] Gabriele Farina, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Clairvoyant regret minimization: Equivalence with Nemirovski's conceptual prox method and extension to general convex games. *arXiv preprint arXiv:2208.14891*, 2022.

[20] Gabriele Farina, Chung-Wei Lee, Haipeng Luo, and Christian Kroer. Kernelized multiplicative weights for 0/1-polyhedral games: Bridging the gap between learning in extensive-form and normal-form games. In *International Conference on Machine Learning*, pages 6337–6357. PMLR, 2022.

[21] Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet, and Michal Valko. Adapting to game trees in zero-sum imperfect information games. In *International Conference on Machine Learning*, pages 10093–10135. PMLR, 2023.

[22] Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2): 494–512, 2010.

[23] Wan Huang and Bernhard von Stengel. Computing an extensive-form correlated equilibrium in polynomial time. In *International Workshop on Internet and Network Economics*, pages 506–513. Springer, 2008.

[24] Albert Xin Jiang and Kevin Leyton-Brown. Polynomial-time computation of exact correlated equilibrium in compact games. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 119–126, 2011.

[25] Wouter M Koolen, Manfred K Warmuth, and Jyrki Kivinen. Hedging structured concepts. In *COLT 2010: Proceedings of the 23rd Annual Conference on Learning Theory*, pages 93–105, 2010.

[26] Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Solving large sequential games with the excessive gap technique. *Advances in neural information processing systems*, 31, 2018.

[27] Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, pages 1–33, 2020.

[28] Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. *Advances in Neural Information Processing Systems*, 34:14293–14305, 2021.

[29] Mingyang Liu, Asuman E Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *The Eleventh International Conference on Learning Representations*.

[30] Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. A policy-gradient approach to solving imperfect-information games with iterate convergence. *arXiv preprint arXiv:2408.00751*, 2024.

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

[32] John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[33] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[34] Yu Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005.

[35] Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 55(3):1–29, 2008.

[36] Georgios Piliouras, Ryann Sim, and Stratis Skoulakis. Beyond time-average convergence: Near-optimal uncoupled online learning via clairvoyant multiplicative weights update. *Advances in Neural Information Processing Systems*, 35:22258–22269, 2022.

[37] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[38] Tuomas Sandholm. The state of solving large incomplete-information games, and application to poker. *Ai Magazine*, 31(4):13–32, 2010.

[39] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

[40] Oskari Tammelin, Neil Burch, Michael Johanson, and Michael Bowling. Solving heads-up limit texas hold'em. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[41] Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

[42] Brian Hu Zhang, Ioannis Anagnostides, Gabriele Farina, and Tuomas Sandholm. Efficient Φ-regret minimization with low-degree swap deviations in extensive-form games. *arXiv preprint arXiv:2402.09670*, 2024.

[43] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

 https://doi.org/10.52202/079017-2804

A Extended Preliminaries

An extensive-form game (EFG) is an n-player game with a sequential structure representable as a tree. Let $\mathcal{T}$ be the set of all nodes in the tree, where each node $z \in \mathcal{T}$ corresponds to a game state. Each node is assigned to either a player $i \in [\![n]\!]$ or the environment for action. The subset $\mathcal{T}^{(i)} \subseteq \mathcal{T}$ comprises nodes assigned to player i. The environment, treated as a special player, acts according to a fixed distribution, modeling stochastic outcomes like card dealing in games. Each node's branches represent possible actions. Upon reaching a node, the assigned player selects an action, moving the game to the next node per the tree structure. Let $\mathcal{Z}$ be the set of terminal nodes. The game concludes when it reaches a terminal node $z \in \mathcal{Z}$, where each player i receives a reward $\mathbf{u}^{(i)}[z]$. Players aim to maximize their expected reward by reaching these terminal nodes. We assume $\mathbf{u}^{(i)}[z] \in [0, 1]$, following Assumption 3.1.

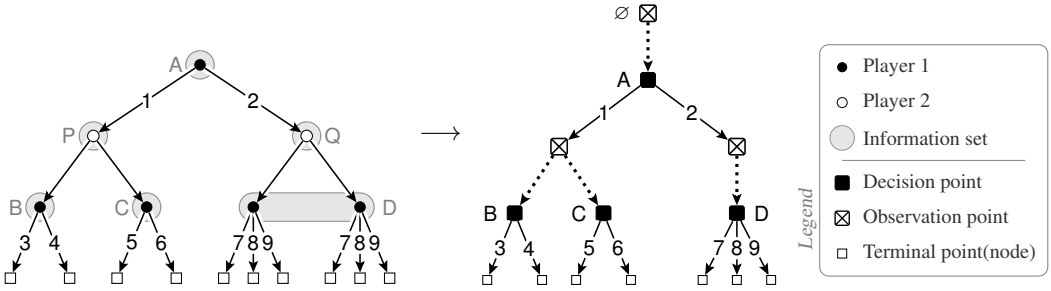

Figure 1: An two-player extensive-form game and the corresponding TFSDP of player 1. The TFSDP has decision point $\mathcal{J} = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}\}$. It has tree size $\|\mathcal{Q}\|_1 = 4$ and leaf count $\|\mathcal{Q}\|_\perp = 2$, both given by the pure strategy $\{\mathsf{A} \to 1, \mathsf{B} \to 3, \mathsf{C} \to 5\}$. Furthermore, The player 1 has $|\mathcal{V}| = 7$ pure strategy profiles in total.

We model imperfect information with information sets. An information set $\mathcal{I} \subseteq \mathcal{T}^{(i)}$ is a subset of nodes assigned to player i, which the player cannot distinguish. The player must act consistently across all nodes within the same information set. For example, in poker, an information set includes all states with identical public cards and bets, with each node representing a different potential hand held by the opponent. Each terminal observation point $\sigma \in \mathcal{E}^{(i)}$ is associated with a set $\mathcal{I}_\sigma \subseteq \mathcal{Z}$ of corresponding terminal nodes in the original EFG. For each terminal node $z \in \mathcal{Z}$, $\sigma_z^{(i)}$ denotes the observation point of player i in the TFSDP.

Let $\pi = \{\mathbf{x}^{(i)}\}_{i=1}^n$ be a joint policy of n players. We denote by $u^{(i)}(\pi)$ the expected reward received by player. Consider the terminal node $z \in \mathcal{Z}$, the reach of probability can be computed by $\mathbf{p}[z] \prod_{j=1}^n \mathbf{x}^{(j)}[\sigma_z^{(j)}]$, where $\mathbf{p}[z]$ is the product of transition probability of the environment actions from the root to z. In this case, the expected reward of player i is given by

$$u^{(i)}(\pi) = \sum_{z \in \mathcal{Z}} \mathbf{u}^{(i)}[z] \cdot \mathbf{p}[z] \prod_{j=1}^{n} \mathbf{x}^{(j)}[\sigma_z^{(j)}].$$

Consider the corresponding reward vector $\mathbf{w}^{(i)} := \partial_i u^{(i)}(\pi) \in \mathbb{R}^{\mathcal{E}^{(i)}}$ of player i when the players agree on joint policy $\pi = \{\mathbf{x}^{(i)}\}_{i=1}^n$. The vector satisfies that $u^{(i)}(\pi) = \langle \mathbf{x}^{(i)}, \mathbf{w}^{(i)} \rangle$. It is clear that the reward vector has the following closed form for each entry:

$$\mathbf{w}^{(i)}[\sigma] = \sum_{z \in \mathcal{I}_\sigma} \mathbf{u}^{(i)}[z] \cdot \mathbf{p}[z] \prod_{j \neq i} \mathbf{x}^{(j)}[\sigma_z^{(j)}] \quad \forall \sigma \in \mathcal{E}^{(i)}.$$

In this work, we examine two problems in EFGs. For the online learning problem, an agent seeks to maximize their expected reward while facing an adversarial environment and other players in the online decision-making process. The agent can observe the reward vector $\mathbf{w}_t$ after committing to the strategy profile $\mathbf{x}_t$ in episode t. We measure the performance of the online learning algorithm using regret. The cumulative (external) regret over T episodes is defined as:

$$\text{Regret}(T) := \max_{\mathbf{x}_* \in \mathcal{Q}} \sum_{t=1}^{T} \langle \mathbf{x}_*, \mathbf{w}_t \rangle - \sum_{t=1}^{T} \langle \mathbf{x}_t, \mathbf{w}_t \rangle.$$

where $\mathbf{x}_t$ is the strategy proposed by the player in episode t. This definition quantifies the cumulative difference between the expected rewards that could have been obtained by the optimal strategy and those achieved under the actual strategy used.

For equilibrium computation, a group of agents aim to jointly learn the coarse correlated equilibrium (CCE) given only oracle access to the game. A joint policy profile π is said an ε-CCE, if for any player i, it satisfies that

$$u^{(i)}(\mathbf{x}^{(i)} \times \pi^{(-i)}) \leq u^{(i)}(\pi) + \varepsilon, \quad \forall \mathbf{x}^{(i)} \in \mathcal{Q}^{(i)},$$

where we denote by $\mathbf{x}^{(i)} \times \pi^{(-i)}$ the joint policy in which player i takes strategy profile $\mathbf{x}^{(i)}$ while other players still follows π. It is known that the equilibrium computation can be reduced to online learning: If every player runs a no-regret learning algorithm simultaneously, then the average joint policy $\bar{\pi}_T$ of the interaction history is an ε-CCE with $\varepsilon \leq \max_{i \in [\![n]\!]} \text{Regret}^{(i)}(T)/T$ [9]. As the players are not adversarial against each other, the regret bound of the learning algorithm can sometimes be improved compared to the online learning setting.

B Pseudocode of Predictive OMD and Clairvoyant OMD

In this section, we list the pseudocode of the algorithms used in the paper.

Algorithm 1: (Predictive) Online Mirror Descent

$\widetilde{\mathbf{x}}_1 \leftarrow \operatorname{argmin}_{\mathbf{x} \in \mathcal{Q}} \varphi(\mathbf{x})$
for $t = 1$ **to** T **do**
  Receive prediction $\mathbf{m}_t$ (set $\mathbf{m}_t = \mathbf{0}$ for the non-predictive variant)
  $\mathbf{x}_t \leftarrow \Pi_\varphi(\eta \mathbf{m}_t, \widetilde{\mathbf{x}}_t)$
  Commit policy $\mathbf{x}_t$, receive reward $\langle \mathbf{x}_t, \mathbf{w}_t \rangle$ and observe reward vector $\mathbf{w}_t$
  $\widetilde{\mathbf{x}}_{t+1} \leftarrow \Pi_\varphi(\eta \mathbf{w}_t, \widetilde{\mathbf{x}}_t)$
end

The (Predictive) OMD framework [12, 39] depicts a family of no-regret learning algorithms. The pseudocode of the algorithm can be found in Algorithm 1. The algorithm takes a learning rate μ and a DGF φ for the decision set $\mathcal{Q}$ of the TFSDP as parameters. It starts from an initial point $\widetilde{\mathbf{x}}_1$ and follows a straightforward structure in each episode t:

- Receive prediction vector $\mathbf{m}_t$ from external logic (set $\mathbf{m}_t = 0$ for the non-predictive variant)
- Take a proximal gradient step from $\widetilde{\mathbf{x}}_t$ according to the prediction $\mathbf{m}_t$ to get the policy $\mathbf{x}_t$.
- Execute policy $\mathbf{x}_t$ and observe reward vector $\mathbf{w}_t$.
- Take another proximal gradient step from $\widetilde{\mathbf{x}}_t$ according to $\mathbf{w}_t$ to get $\widetilde{\mathbf{x}}_{t+1}$ for the next episode.

It can be shown from Theorem 5.3 that the algorithm achieves a sub-linear regret rate of $\sqrt{T}$ in general. The performance of the algorithm can be further improved by selecting more accurate $\mathbf{m}_t$. For example, if we can ensure $\mathbf{m}_t = \mathbf{w}_t$, then the algorithm suffers only constant regret, upper bounded by $\mathcal{D}_\varphi(\mathbf{x}_*, \mathbf{x}_1)/\eta$.

Although the reward vector $\mathbf{w}_t$ in the adversarial setting is given by the environment which is generally unpredictable, in self-playing, it is determined by other agents, which generally follow a smooth dynamic and thus can be predictable. Syrgkanis et al. [39] introduced Optimistic OMD, which sets the prediction $\mathbf{m}_t \leftarrow \mathbf{w}_{t-1}$ as the reward vector given from the previous play. This allows them to establish a regret bound of $\mathcal{O}(T^{1/4})$. By analyzing the higher-order derivatives, Daskalakis et al. [13] show $\|\mathbf{m}_t - \mathbf{w}_t\|_*$ is small in Optimistic OMD, which finally leads to a logarithmic regret bound of $\mathcal{O}(\log^4 T)$.

Note that the desired prediction $\mathbf{m}_t = \mathbf{w}_t$ is a solution to $\mathbf{w}_t^{(i)} = \partial_i u^{(i)}(\pi_t)$ for every player i, where the joint policy $\pi_t = \{\mathbf{x}_t^{(i)}\}_{i=1}^n$ is given by proximal step $\mathbf{x}_t^{(i)} = \Pi_{\varphi^{(i)}}(\eta \mathbf{w}_t^{(i)}, \widetilde{\mathbf{x}}_t^{(i)})$ for each player i.

https://doi.org/10.52202/079017-2804

Algorithm 2: Clairvoyant OMD, Decentralized

for *each player* $i \in [\![n]\!]$, *in parallel* **do**
  $\widetilde{\mathbf{x}}_1^{(i)} \leftarrow \operatorname{argmin}_{\mathbf{x}\in\mathcal{Q}^{(i)}} \varphi^{(i)}(\mathbf{x})$
  for $t = 1$ **to** K **do**
    $\mathbf{x}_{t,1}^{(i)} \leftarrow \widetilde{\mathbf{x}}_t^{(i)}$
    for $l = 1$ **to** L **do**
      Synchronous with other players and commit joint policy $\pi_{t,l} \leftarrow \{\mathbf{x}_{t,l}^{(i)}\}_{i=1}^n$
      Observe reward vector $\mathbf{w}_{t,l}^{(i)} \leftarrow \partial_i u^{(i)}(\pi_{t,l})$
      $\mathbf{x}_{t,l+1}^{(i)} \leftarrow \Pi_{\varphi^{(i)}}(\eta \mathbf{w}_{t,l}^{(i)}, \widetilde{\mathbf{x}}_t^{(i)})$
    end
    $\mathbf{m}_t^{(i)} \leftarrow \mathbf{w}_{t,L}^{(i)}$
    $\mathbf{x}_t^{(i)} \leftarrow \mathbf{x}_{t,L}^{(i)}; \pi_t \leftarrow \pi_{t,L}$ // $\mathbf{x}_t^{(i)} = \Pi_{\varphi^{(i)}}(\eta \mathbf{m}_t^{(i)}, \widetilde{\mathbf{x}}_t^{(i)})$
    $\mathbf{w}_t^{(i)} \leftarrow \mathbf{w}_{t,L+1}^{(i)}$
    $\widetilde{\mathbf{x}}_{t+1}^{(i)} \leftarrow \mathbf{x}_{t,L+1}^{(i)}$ // $\widetilde{\mathbf{x}}_{t+1}^{(i)} = \Pi_{\varphi^{(i)}}(\eta \mathbf{w}_t^{(i)}, \widetilde{\mathbf{x}}_t^{(i)})$
  end
end
Report average joint policy $\bar{\pi}_K$ of π_t among $t \in [\![K]\!]$

Note this is a fixed-point to dynamics

$$\begin{cases} \pi_t & \leftarrow \{\mathbf{x}_t^{(i)}\}_{i=1}^n \\ \mathbf{w}_t^{(i)} & \leftarrow \partial_i u^{(i)}(\pi_t) \\ \mathbf{x}_t^{(i)} & \leftarrow \Pi_{\varphi^{(i)}}(\eta \mathbf{w}_t^{(i)}, \widetilde{\mathbf{x}}_t^{(i)}) \end{cases}$$

Clairvoyant OMD [36, 19] uses these update rules to find $\mathbf{m}_t$ through the fixed-point iteration. By showing that all these updating rules are Lipschitz, one can see that this fixed-point iteration achieves a linear convergence rate when the learning rate η is small. Let K be the number of episodes used for aggravating the average joint policy. The linear convergence rate allows one to find $\mathbf{m}_t$ where $\|\mathbf{m}_t - \mathbf{w}_t\|_*$ is polynomially small with $\mathcal{O}(1/K)$ in only $L = \mathcal{O}(\log K)$ steps, indicating that the corresponding OMD dynamics only suffer from a constant regret bound. This finally establishes that the algorithm has an almost-optimal convergence rate of $\log T/T$ where $T = KL$ is the number of oracle access to the game.

C Proof of Treeplex Norm

C.1 Proof of Lemma 4.1

Lemma 4.1 (restatement)**.** The functions $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ are norms defined on the space $\mathbb{R}^{\mathcal{E}}$.

Proof. We verify that $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ are norms as follows:

Positive definiteness: It is clear from the definition that $\|\mathbf{u}\|_{\mathcal{H},1} = 0$ and $\|\mathbf{u}\|_{\mathcal{H},\infty} = 0$ when $\mathbf{u} = \mathbf{0}$. When $\mathbf{u} \neq \mathbf{0}$, there exists $\sigma_{\mathbf{u}} \in \mathcal{E}$ such that $|\mathbf{u}[\sigma_{\mathbf{u}}]| > 0$. From the definition of $\mathcal{Y}$, we can always find some transition kernel $\mathbf{y}_{\mathbf{u}} \in \mathcal{Y}$ such that $\sigma_{\mathbf{u}}$ is reachable. In this case, $\mathbf{y}_{\mathbf{u}}[\sigma_{\mathbf{u}}] > 0$ and we have

$$\|\mathbf{u}\|_{\mathcal{H},1} \geq \sum_{\sigma\in\mathcal{E}} |\mathbf{u}[\sigma]| \cdot \mathbf{y}_{\mathbf{u}}[\sigma] \geq |\mathbf{u}[\sigma_{\mathbf{x}}]| \cdot \mathbf{y}_{\mathbf{u}}[\sigma_{\mathbf{u}}] > 0.$$

Similarly, we can always find some strategy profile $\mathbf{x}_{\mathbf{u}}$ with $\mathbf{x}_{\mathbf{u}}[\sigma_{\mathbf{u}}] > 0$, which implies that

$$\|\mathbf{u}\|_{\mathcal{H},\infty} \geq \sum_{\sigma\in\mathcal{E}} |\mathbf{u}[\sigma]| \cdot \mathbf{x}_{\mathbf{u}}[\sigma] \geq |\mathbf{u}[\sigma_{\mathbf{x}}]| \cdot \mathbf{x}_{\mathbf{u}}[\sigma_{\mathbf{u}}] > 0.$$

This verifies that both functions are strictly positive on non-zero vectors.

Homogeneity: For any $k \in \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^{\mathcal{E}}$, it holds that

$$\|k\mathbf{u}\|_{\mathcal{H},1} = \sup_{\mathbf{y}\in\mathcal{Y}} \sum_{\sigma\in\mathcal{E}} |k\mathbf{u}[\sigma]| \cdot \mathbf{y}[\sigma] = |k| \sup_{\mathbf{y}\in\mathcal{Y}} \sum_{\sigma\in\mathcal{E}} |\mathbf{u}[\sigma]| \cdot \mathbf{y}[\sigma] = |k| \cdot \|\mathbf{u}\|_{\mathcal{H},1}$$
$$\|k\mathbf{u}\|_{\mathcal{H},\infty} = \sup_{\mathbf{x}\in\mathcal{Q}} \sum_{\sigma\in\mathcal{E}} |k\mathbf{u}[\sigma]| \cdot \mathbf{x}[\sigma] = |k| \sup_{\mathbf{x}\in\mathcal{Q}} \sum_{\sigma\in\mathcal{E}} |\mathbf{u}[\sigma]| \cdot \mathbf{x}[\sigma] = |k| \cdot \|\mathbf{u}\|_{\mathcal{H},\infty},$$

which verifies absolute homogeneity.

Triangle inequality: We verify the triangle inequality for any $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^{\mathcal{E}}$ by

$$\begin{aligned}\|\mathbf{u}_1 + \mathbf{u}_2\|_{\mathcal{H},1} &= \sup_{\mathbf{y}\in\mathcal{Y}} \sum_{\sigma\in\mathcal{E}} |\mathbf{u}_1[\sigma] + \mathbf{u}_2[\sigma]| \cdot \mathbf{y}[\sigma] \\ &\le \sup_{\mathbf{y}\in\mathcal{Y}} \sum_{\sigma\in\mathcal{E}} |\mathbf{u}_1[\sigma]| \cdot \mathbf{y}[\sigma] + \sup_{\mathbf{y}\in\mathcal{Y}} \sum_{\sigma\in\mathcal{E}} |\mathbf{u}_2[\sigma]| \cdot \mathbf{y}[\sigma] = \|\mathbf{u}_1\|_{\mathcal{H},1} + \|\mathbf{u}_2\|_{\mathcal{H},1}.\end{aligned}$$

With a similar calculation, it is straightforward to check $\|\cdot\|_{\mathcal{H},\infty}$ also satisfies triangle inequality.

To conclude, both $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ are norms. □

C.2 Proof of Lemma 4.2

Lemma 4.2 (restatement)**.** Let $\mathbf{u} \in \mathbb{R}^{\mathcal{E}_h}$ be a vector with respect to some point $h \in \mathcal{H}$. The treeplex ℓ_1 norm and the treeplex ℓ_∞ norm of vector $\mathbf{u}$ over $\mathcal{H}_h$ can be computed recursively as follows.

- If $h = \sigma \in \mathcal{E}$ is a terminal observation point, then:
$$\|\mathbf{u}\|_{\mathcal{H}_\sigma,1} := |\mathbf{u}[\sigma]|, \qquad \|\mathbf{u}\|_{\mathcal{H}_\sigma,\infty} := |\mathbf{u}[\sigma]|.$$
- If $h = j \in \mathcal{J}$ is a decision point, then:
$$\|\mathbf{u}\|_{\mathcal{H}_j,1} := \textstyle\sum_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},1}, \qquad \|\mathbf{u}\|_{\mathcal{H}_j,\infty} := \max_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},\infty}.$$
- If $h = \sigma \in \Sigma \setminus \mathcal{E}$ is a non-terminal observation point, then:
$$\|\mathbf{u}\|_{\mathcal{H}_j,1} := \max_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},1}, \qquad \|\mathbf{u}\|_{\mathcal{H}_j,\infty} := \textstyle\sum_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},\infty}.$$

Proof. Consider the statement for treeplex ℓ_1 norm. The plan is to prove the statement by induction on $\mathcal{H}$ from the bottom up on TFSDP. We will show that the recursive definition ensures $\|\mathbf{u}\|_{\mathcal{H}_h,1}$ represents the treeplex ℓ_1 norm restricted to the subtree of h. Specifically, we will demonstrate that for every point $h \in \mathcal{H}$ and vector $\mathbf{u} \in \mathbb{R}^{\mathcal{E}_h}$, the recursive formula at point h matches the original definition of treeplex ℓ_1 norm with restricted in the subtree $\mathcal{H}_h$, which is:

$$\|\mathbf{u}\|_{\mathcal{H}_h,1} = \sup_{\mathbf{y}\in\mathcal{Y}_h} \langle |\mathbf{u}|, \mathbf{y} \rangle.$$

Case 1: The base case for the induction occurs when $h = \sigma$, where $\sigma \in \mathcal{E}$ is a terminal point. We verify the induction basis by

$$\|\mathbf{u}\|_{\mathcal{H}_\sigma,1} = |\mathbf{u}[\sigma]| \cdot 1 = \sup_{\mathbf{y}\in\mathcal{Y}_\sigma} \langle |\mathbf{u}|, \mathbf{y} \rangle,$$

where the last equality is given by the fact that $\mathbf{y}[\mathcal{E}_\sigma] = 1$ for $\mathbf{y} \in \mathcal{Y}_\sigma$.

Case 2: For any decision point $h = j \in \mathcal{J}$, it holds that

$$\begin{aligned}\|\mathbf{u}\|_{\mathcal{H}_j,1} &= \sum_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},1} \\ &= \sum_{a\in\mathcal{A}_j} \sup_{\mathbf{y}_{ja}\in\mathcal{Y}_{ja}} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, \mathbf{y}_{ja}[\mathcal{E}_{ja}] \rangle \\ &= \sup_{\{\mathbf{y}_{ja}\in\mathcal{Y}_{ja}\}_{a\in\mathcal{A}_j}} \sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, \mathbf{y}_{ja}[\mathcal{E}_{ja}] \rangle \\ &= \sup_{\mathbf{y}\in\mathcal{Y}_j} \sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, \mathbf{y}[\mathcal{E}_{ja}] \rangle \\ &= \sup_{\mathbf{y}\in\mathcal{Y}_j} \langle |\mathbf{u}|, \mathbf{y} \rangle,\end{aligned}$$

https://doi.org/10.52202/079017-2804

where the first equality holds according to the recursive definition of $\|\mathbf{u}\|_{\mathcal{H}_j,1}$, the second equality follows from the induction hypothesis, the third equality holds as the set of terminal points $\mathcal{E}_{ja}$ are disjoint for each $a \in \mathcal{A}_j$, the fourth equality follows from the fact that $\mathcal{Y}_j$ can be decomposed into the Cartesian product of $\mathcal{Y}_{ja}$ among all $a \in \mathcal{A}_j$ for any $j \in \mathcal{J}$, and the last equality holds as each term of the ℓ_1 norm is positive.

Case 3: For some non-terminal observation point $h = \sigma \in \Sigma \setminus \mathcal{E}$, we will show that quantity $\|\mathbf{u}\|_{\mathcal{H}_\sigma,1} = \max_{j\in\mathcal{C}_\sigma} \|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1}$ is neither less than nor greater than $\sup_{\mathbf{y}\in\mathcal{Y}_\sigma}\langle|\mathbf{u}|,\mathbf{y}[\mathcal{E}_\sigma]\rangle$. Firstly, for decision point $j \in \mathcal{C}_\sigma$ that is a successor of σ, it satisfies that

$$\|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1} = \sup_{\mathbf{y}\in\mathcal{Y}_j}\langle|\mathbf{u}[\mathcal{E}_j]|,\mathbf{y}[\mathcal{E}_j]\rangle \le \sup_{\mathbf{y}\in\mathcal{Y}_\sigma}\langle|\mathbf{u}|,\mathbf{y}\rangle,$$

where the equality holds from the induction hypothesis and the inequality holds since the inner product can be upper bounded according to $\langle|\mathbf{u}[\mathcal{E}_j]|,\mathbf{y}[\mathcal{E}_j]\rangle \le \langle|\mathbf{u}|,\mathbf{y}\rangle$ and $\mathcal{Y}_j$ is a subset of $\mathcal{Y}_\sigma$. By taking the maximal among all successor $j \in \mathcal{C}_\sigma$, this inequality immediately establishes an upper bound for $\|\mathbf{u}\|_{\mathcal{H}_\sigma,1}$:

$$\|\mathbf{u}\|_{\mathcal{H}_\sigma,1} = \max_{j\in\mathcal{C}_\sigma}\|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1} \le \sup_{\mathbf{y}\in\mathcal{Y}_\sigma}\langle|\mathbf{u}|,\mathbf{y}\rangle. \tag{C.1}$$

Moreover, fix some transition kernel $\mathbf{y} \in \mathcal{Y}_\sigma$. According to the tree-structure of the TFSDP, we have that the vector $\mathbf{y}[\mathcal{E}_j]/\mathbf{y}[j]$ with respect to some successor $j \in \mathcal{C}_\sigma$ is a valid transition kernel with in the subtree of j. In other words, we have vector $\mathbf{y}[\mathcal{E}_j]/\mathbf{y}[j] \in \mathcal{Y}_j$. Thus, we can write

$$\begin{aligned}
\langle|\mathbf{u}|,\mathbf{y}\rangle &= \sum_{j\in\mathcal{C}_\sigma}\mathbf{y}[j]\cdot\langle|\mathbf{u}[\mathcal{E}_j]|,\mathbf{y}[\mathcal{E}_j]/\mathbf{y}[j]\rangle \\
&\le \sum_{j\in\mathcal{C}_\sigma}\mathbf{y}[j]\cdot\sup_{\mathbf{y}_j\in\mathcal{Y}_j}\langle|\mathbf{u}[\mathcal{E}_j]|,\mathbf{y}_j[\mathcal{E}_j]\rangle \\
&= \sum_{j\in\mathcal{C}_\sigma}\mathbf{y}[j]\cdot\|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1} \\
&\le \max_{j\in\mathcal{C}_\sigma}\|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1}
\end{aligned}$$

where the first equality follows from the fact that $\mathcal{E}_j$, for all successor $j \in \mathcal{C}_\sigma$, forms a partition of $\mathcal{E}_\sigma$, the first inequality holds since $\mathbf{y}[\mathcal{E}_j]/\mathbf{y}[j] \in \mathcal{Y}_j$, the second equality follows from induction hypothesis, and the last inequality holds since $\sum_{j\in\mathcal{C}_\sigma}\mathbf{y}[j] = 1$ and $\mathbf{y}[j] \ge 0$ for transition kernel $\mathbf{y} \in \mathcal{Y}_\sigma$. By taking supremum among all transition kernel $\mathbf{y} \in \mathcal{Y}_\sigma$, we establishes an lower bound $\|\mathbf{u}\|_{\mathcal{H}_\sigma,1}$

$$\sup_{\mathbf{y}\in\mathcal{Y}_\sigma}\langle|\mathbf{u}|,\mathbf{y}\rangle \le \max_{j\in\mathcal{C}_\sigma}\|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1} = \|\mathbf{u}\|_{\mathcal{H}_\sigma,1}. \tag{C.2}$$

As the upper bound in (C.1) and the lower bound in (C.2) agrees on the same quantity, we immediately reach the following equation

$$\|\mathbf{u}\|_{\mathcal{H}_\sigma,1} = \max_{j\in\mathcal{C}_\sigma}\|\mathbf{u}[\mathcal{E}_j]\|_{\mathcal{H}_j,1} = \sup_{\mathbf{y}\in\mathcal{Y}_\sigma}\langle|\mathbf{u}|,\mathbf{y}\rangle$$

which proves the induction statement on observation point σ.

In general, the induction hypothesis always holds. By inspecting $h = \varnothing$, we reach the desired statement in which $\|\mathbf{u}\|_{\mathcal{H}_\varnothing,1} = \sup_{\mathbf{y}\in\mathcal{Y}}\langle|\mathbf{u}|,\mathbf{y}\rangle = \|\mathbf{u}\|_{\mathcal{H},1}$.

Finally, since the treeplex ℓ_∞ norm closely mirrors treeplex ℓ_1 norm, the result for treeplex ℓ_∞ norm can be directly reached with the only modification being the interchange of cases 2 and 3. □

C.3 Proof of Theorem 4.3

Theorem 4.3 (restatement)**.** We have $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ is a pair of primal-dual norms for a given TFSDP with point set $\mathcal{H}$. Specifically, for any vector $\mathbf{u} \in \mathbb{R}^{\mathcal{E}}$,

$$\|\mathbf{u}\|_{\mathcal{H},1}^* := \sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}}}\frac{\langle\mathbf{u},\mathbf{v}\rangle}{\|\mathbf{v}\|_{\mathcal{H},1}} = \|\mathbf{u}\|_{\mathcal{H},\infty}.$$

Proof. Firstly, from the definition of the treeplex ℓ_1 norm and the treeplex ℓ_∞ norm, the norm of any vector is equal to the norm of the vector that takes the absolute value at each index, that is, we have $\|\mathbf{v}\|_{\mathcal{H},1} = \||\mathbf{v}|\|_{\mathcal{H},1}$ and $\|\mathbf{u}\|_{\mathcal{H},\infty} = \||\mathbf{u}|\|_{\mathcal{H},\infty}$ holds for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathcal{E}}$. Therefore, we can always choose $\mathbf{v}$ such that it matches the sign of $\mathbf{u}$ to maximize the inner product in the dual norm, which implies

$$\|\mathbf{u}\|^*_{\mathcal{H},1} = \sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}}} \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\|\mathbf{v}\|_{\mathcal{H},1}} = \sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H},1}}$$

We will use induction on $\mathcal{H}$ from the bottom up, demonstrating that $\|\mathbf{u}\|^*_{\mathcal{H}_h,1} = \|\mathbf{u}\|_{\mathcal{H}_h,\infty}$ always holds for every $h \in \mathcal{H}$. Specifically, we will show that for every $h \in \mathcal{H}$ and $\mathbf{u} \in \mathbb{R}^{\mathcal{E}_h}$:

$$\sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_h}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H}_h,1}} = \|\mathbf{u}\|_{\mathcal{H}_h,\infty}.$$

Case 1: The induction basis occurs at $h = \sigma \in \mathcal{E}$, where the statement can be verified from

$$\sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_\sigma}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H}_\sigma,1}} = \sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_\sigma}} \frac{|\mathbf{u}[\sigma]| \cdot |\mathbf{v}[\sigma]|}{|\mathbf{v}[\sigma]|} = |\mathbf{u}[\sigma]| = \|\mathbf{u}\|_{\mathcal{H}_\sigma,\infty}.$$

where the first equality is given by Lemma 4.2 and the last equality is given by Lemma 4.2.

Case 2: Consider some decision point $h = j \in \mathcal{J}$. It satisfies that

$$\sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_j}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H}_j,1}} = \sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_j}} \frac{\sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}[\mathcal{E}_{ja}]|\rangle}{\sum_{a\in\mathcal{A}_j} \|\mathbf{v}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},1}} = \sup_{\{\mathbf{v}_{ja}\in\mathbb{R}^{\mathcal{E}_{ja}}\}_{a\in\mathcal{A}_j}} \frac{\sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\sum_{a\in\mathcal{A}_j} \|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}}, \tag{C.3}$$

where the expression for the numerator in the first equality is valid since $\mathcal{E}_{ja}$ for $a \in \mathcal{A}_j$ is a partition of $\mathcal{E}_j$, while the expression for the denominator follows from Lemma 4.2. The fraction can be interpreted as a weighted average of $\langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle / \|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}$, indicating that

$$\frac{\sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\sum_{a\in\mathcal{A}_j} \|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}} \le \max_{a\in\mathcal{A}_j} \frac{\langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}}.$$

Additionally, by choosing some specific $a \in \mathcal{A}_j$ and assigning $\mathbf{v}_{ja'} = \mathbf{0}$ for any $a' \neq a$, we have that

$$\sup_{\{\mathbf{v}_{ja}\in\mathbb{R}^{\mathcal{E}_{ja}}\}_{a\in\mathcal{A}_j}} \frac{\sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\sum_{a\in\mathcal{A}_j} \|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}} \ge \sup_{\mathbf{v}_{ja}\in\mathbb{R}^{\mathcal{E}_{ja}}} \frac{\langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}}.$$

These inequalities define the upper and lower bounds for the same quantity, leading to the equation:

$$\sup_{\{\mathbf{v}_{ja}\in\mathbb{R}^{\mathcal{E}_{ja}}\}_{a\in\mathcal{A}_j}} \frac{\sum_{a\in\mathcal{A}_j} \langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\sum_{a\in\mathcal{A}_j} \|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}} = \max_{a\in\mathcal{A}_j} \sup_{\mathbf{v}_{ja}\in\mathbb{R}^{\mathcal{E}_{ja}}} \frac{\langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}}. \tag{C.4}$$

Moreover, according to the induction hypothesis, we can replace the inner supremum by

$$\sup_{\mathbf{v}_{ja}\in\mathbb{R}^{\mathcal{E}_{ja}}} \frac{\langle |\mathbf{u}[\mathcal{E}_{ja}]|, |\mathbf{v}_{ja}|\rangle}{\|\mathbf{v}_{ja}\|_{\mathcal{H}_{ja},1}} = \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},\infty}. \tag{C.5}$$

By combining (C.3), (C.4), (C.5), and Lemma 4.2, we establish the induction statement on $j \in \mathcal{J}$:

$$\sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_j}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H}_j,1}} = \max_{a\in\mathcal{A}_j} \|\mathbf{u}[\mathcal{E}_{ja}]\|_{\mathcal{H}_{ja},\infty} = \|\mathbf{u}\|_{\mathcal{H}_j,\infty}.$$

Case 3: When $h = \sigma \in \Sigma \setminus \mathcal{E}$, it follows from the symmetric relationship between treeplex ℓ_1 norm and treeplex ℓ_∞ norm that, similarly to the previous arguments, we can conclude that

$$\sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_\sigma}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H}_\sigma,\infty}} = \|\mathbf{u}\|_{\mathcal{H}_\sigma,1}.$$

This suggests that the norm $\|\cdot\|_{\mathcal{H}_\sigma,\infty}$ is a dual norm of $\|\cdot\|_{\mathcal{H}_\sigma,1}$, which immediately leads to the desired statement, in which

$$\sup_{\mathbf{v}\in\mathbb{R}^{\mathcal{E}_\sigma}} \frac{\langle |\mathbf{u}|, |\mathbf{v}|\rangle}{\|\mathbf{v}\|_{\mathcal{H}_\sigma,1}} = \|\mathbf{u}\|_{\mathcal{H}_\sigma,\infty}.$$

□

https://doi.org/10.52202/079017-2804

C.4 Proof of Lemma 4.4

Lemma 4.4 (restatement)**.** We have $\|\mathbf{x}\|_{\mathcal{H},1} = 1$ for any strategy profile $\mathbf{x} \in \mathcal{Q}$, $\|\mathbf{y}\|_{\mathcal{H},\infty} = 1$ for any transition kernel $\mathbf{y} \in \mathcal{Y}$, and $\|\mathbf{w}\|_{\mathcal{H},\infty} \leq 1$ for any feasible reward vector $\mathbf{w}$ under Assumption 3.1.

Proof. According to the definition of $\mathcal{Q}$ and $\mathcal{Y}$, for any element $\mathbf{x} \in \mathcal{Q}$ and $\mathbf{y} \in \mathcal{Y}$, it is established that $\|\mathbf{x}\|_{\mathcal{H}_h,1}$, $\|\mathbf{y}\|_{\mathcal{H}_h,\infty}$, and their respective $\mathbf{x}[h]$ and $\mathbf{y}[h]$ agree on the same recursive formula across all $h \in \mathcal{H}$ given by Lemmas 4.2. Consequently, it is always true that $\|\mathbf{x}\|_{\mathcal{H}_h,1} = \mathbf{x}[h]$ and $\|\mathbf{y}\|_{\mathcal{H}_h,\infty} = \mathbf{y}[h]$, thereby implying $\|\mathbf{x}\|_{\mathcal{H},1} = 1$ and $\|\mathbf{y}\|_{\mathcal{H},\infty} = 1$. Finally, we have $\mathbf{w} = \mathbf{y} \circ \mathbf{r}$ for $\mathbf{r} \in [0,1]^{\mathcal{E}}$ under Assumption 3.1. Thus, we have

$$\|\mathbf{w}\|_{\mathcal{H},\infty} = \sup_{\mathbf{x}\in\mathcal{Q}} \langle |\mathbf{w}|, \mathbf{x}\rangle = \sup_{\mathbf{x}\in\mathcal{Q}} \langle |\mathbf{r}| \odot |\mathbf{y}|, \mathbf{x}\rangle \leq \sup_{\mathbf{x}\in\mathcal{Q}} \langle |\mathbf{y}|, \mathbf{x}\rangle = \|\mathbf{y}\|_{\mathcal{H},\infty} = 1.$$

where the inequality is given by the $\mathbf{x}$ is always non-negative as well as $\mathbf{r} \leq 1$. □

D Proof of Regret Upper Bounds

D.1 Proof of Lemma 5.1

Lemma 5.1 (restatement)**.** The weight-one dilated entropy (DilEnt) is 1-strongly convex within the extensive-form decision space $\mathcal{Q}$ with respect to the treeplex ℓ_1 norm $\|\cdot\|_{\mathcal{H},1}$. Specifically, for any vector $\mathbf{z} \in \mathbb{R}^{\mathcal{E}}$ and strategy profile $\mathbf{x} \in \mathcal{Q}$, we have that $\|\mathbf{z}\|^2_{\nabla^2\varphi_1(\mathbf{x})} \geq \|\mathbf{z}\|^2_{\mathcal{H},1}$.

Proof. For some transition kernel $\mathbf{y} \in \mathcal{Y}$, we define the $\mathbf{y}$-weighted dilated entropy:

$$\varphi_{\mathbf{y}}(\mathbf{x}) := \sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{y}[ja]\mathbf{x}[ja] \ln\left(\frac{\mathbf{x}[ja]}{\mathbf{x}[p_j]}\right).$$

By decomposing the logarithm term in the summation, we get

$$\varphi_{\mathbf{y}}(\mathbf{x}) = \underbrace{\sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{y}[ja]\mathbf{x}[ja] \ln \mathbf{x}[ja]}_{\mathcal{I}_1} - \underbrace{\sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{y}[ja]\mathbf{x}[ja] \ln \mathbf{x}[p_j]}_{\mathcal{I}_2}. \tag{D.1}$$

We can rewrite the first term as

$$\mathcal{I}_1 = \sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{y}[ja]\mathbf{x}[ja] \ln \mathbf{x}[ja] = \sum_{\sigma\in\Sigma^+} \mathbf{y}[\sigma]\mathbf{x}[\sigma] \ln \mathbf{x}[\sigma]. \tag{D.2}$$

For the second term, we can further write:

$$\begin{aligned}
\mathcal{I}_2 &= \sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} \mathbf{y}[ja]\mathbf{x}[ja] \ln \mathbf{x}[p_j] \\
&= \sum_{j\in\mathcal{J}} \mathbf{y}[j]\mathbf{x}[p_j] \ln \mathbf{x}[p_j] \\
&= \sum_{\sigma\in\Sigma\setminus\mathcal{E}} \sum_{j\in\mathcal{C}_\sigma} \mathbf{y}[j]\mathbf{x}[\sigma] \ln \mathbf{x}[\sigma] \\
&= \sum_{\sigma\in\Sigma\setminus\mathcal{E}} \mathbf{y}[\sigma]\mathbf{x}[\sigma] \ln \mathbf{x}[\sigma]
\end{aligned} \tag{D.3}$$

where the second equality is given by $\mathbf{y}[j] = \mathbf{y}[ja]$ for transition kernel $\mathbf{y} \in \mathcal{Y}$ and $\mathbf{x}[p_j] = \sum_{a\in\mathcal{A}_j} \mathbf{x}[ja]$ for strategy profile $\mathbf{x} \in \mathcal{Q}$, the third equality is derived from the fact that $\mathcal{C}_\sigma$, for all $\sigma \in \Sigma \setminus \mathcal{E}$, forms a partition of $\mathcal{J}$, and the last equality follows from $\mathbf{y}[\sigma] = \sum_{j\in\mathcal{C}_\sigma} \mathbf{y}[j]$ for any non-terminal observation point $\sigma \in \Sigma \setminus \mathcal{E}$ over $\mathbf{y} \in \mathcal{Y}$.

Plugging (D.2) and (D.3) into (D.1), we obtain the following result, indicating that function $\varphi_{\mathbf{y}}$ can be expressed as the weighted negative entropy over all terminal observation points $\sigma \in \mathcal{E}$:

$$\varphi_{\mathbf{y}}(\mathbf{x}) = \sum_{\sigma\in\mathcal{E}} \mathbf{y}[\sigma]\mathbf{x}[\sigma] \ln \mathbf{x}[\sigma].$$

We are ready to show the strong convexity of weight-one dilated entropy $\varphi_{\mathbf{1}}$. Consider a vector $\mathbf{z} \in \mathbb{R}^{\mathcal{E}}$. Since function $\varphi_{\mathbf{y}}$ is additively separable over all variables $\mathbf{x}[\sigma]$, the Hessian matrix of $\varphi_{\mathbf{y}}(\mathbf{x})$ is diagonal. Thus, the squared norm of $\mathbf{z}$ over $\nabla^2\varphi_{\mathbf{y}}(\mathbf{x})$ can be interpreted according to

$$\begin{aligned}
\mathbf{z}^\top \nabla^2 \varphi_{\mathbf{y}}(\mathbf{x})\mathbf{z} &= \sum_{\sigma\in\mathcal{E}} \mathbf{z}[\sigma]^2 \cdot \nabla^2_{\mathbf{x}[\sigma]} \varphi_{\mathbf{y}}(\mathbf{x}) \\
&= \sum_{\sigma\in\mathcal{E}} \mathbf{z}[\sigma]^2 \cdot \frac{\mathbf{y}[\sigma]}{\mathbf{x}[\sigma]} \\
&\geq \Big(\sum_{\sigma\in\mathcal{E}} \mathbf{z}[\sigma]^2 \cdot \frac{\mathbf{y}[\sigma]}{\mathbf{x}[\sigma]}\Big) \cdot \Big(\sum_{\sigma\in\mathcal{E}} \mathbf{y}[\sigma]\mathbf{x}[\sigma]\Big) \\
&\geq \Big(\sum_{\sigma\in\mathcal{E}} |\mathbf{z}[\sigma]| \cdot \sqrt{\frac{\mathbf{y}[\sigma]}{\mathbf{x}[\sigma]}} \cdot \sqrt{\mathbf{y}[\sigma]\mathbf{x}[\sigma]}\Big)^2 \\
&= \Big(\sum_{\sigma\in\mathcal{E}} |\mathbf{z}[\sigma]| \cdot \mathbf{y}[\sigma]\Big)^2 = \langle |\mathbf{z}|, \mathbf{y}\rangle^2. \qquad \text{(D.4)}
\end{aligned}$$

where the first inequality follows from Lemma 4.4 which implies $\sum_{\sigma\in\mathcal{E}} \mathbf{y}[\sigma]\mathbf{x}[\sigma] \leq \|\mathbf{y}\|_{\mathcal{H},\infty} = 1$ as $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ and the second inequality is given by the Cauchy–Schwarz inequality.

Moreover, consider the difference between function $\varphi_{\mathbf{1}}(\cdot)$ and $\varphi_{\mathbf{y}}(\cdot)$, we have

$$\varphi_{\mathbf{1}-\mathbf{y}}(\mathbf{x}) := \varphi_{\mathbf{1}}(\mathbf{x}) - \varphi_{\mathbf{y}}(\mathbf{x}) = \sum_{j\in\mathcal{J}} \sum_{a\in\mathcal{A}_j} (1 - \mathbf{y}[ja])\mathbf{x}[ja] \ln\Big(\frac{\mathbf{x}[ja]}{\mathbf{x}[p_j]}\Big).$$

From transition kernel $\mathbf{y} \in \mathcal{Y}$, we always have $1 - \mathbf{y}[ja] \geq 0$ for any $j \in \mathcal{J}$ and $a \in \mathcal{A}_j$. Together with the fact that $a\ln(a/b)$ is convex for $a, b \in \mathbb{R}_{\geq 0}$, we have that the difference $\varphi_{\mathbf{1}-\mathbf{y}}$ is a positive combination of convex functions. Thus, function $\varphi_{\mathbf{1}-\mathbf{y}}$ is a convex function, which implies

$$\mathbf{z}^\top \nabla^2 \varphi_{\mathbf{1}-\mathbf{y}}(\mathbf{x})\mathbf{z} \geq 0. \qquad \text{(D.5)}$$

Using additivity of the Hessian and inequalities (D.5) and (D.4), we obtain

$$\mathbf{z}^\top \nabla^2 \varphi_{\mathbf{1}}(\mathbf{x})\mathbf{z} = \mathbf{z}^\top \nabla^2 \varphi_{\mathbf{y}}(\mathbf{x})\mathbf{z} + \mathbf{z}^\top \nabla^2 \varphi_{\mathbf{1}-\mathbf{y}}(\mathbf{x})\mathbf{z} \geq \langle |\mathbf{z}|, \mathbf{y}\rangle^2.$$

By taking the supremum among all transition kernels $\mathbf{y} \in \mathcal{Y}$, we reach the final statement

$$\|\mathbf{z}\|^2_{\nabla^2\varphi_{\mathbf{1}}(\mathbf{x})} = \mathbf{z}^\top \nabla^2 \varphi_{\mathbf{1}}(\mathbf{x})\mathbf{z} \geq \max_{\mathbf{y}\in\mathcal{Y}} \langle |\mathbf{z}|, \mathbf{y}\rangle^2 = \|\mathbf{z}\|^2_{\mathcal{H},1}.$$

This concludes that the weight-one dilated entropy $\varphi_{\mathbf{1}}$ is 1-strongly convex with respect to the treeplex ℓ_1 norm $\|\cdot\|_{\mathcal{H},1}$. □

D.2 Proof of Lemma 5.2

Lemma D.1. The value of the DilEnt regularizer for some strategy profile $\mathbf{x} \in \mathcal{Q}$ can be bounded by $-\ln|\mathcal{V}| \leq \varphi_{\mathbf{1}}(\mathbf{x}) \leq 0$, where $|\mathcal{V}|$ is the number of reduced normal-form strategies.

Proof. Since the DilEnt regularizer is the weighted summation of negative entropy, and the fact that negative entropy is always non-positive, we directly get $\varphi_{\mathbf{1}}(\mathbf{x}) \leq 0$. We will prove $\varphi_{\mathbf{1}}(\mathbf{x}) \geq -\ln|\mathcal{V}|$ by induction on the TFSDP from the bottom up. In specific, we will show that for any point $h \in \mathcal{H}$ and a corresponding strategy profile $\mathbf{x}_h \in \mathcal{Q}_h$, the DilEnt with retricted to $\mathcal{H}_h$ satisfies

$$\varphi_{\mathbf{1},h}(\mathbf{x}_h) := \sum_{j\in\mathcal{J}_h} \sum_{a\in\mathcal{A}_j} \mathbf{x}_h[ja] \ln\Big(\frac{\mathbf{x}_h[ja]}{\mathbf{x}_h[p_j]}\Big) \geq -\ln|\mathcal{V}_h|.$$

Case 1: The induction basis occurs at $h = \sigma \in \mathcal{E}$, where the statement holds since for any $\mathbf{x}_\sigma \in \mathcal{Q}_\sigma$,

$$\varphi_{\mathbf{1},\sigma}(\mathbf{x}_\sigma) = 0 = -\ln|\mathcal{V}_\sigma|,$$

where we have $|\mathcal{V}_\sigma| = 1$ from Lemma F.2.

https://doi.org/10.52202/079017-2804

Case 2: Consider some decision point $h = j \in \mathcal{J}$. We can decompose the DilEnt regularizer over $\mathbf{x}_j \in \mathcal{Q}_j$ according to

$$\varphi_{\mathbf{1},j}(\mathbf{x}_j) = \sum_{a \in \mathcal{A}_j} \mathbf{x}_j[ja] \ln \mathbf{x}_j[ja] + \sum_{a \in \mathcal{A}_j} \sum_{j' \in \mathcal{J}_{ja}} \sum_{a' \in \mathcal{A}_{j'}} \mathbf{x}_j[j'a'] \ln \left(\frac{\mathbf{x}_j[j'a']}{\mathbf{x}_j[p_{j'}]} \right) \tag{D.6}$$

Consider the vector $\mathbf{x}_{ja} := \mathbf{x}[\mathcal{E}_{ja}]/\mathbf{x}[ja]$ for some action $a \in \mathcal{A}_j$. If $\mathbf{x}[ja] = 0$, we have

$$\sum_{j' \in \mathcal{J}_{ja}} \sum_{a' \in \mathcal{A}_{j'}} \mathbf{x}_j[j'a'] \ln \left(\frac{\mathbf{x}_j[j'a']}{\mathbf{x}_j[p_{j'}]} \right) = 0 \geq -\mathbf{x}_j[ja] \ln |\mathcal{V}_{ja}|.$$

Otherwise, it satisfies that $\mathbf{x}_{ja} \in \mathcal{Q}_{ja}$ according to the tree-structure of TFSDP. Thus, we can write

$$\begin{aligned}
\sum_{j' \in \mathcal{J}_{ja}} \sum_{a' \in \mathcal{A}_{j'}} \mathbf{x}_j[j'a'] \ln \left(\frac{\mathbf{x}_j[j'a']}{\mathbf{x}_j[p_{j'}]} \right) &= \mathbf{x}_j[ja] \sum_{j' \in \mathcal{J}_{ja}} \sum_{a' \in \mathcal{A}_{j'}} \left(\frac{\mathbf{x}_j[j'a']}{\mathbf{x}_j[ja]} \right) \ln \left(\frac{\mathbf{x}_j[j'a']}{\mathbf{x}_j[p_{j'}]} \right) \\
&= \mathbf{x}_j[ja] \sum_{j' \in \mathcal{J}_{ja}} \sum_{a' \in \mathcal{A}_{j'}} \mathbf{x}_{ja}[j'a'] \ln \left(\frac{\mathbf{x}_{ja}[j'a']}{\mathbf{x}_{ja}[p_{j'}]} \right) \\
&= \mathbf{x}_j[ja] \varphi_{\mathbf{1},ja}(\mathbf{x}_{ja}) \\
&\geq -\mathbf{x}_j[ja] \ln |\mathcal{V}_{ja}|,
\end{aligned}$$

where the last inequality is given by induction hypothesis.

In general, it always satisfies that

$$\sum_{j' \in \mathcal{J}_{ja}} \sum_{a' \in \mathcal{A}_{j'}} \mathbf{x}_j[j'a'] \ln \left(\frac{\mathbf{x}_j[j'a']}{\mathbf{x}_j[p_{j'}]} \right) \geq -\mathbf{x}_j[ja] \ln |\mathcal{V}_{ja}|.$$

Plugging this inequality into (D.6) gives

$$\begin{aligned}
\varphi_{\mathbf{1},j}(\mathbf{x}_j) &\geq \sum_{a \in \mathcal{A}_j} \mathbf{x}_j[ja] \ln \mathbf{x}_j[ja] - \sum_{a \in \mathcal{A}_j} \mathbf{x}_j[ja] \ln |\mathcal{V}_{ja}| \\
&\geq \ln \Big(\sum_{a \in \mathcal{A}_j} \exp(-\ln |\mathcal{V}_{ja}|) \Big) \\
&= -\ln \Big(\sum_{a \in \mathcal{A}_j} |\mathcal{V}_{ja}| \Big) \\
&= -\ln |\mathcal{V}_j|.
\end{aligned}$$

where the first inequality holds since the minimizer is given by $\mathbf{x}_j[ja] \propto \exp(-\ln |\mathcal{V}_{ja}|)$, and the last equality is given by Lemma F.2.

Case 3: Consider some decision point $h = \sigma \in \Sigma$. We can decompose the weight-one dilated entropy according to the tree-structure of TFSDP. Thus, we can write

$$\begin{aligned}
\varphi_{\mathbf{1},\sigma}(\mathbf{x}_\sigma) &= \sum_{j \in \mathcal{C}_\sigma} \sum_{j' \in \mathcal{J}_j} \sum_{a' \in \mathcal{A}_{j'}} \mathbf{x}_\sigma[j'a'] \ln \left(\frac{\mathbf{x}_\sigma[j'a']}{\mathbf{x}_\sigma[p_{j'}]} \right) \\
&= \sum_{j \in \mathcal{C}_\sigma} \varphi_{\mathbf{1},j}(\mathbf{x}_\sigma[\mathcal{E}_j]) \\
&\geq \sum_{j \in \mathcal{C}_\sigma} -\ln |\mathcal{V}_j| \\
&= -\ln |\mathcal{V}_\sigma|,
\end{aligned}$$

where the second inequality is given by induction hypothesis, and the last equality is given by Lemma F.2.

In general, it always holds that $\varphi_{\mathbf{1},h}(\mathbf{x}_h) \geq -\ln |\mathcal{V}_h|$, which concludes the proof. Substituting this with $h = \varnothing$ reaches the desired result. □

Lemma 5.2 (restatement)**.** Let $\mathbf{x}_1 := \operatorname{argmin}_{\mathbf{x}\in\mathcal{Q}} \varphi_{\mathbf{1}}(\mathbf{x})$ be the strategy profile minimize the DilEnt regularizer. The Bregman divergence generated by the DilEnt regularizer between $\mathbf{x}_1$ and any $\mathbf{x}_* \in \mathcal{Q}$ can be upper bounded by $\mathcal{D}_{\varphi_{\mathbf{1}}}(\mathbf{x}_*, \mathbf{x}_1) \le \ln|\mathcal{V}|$.

Proof. From the definition of Bregman divergence $\mathcal{D}_{\varphi_{\mathbf{1}}}$, we can write

$$\mathcal{D}_{\varphi_{\mathbf{1}}}(\mathbf{x}_*, \mathbf{x}_1) = \varphi_{\mathbf{1}}(\mathbf{x}_*) - \varphi_{\mathbf{1}}(\mathbf{x}_1) - \langle \nabla\varphi_{\mathbf{1}}(\mathbf{x}_1), \mathbf{x}_* - \mathbf{x}_1\rangle \tag{D.7}$$

According to Lemma D.1, we have $\varphi_{\mathbf{1}}(\mathbf{x}_*) - \varphi_{\mathbf{1}}(\mathbf{x}_1) \le \ln|\mathcal{V}|$. From the chosen of $\mathbf{x}_1$, we have $\langle \nabla\varphi_{\mathbf{1}}(\mathbf{x}_1), \mathbf{x}_* - \mathbf{x}_1\rangle = 0$. By plugging both inequalities into (D.7), we reach the desired result

$$\mathcal{D}_{\varphi_{\mathbf{1}}}(\mathbf{x}_*, \mathbf{x}_1) \le \ln|\mathcal{V}|.$$

□

D.3 Proof of Theorem 5.4

Theorem 5.4 (restatement)**.** Let φ be a regularizer for extensive-form decision space $\mathcal{Q}$ which is μ-strongly convex on $\|\cdot\|_{\mathcal{H},1}$ and has a diameter $|\mathcal{D}| := \sup_{x_*\in\mathcal{Q}} \mathcal{D}_\varphi(x_*, x_1)$. Under Assumption 3.1, the cumulative regret of running OMD with regularizer φ and learning rate $\eta := \sqrt{2|\mathcal{D}|/(\mu T)}$ is upper bounded by

$$\text{Regret}(T) \le \sqrt{2|\mathcal{D}|/\mu}\sqrt{T}.$$

Moreover, if we use the DilEnt regularizer $\varphi_{\mathbf{1}}$ in proximal steps, the result can be specified as

$$\text{Regret}(T) \le \sqrt{2\ln|\mathcal{V}|}\sqrt{T}.$$

Proof. We will apply Theorem 5.3 with $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{\mathcal{H},\infty}$ be the desired primal-dual pair. In the context of the theorem, we have $\|\mathbf{w}_t\|_{\mathcal{H},\infty} \le 1$ for any $t \in [\![T]\!]$ from Lemma 4.4. Together with $\mathcal{D}_\varphi(\mathbf{x}_*, \mathbf{x}_1) \le |\mathcal{D}|$, when choosing learning rate $\eta := \sqrt{2|\mathcal{D}|/(\mu T)}$, we have the regret can be upper bounded by

$$\text{Regret}(T) \le \frac{1}{\eta}\cdot|\mathcal{D}| + \frac{\eta}{2}\cdot T = \sqrt{2|\mathcal{D}|/\mu}\sqrt{T}.$$

When selecting DilEnt as the regularizer, we have $|\mathcal{D}| \le \ln|\mathcal{V}|$ from Lemma 5.2 and $\mu \ge 1$ from Lemma 5.1. Plugging these results indicates

$$\text{Regret}(T) \le \sqrt{2\ln|\mathcal{V}|}\sqrt{T}.$$

□

D.4 Proof of Lemma 5.5

We first prove the lemma starts from the following locally Lipshitz.

Lemma D.2. In Algorithm 2, consider two joint policies that agree on all strategy profile expect for player $j \in [\![n]\!]$, $\pi_1 := \{\mathbf{x}_0^{(1)}, \cdots, \mathbf{x}_1^{(j)}, \cdots, \mathbf{x}_0^{(n)}\}$ and $\pi_2 := \{\mathbf{x}_0^{(1)}, \cdots, \mathbf{x}_2^{(j)}, \cdots, \mathbf{x}_0^{(n)}\}$. Under Assumption 3.1, we have the reward vector is locally Lipschitz under the treeplex ℓ_1 norm, that is, denote by $\mathbf{w}_1^{(i)} := \partial_i u^{(i)}(\pi_1)$ and $\mathbf{w}_2^{(i)} := \partial_i u^{(i)}(\pi_2)$ the reward vector of player i, it satisfies that

$$\|\mathbf{w}_1^{(i)} - \mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty} \le \|\mathbf{x}_1^{(j)} - \mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(i)},1}.$$

Proof. If $j = i$, we have $\mathbf{w}_1^{(i)} = \mathbf{w}_2^{(i)}$ and the statement holds true from

$$\|\mathbf{w}_1^{(i)} - \mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty} = 0 \le \|\mathbf{x}_1^{(j)} - \mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(i)},1}.$$

Otherwise it satisfies that $j \ne i$. According to the definition of treeplex ℓ_∞ norm, we have that

$$\|\mathbf{w}_1^{(i)} - \mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty} = \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}} \langle|\mathbf{w}_1^{(i)} - \mathbf{w}_2^{(i)}|, \mathbf{x}\rangle = \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}} \sum_{\sigma\in\mathcal{E}^{(i)}} |\mathbf{w}_1^{(i)}[\sigma] - \mathbf{w}_2^{(i)}[\sigma]|\cdot\mathbf{x}[\sigma]. \tag{D.8}$$

https://doi.org/10.52202/079017-2804

By expressing the reward vector using the strategy of other players, we have that

$$\mathbf{w}_1^{(i)}[\sigma] = \sum_{z\in\mathcal{I}_\sigma} \mathbf{u}^{(i)}[z]\cdot\mathbf{p}[z]\cdot\mathbf{x}_1^{(j)}[\sigma_z^{(j)}]\prod_{k\neq i,j}\mathbf{x}_0^{(k)}[\sigma_z^{(k)}]. \tag{D.9}$$

Plugging (D.9) into (D.8) gives

$$\begin{aligned}
&\|\mathbf{w}_1^{(i)}-\mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty} \qquad\qquad (D.10)\\
&= \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}}\sum_{\sigma\in\mathcal{E}^{(i)}}\mathbf{x}[\sigma]\cdot\Big|\sum_{z\in\mathcal{I}_\sigma}\mathbf{u}^{(i)}[z]\cdot\mathbf{p}[z]\cdot(\mathbf{x}_1^{(j)}[\sigma_z^{(j)}]-\mathbf{x}_2^{(j)}[\sigma_z^{(j)}])\prod_{k\neq i,j}\mathbf{x}_0^{(k)}[\sigma_z^{(k)}]\Big|\\
&\le \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}}\sum_{\sigma\in\mathcal{E}^{(i)}}\mathbf{x}[\sigma]\sum_{z\in\mathcal{I}_\sigma}\mathbf{p}[z]\cdot\big|\mathbf{x}_1^{(j)}[\sigma_z^{(j)}]-\mathbf{x}_2^{(j)}[\sigma_z^{(j)}]\big|\prod_{k\neq i,j}\mathbf{x}_0^{(k)}[\sigma_z^{(k)}]\\
&= \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}}\sum_{z\in\mathcal{Z}}\mathbf{x}[\sigma_z^{(i)}]\cdot\mathbf{p}[z]\cdot\big|\mathbf{x}_1^{(j)}[\sigma_z^{(j)}]-\mathbf{x}_2^{(j)}[\sigma_z^{(j)}]\big|\prod_{k\neq i,j}\mathbf{x}_0^{(k)}[\sigma_z^{(k)}], \qquad (D.11)
\end{aligned}$$

where the inequality holds since the reward $\mathbf{u}^{(i)}[z]\in[0,1]$ and the last equality is given by permuting the summation.

According to the definition of treeplex ℓ_1 norm, we have

$$\|\mathbf{x}_1^{(j)}-\mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(j)},1} = \sup_{\mathbf{y}\in\mathcal{Y}^{(j)}}\langle|\mathbf{x}_1^{(j)}-\mathbf{x}_2^{(j)}|,\mathbf{y}\rangle = \sup_{\mathbf{y}\in\mathcal{Y}^{(j)}}\sum_{\sigma\in\mathcal{E}^{(j)}}|\mathbf{x}_1^{(j)}[\sigma]-\mathbf{x}_2^{(j)}[\sigma]|\cdot\mathbf{y}[\sigma]. \tag{D.12}$$

By expressing the transition kernel using the strategy of other players, we can write

$$\mathbf{y}[\sigma] = \sum_{z\in\mathcal{I}_\sigma}\mathbf{x}^{(i)}[\sigma_z^{(i)}]\cdot\mathbf{p}[z]\prod_{k\neq i,j}\mathbf{x}^{(k)}[\sigma_z^{(k)}] \tag{D.13}$$

such that $\mathbf{x}^{(i)}\in\mathcal{Q}^{(i)}$ and $\mathbf{x}^{(k)}\in\mathcal{Q}^{(k)}$. Plugging (D.13) into (D.12) gives

$$\begin{aligned}
\|\mathbf{x}_1^{(j)}-\mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(j)},1} &= \sup_{\{\mathbf{x}^{(k)}\in\mathcal{Q}^{(k)}\}_{k\neq j}}\sum_{\sigma\in\mathcal{E}^{(j)}}|\mathbf{x}_1^{(j)}[\sigma]-\mathbf{x}_2^{(j)}[\sigma]|\sum_{z\in\mathcal{I}_\sigma}\mathbf{x}^{(i)}[\sigma_z^{(i)}]\cdot\mathbf{p}[z]\prod_{k\neq i,j}\mathbf{x}^{(k)}[\sigma_z^{(k)}]\\
&\ge \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}}\sum_{\sigma\in\mathcal{E}^{(j)}}|\mathbf{x}_1^{(j)}[\sigma]-\mathbf{x}_2^{(j)}[\sigma]|\sum_{z\in\mathcal{I}_\sigma}\mathbf{x}[\sigma_z^{(i)}]\cdot\mathbf{p}[z]\prod_{k\neq i,j}\mathbf{x}_0^{(k)}[\sigma_z^{(k)}]\\
&= \sup_{\mathbf{x}\in\mathcal{Q}^{(i)}}\sum_{z\in\mathcal{Z}}\mathbf{x}[\sigma_z^{(i)}]\cdot\mathbf{p}[z]\cdot\big|\mathbf{x}_1^{(j)}[\sigma_z^{(j)}]-\mathbf{x}_2^{(j)}[\sigma_z^{(j)}]\big|\prod_{k\neq i,j}\mathbf{x}_0^{(k)}[\sigma_z^{(k)}], \qquad (D.14)
\end{aligned}$$

where the inequality holds as $\mathbf{x}_0^{(k)}\in\mathcal{Q}^{(k)}$ and the last equality is given by permuting the summation. By combining (D.10) and (D.14), we can reach that

$$\|\mathbf{w}_1^{(i)}-\mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty}\le\|\mathbf{x}_1^{(j)}-\mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(j)},1}. \tag{D.15}$$

□

Lemma D.3. In Algorithm 2, for two joint policy $\pi_1=\{\mathbf{x}_1^{(j)}\}_{j=1}^n$ and $\pi_2=\{\mathbf{x}_2^{(j)}\}_{j=1}^n$ where $\mathbf{x}_1^{(j)},\mathbf{x}_2^{(j)}\in\mathcal{Q}^{(j)}$ are the strategy profiles for player $j\in[\![n]\!]$. Consider the corresponding reward vectors $\mathbf{w}_1^{(i)}:=\partial_i u^{(i)}(\pi_1)$ and $\mathbf{w}_2^{(i)}:=\partial_i u^{(i)}(\pi_2)$ of player $i\in[\![n]\!]$ when other players follow the joint policy. Under Assumption 3.1, we have the reward vector is Lipschitz under the treeplex ℓ_1 norm, that is,

$$\|\mathbf{w}_1^{(i)}-\mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty}\le\sum_{j=1}^n\|\mathbf{x}_1^{(j)}-\mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(j)},1}.$$

Proof. Consider a series of reward vector policy $\pi_{(j)}:=\{\mathbf{x}_2^{(1)},\cdots,\mathbf{x}_2^{(j)},\mathbf{x}_1^{(j+1)}\cdots,\mathbf{x}_1^{(n)}\}$ which are generated by the joint policy that aligns with joint policy π_2 on the first j players while aligns with π_1 on the rest. Denote by $\mathbf{w}_{(j)}^{(i)}:=\partial_i u^{(i)}(\pi_{(j)})$ the reward vector. Under this definition, it satisfies that $\mathbf{w}_{(0)}^{(i)}=\mathbf{w}_1^{(i)}$ and $\mathbf{w}_{(n)}^{(i)}=\mathbf{w}_2^{(i)}$. Therefore, we can write

$$\|\mathbf{w}_1^{(i)}-\mathbf{w}_2^{(i)}\|_{\mathcal{H}^{(i)},\infty}\le\sum_{j=1}^n\|\mathbf{w}_{(j-1)}^{(i)}-\mathbf{w}_{(j)}^{(i)}\|_{\mathcal{H}^{(i)},\infty}\le\sum_{j=1}^n\|\mathbf{x}_1^{(j)}-\mathbf{x}_2^{(j)}\|_{\mathcal{H}^{(j)},1},$$

where the first inequality follows from the triangle inequality and the second inequality is given by Lemma D.2. □

Lemma 5.5 (restatement)**.** Under Assumption 3.1, when running COMD with DilEnt, the reward vector $\mathbf{w}_{t,l}^{(i)}$ received by player i in any $(t,l) \in [\![K]\!] \times [\![L]\!]$ satisfies $\|\mathbf{w}_{t,l+1}^{(i)} - \mathbf{w}_{t,l}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq 2(n\eta)^{l-1}$.

Proof. By applying Lemma 3.2 to the proximal steps $\mathbf{x}_{t,l}^{(i)} \leftarrow \Pi_{\varphi_1^{(i)}}(\eta \mathbf{w}_{t,l-1}^{(i)}, \widetilde{\mathbf{x}}_t^{(i)})$ and $\mathbf{x}_{t,l+1}^{(i)} \leftarrow \Pi_{\varphi_1^{(i)}}(\eta \mathbf{w}_{t,l}^{(i)}, \widetilde{\mathbf{x}}_t^{(i)})$, we have that for every $l \geq 1$,

$$\|\mathbf{x}_{t,l+1}^{(i)} - \mathbf{x}_{t,l}^{(i)}\|_{\mathcal{H}^{(i)},1} \leq \eta \|\mathbf{w}_{t,l}^{(i)} - \mathbf{w}_{t,l-1}^{(i)}\|_{\mathcal{H}^{(i)},\infty}, \tag{D.16}$$

where the strongly convex modulus $\mu \geq 1$ is given by Lemma 5.1. From Lemma D.3, we have that

$$\|\mathbf{w}_{t,l+1}^{(i)} - \mathbf{w}_{t,l}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq \sum_{j=1}^{n} \|\mathbf{x}_{t,l+1}^{(j)} - \mathbf{x}_{t,l}^{(j)}\|_{\mathcal{H}^{(i)},1}. \tag{D.17}$$

In addition, the difference between the initial steps can be upper bounded according to

$$\|\mathbf{w}_{t,2}^{(i)} - \mathbf{w}_{t,1}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq \|\mathbf{w}_{t,2}^{(i)}\|_{\mathcal{H}^{(i)},\infty} + \|\mathbf{w}_{t,1}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq 2, \tag{D.18}$$

where the first inequality follows from the triangle inequality and the second inequality is given by Lemma 4.4. By combining (D.16), (D.17), and (D.18), we reach the desired statement:

$$\|\mathbf{w}_{t,l+1}^{(i)} - \mathbf{w}_{t,l}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq 2(n\eta)^{l-1}.$$

□

D.5 Proof of Theorem 5.6

Theorem 5.6 (restatement)**.** Under Assumption 3.1, if every player runs Clairvoyant OMD with DilEnt regularizer and learning rate $\eta = 1/(2n)$ for K episodes. With $L = \lceil \log K \rceil$ steps of inner iterations, the average joint policy $\bar{\pi}_K$ is an ε-CCE for $\varepsilon \leq \mathcal{O}(n \ln |\mathcal{V}|/K)$. This implies the algorithm converge to a CCE at rate $\mathcal{O}(n \log |\mathcal{V}| \log T/T)$ where $T = KL$ is the number of oracle access to the game.

Proof. When $\eta \leq 1/(2n)$, we have $n\eta \leq 1/2$. According to Lemma 5.5, the difference between the actual reward vector and the prediction can be upper bounded by

$$\|\mathbf{w}_t^{(i)} - \mathbf{m}_t^{(i)}\|_{\mathcal{H}^{(i)},\infty} = \|\mathbf{w}_{t,L+1}^{(i)} - \mathbf{w}_{t,L}^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq 2(n\eta)^{L-1} \leq 2^{2-L}.$$

Therefore, with $L = \lceil \log K \rceil$ steps of fixed-point iterations, the difference can be as small as $\|\mathbf{w}_t^{(i)} - \mathbf{m}_t^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq 4/K$.

Let $\|\cdot\|_{\mathcal{H}^{(i)},1}$ and $\|\cdot\|_{\mathcal{H}^{(i)},\infty}$ be the pair of primal-dual norms required by Theorem 5.3. In the context of the theorem, we have $\|\mathbf{w}_t\|_{\mathcal{H}^{(i)},\infty} \leq 1$ for any $t \in [\![K]\!]$ given by the definition of treeplex ℓ_∞ norm, $\mathcal{D}_{\varphi_1}(\mathbf{x}_*, \mathbf{x}_1) \leq \ln |\mathcal{V}|$ according to Lemma 5.2, $\mu \geq 1$ according to Lemma 5.1, and $\|\mathbf{w}_t^{(i)} - \mathbf{m}_t^{(i)}\|_{\mathcal{H}^{(i)},\infty} \leq 1/K$ from the reasoning above. Plugging these results into the statement of Theorem 5.3 gives:

$$\text{Regret}(K) \leq \frac{1}{\eta} \cdot \ln |\mathcal{V}| + \frac{\eta K}{2} \cdot \left(\frac{4}{K}\right)^2.$$

With $\eta = 1/(2n)$, we get that

$$\text{Regret}(K) \leq \mathcal{O}(n \log |\mathcal{V}|).$$

According to the online-to-batch conversion [see e.g. 36], we can conclude that the average joint policy $\bar{\pi}_K$ is an ε-CCE with $\varepsilon \leq \mathcal{O}(n \log |\mathcal{V}|/K)$. Given oracle access budget T, we can select $K = \lfloor T/\log T \rfloor$ satisfying $KL \leq T$. This indicates that the algorithm converge to a CCE with approximation rate $\mathcal{O}(n \log |\mathcal{V}| \log T/T)$. □

https://doi.org/10.52202/079017-2804

E Proof of Regret Lower Bounds

E.1 Proof of Theorem 6.1

Theorem 6.1 (restatement)**.** Given a TFSDP with decision space $\mathcal{Q}$, there is an EFG satisfying Assumption 3.1 such that: when the other players are controlled by the adversary, any algorithm `Alg` incurs an expected regret of at least $\Omega(\sqrt{\|\mathcal{Q}\|_\perp \log |\mathcal{A}_0| \sqrt{T}})$ for a given of episode number $T \geq \|\mathcal{Q}\|_\perp$, where $|\mathcal{A}_0| := \min_{j \in \mathcal{J}} |\mathcal{A}_j|$ is the size of the minimum action set.

Proof. We will prove the statement by induction on $\mathcal{H}$ from the bottom up. For each point $h \in \mathcal{H} \setminus \mathcal{E}$, we construct a hard instance $\mathcal{I}_h(T)$ such that any algorithm $\mathcal{H}$ playing in subgame $\mathcal{H}_h$ incurs an expected regret of at least $\Omega(\sqrt{\|\mathcal{Q}_h\|_\perp \log |\mathcal{A}_0| \sqrt{T}})$. We construct hard instance $\mathcal{I}_h(T)$ to be a two-player perfect-information EFG that has the same structure as the given TFSDP, where all the observation points are the decision points of an adversarial opponent. It can be represented using a set of random variables $\{\mathbf{y}_{h,t}, \mathbf{r}_{h,t}\}_{t=1}^T$, where $\mathbf{y}_{h,t} \in \mathcal{Y}_h$ is the transition kernel and $\mathbf{r}_{h,t} \in [0,1]^{\mathcal{E}_h}$ encodes the expected reward of player conditional on each terminal observation point. Note that the transition kernel $\mathbf{y}_{h,t}$ is also the strategy profile of the opponent in his extensive-form decision space. In this case, the expected reward of playing strategy profile $\mathbf{x}_{h,t}$ on $\mathcal{H}_h$ on episode t can be computed by $\langle \mathbf{x}_{h,t}, \mathbf{w}_{h,t} \rangle$ where $\mathbf{w}_{h,t} := \mathbf{r}_{h,t} \odot \mathbf{y}_{h,t}$ is the reward vector.

Case 1: The base case of the induction is that $h = j \in \mathcal{J}$ and $ja \in \mathcal{E}$ holds for every $a \in \mathcal{A}_j$. In this case, the TFSDP with point set $\mathcal{H}_h$ is equivalent to a full-information multi-arm bandit problem (a.k.a. expert problem). We construct the hard instance $\mathcal{I}_j(T)$ by assigning

$$\mathbf{y}_{j,t}[ja] = 1, \mathbf{r}_{j,t}[ja] = \mathrm{Unif}(\{0,1\})$$

for every episode $t \in [\![T]\!]$, where $\mathrm{Unif}(\{0,1\})$ is the Bernoulli random variable with $p = 0.5$. In this case, the entries of the reward vector $\mathbf{w}_{j,t} := \mathbf{r}_{j,t} \odot \mathbf{y}_{j,t}$ are independently random variables. Thus, the expected cumulative reward among T episodes of any algorithm `Alg` can be computed by

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{x}_{j,t}, \mathbf{w}_{j,t} \rangle\right] = \sum_{t=1}^T \frac{1}{2} = \frac{T}{2}. \tag{E.1}$$

Additionally, the cumulative reward of the optimal policy can be computed according to

$$\begin{aligned}
\mathbb{E}\left[\max_{a \in \mathcal{A}_j} \sum_{t=1}^T \langle \mathbf{e}_{ja}, \mathbf{w}_{j,t} \rangle\right] &= \mathbb{E}\left[\max_{a \in \mathcal{A}_j} \sum_{t=1}^T \left(\frac{1}{2}\sigma_{a,t} + \frac{1}{2}\right)\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[\max_{a \in \mathcal{A}_j} \sum_{t=1}^T \sigma_{a,t}\right] + \frac{T}{2} \\
&\geq \Omega(\sqrt{T \log |\mathcal{A}_j|}) + \frac{T}{2},
\end{aligned} \tag{E.2}$$

where $\mathbf{e}_{ja}$ is the pure strategy profile that always executes action a at decision point j and $\sigma_{a,t} \sim \mathrm{Unif}(\{-1,1\})$ are independent Rademacher random variables for $a \in \mathcal{A}_j$ and $t \in [\![T]\!]$. The last inequality follows from Lemma A.11 in Cesa-Bianchi and Lugosi [9]. Combining (E.1) and (E.2) indicates that any algorithm suffer an expected regret of at least

$$\mathrm{Regret}_j(T) = \mathbb{E}\left[\max_{a \in \mathcal{A}_j} \sum_{t=1}^T \langle \mathbf{e}_{ja}, \mathbf{w}_{j,t} \rangle\right] - \mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{x}_{j,t}, \mathbf{w}_{j,t} \rangle\right] \geq \Omega\Big(\sqrt{\log |\mathcal{A}_j|}\sqrt{T}\Big).$$

Note that for the $j \in \mathcal{J}$, we have $\|\mathcal{Q}_j\|_\perp = 1$ from Lemma F.1. Together with $|\mathcal{A}_j| \geq |\mathcal{A}_0|$ from the definition of $|\mathcal{A}_0|$, we conclude that

$$\mathrm{Regret}_j(T) \geq \Omega\Big(\sqrt{\|\mathcal{Q}_j\|_\perp \log |\mathcal{A}_0|}\sqrt{T}\Big),$$

establishing the induction basis.

Case 2: For any other decision point $h = j \in \mathcal{J}$ that is non-terminal, let $a_* = \mathrm{argmax}_{a \in \mathcal{A}} \|\mathcal{Q}_{ja}\|_\perp$ be the action that maximizes the leaf count in the subtree (breaking ties arbitrarily). Let $\mathcal{I}_{ja_*}(T) =$

$\{\mathbf{y}_{ja_*,t}, \mathbf{r}_{ja_*,t}\}_{t=1}^T$ be the hard instance built in the subtree satisfying the induction hypothesis. We construct the hard instance $\mathcal{I}_j(T) = \{\mathbf{y}_{ja_*,t}, \mathbf{r}_{ja_*,t}\}_t$ according to

$$\mathbf{y}_{j,t}[h] = \begin{cases} \mathbf{y}_{ja_*,t}[\sigma] & \text{if } \sigma \in \mathcal{E}(ja_*) \\ \perp & \text{otherwise} \end{cases}, \mathbf{r}_{j,t}[\sigma] = \begin{cases} \mathbf{r}_{ja_*,t}[\sigma] & \text{if } \sigma \in \mathcal{E}(ja_*) \\ 0 & \text{otherwise} \end{cases}$$

where $\perp$ refers to any valid transition kernel. In other words, the construction ensures taking action a_* leads to the hard instance $\mathcal{I}_{ja*}(T)$ while taking any other actions always results in a reward of 0.

Since any action different from a_* will result in a reward of 0, any algorithm `Alg` that ever assigned weight outside ja_* end up in a lower cumulative reward compared to the other algorithm `Alg`$'$ that consistently play action a_* on decision point j. Together with the fact that playing a_* will reduce the problem to $\mathcal{I}_{ja_*}(T)$, no algorithm can achieve a higher reward on the hard instance $\mathcal{I}_j(T)$ comparing to $\mathcal{I}_{ja_*}(T)$. Moreover, the strategy profile given by choosing a_* then following the optimal strategy profile on $\mathcal{I}_{ja_*}(T)$ leads to the same cumulative reward as the optimal strategy profile on $\mathcal{I}_{ja_*}(T)$. This indicates that the cumulative reward of the optimal strategy profile on $\mathcal{I}_j(T)$ is no less than its counterpart on $\mathcal{I}_{ja_*}(T)$. Combining the two statements, we have that the regret lower bound on $\mathcal{I}_j(T)$ is no less than the regret lower bound on $\mathcal{I}_{ja_*}(T)$. This implies the regret of any algorithm `Alg` can be lower bounded by

$$\text{Regret}_j(T) \geq \Omega\Big(\sqrt{\|\mathcal{Q}_{ja_*}\|_\perp \log|\mathcal{A}_0|}\sqrt{T}\Big) = \Omega\Big(\sqrt{\|\mathcal{Q}_j\|_\perp \log|\mathcal{A}_0|}\sqrt{T}\Big).$$

where the last equality follows from Lemma F.1, which indicates $\|\mathcal{Q}_j\|_\perp = \max_{a\in\mathcal{A}_j}\|\mathcal{Q}_{ja}\|_\perp = \|\mathcal{Q}_{ja_*}\|_\perp$, where the last equality follows from the choice of action a_*.

Case 3: If $h = \sigma \in \Sigma \setminus \mathcal{E}$ is some non-terminal observation point, we construct the hard instance $\mathcal{I}_\sigma(T)$ by concatenating several hard instance blocks, each block for one observation outcome $j \in \mathcal{C}_\sigma$. For $j \in \mathcal{C}_\sigma$, taking observation at point σ always leads to decision point j. Let $\{T_j \in \mathbb{Z}_{\geq 0}\}_{j\in\mathcal{C}_\sigma}$ be a partition which maximizes $\sum_j T_j\|\mathcal{Q}_j\|_\perp$ under the constraint $\sum_j T_j = T$. This ensures $T_j \approx T\|\mathcal{Q}_j\|_\perp/\|\mathcal{Q}_\sigma\|_\perp$. We construct the hard instance block $\mathcal{I}_{\sigma\to j}(T_j) = \{\mathbf{r}_{\sigma,t}, \mathbf{y}_{\sigma,t}\}_{t=1}^{T_j}$ using the hard instance $\mathcal{I}_j(T_j) = \{\mathbf{r}_{j,t}, \mathbf{y}_{j,t}\}_{t=1}^{T_j}$ in $\mathcal{H}_j$ given by the induction, by assigning

$$\mathbf{y}_{\sigma,t}[\sigma'] = \begin{cases} \mathbf{y}_{j,t}[\sigma'] & \text{if } \sigma' \in \mathcal{E}(j) \\ \perp & \text{otherwise} \end{cases}, \mathbf{r}_{\sigma,t}[\sigma'] = \begin{cases} \mathbf{r}_{j,t}[\sigma'] & \text{if } \sigma' \in \mathcal{E}(j) \\ 0 & \text{otherwise} \end{cases}.$$

We construct the hard instance $\mathcal{I}_\sigma(T)$ by concatenating $\mathcal{I}_{\sigma\to j}(T_j)$ over episodes.

From the property of observation point, the cumulative regret of any algorithm on hard instance $\mathcal{I}_{\sigma\to j}(T_j)$ is equal to that on $\mathcal{I}_j(T_j)$. Since hard instances $\mathcal{I}_j(T_j)$ are independent on the decision space, the cumulative regret on $\mathcal{I}_\sigma(T)$ is the summation of the cumulative regret among all $\mathcal{I}_{\sigma\to}(T_j)$. This indicates that any algorithm Alg will suffer a regret of at least

$$\begin{aligned}
\text{Regret}_\sigma(T) &= \sum_{j\in\mathcal{C}_\sigma} \text{Regret}_{\sigma\to j}(T_j) \\
&= \sum_{j\in\mathcal{C}_\sigma} \text{Regret}_j(T_j) \\
&\geq \sum_{j\in\mathcal{C}_\sigma} \Omega\Big(\sqrt{\|\mathcal{Q}_j\|_\perp \log|\mathcal{A}_0|}\sqrt{T_j}\Big) \\
&\geq \sum_{j\in\mathcal{C}_\sigma} \Omega\Big(\sqrt{\|\mathcal{Q}_j\|_\perp \log|\mathcal{A}_0|}\sqrt{T\|\mathcal{Q}_j\|_\perp/\|\mathcal{Q}_\sigma\|_\perp}\Big) \\
&= \Omega(\sqrt{\|\mathcal{Q}_\sigma\|_\perp \log|\mathcal{A}_0|}\sqrt{T}),
\end{aligned}$$

where the second inequality is given by the assignment of T_j and the last equality is given by Lemma F.1 in which $\|\mathcal{Q}_\sigma\|_\perp = \sum_{j\in\mathcal{C}_\sigma}\|\mathcal{Q}_j\|_\perp$.

In general, the induction hypothesis always holds, indicating that for any algorithm `Alg`, it suffers an expected regret of at least $\Omega(\sqrt{\|\mathcal{Q}_h\|_\perp \log|\mathcal{A}_0|}\sqrt{T})$ in subtree h. The desired result can be reached by inspecting $h = \varnothing$. □

https://doi.org/10.52202/079017-2804

E.2 Proof of Lemma 6.2

We first establish a structural result for TFSDPs where no non-root observation point yield exactly one outcome.

Lemma E.1. For a TFSDP with a given point set $\mathcal{H}$, if no non-root observation point yield exactly one outcome, that is, $|\mathcal{C}_\sigma| \geq 2$ for any $\sigma \in \Sigma^+ \setminus \mathcal{E}$, then it follows that $\|\mathcal{Q}\|_1 \leq 2\|\mathcal{Q}\|_\perp$.

Proof. We prove the statement by induction on $\mathcal{H}$ from the bottom up, showing that for any strategy profile $\mathbf{x} \in \mathcal{Q}$, and any non-root point $h \in \mathcal{H} \setminus \{\varnothing\}$, it satisfies that

$$\|\mathbf{x}[\Sigma_h]\|_1 := \sum_{\sigma \in \Sigma_h} \mathbf{x}[\sigma] \leq \mathbf{x}[h] \cdot (2\|\mathcal{Q}_h\|_\perp - 1).$$

Case 1: The base case for the induction occurs when $h = \sigma \in \mathcal{E}$ is a terminal node. In this scenario, the statement holds true since

$$\sum_{\sigma \in \Sigma_h} \mathbf{x}[\sigma] = \mathbf{x}[h] \leq \mathbf{x}[\sigma] \cdot (2\|\mathcal{Q}_\sigma\|_\perp - 1).$$

where the last inequality is given by Lemma F.1 in which $2\|\mathcal{Q}_\sigma\|_\perp = 1$.

Case 2: For any decision point $h = j \in \mathcal{J}$, it holds that

$$\begin{aligned}
\sum_{\sigma \in \Sigma_j} \mathbf{x}[\sigma] &= \sum_{a \in \mathcal{A}_j} \sum_{\sigma \in \Sigma_{ja}} \mathbf{x}[\sigma] \\
&\leq \sum_{a \in \mathcal{A}_j} \mathbf{x}[ja] \cdot (2\|\mathcal{Q}_{ja}\|_\perp - 1) \\
&\leq \sum_{a \in \mathcal{A}_j} \mathbf{x}[ja] \cdot (2\|\mathcal{Q}_j\|_\perp - 1) \\
&= \mathbf{x}[j] \cdot (2\|\mathcal{Q}_j\|_\perp - 1),
\end{aligned}$$

where the first equality follows from to the tree hierarchy of Σ, the first inequality follows the induction hypothesis, the second inequality holds since $\|\mathcal{Q}_{ja}\|_\perp \leq \|\mathcal{Q}_j\|_\perp$ implied by Lemma F.1, and the last equality follows from $\mathbf{x}[j] = \sum_{a \in \mathcal{A}_j} \mathbf{x}[ja]$ as $\mathbf{x} \in \mathcal{Q}$. This indicates that $\sum_{\sigma \in \Sigma(h)} \mathbf{x}[\sigma] \leq (2\|\mathcal{Q}_h\|_\perp - 1) \cdot \mathbf{x}[h]$ holds in this case.

Case 3: For any non-root-non-terminal observation point $h = \sigma \in \Sigma^+ \setminus \mathcal{E}$, it holds that

$$\begin{aligned}
\sum_{\sigma' \in \Sigma_\sigma} \mathbf{x}[\sigma'] &= \mathbf{x}[\sigma] + \sum_{j \in \mathcal{C}_\sigma} \sum_{\sigma' \in \Sigma_j} \mathbf{x}[\sigma'] \\
&\leq \mathbf{x}[\sigma] + \sum_{j \in \mathcal{C}_\sigma} \mathbf{x}[j] \cdot (2\|\mathcal{Q}_j\|_\perp - 1) \\
&= \mathbf{x}[\sigma] \cdot (1 + 2\|\mathcal{Q}_\sigma\|_\perp - |\mathcal{C}_\sigma|) \\
&\leq \mathbf{x}[\sigma] \cdot (2\|\mathcal{Q}_\sigma\|_\perp - 1),
\end{aligned}$$

where the the first equality holds due to the tree hierarchy of Σ, the first inequality follows the induction hypothesis, the second equality holds since $\mathbf{x}[j'] = \mathbf{x}[\sigma]$ for $j' \in \mathcal{C}_\sigma$ from $\mathbf{x} \in \mathcal{Q}$ as well as $\sum_{j' \in \mathcal{C}_\sigma} \|\mathcal{Q}_{j'}\|_\perp = \|\mathcal{Q}_\sigma\|_\perp$ from Lemma F.1, and the last inequality follows from the assumption that $|\mathcal{C}_\sigma| \geq 2$.

In general, for any non-root point $h \in \Sigma^+$, it satisfies that

$$\sum_{\sigma \in \Sigma_h} \mathbf{x}[\sigma] \leq (2\|\mathcal{Q}_h\|_\perp - 1) \cdot \mathbf{x}[h].$$

Now consider the root point $h = \varnothing$, we have that

$$\begin{aligned}\sum_{\sigma\in\Sigma}\mathbf{x}[\sigma] &= \mathbf{x}[\varnothing] + \sum_{j\in\mathcal{C}_\varnothing}\sum_{\sigma\in\Sigma_j}\mathbf{x}[\sigma]\\ &\le \mathbf{x}[\varnothing] + \sum_{j\in\mathcal{C}_\varnothing}\mathbf{x}[j]\cdot(2\|\mathcal{Q}_j\|_\perp - 1)\\ &= 1 + 2\|\mathcal{Q}_\sigma\|_\perp - |\mathcal{C}_\sigma|\\ &\le 2\|\mathcal{Q}_\sigma\|_\perp,\end{aligned}$$

where the the first equality holds due to the tree hierarchy of Σ, the first inequality follows the induction hypothesis, the second equality holds since $\mathbf{x}[j'] = \mathbf{x}[\varnothing] = 1$ for $j' \in \mathcal{C}_\varnothing$ from $\mathbf{x} \in \mathcal{Q}$ as well as $\sum_{j'\in\mathcal{C}_\varnothing}\|\mathcal{Q}_{j'}\|_\perp = \|\mathcal{Q}_\varnothing\|_\perp$ from Lemma F.1, and the last inequality follows from $|\mathcal{C}_\varnothing| \ge 1$. According to the definition that $\|\mathcal{Q}\|_\perp = \max\sum_{\sigma\in\Sigma}\mathbf{x}[\sigma]$, we conclude that $\|\mathcal{Q}\|_\perp \le 2\|\mathcal{Q}\|_1$ if $|\mathcal{C}_\sigma| \ge 2$ for any $\sigma \in \Sigma^+ \setminus \mathcal{E}$. □

The next lemma shows any TFSDP can be transformed into a TFSDP where no non-root observation point yield exactly one outcome.

Lemma E.2. Given TFSDP with a given point set $\mathcal{H}_0$, it can always be represented using another TFSDP $\mathcal{H}$ with the same compressed extensive-form decision space $\mathcal{Q}$ such that no non-root observation point yield exactly one outcome. Furthermore, the leaf count $\|\mathcal{Q}\|_\perp$ remains unchanged after the transformation, while the total number of actions $\sum_{j\in\mathcal{J}}|\mathcal{A}_j|$ does not increase.

Proof. We first present a transformation which removes each non-root observation point that yields only one outcome while makes the compressed extensive-form decision space $\mathcal{Q}$ remain unchanged.

Let $\mathcal{H}_0$ be a TFSDP, where is some non-root observation point $\sigma = ja \in \Sigma^+$ yields only one outcome, say $\mathcal{C}_\sigma = \{j'\}$ for a single j'. The transformation is achieved by examining the local reduced normal-form strategy at decision point j. As taking action a at decision point j invariably leads to the state transition to decision point j', we can dictate the agent's actions at j' contingent on the choice of action a at point j. This eliminates the observation point σ while the compressed extensive-form decision space $\mathcal{Q}$ remains unchanged, and the number of actions is reduced by 1 since ja is eliminated. We present an example of this transformation in Figure 2.

Since the number of actions is bounded, this process will always terminate. At this stage, there is no non-root observation point yield exactly one outcome, while the compressed extensive-form decision space $\mathcal{Q}$ remains unchanged. The total number of actions $\sum_{j\in\mathcal{J}}|\mathcal{A}_j|$ does not increase. Furthermore, since the leaf count $\|\mathcal{Q}\|_\perp$ can be determined by $\mathcal{Q}$ alone, this suggests the leaf count $\|\mathcal{Q}\|_\perp$ also remains unchanged after the transformation.

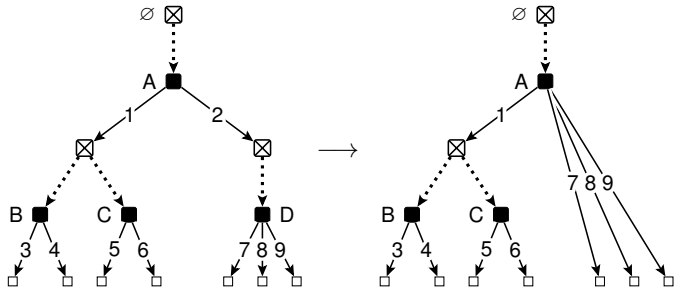

Figure 2: Eliminating observation point A2 from the TFSDP. The compressed extensive-form decision space $\mathcal{Q}$ remains unchanged and still has support $\{3, 4, 5, 6, 7, 8, 9\}$. Thus, the leaf count for the new TFSDP remains $\|\mathcal{Q}\|_\perp = 2$. Furthermore the total number of actions is reduced by one.

□

Lemma 6.2 (restatement)**.** Consider a TFSDP with a given point set $\mathcal{H}$. Define $|\mathcal{A}| := \max_{j\in\mathcal{J}}|\mathcal{A}_j|$ as the size of the largest action set. If there is no non-root observation point yields exactly one observation outcome, that is, $|\mathcal{C}_\sigma| \ge 2$ for any $\sigma \in \Sigma^+ \setminus \mathcal{E}$, then it follows that $\ln|\mathcal{V}| \le \mathcal{O}(\|\mathcal{Q}\|_\perp \log|\mathcal{A}|)$. Without this structural condition, we have $\ln|\mathcal{V}| \le \mathcal{O}(\|\mathcal{Q}\|_\perp \log|\mathcal{J}\times\mathcal{A}|)$ in general.

https://doi.org/10.52202/079017-2804

Proof of Lemma 6.2. According to Proposition 5.1 in Farina et al. [20], we have $|\mathcal{V}| \leq |\mathcal{A}|^{\|\mathcal{Q}\|_1}$. When there is no non-root observation point yields exactly one outcome, we have $\|\mathcal{Q}\|_1 \leq 2\|\mathcal{Q}\|_\perp$ from Lemma E.1. This implies $\ln|\mathcal{V}| \leq \mathcal{O}(\|\mathcal{Q}\|_1 \log|\mathcal{A}|)$.

In the general cases, we can use Lemma E.2 to transform the TFSDP into the desired representation. In this case, neither the leaf count $\|\mathcal{Q}\|_\perp$ nor the total number of actions $\sum_{j\in\mathcal{J}}|\mathcal{A}_j|$ increases. The size of the largest action set can be upper bounded using the total number of actions, which can be further upper bounded from $\sum_{j\in\mathcal{J}}|\mathcal{A}_j| \leq |\mathcal{J}\times\mathcal{A}|$. This implies $\ln|\mathcal{V}| \leq \mathcal{O}(\|\mathcal{Q}\|_1 \log|\mathcal{J}\times\mathcal{A}|)$ always holds.

□

F Auxiliary Lemmas

Lemma F.1. The leaf count $\|\mathcal{Q}\|_\perp = \|\mathcal{Q}_\varnothing\|_\perp$ can be computed recursively as:

- If $h = \sigma \in \mathcal{E}$ is a terminal observation point, then:
$$\|\mathcal{Q}_\sigma\|_\perp := 1.$$
- If $h = j \in \mathcal{J}$ is a decision point, then:
$$\|\mathcal{Q}_j\|_\perp := \max_{a\in\mathcal{A}_j}\|\mathcal{Q}_{ja}\|_\perp.$$
- If $h = \sigma \in \Sigma\setminus\mathcal{E}$ is a non-terminal observation point, then:
$$\|\mathcal{Q}_\sigma\|_\perp := \sum_{j\in\mathcal{C}_\sigma}\|\mathcal{Q}_j\|_\perp.$$

Proof. According to the definition of leaf count and the treeplex ℓ_∞ norm, we have

$$\|\mathcal{Q}\|_\perp = \sup_{\mathbf{x}\in\mathcal{Q}}\|\mathbf{x}\|_1 = \sup_{\mathbf{x}\in\mathcal{Q}}\langle|\mathbf{1}|,\mathbf{x}\rangle = \|\mathbf{1}\|_{\mathcal{H},\infty}.$$

This establishes a connection between leaf count with the treeplex ℓ_∞ norm. According to Lemma 4.2, we immediately reach the desired statement. □

Lemma F.2. The number of pure strategy profiles $|\mathcal{V}| = |\mathcal{V}_\varnothing|$ can be computed recursively as

- If $h = \sigma \in \mathcal{E}$ is a terminal observation point, then:
$$|\mathcal{V}_\sigma| := 1.$$
- If $h = j \in \mathcal{J}$ is a decision point, then:
$$|\mathcal{V}_j| := \sum_{a\in\mathcal{A}_j}|\mathcal{V}_{ja}|.$$
- If $h = \sigma \in \Sigma\setminus\mathcal{E}$ is a non-terminal observation point, then:
$$|\mathcal{V}_\sigma| := \prod_{j\in\mathcal{C}_\sigma}|\mathcal{V}_j|.$$

Proof. According to the definition of reduced normal-form strategies, the statement can be immediately reached by inspecting the vertices of each extensive-form strategy space $\mathcal{Q}_h$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We study the optimality of weight-one dilated entropy from a theoretical perceptive. The contribution, assumptions and scope are clearly claimed in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitation in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-2804

Answer: [Yes]

Justification: We provide detailed proof for all theorems in the Appendix C, D, and E.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper focuses on theoretical understanding and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper focuses on theoretical understanding and does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper focuses on theoretical understanding and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper focuses on theoretical understanding and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-2804

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper focuses on theoretical understanding and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work serves as foundational research for algorithmic game theory. Although there might be some potential social impacts on solving large-scale games, according to the guidelines, we believe our result does not have a direct connection with these issues.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Does not apply.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/079017-2804

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Does not apply.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects nor crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects nor crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2804