Counterfactual-Free Regret Minimization for Sequential Decision Making and Extensive-Form Games

Gabriele Farina[1] and **Robin Schmucker**[2] and **Tuomas Sandholm**[1,2,3]
[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Strategic Machine, Inc., Strategy Robot, Inc., and Optimized Markets, Inc., Pittsburgh, PA 15213, USA
{gfarina, rschmuck, sandholm}@cs.cmu.edu

Abstract

Sequential decision processes (SDPs) model the multi-stage online decision-making problems that each player faces in an extensive-form game, as well as MDPs and POMDPs where the agent conditions on observed history. Prior regret minimization approaches for sequential decision processes typically rely heavily on having access to *counterfactuals*, that is, information on what *would have happened* had the agent chosen a different action at any decision point. While this assumption is reasonable when regret minimization algorithms are used in self-play (for instance, as a way to converge to a Nash equilibrium in an extensive-form game), it is unrealistic in online decision-making settings, where the algorithm is deployed to learn strategies against an unknown environment. In this paper, we give the first efficient algorithm for the bandit linear optimization problem on SDPs—and therefore also extensive-form games—and show that it achieves $O(\sqrt{T})$ cumulative regret in expectation against any strategy.

1 Introduction

Sequential decision processes (SDPs) are multi-stage online decision-making problems. In an SDP, an agent interacts sequentially with a potentially adversarial environment in two ways: (i) decision points, in which an action must be selected by the agent; and (ii) observation points, in which the environment reveals a signal to the agent. Decision points and observation points alternate along a tree-like structure. SDPs model the online decision process that each player faces in an extensive-form game, as well as MDPs and POMDPs where the agent conditions on observed history.

Regret minimization, one of the main mathematical abstractions in the field of online learning, has proved to be an extremely versatile tool for decision-making over SDPs. In fact, over the past decade regret minimization algorithms for SDPs, such as counterfactual regret minimization (CFR) Zinkevich et al. (2007) and its later variants Tammelin et al. (2015); Brown and Sandholm (2019a), has become the state of the art technique for computing strong strategies in SDPs. In the particular case of extensive-form games, CFR was also a critical component that enabled several recent milestones in computing superhuman strategies in the game of heads-up Limit and No-Limit poker (Bowling et al., 2015; Brown and Sandholm, 2017; Moravčík et al., 2017; Brown and Sandholm, 2019b). However, these methods rely on having access to *counterfactuals*, that is information on what would have happened had the agent chosen a different action at any decision point. This makes their applicability limited in online decision-making settings, where the algorithm is deployed to learn strategies (for instance, exploitative strategies) against an opponent.

In this paper we introduce a new and efficient regret minimization algorithm for sequential decision making and extensive-form games that does not use any counterfactual information and yet enjoys the same (asymptotic) expected regret bound of $O(\sqrt{T})$ as CFR. Our regret minimizer runs in linear time per iteration unlike the only other prior approach by Abernethy, Hazan, and Rakhlin (2008), which requires that an eigendecomposition of a Hessian matrix be computed at each iteration. More precisely, we give an efficient algorithm for the bandit linear optimization problem on SDPs—and therefore for extensive-form games as well—and show that it achieves $O(\sqrt{T})$ cumulative regret in expectation against any fixed strategy.

Overview of Our Approach

In this subsection we give an overview of the key ideas behind our method. We assume some basic familiarity with the concept of full-information and bandit regret minimizers; both concepts are recalled in Section 2.

At a high level, we construct a bandit (that is, counterfactual-free) regret minimizer $\mathcal{R}$ starting from a *full-information* regret minimizer $\tilde{\mathcal{R}}$. Our bandit regret minimizer $\mathcal{R}$ works as follows:

- The next strategy $\boldsymbol{y}^t$ for $\mathcal{R}$ is computed starting from the strategy $\bar{\boldsymbol{x}}^t$ output by $\tilde{\mathcal{R}}$. We employ a specific unbiased *sampling scheme* to sample $\boldsymbol{y}^t$ from $\bar{\boldsymbol{x}}^t$.
- Each loss evaluation $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \in \mathbb{R}$ is used by $\mathcal{R}$ to constructs an artificial loss vector $\tilde{\boldsymbol{\ell}}^t$. This artificial loss vector is then passed to $\tilde{\mathcal{R}}$. The artificial loss is an unbiased estimator of $\boldsymbol{\ell}^t$. This construction is possible even if only $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$ but not $\boldsymbol{\ell}^t$ is observed by $\mathcal{R}$.

We implement $\tilde{\mathcal{R}}$ using the *online mirror descent* algorithm paired with the *dilated entropy distance-generating function* (DGF). The reason behind this particular choice of distance-generating function is twofold. First, it enables an efficient implementation of $\tilde{\mathcal{R}}$, since projections onto sequential strategy spaces based on the dilated entropy DGF amount

to a (linear-time) traversal of the sequential decision tree. Second, it serves as the basis for defining a notion of *local, time-dependent* norms $\|\cdot\|_t$ that play well with the regret bound of online mirror descent. In particular, we prove that the regret cumulated by $\tilde{\mathcal{R}}$ against any given strategy $\boldsymbol{z}$ up to time T asymptotically grows as $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\|\tilde{\boldsymbol{\ell}}^t\|_{*,t}^2$, where $\|\cdot\|_{*,t}$ is the (time-dependent) *dual* norm of $\|\cdot\|_t$.

Two steps are critical in the proof of the regret bound for overall regret minimizer $\mathcal{R}$. First, we show that, in expectation, $\|\tilde{\boldsymbol{\ell}}\|_{*,t}$ is upper bounded by a small (time-independent) constant c (the same property would not hold for a generic time-independent norm). This, combined with the local-norm regret bound mentioned above, guarantees that the regret cumulated by $\tilde{\mathcal{R}}$ asymptotically grows as $O(\sqrt{T})$ in expectation. Second, we use the unbiasedness of $\boldsymbol{y}^t$ and $\tilde{\boldsymbol{\ell}}^t$ to conclude that the expected regret accumulated by $\mathcal{R}$ matches that accumulated by $\tilde{\mathcal{R}}$. Combining the two steps, we obtain a $O(\sqrt{T})$ bound on the expected regret of $\mathcal{R}$.

Related Work

The idea of constructing a bandit regret minimizer starting from a full-information regret minimizer already appeared in Abernethy and Rakhlin (2009). In that paper, the authors give a general framework for constructing bandit regret minimizers with high-probability regret bounds and show how that framework can be instantiated in the case of simplex domains and Euclidean balls. The construction of an unbiased estimator $\tilde{\boldsymbol{\ell}}^t$ of $\boldsymbol{\ell}^t$ starting from the loss evaluation $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$ appears in the seminal paper of Auer et al. (2003) in the case of simplex domains. A more general construction appeared in Bartlett et al. (2008). We generalize the argument of Bartlett et al. (2008) to handle strategy domains where the vector space spanned by all decision vectors is rank-deficient (this is the case for sequential strategy spaces). The idea of using time-dependent norms to obtain a tighter regret analysis than time-independent norms already appeared several times, for example in Abernethy, Hazan, and Rakhlin (2008); Abernethy and Rakhlin (2009); Shalev-Shwartz (2012). The use of the dilated entropy regularizer in the context of sequential decision making and extensive-form games goes back to the original work of Hoda et al. (2010), with important practical observations in the work of Kroer et al. (2018).

Other approaches to bandit regret minimization are known in the literature. EXP3 (Auer et al., 2003) is credited to be the first bandit regret minimizer for simplex domains. GEOMETRICHEDGE (Dani, Kakade, and Hayes, 2008) is a general-purpose bandit regret minimizer that can be applied to any set of decisions, not just simplex domains. However, it requires one to compute a barycentric spanner (Awerbuch and Kleinberg, 2004) for our domain, which is a significant preprocessing cost. Furthermore, it runs in exponential time per iteration in the general case. Abernethy, Hazan, and Rakhlin (2008) gave the first bandit regret minimizer that runs in theoretical polynomial (in the dimension of the decision space) time per iteration and can handle any set of feasible decisions (as opposed to only a simplex domain). However, it requires to compute and sample from the eigenvectors of a Dikin ellipsoid centered at every iterate $\bar{\boldsymbol{x}}^t$ produced by $\tilde{\mathcal{R}}$, an operation that does not seem practical on the strategy polytopes we consider. Rather, the sampling scheme we use to implement $\mathcal{R}$.NEXTSTRATEGY() is extremely simple and can be implemented efficiently via a simple linear-time traversal of the decision tree.

Finally, we point out two key differences that set our method apart from Monte Carlo CFR (MCCFR), a popular stochastic method for computing equilibria in extensive-form games Lanctot et al. (2009). First, our method only requires that the loss evaluation $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$ be given as input, and no assumptions are made as to how $\boldsymbol{\ell}^t$ is chosen or computed by the environment. In contrast, MCCFR always assumes that $\boldsymbol{\ell}^t$ be in the form $\boldsymbol{A}\boldsymbol{z}$, where $\boldsymbol{A}$ is the *payoff matrix* of the game and $\boldsymbol{z}$ is a strategy vector of the opponent. In other words, MCCFR can be used as a bandit regret minimization method only if additional structure is enforced on the loss vectors. This limitation is significant, as it prevents one from using MCCFR to compute, for example, quantal-response equilibria and some types of exploitative strategies that require that more complex losses be used in the process (Farina, Kroer, and Sandholm, 2018). In this sense, MCCFR is strictly speaking not a general-purpose bandit regret minimizer. Second, in order to use MCCFR in an online setting, the algorithm must use the *outcome sampling* gradient-estimation variant, where at each iteration the sampling profile is chosen to be the strategy returned by CFR. However, this makes it impossible to provide a uniform lower bound on the probability of reaching every leaf in the game, which makes the theoretical guarantee on the regret cumulated by MCCFR inapplicable. Obtaining guarantees on the theoretical performance of MCCFR when the algorithm is used in an online setting is an interesting open question in the literature.

2 Preliminaries

In this section we briefly recall some important concepts about sequential decision processes and regret minimization.

Sequential Decision Processes

A sequential decision process (SDP) describes a sequential (that is, multi-stage) interaction between an agent and a—possibly adversarial—environment. SDPs provide a general formalism which captures the interaction model of extensive-form games with perfect recall, as well as POMDPs and MDPs for which the agent conditions its policy on the entire history of observations and actions (Farina, Kroer, and Sandholm, 2019). An SDP is structured as a tree of decision points—in which an action must be selected by the agent—and observation points—in which the environment reveals a signal to the agent. As an example, consider the SDP in Figure 1, corresponding to a game of Kuhn poker—a simplified version of poker played with a three-card deck, as introduced by Kuhn (1950). The process starts at decision

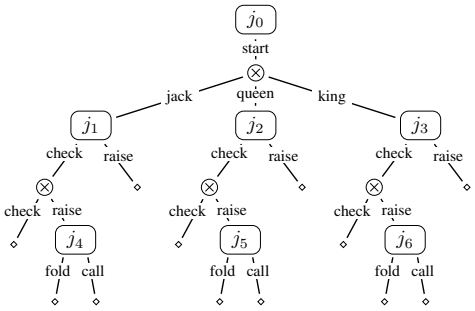

Figure 1: Sample sequential decision process. The decision process corresponds to the game of Kuhn Poker.

point j_0, where the agent can only take action 'start'. After

taking that action, the process moves to an observation point (denoted as $\otimes$), where the agent observes their private card. Assuming that the agent observes the signal 'queen', the process moves to decision point j_2, where the agent can either 'check' or 'raise'. If the agent 'check's, the process moves to another observation point, where the agent gets informed about the environment's action—either a 'check' or a 'raise'. If the observed signal is 'check', the decision process ends.

Notation for SDPs. We denote the set of decision points in the process as $\mathcal{J}$, and the set of observation points as $\mathcal{K}$. At each decision point $j \in \mathcal{J}$, the agent selects an action from the set A_j of available actions. At each observation point $k \in \mathcal{K}$, the agent observes a signal s_k from the environment out a set of possible signals S_k. We denote with ρ the transition function of the process. Picking action $a \in A_j$ at decision point $j \in \mathcal{J}$ results in the process transitioning to $\rho(j, a) \in \mathcal{J} \cup \mathcal{K} \cup \{\diamond\}$, where $\diamond$ denotes the end of the process. Similarly, the process transitions to $\rho(k, s) \in \mathcal{J} \cup \mathcal{K} \cup \{\diamond\}$ after the agent observes signal $s \in S_k$ at observation point $k \in \mathcal{K}$. In line with the game theory literature, we call a pair (j, a) where $j \in \mathcal{J}$ and $a \in A_j$ a *sequence*. The set of all sequences is denoted as $\Sigma := \{(j, a) : j \in \mathcal{J}, a \in A_j\}$. For notational convenience, we will often denote an element (j, a) in Σ as ja without using parentheses. Given a sequence $ja \in \Sigma$, we denote with $\boldsymbol{u}_{ja}$ the vector such that $(\boldsymbol{u}_{ja})_{j'a'} = 1$ if the (unique) path from the root node of the SDP to action a' at decision point j' passes through action a at decision point j, and $(\boldsymbol{u}_{ja})_{j'a'} = 0$ otherwise. Finally, given a decision point $j \in \mathcal{J}$, we denote with p_j its *parent sequence*, defined as the last sequence (that is, decision point-action pair) encountered on the path from the root of the SDP to j. If the agent does not act before j (i.e., j is the root of the SDP or only observation points are encountered on the path from the root to j), we let $p_j = \varnothing$.

Strategies in SDPs. Conceptually, a strategy for the agent in a sequential decision process is a choice of distribution over the set of actions A_j at each decision point $j \in \mathcal{J}$ in the process. We represent a strategy using the *sequence-form representation*, that is as a vector $\bar{\boldsymbol{x}} \in \mathbb{R}^{|\Sigma|}_{\geq 0}$ whose entries are indexed by Σ. The entry $\bar{x}_{ja}$ contains the product of the probabilities of all actions at all decision points on the path from the root of the SDP down to action a at decision point $j \in \mathcal{J}$. Clearly, in order to be a valid sequence-form strategy, the entries in $\bar{\boldsymbol{x}}$ must satisfy the following consistency constraints (Romanovskii, 1962; Koller, Megiddo, and von Stengel, 1994; von Stengel, 1996):

$$\begin{aligned} \textstyle\sum_{a \in A_j} \bar{x}_{ja} &= \bar{x}_{p_j} && \forall j \in \mathcal{J} \text{ such that } p_j \neq \varnothing, \\ \textstyle\sum_{a \in A_j} \bar{x}_{ja} &= 1 && \forall j \in \mathcal{J} \text{ such that } p_j = \varnothing. \end{aligned} \tag{1}$$

Since $\varnothing$ is not an element in Σ, there is no entry in $\bar{\boldsymbol{x}}$ that corresponds to $\varnothing$, and the notation $\bar{x}_{\varnothing}$ is invalid. However, we will abuse notation and refer to $\bar{x}_{\varnothing}$ to mean the constant value 1. This allows us to write the consistency constraints (1) as $\sum_{a \in A_j} \bar{x}_{ja} = \bar{x}_{p_j}$ even when $p_j = \varnothing$. With this convention, it is also valid to say that the probability of the agent picking action $a \in A_j$ conditioned to the agent being at decision point j is x_{ja}/x_{p_j}, provided $x_{p_j} \neq 0$.

Finally, we let $\mathcal{T} \subseteq \mathbb{R}^{|\Sigma|}_{\geq 0}$ be the finite set of all sequence-form strategies that correspond to *pure* (also known as *deterministic*) strategies, that is strategies that assign probability 1 to exactly one action at each decision point. It is well-known that the set of all sequence-form strategies is the convex hull $\operatorname{co} \mathcal{T}$ of the set of pure strategies $\mathcal{T}$.

Regret Minimization

A regret minimizer is an abstraction for a repeated decision maker. The decision maker repeatedly interacts with an unknown (possibly adversarial) environment by choosing points $\boldsymbol{x}^1, \dots, \boldsymbol{x}^T$ from a set $\mathcal{X} \subseteq \mathbb{R}^n$ of feasible decisions and incurring a linear loss $(\boldsymbol{\ell}^1)^\top \boldsymbol{x}^1, \dots, (\boldsymbol{\ell}^T)^\top \boldsymbol{x}^T$ after each iteration. The quality metric for a regret minimizer is its *regret*, which measures the difference in loss against the best *fixed* (that is, time-independent) decision in hindsight. Formally, given a decision $\boldsymbol{z} \in \mathcal{X}$, the regret cumulated against $\boldsymbol{z}$ up to time T is defined as $R^T(\boldsymbol{z}) := \sum_{t=1}^{T} (\boldsymbol{\ell}^t)^\top (\boldsymbol{x}^t - \boldsymbol{z})$. A "good" regret minimizer (also called a *Hannan consistent* regret minimizer) is such that the regret against *any* decision $\boldsymbol{z}$ grows sublinearly as a function of T.

In this paper, we will be interested in two types of regret minimizers, which differ in the feedback that is received by the regret minimizer.

Full-Information Setting. In the full-information setting, at all time steps $t = 1, \dots, T$ the regret minimizer interacts with the environment as follows:

- NEXTSTRATEGY(): the agent outputs the next decision $\boldsymbol{x}^t \in \mathcal{X} \subseteq \mathbb{R}^n$.
- OBSERVELOSS($\boldsymbol{\ell}^t$): the environment selects a loss vector $\boldsymbol{\ell}^t \in \mathbb{R}^n$ and the agent observes $\boldsymbol{\ell}^t$. The loss vector can depend on the decisions $\boldsymbol{x}^1, \dots, \boldsymbol{x}^t$ that were output by the regret minimizer in the past.

Our construction of $\tilde{\mathcal{R}}$ (Section 4) provides a full-information regret minimizer for the set $\mathcal{X} = \operatorname{co} \mathcal{T}$.

Bandit Setting. In the bandit setting the environment does *not* reveal the selected loss vector $\boldsymbol{\ell}^t$ at each iteration, but only the evaluation $(\boldsymbol{\ell}^t)^\top \boldsymbol{x}^t$ of the loss function for the last decision. Formally, at all time steps $t = 1, \dots, T$ the regret minimizer interacts with the environment as follows:

- NEXTSTRATEGY(): the agent outputs the next decision $\boldsymbol{x}^t \in \mathcal{X} \subseteq \mathbb{R}^n$.
- OBSERVELOSSEVALUATION($(\boldsymbol{\ell}^t)^\top \boldsymbol{x}^t$): the environment selects a loss vector $\boldsymbol{\ell}^t \in \mathbb{R}^n$ and the agent observes $(\boldsymbol{\ell}^t)^\top \boldsymbol{x}^t$. The loss vector can depend on the decisions $\boldsymbol{x}^1, \dots, \boldsymbol{x}^{t-1}$ that were output by the regret minimizer before time t.

Since the regret minimizer only observes $(\boldsymbol{\ell}^t)^\top \boldsymbol{x}^t$, it cannot compute any *counterfactual* information (that is, compute the value of the loss at a decision other than the one that was output). The main contribution of this paper is an efficient bandit regret minimizer $\mathcal{R}$ for the set $\mathcal{X} = \mathcal{T}$ whose expected regret is $R^T(\boldsymbol{z}) = O(\sqrt{T})$ for all $\boldsymbol{z} \in \operatorname{co} \mathcal{T}$.

3 Dilated Entropic Regularization and Local Norms

The dilated entropy distance-generating function is a regularizer that induces a notion of distance that is suitable for the space of sequence-form strategies in a sequential decision process. The dilated entropy DGF was first introduced in the context of extensive-form games by Hoda et al. (2010). Kroer et al. (2018)—with earlier and weaker results by Kroer et al. (2015)—analyzed several properties of this function, including its strong convexity modulus with respect to the ℓ_1 and ℓ_2 norms. They also showed that the dilated entropy DGF

leads to state-of-the-art convergence guarantees in iterative methods for computing Nash equilibrium in two-player zero-sum extensive-form games of perfect recall. In Definition 1 we state the definition of the dilated entropy DGF:

Definition 1 (Dilated entropy DGF)**.** *Let $\operatorname{co}\mathcal{T}$ be the set of sequence-form strategies for the SDP. The* dilated entropy *distance-generating function for $\operatorname{co}\mathcal{T}$ is the function $\varphi : \mathbb{R}^{|\Sigma|}_{>0} \to \mathbb{R}_{\geq 0}$ defined as*

$$\varphi : \bar{\boldsymbol{x}} \mapsto \sum_{j \in \mathcal{J}} w_j \left(\bar{x}_{p_j} \log |A_j| + \sum_{a \in A_j} \bar{x}_{ja} \log \frac{\bar{x}_{ja}}{\bar{x}_{p_j}} \right),$$

where the weights w_j are defined recursively according to:

$$w_j = 2 + 2 \max_{a \in A_j} \{w_{\rho(j,a)}\}; \quad w_k = \sum_{s \in S_k} w_{\rho(j,s)}. \tag{2}$$

(In particular, $w_k = 0$ for any observation point k such that $\mathcal{C}_k = \emptyset$.)

Our definition differs from that of Kroer et al. (2018) in that we add the "shifting" terms $\bar{x}_{p_j} \log |A_j|$, where the minimum of φ is 0. This unique minimum is attained by the sequence-form strategy that at each decision point uniformly randomizes among all available actions (that is, $x_{ja} = x_{p_j}/|A_j|$ for all $j \in \mathcal{J}, a \in A_j$). The idea of adding shifting terms to make the regularizer non-negative over its domain already appears in Hoda et al. (2010) and Kroer, Farina, and Sandholm (2018). Since those additional terms amount to adding a linear function to the definition of the dilated entropy DGF found in Kroer et al. (2018), the analysis of the strong-convexity modulus of Kroer et al. (2018) holds verbatim, and in particular:

Lemma 1. *(Kroer et al., 2018, Theorems 2 and 3) The dilated-entropy DGF of Definition 1 is 1-strongly-convex on $\operatorname{co}\mathcal{T}$ with respect to both the ℓ_1 and the ℓ_2 norm.*

The dilated entropy DGF has the advantage that its gradient and its Fenchel conjugate function can be evaluated efficiently via a linear-time pass on the decision space (Hoda et al., 2010). Specifically:

Observation 1. *For all $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{>0}$, there exists an exact algorithm, denoted* GRADIENT*, to compute $\nabla\varphi(\boldsymbol{z})$ in linear time in $|\Sigma|$.*

Observation 2. *For all $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{>0}$, there exists an exact algorithm, denoted* ARGCONJUGATE*, to compute*

$$\nabla\varphi^*(\boldsymbol{z}) = \arg\max_{\hat{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}} \{\boldsymbol{z}^\top \hat{\boldsymbol{x}} - \varphi(\hat{\boldsymbol{x}})\}.$$

in linear time in $|\Sigma|$.

The two properties above make φ an appealing candidate regularizer in many optimization algorithms that operate on sequential decision making domains, including the full-information regret minimizer $\tilde{\mathcal{R}}$ that we use in this paper. Pseudocode for the algorithms mentioned in Observation 1 and Observation 2 can be found in Appendix B.

Local Norms Induced by the Dilated Entropy DGF

At each point $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ in the sequence-form strategy space, the dilated entropy DGF induces a pair of primal-dual *local* norms $(\|\cdot\|_{\bar{\boldsymbol{x}}}, \|\cdot\|_{*,\bar{\boldsymbol{x}}})$ defined for all $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$ as

$$\|\boldsymbol{z}\|_{\bar{\boldsymbol{x}}} := \sqrt{\boldsymbol{z}^\top \nabla^2\varphi(\bar{\boldsymbol{x}})\, \boldsymbol{z}}; \;\; \|\boldsymbol{z}\|_{*,\bar{\boldsymbol{x}}} := \sqrt{\boldsymbol{z}^\top (\nabla^2\varphi(\bar{\boldsymbol{x}}))^{-1} \boldsymbol{z}},$$

where $\nabla^2\varphi(\bar{\boldsymbol{x}})$ denotes the Hessian matrix of φ at $\bar{\boldsymbol{x}}$. Since $\nabla^2\varphi(\bar{\boldsymbol{x}})$ is positive-definite, it is well-known that $\|\cdot\|_{*,\bar{\boldsymbol{x}}}$ is well-defined and that it is indeed dual to $\|\cdot\|_{\bar{\boldsymbol{x}}}$, in the sense that $\|\boldsymbol{z}\|_{*,\bar{\boldsymbol{x}}} = \max\{\boldsymbol{z}^\top \boldsymbol{w} : \|\boldsymbol{w}\|_{\bar{\boldsymbol{x}}} \leq 1\}$ for all $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$.

To our knowledge, we are the first to explore the local norms induced by the dilated entropy DGF. We start by giving a convenient bound for the norm of a generic vector $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{\geq 0}$ as measured with respect to the local norm at $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$:

Lemma 2. *Let $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ and $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{\geq 0}$. Then,*

$$\|\boldsymbol{z}\|^2_{\bar{\boldsymbol{x}}} \leq \frac{3}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \frac{w_j}{\bar{x}_{ja}} z^2_{ja}.$$

The analysis of the dual norm is more complicated, as the inverse of the Hessian matrix is significantly more involved. We start by giving a characterization of the inverse Hessian matrix of the DGF d at a generic strategy $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ in terms of sum of dyadics:

Lemma 3. *Let $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ be a sequence-form strategy. The inverse Hessian $(\nabla^2\varphi)^{-1}(\bar{\boldsymbol{x}})$ at $\bar{\boldsymbol{x}}$ can be expressed as:*

$$(\nabla^2\varphi)^{-1}(\bar{\boldsymbol{x}}) = \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \frac{(\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})(\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})^\top}{w_j \bar{x}_{ja}}, \tag{3}$$

where $\circ$ denotes componentwise product of vectors.

Lemma 3 immediately implies the following corollary, which gives an alternative way of computing the dual norm of any vector $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$:

Corollary 1. *Let $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ be a sequence-form strategy, and let $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$. The local dual norm of $\boldsymbol{z}$ satisfies*

$$\|\boldsymbol{z}\|^2_{*,\bar{\boldsymbol{x}}} = \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \frac{(\boldsymbol{u}^\top_{ja}(\boldsymbol{z} \circ \bar{\boldsymbol{x}}))^2}{w_j \bar{x}_{ja}}. \tag{4}$$

4 Construction of $\tilde{\mathcal{R}}$

In this section, we describe the full-information regret minimizer $\tilde{\mathcal{R}}$ that we use in our construction (Section 1). The pseudocode for $\tilde{\mathcal{R}}$ is given in Algorithm 1, while its analysis is given in Theorem 1. Our analysis fundamentally relies on the notion of local norms induced by the dilated entropy DGF (Section 3).

Online Mirror Descent with Dilated Entropy DGF

Online mirror descent is one of the most well-studied full-information regret minimization algorithms in online learning. In its general form, given a strongly-convex regularizer d and a convex and compact domain $\mathcal{X} \subseteq \mathbb{R}^n$, each decision

is computed according to

$$\bar{\boldsymbol{x}}^1 = \arg\min_{\hat{\boldsymbol{x}} \in \mathcal{X}} d(\hat{\boldsymbol{x}}); \tag{5}$$

$$\bar{\boldsymbol{x}}^{t+1} = \arg\min_{\hat{\boldsymbol{x}} \in \mathcal{X}} \Big\{ (\eta \tilde{\boldsymbol{\ell}}^t - \nabla d(\bar{\boldsymbol{x}}^t))^\top \hat{\boldsymbol{x}} + d(\hat{\boldsymbol{x}}) \Big\}. \tag{6}$$

Our full-information regret minimizer $\tilde{\mathcal{R}}$ (Algorithm 1) is constructed using online mirror descent instantiated with the dilated entropy DGF φ (Definition 1) as the regularizer d and the set $\operatorname{co} \mathcal{T} \subseteq \mathbb{R}^{|\Sigma|}$ of sequence-form strategies in the game as the domain of feasible iterates $\mathcal{X}$. In that setting, the initial point (Equation 5) is attained by the strategy that at each decision point uniformly randomized among all available actions and therefore $\bar{\boldsymbol{x}}_1$ can be computed efficiently. Furthermore, each proximal step (6) can be implemented efficiently via a call to GRADIENT (Observation 1) followed by one to ARGCONJUGATE (Observation 2).

Algorithm 1: Full-information regret minimizer $\tilde{\mathcal{R}}$

Data: η is a step-size parameter.

function SETUP()
  for $j \in \mathcal{J}$ *in top-down order* **do**
    for $a \in A_j$ **do** $\bar{x}^1_{ja} \leftarrow \frac{\bar{x}_{p_j}}{|A_j|}$

function NEXTSTRATEGY()
  return $\bar{\boldsymbol{x}}^t$

function OBSERVELOSS($\tilde{\boldsymbol{\ell}}^t$)
  $\boldsymbol{g} \leftarrow \eta \tilde{\boldsymbol{\ell}}^t - \text{GRADIENT}(\bar{\boldsymbol{x}}^t)$ [▷ Observ. 1]
  $\bar{\boldsymbol{x}}^{t+1} \leftarrow \text{ARGCONJUGATE}(-\boldsymbol{g})$ [▷ Observ. 2]

Observation 3. *At all times t the decision produced by Algorithm 1 satisfies $\bar{\boldsymbol{x}}^t \in \mathbb{R}^{|\Sigma|}_{>0}$.*

Analysis

The analysis of the regret cumulated by Algorithm 1 as a function fo the local dual norms of the loss vectors $\tilde{\boldsymbol{\ell}}^t$ is rather lengthy and is deferred to Appendix C. Here, we only state the central result of this section, which builds on Lemma 2 and Corollary 1:

Theorem 1. *Let D be the maximum depth of any node in the SDP, and let $\boldsymbol{z} \in \operatorname{co} \mathcal{T}$. If $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}^{|\Sigma|}_{\geq 0}$ for all times t, then the regret $\tilde{R}^T(\boldsymbol{z})$ cumulated by $\tilde{\mathcal{R}}$ satisfies, at all times T:*

$$\tilde{R}^T(\boldsymbol{z}) \leq \frac{\varphi(\boldsymbol{z})}{\eta} + \eta \sqrt{3D} \cdot \sum_{t=1}^{T} \|\tilde{\boldsymbol{\ell}}^t\|^2_{*, \bar{\boldsymbol{x}}^t}. \tag{7}$$

In the rest of the paper, we will give guarantees about the expected magnitude of the right-hand side.

5 Construction of $\mathcal{R}$

In this section, we describe the bandit regret minimizer $\mathcal{R}$. As mentioned in Section 1, two different components are important in the algorithm: the sampling scheme, which we describe in Section 5, and the construction of the unbiased loss estimates, which we give in Section 5.

Here, we give an overview of the interaction between the sampling scheme and the construction of the loss estimates. Our construction of the unbiased loss estimates extends and generalizes that of Dani, Kakade, and Hayes (2008), in that it can be applied even when the set of strategies is rank-deficient—as is the case for our set of pure strategies $\mathcal{T}$. In particular, we relax the notion of unbiasedness to mean the weaker condition that the projection $\tilde{\boldsymbol{\ell}}^t$ onto the direction[1] $\operatorname{dir} \mathcal{T}$ of $\mathcal{T}$ be an unbiased estimator of the projection of the original (and unknown) $\boldsymbol{\ell}^t$ onto $\operatorname{dir} \mathcal{T}$:

$$\mathbb{E}_t[\tilde{\boldsymbol{\ell}}^t]^\top \boldsymbol{w} = (\boldsymbol{\ell}^t)^\top \boldsymbol{w} \qquad \forall\, \boldsymbol{w} \in \operatorname{dir} \mathcal{T}, \tag{$\star$}$$

where $\mathbb{E}_t[\cdot]$ is an abbreviation for $\mathbb{E}_t[\cdot | \boldsymbol{y}^1, \dots, \boldsymbol{y}^{t-1}]$, that is the expectation conditional on the previous decisions that were output by $\mathcal{R}$.

The main technical tool in our construction is to use a generalized inverse of the autocorrelation matrix of $\boldsymbol{y}^t$:

Proposition 1. *Let π^t be the conditional distribution over $\mathcal{T}$, given the previous decisions $\boldsymbol{y}^1, \dots, \boldsymbol{y}^{t-1}$, and suppose that the support of π^t is full-rank (that is, $\operatorname{span} \operatorname{supp} \pi^t = \operatorname{span} \mathcal{T}$). Let $\boldsymbol{C}^t := \mathbb{E}_t[\boldsymbol{y}^t (\boldsymbol{y}^t)^\top]$ be the autocorrelation matrix of $\boldsymbol{y}^t$, and let $\boldsymbol{C}^{t-}$ be any generalized inverse of $\boldsymbol{C}^t$, that is any matrix such that $\boldsymbol{C}^t \boldsymbol{C}^{t-} \boldsymbol{C}^t = \boldsymbol{C}^t$. Then, for any $\boldsymbol{b}^t \perp \operatorname{dir} \mathcal{T}$, the random variable*

$$\tilde{\boldsymbol{\ell}}^t := [(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t] \cdot (\boldsymbol{C}^{t-} \boldsymbol{y}^t + \boldsymbol{b}^t), \tag{8}$$

satisfies ($\star$).

Crucially, the loss estimate $\tilde{\boldsymbol{\ell}}^t$ in (8) can be constructed using only the bandit information (that is, loss evaluation) $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$ that was received at time t after the regret minimizer output $\boldsymbol{y}^t$ as its decision. A proof of Proposition 1 can be found in Appendix D.

Sampling Scheme for Sequential Decision Spaces

At every time step t, the bandit regret minimizer $\mathcal{R}$ internally calls into $\tilde{\mathcal{R}}$.NEXTSTRATEGY() and receives a sequence-form strategy $\bar{\boldsymbol{x}}^t \in \operatorname{co} \mathcal{T}$. After that, $\mathcal{R}$ samples and returns a *pure* sequence-form strategy $\boldsymbol{y}^t \in \mathcal{T}$ such that $\mathbb{E}_t[\boldsymbol{y}^t] = \bar{\boldsymbol{x}}^t$. Our sampling scheme (Algorithm 2) is natural: at each decision point j we randomly pick an action $a \in A_j$ according to the distribution $\bar{x}^t_{ja} / \bar{x}^t_{p_j}$ induced by the sequence-form strategy $\bar{\boldsymbol{x}}^t$.

It is straightforward to verify that (see Appendix D):

Lemma 4. *The sampling scheme given by Algorithm 2 is unbiased, that is, $\mathbb{E}_t[\boldsymbol{y}^t] = \bar{\boldsymbol{x}}^t$.*

The study of the autocorrelation matrix $\boldsymbol{C}^t$ of the sampling scheme—a key ingredient in Proposition 1—is more complicated, and we defer the full analysis to Appendix D.

Computation of the Loss Estimate ($\tilde{\boldsymbol{\ell}}^t$)

At each time t, we use Proposition 1 to construct the loss estimate $\tilde{\boldsymbol{\ell}}^t$. The main conceptual leap is to identify (i) a choice of generalized inverse $\boldsymbol{C}^{t-}_*$ for the autocorrelation matrix $\boldsymbol{C}^t$ of $\boldsymbol{y}^t$ returned by Algorithm 2 and (ii) a particular choice of

[1] The direction $\operatorname{dir} \mathcal{X}$ of a set $\mathcal{X}$ is the subspace defined as $\operatorname{dir} \mathcal{X} := \operatorname{span}\{\boldsymbol{u} - \boldsymbol{v} : \boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}\}$.

Algorithm 2: SAMPLE($\bar{\boldsymbol{x}}^t$)

Input: $\bar{\boldsymbol{x}}^t \in \operatorname{co}\mathcal{T}$ sequence-form strategy
Output: $\boldsymbol{y}^t \in \mathcal{T}$ such that $\mathbb{E}[\boldsymbol{y}^t] = \bar{\boldsymbol{x}}^t$

$\boldsymbol{y}^t \leftarrow \mathbf{0}$
subroutine RECURSIVESAMPLE(v)
  if $v \in \mathcal{J}$ **then**
    Sample an action $a \sim (\bar{x}^t_{va}/\bar{x}_{p_v})_{a \in A_v}$
    $y^t_{va} \leftarrow 1$
    RECURSIVESAMPLE($\rho(v, a)$)
  else if $v \in \mathcal{K}$ **then**
    for $s \in S_k$ **do** RECURSIVESAMPLE($\rho(v, s)$)
RECURSIVESAMPLE(r) [▷ r: root of the SDP]
return $\boldsymbol{y}^t$

vector $\boldsymbol{b}^t_* \perp \operatorname{dir}\mathcal{T}$ so that (a) the product $\boldsymbol{C}^{t-}_*(\boldsymbol{y}^t + \boldsymbol{b}^t_*)$ can be carried out in $O(|\Sigma|)$ time and (b) the resulting loss function $\tilde{\boldsymbol{\ell}}^t$ is nonnegative, as required by $\tilde{\mathcal{R}}$ (see Theorem 1). At a high level, the particular construction that we use generates $\boldsymbol{C}^{t-}_*$ and $\boldsymbol{b}^t_*$ inductively in a bottom-up fashion by traversing the SDP, and heavily relies on the combinatorial structure of the autocorrelation matrix $\boldsymbol{C}^t$ induced by Algorithm 2. All details and proofs can be found in Appendix D; here, we only show the algorithm that carries out the multiplication in (8) for the particular choice of generalized inverse and orthogonal vector:

Algorithm 3: LOSSESTIMATE($l := (\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t, \bar{\boldsymbol{x}}^t, \boldsymbol{y}^t$)

Input: Loss evaluation (bandit input) $l = (\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$
$\bar{\boldsymbol{x}}^t \in \operatorname{co}\mathcal{T}$ strategy output by $\tilde{\mathcal{R}}$
$\boldsymbol{y}^t \in \mathcal{T}$ *pure* strategy output by $\mathcal{R}$
Output: $\tilde{\boldsymbol{\ell}}^t = l \cdot \boldsymbol{C}^{t-}_*(\boldsymbol{y}^t + \boldsymbol{b}^t_*)$ such that ($\star$) holds

1 $\tilde{\boldsymbol{\ell}}^t \leftarrow \mathbf{0}$
2 **for** $j \in \mathcal{J}$ **do**
3 **for** $a \in A_j$ **do**
4 **if** $y^t_{ja} = 1$ **and** $\rho(j, a) = \diamond$ **then** $\tilde{\ell}^t_{ja} \leftarrow l/\bar{x}^t_{ja}$
5 **return** $\boldsymbol{\ell}^t$

Algorithm 3 is trivial to implement and looks deceptively simple. Furthermore, the loss estimate returned by Algorithm 3 coincides with the loss estimate constructed by EXP3 (Auer et al., 2003) if the SDP only has one decision point (that is, the strategy space is a simplex). Finally, we remark that when $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \geq 0$ (which can be assumed without loss of generality), the loss estimate constructed by Algorithm 3 has non-negative entries: $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}^{|\Sigma|}_{\geq 0}$ so that Theorem 1 is applicable.

Norm of the Loss Estimate

In theory, each entry of $\tilde{\boldsymbol{\ell}}^t$ (Line 4 in Algorithm 3) can be arbitrarily large, since $\bar{x}^t_{ja}$ can be arbitrarily small. As a consequence, the Euclidean norm $\|\tilde{\boldsymbol{\ell}}^t\|_2$ of the loss estimate can be arbitrarily large, even in expectation. This shows the importance of having Equation (7) be expressed in terms of the *local* norms $\|\cdot\|_{*,\bar{\boldsymbol{x}}^t}$ instead of a generic time-invariant norm. Indeed, it is possible to give guarantees on the expectation of the local dual norm of $\tilde{\boldsymbol{\ell}}^t$ returned by Algorithm 3:

Theorem 2. *Assume that the bandit information $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \in [0, 1]$ at all times t. Then, at all times t, the loss estimate $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}^{|\Sigma|}_{\geq 0}$ returned by Algorithm 3 satisfies*

$$\mathbb{E}_t\left[\|\tilde{\boldsymbol{\ell}}^t\|^2_{*,\bar{\boldsymbol{x}}^t}\right] \leq 2 \cdot |\Sigma|^2.$$

Theorem 2 is proved in Appendix D. Theorem 2 is one of the deepest results in this paper: it ties together the sampling scheme (Section 5), the construction of the loss estimates (Section 5), and the geometry of local norms (Section 3) induced by the dilated entropy DGF.

6 The Full Algorithm

As foretold in Section 1, we construct our bandit regret minimizer $\mathcal{R}$ starting from the full-information regret minimizers $\tilde{\mathcal{R}}$ of Algorithm 1, as in Algorithm 4. The resulting algorithm is surprisingly easy to implement, and requires only two linear traversals of the SDP per iteration.

Algorithm 4: Bandit regret minimizer $\mathcal{R}$

function NEXTSTRATEGY()
  $\bar{\boldsymbol{x}}^t \leftarrow \tilde{\mathcal{R}}$.NEXTSTRATEGY() [▷ Algorithm 1]
  $\boldsymbol{y}^t \leftarrow$ SAMPLE($\bar{\boldsymbol{x}}^t$) [▷ Algorithm 2]
  return $\boldsymbol{y}_t$
function OBSERVELOSSEVALUATION($l := (\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$)
  $\tilde{\boldsymbol{\ell}}^t \leftarrow$ LOSSESTIMATE($l, \bar{\boldsymbol{x}}^t, \boldsymbol{y}^t$) [▷ Algorithm 3]
  $\tilde{\mathcal{R}}$.OBSERVELOSS($\tilde{\boldsymbol{\ell}}^t$) [▷ Algorithm 1]

The regret $R^T(\boldsymbol{z})$ of $\mathcal{R}$ is linked to the regret $\tilde{R}^T(\boldsymbol{z})$ of $\tilde{\mathcal{R}}$: using the definition of regret and the law of total expectation,

$$\mathbb{E}[R^T(\boldsymbol{z})] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t\left[(\boldsymbol{\ell}^t)^\top(\boldsymbol{y}^t - \boldsymbol{z})\right]\right] = \mathbb{E}[\tilde{R}^T(\boldsymbol{z})],$$

where we used the hypothesis that $\boldsymbol{\ell}^t$ is independent from $\boldsymbol{y}^t$, as well as Lemma 4 and ($\star$). Theorem 1 gives an upper bound for the regret $\tilde{R}^T(\boldsymbol{z})$ of $\tilde{\mathcal{R}}$ as a function of the sequence of the loss estimates $\tilde{\boldsymbol{\ell}}^1, \ldots, \tilde{\boldsymbol{\ell}}^T$. In particular, taking expectations in Equation (7) and using the law of total expectation as well as Theorem 2, we have

$$\mathbb{E}[\tilde{R}^T(\boldsymbol{z})] \leq \frac{\varphi(\boldsymbol{z})}{\eta} + \eta\sqrt{3D} \cdot \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t\left[\|\tilde{\boldsymbol{\ell}}^t\|^2_{*,\bar{\boldsymbol{x}}^t}\right]\right].$$
$$\leq \frac{\varphi(\boldsymbol{z})}{\eta} + 2\eta\,|\Sigma|^2\sqrt{3D} \cdot T.$$

Hence, by picking $\eta = 1/(|\Sigma|\sqrt{T})$, we obtain the following theorem, which is the central result of this paper:

Theorem 3. *Let D be the maximum depth of any node in the SDP. Then, assuming $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \in [0, 1]$ at all times $t =$*

$1, \dots, T$, *the regret* $R^T(\boldsymbol{z})$ *cumulated by Algorithm 4 against any* $\boldsymbol{z} \in \operatorname{co} \mathcal{T}$ *satisfies*

$$\mathbb{E}[R^T(\boldsymbol{z})] \leq (\varphi(\boldsymbol{z}) + 2\sqrt{3D})\,|\Sigma| \cdot \sqrt{T}.$$

Theorem 3 shows that the regret cumulated by our bandit regret minimizer $\mathcal{R}$ grows as $O(\sqrt{T})$. This is better than the algorithm of Abernethy, Hazan, and Rakhlin (2008), which grows as $O(\sqrt{T \log T})$, and only provided $T = \Omega(|\Sigma| \log T)$.

Finally, we conclude with a word of caution. Our algorithm, just like the one of Abernethy, Hazan, and Rakhlin (2008), only guarantees that $\max_{\boldsymbol{z} \in \operatorname{co} \mathcal{T}} \mathbb{E}[R^T(\boldsymbol{z})]$ is small, and *not* that $\mathbb{E}[\max_{\boldsymbol{z} \in \operatorname{co} \mathcal{T}} R^T(\boldsymbol{z})]$ is small. Depending on the application, this might or might not be strong enough a property. This limitation is well-known (see, e.g., Abernethy and Rakhlin (2009)) and is one of the main conceptual drives behind the research of regret minimizers that give high-probability bounds on regret. Section 8 briefly discusses how the techniques of this paper are relevant towards that effort.

7 Experimental Evaluation

We implemented our bandit regret minimizer (Algorithm 4), as well as the bandit regret minimizer of Abernethy, Hazan, and Rakhlin (2008) and MCCFR (instantiated as a bandit regret minimizer), and tested them on the game of Kuhn poker Kuhn (1950). The SDP corresponding to the player that acts first in the game is given in Figure 1. All three algorithms face a strong opponent that at each iteration plays according to a precomputed strategy $\bar{\boldsymbol{s}}$ that is part of a Nash equilibrium of the game (in other words, the opponent is playing optimally). Correspondingly, the sequence of loss functions is in the form $\boldsymbol{\ell}^t = \boldsymbol{A}\boldsymbol{s}^t$, where $\boldsymbol{A}$ is the payoff matrix of the game, and $\boldsymbol{s}^t$ is a (pure) strategy of the opponent sampled so that $\mathbb{E}[\boldsymbol{s}^t] = \bar{\boldsymbol{s}}$. This loss structure is necessary for MCCFR to be applicable, as discussed in Section 1. Figure 2 shows the evolution of the regret of all three algorithms against the best response strategy for $\bar{\boldsymbol{s}}$. We ran each algorithm 100 times, and for each algorithm we draw the average regret and shade one standard deviation around that average. For our method, we use the theoretical step size multiplied by 1000.0, while for the method of Abernethy, Hazan, and Rakhlin (2008) we divide their step size parameter by 10; these changes do not affect the theoretical guarantees but improved the practical performances of both algorithms.

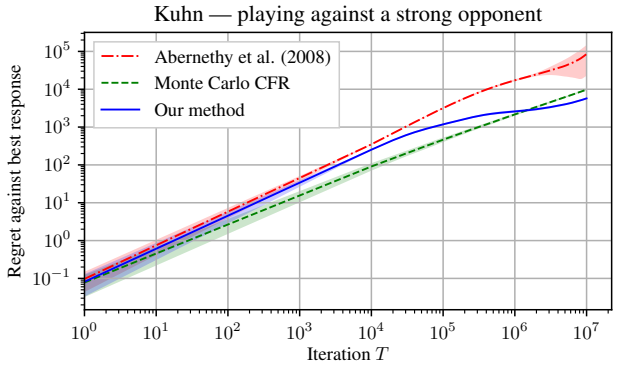

Figure 2: Evolution of the regret of all three algorithms against the best response of $\bar{\boldsymbol{s}}$.

In this setting, all algorithms but MCCFR guarantee $\tilde{O}(\sqrt{T})$ regret, and the plot experimentally confirms this bound. As discussed in Section 1, in order to use MCCFR as a bandit regret minimizer, we needed to instantiate it in its outcome sampling variant in a way that necessarily invalidates the hypotheses needed for its theoretical analysis to hold. However, we observe that in practice its performance aligns well with that of the other algorithms.

In Figure 3, we also looked at the empirical distribution of the ℓ_2 and the local dual norm $\|\cdot\|_{*,\bar{\boldsymbol{x}}^t}$ of each loss estimate $\tilde{\boldsymbol{\ell}}^t$ ever computed by our bandit regret minimizer $\mathcal{R}$ across the 100 runs of the experiment. As predicted by Theorem 2, one of the core results of this paper, the average of $\|\tilde{\boldsymbol{\ell}}^t\|^2_{*,\bar{\boldsymbol{x}}^t}$ is below the theoretical value $2|\Sigma|^2 = 288$. In fact, it is experimentally much lower, as it hovers around the value 1.1. The standard deviation of the squared dual local norms was close to 18.2, and the maximum ever observed squared dual local norm was approximately 26 294.

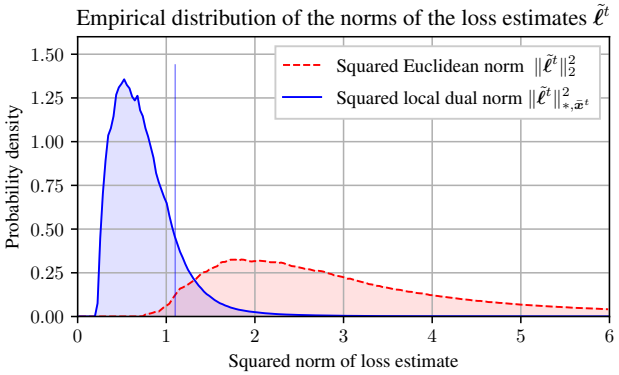

Figure 3: Empirical distribution of the squared norms of the loss estimates $\tilde{\boldsymbol{\ell}}^t$ across 100 runs of $\mathcal{R}$. The thin vertical line denotes the empirical average of the squared dual local norms.

On the other hand, the empirical distribution of the squared ℓ_2 norms has an extremely heavy tail, and the empirical average is slightly larger than 5 050, with a standard deviation in the order of 10^6 and a maximum observed squared ℓ_2 norm in the order of 10^{10}. This reinforces the theoretical observation that started Section 5: an analysis of $\mathcal{R}$ based on ℓ_2 norm would be insufficient.

8 Conclusion and Future Work

In this paper, we gave a practical algorithm for the bandit linear optimization problem on sequential decision spaces. Our method combines a number of ideas and tools. For one, we gave several new results concerning the properties of the dilated entropy regularizer that are of interest beyond the goal of this paper. Another contribution of this paper is an efficient way of constructing an unbiased estimator of the loss vector $\boldsymbol{\ell}^t$ starting from the loss evaluation $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$ at a pure strategy $\boldsymbol{y}^t$. In order to construct the unbiased estimator, we extended and generalized an argument by Bartlett et al. (2008) and showed how it can be applied successfully in the context of sequential decision processes. Finally, we combined the regret bound for $\tilde{\mathcal{R}}$ based on time-dependent local norms to-

gether with the unbiased loss estimator to construct our bandit regret minimizer, by showing that the unbiased loss estimator has a time-dependent dual norm that is upper-bounded by a small time-independent constant.

Our bandit regret minimizer $\mathcal{R}$ is superior to that of Abernethy, Hazan, and Rakhlin (2008) both computationally (each iteration runs in linear time in the SDP size) and in terms of cumulated regret (the regret grows as $O(\sqrt{T})$ instead of $O(\sqrt{T \log T})$. However, our algorithm, just like the one of Abernethy, Hazan, and Rakhlin (2008), only gives a regret bound that (i) only holds in expectation, and (ii) predicates on $\max_{\boldsymbol{z} \in \text{co}\, \mathcal{T}} \mathbb{E}[R^T(\boldsymbol{z})]$ and *not* $\mathbb{E}[\max_{\boldsymbol{z} \in \text{co}\, \mathcal{T}} R^T(\boldsymbol{z})]$. We believe that the techniques presented in this paper can be extended to overcome both shortcomings and yield a high-probability bound on $\max_{\boldsymbol{z} \in \text{co}\, \mathcal{T}} R^T(\boldsymbol{z})$. Indeed, our regularizer, the sampling scheme, the construction of the loss estimates, and the use of local norms might be used within the general framework of Abernethy and Rakhlin (2009) to provide high-probability results. We are interested in pursuing this direction in the future.

References

Abernethy, J. D., and Rakhlin, A. 2009. Beating the adaptive bandit with high probability. *2009 Information Theory and Applications Workshop*.

Abernethy, J.; Hazan, E.; and Rakhlin, A. 2008. Competing in the dark: An efficient algorithm for bandit linear optimization. In *In Proceedings of the 21st Annual Conference on Learning Theory (COLT)*.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2003. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*.

Awerbuch, B., and Kleinberg, R. D. 2004. Adaptive routing with end-to-end feedback: distributed learning and g eometric approaches. In *Proceedings of the Annual Symposium on Theory of Computing (STOC)*. ACM.

Bartlett, P. L.; Dani, V.; Hayes, T.; Kakade, S.; Rakhlin, A.; and Tewari, A. 2008. High-probability regret bounds for bandit online linear optimization. In *Conference on Learning Theory (COLT)*.

Bowling, M.; Burch, N.; Johanson, M.; and Tammelin, O. 2015. Heads-up limit hold'em poker is solved. *Science* 347(6218).

Brown, N., and Sandholm, T. 2017. Safe and nested subgame solving for imperfect-information games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Brown, N., and Sandholm, T. 2019a. Solving imperfect-information games via discounted regret minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Brown, N., and Sandholm, T. 2019b. Superhuman AI for multiplayer poker. *Science*.

Dani, V.; Kakade, S. M.; and Hayes, T. P. 2008. The price of bandit information for online optimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Farina, G.; Kroer, C.; and Sandholm, T. 2018. Online convex optimization for sequential decision processes and extensive-form games. In *arXiv*.

Farina, G.; Kroer, C.; and Sandholm, T. 2019. Online convex optimization for sequential decision processes and extensive-form games. In *AAAI Conference on Artificial Intelligence*.

Hoda, S.; Gilpin, A.; Peña, J.; and Sandholm, T. 2010. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research* 35(2).

Koller, D.; Megiddo, N.; and von Stengel, B. 1994. Fast algorithms for finding randomized strategies in game trees. In *Proceedings of the 26th ACM Symposium on Theory of Computing (STOC)*.

Kroer, C.; Waugh, K.; Kılınç-Karzan, F.; and Sandholm, T. 2015. Faster first-order methods for extensive-form game solving. In *Proceedings of the ACM Conference on Economics and Computation (EC)*.

Kroer, C.; Waugh, K.; Kılınç-Karzan, F.; and Sandholm, T. 2018. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*.

Kroer, C.; Farina, G.; and Sandholm, T. 2018. Solving large sequential games with the excessive gap technique. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Kuhn, H. W. 1950. A simplified two-person poker. In Kuhn, H. W., and Tucker, A. W., eds., *Contributions to the Theory of Games*, volume 1 of *Annals of Mathematics Studies, 24*. Princeton, New Jersey: Princeton University Press. 97–103.

Lanctot, M.; Waugh, K.; Zinkevich, M.; and Bowling, M. 2009. Monte Carlo sampling for regret minimization in extensive games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Ling, C. K.; Fang, F.; and Kolter, J. Z. 2019. Large scale learning of agent rationality in two-player zero-sum games. *CoRR* abs/1903.04101. Available at https://arxiv.org/abs/1903.04101.

Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*.

Rakhlin, A. 2009. Lecture notes on online learning. Available at http://www-stat.wharton.upenn.edu/~rakhlin/courses/stat991/papers/lecture_notes.pdf.

Romanovskii, I. 1962. Reduction of a game with complete memory to a matrix game. *Soviet Mathematics* 3.

Shalev-Shwartz, S. 2012. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning* 4(2).

Tammelin, O.; Burch, N.; Johanson, M.; and Bowling, M. 2015. Solving heads-up limit Texas hold'em. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

von Stengel, B. 1996. Efficient computation of behavior strategies. *Games and Economic Behavior* 14(2):220–246.

Zinkevich, M.; Bowling, M.; Johanson, M.; and Piccione, C. 2007. Regret minimization in games with incomplete information. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Appendix

In order to simplify the proofs, we will make the following assumptions about the structure of the SDP. All assumptions can be made without loss of generality, as the structure of an SDP can always be transformed in polynomial time into an equivalent SDP that fulfills our assumptions:

- No two nodes of the same type immediately follow each other. This assumption does not come at the cost of generality, since two consecutive decision points can always be consolidated into an equivalent one by combining their actions. Similarly, two consecutive observation points can be consolidated into an equivalent one by combining the available signals.
- Any action $a \in A_j$ at decision point $j \in \mathcal{J}$ such that $\rho(j, a) = \diamond$ is called a *terminal action*. A decision point j is called a *terminal decision point* if all actions in A_j are terminal, and *non-terminal decision point* if no action is terminal. We assume that decision points are either terminal or non-terminal. In other words, we assume that either all actions at a generic decision point j are terminal, or none of them is. One can always convert a SDP that contains a decision point j that are neither terminal nor non-terminal into an equivalent SDP by adding artificial observation points.

Given the first assumption, we denote the set of all observation points that are immediately reachable after decision point j as $\mathcal{C}_j := \{\rho(j, a) : a \in A_j\} \setminus \{\diamond\}$. Similarly, the set of all decision points that are immediately reachable after observation point k is $\mathcal{C}_k := \{\rho(k, s) : s \in S_k\} \setminus \{\diamond\}$.

A Additional Notation for Sequential Decision Processes

We now introduce additional notation:

Sequences (Σ). In line with the game theory literature, we call a pair (j, a) where $j \in \mathcal{J}$ and $a \in A_j$ a *sequence*. The set of all sequences is denoted as $\Sigma := \{(j, a) : j \in \mathcal{J}, a \in A_j\}$. For notational convenience, we will often denote an element (j, a) in Σ as ja without using parentheses.

Descendants ($\succeq$). A partial order $\succeq$ can be established on Σ as follows: given two sequences ja and $j'a'$ in Σ, $j'a' \succeq ja$ if and only if the (unique) path from the root node of the SDP to action a' at decision point j' passes through action a at decision point j. Whenever $j'a' \succeq ja$, we say that $j'a'$ is a *descendant* of ja.

Subtree indicator ($\boldsymbol{u}_{ja}$). Given a sequence $ja \in \Sigma$, we denote with $\boldsymbol{u}_{ja}$ the vector such that $(\boldsymbol{u}_{ja})_{j'a'} = 1$ if $j'a' \succeq ja$, and $(\boldsymbol{u}_{ja})_{j'a'} = 0$ otherwise.

Parent sequence (p_j). Given a decision point $j \in \mathcal{J}$, we denote with p_j its *parent sequence*, defined as the last sequence (that is, decision point-action pair) encountered on the path from the root of the SDP to decision point j. If the agent does not act before j (i.e., j is the root of the SDP or only observation points are encountered on the path from the root to j), we let $p_j = \varnothing$.

Inductive Definition of $\mathcal{T}$

The set $\mathcal{T}$ can be constructed recursively in a bottom-up fashion, as follows:

- At each terminal decision point $j \in \mathcal{J}$, the set of pure strategies is the set

$$\mathcal{T}_j := \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_{|A_j|}\} \subseteq \mathbb{R}^{|A_j|}, \tag{9}$$

  where $\boldsymbol{e}_i$ the i-th canonical basis vector.
- At each observation point $k \in \mathcal{K}$, the set of pure strategies is simply the Cartesian product of strategies for each of the child subtrees:

$$\mathcal{T}_k := \mathcal{T}_{j_1} \times \cdots \times \mathcal{T}_{j_n}, \tag{10}$$

  where $\{j_1, \dots, j_n\} = \mathcal{C}_k$ are the decision points immediately reachable after k.
- At each non-terminal decision point $j \in \mathcal{J}$, we put probability 1 on exactly one action and recurse on the subtree rooted at that action:

$$\mathcal{T}_j := \{(\boldsymbol{e}_1, \boldsymbol{x}_{k_1}, \boldsymbol{0}, \dots, \boldsymbol{0}) : \boldsymbol{x}_{k_1} \in \mathcal{T}_{k_1}\} \cup \cdots \cup \{(\boldsymbol{e}_n, \boldsymbol{0}, \dots, \boldsymbol{0}, \boldsymbol{x}_{k_n}) : \boldsymbol{x}_{k_n} \in \mathcal{T}_{k_n}\} \tag{11}$$

  where $\{k_1, \dots, k_n\} = \mathcal{C}_j$ are the observation points immediately reachable after j and $\boldsymbol{e}_i$ the i-th canonical basis vector.

Inductive Definition of $\operatorname{co} \mathcal{T}$

The set of mixed sequence-form strategies can also equivalently constructed inductively along the tree structure:

- At each terminal decision point $j \in \mathcal{J}$, the set of mixed strategies is the set

$$\operatorname{co} \mathcal{T}_j := \Delta^{|A_j|}. \tag{12}$$

- At each observation point $k \in \mathcal{K}$, the set of mixed strategies is the Cartesian product of mixed strategies for each of the child subtrees:
$$\operatorname{co} \mathcal{T}_k := \operatorname{co} \mathcal{T}_{j_1} \times \cdots \times \operatorname{co} \mathcal{T}_{j_n}, \tag{13}$$
where $\{j_1, \ldots, j_n\} = \mathcal{C}_k$ are the decision points immediately reachable after k.
- At each non-terminal decision point $j \in \mathcal{J}$, we first fix a distributions over the actions in A_j and then recurse:
$$\operatorname{co} \mathcal{T}_j := \{(\lambda_1, \ldots, \lambda_n, \lambda_1 \boldsymbol{x}_{k_1}, \ldots, \lambda_n \boldsymbol{x}_{k_n}) : (\lambda_1, \ldots, \lambda_n) \in \Delta^n, \boldsymbol{x}_i \in \operatorname{co} \mathcal{T}_{k_i} \ \forall i = 1, \ldots, n\}. \tag{14}$$
where $\{k_1, \ldots, k_n\} = \mathcal{C}_j$ are the observation points immediately reachable after j.

B Properties of the Dilated Entropy Distance-Generating Function

Preliminaries

We look at the computation of the gradient of φ at a generic point $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{>0}$. Some elementary algebra reveals that

$$\frac{\partial \varphi}{\partial z_{ja}}(\boldsymbol{z}) = w_j \left(1 + \log \frac{z_{ja}}{z_{p_j}}\right) + \sum_{j' \in \mathcal{C}_{\rho(j,a)}} w_{j'} \left(\log |A_{j'}| - \sum_{a' \in A_{j'}} \frac{z_{j'a'}}{z_{ja}}\right) \tag{15}$$

for every decision point-action pair $ja \in \Sigma$. Hence, we can compute $\nabla\varphi(\boldsymbol{z})$ at any $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{>0}$ in one linear-time traversal of the sequential decision tree as in Algorithm 5.

Algorithm 5: GRADIENT($\boldsymbol{z}$)

Input: $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$
Output: The value of $\nabla\varphi(\boldsymbol{z})$

$\boldsymbol{g} \leftarrow \boldsymbol{0} \in \mathbb{R}^{|\Sigma|}$
for $j \in \mathcal{J}$ *in bottom-up order* **do**
  for $a \in A_j$ **do**
    $g_{ja} \leftarrow g_{ja} + w_j \left(1 + \log \frac{z_{ja}}{z_{p_j}}\right)$
    if $p_j \neq \varnothing$ **then** $g_{p_j} \leftarrow g_{p_j} - w_j \frac{z_{j'a'}}{z_{ja}}$
  if $p_j \neq \varnothing$ **then** $g_{p_j} \leftarrow g_{p_j} + w_j \log |A_j|$
return $\boldsymbol{g}$

The Fenchel conjugate of φ on $\operatorname{co}\mathcal{T}$ is defined as

$$\varphi^* : \boldsymbol{z} \mapsto \max_{\hat{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}} \{\boldsymbol{z}^\top \hat{\boldsymbol{x}} - \varphi(\hat{\boldsymbol{x}})\}$$

for any $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$. It is well-known (and easy to check via a straightforward application of Danskin's theorem) that

$$\nabla\varphi^* : \boldsymbol{z} \mapsto \arg\max_{\hat{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}} \{\boldsymbol{z}^\top \hat{\boldsymbol{x}} - \varphi(\hat{\boldsymbol{x}})\}. \tag{16}$$

For this reason, we call $\nabla\varphi^*$ the *Fenchel arg-conjugate* function of φ on $\operatorname{co}\mathcal{T}$. Of course, for any $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$ one can efficiently compute the value of $\varphi^*(\boldsymbol{z})$ given $\boldsymbol{x}^* := \nabla\varphi^*(\boldsymbol{z})$ (which is guaranteed to be in $\in \mathbb{R}^{|\Sigma|}_{>0}$) by direct substitution as $\boldsymbol{z}^\top \boldsymbol{x}^* - \varphi(\boldsymbol{x}^*)$. In Algorithm 6 we give a linear-time algorithm for computing $\nabla\varphi^*(\boldsymbol{z})$. We refer the reader to the original work by Hoda et al. (2010) for a proof of correctness.

Local Norm

Lemma 5 (Ling, Fang, and Kolter (2019))**.** *The Hessian $\nabla^2\varphi(\boldsymbol{z})$ of the dilated entropy DGF at $\boldsymbol{z} \in \operatorname{co}\mathcal{T}$ is given by:*

$$\frac{\partial^2}{\partial z_{ja} \partial z_{j'a'}} \varphi(\boldsymbol{z}) = \begin{cases} \dfrac{w_j + w_{\rho(j,a)}}{z_{ja}} & \text{if } ja = j'a' \\ -\dfrac{w_{j'}}{z_{ja}} & \text{if } ja = p_{j'} \text{ and } p_{j'} \neq \varnothing \\ -\dfrac{w_j}{z_{p_j}} & \text{if } j'a' = p_j \text{ and } p_j \neq \varnothing \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 2. *Let $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ and $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{\geq 0}$. Then,*

$$\|\boldsymbol{z}\|^2_{\bar{\boldsymbol{x}}} \leq \frac{3}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \frac{w_j}{\bar{x}_{ja}} z^2_{ja}.$$

Algorithm 6: ARGCONJUGATE($\boldsymbol{z}$)

Input: $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}$
Output: The value of $\nabla\varphi^*(\boldsymbol{z})$

$\boldsymbol{x}^* \leftarrow \mathbf{0} \in \mathbb{R}^{|\Sigma|}$
for $j \in \mathcal{J}$ *in bottom-up order* **do**
  $s \leftarrow 0$
  for $a \in A_j$ **do**
    $x^*_{ja} \leftarrow \exp\{\frac{z_{ja}}{w_j}\}$
    $s \leftarrow s + x^*_{ja}$
  $v \leftarrow w_j \log |A_j|$
  for $a \in A_j$ **do**
    $x^*_{ja} \leftarrow \frac{x^*_{ja}}{s}$ [▷ Normalization step
    $v \leftarrow v + z_{ja} x^*_{ja} - x^*_{ja} \log x^*_{ja}$
  $z_{ja} \leftarrow z_{ja} + v$
for $j \in \mathcal{J}$ *in top-down order* **do**
  for $a \in A_j$ **do**
    $x^*_{ja} \leftarrow x^*_{ja} \cdot x^*_{p_j}$
return $\boldsymbol{x}^*$

Proof. Using the explicit expression of the Hessian of the dilated entropy regularizer (Lemma 5) we can write

$$\|\boldsymbol{z}\|^2_{\bar{\boldsymbol{x}}} = \sum_{j\in\mathcal{J}} \sum_{a\in A_j} \frac{w_j + w_{\rho(j,a)}}{\bar{x}_{ja}} z^2_{ja} - 2 \sum_{j\in\mathcal{J}} \sum_{a\in A_j} \sum_{j'\in\mathcal{C}_{\rho(j,a)}} \sum_{a'\in A_{j'}} \frac{w_{j'}}{\bar{x}_{ja}} z_{j'a'} z_{ja} \le \sum_{j\in\mathcal{J}} \sum_{a\in A_j} \frac{w_j + w_{\rho(j,a)}}{\bar{x}_{ja}} z^2_{ja}, \tag{17}$$

where the inequality holds since $\boldsymbol{z} \in \mathbb{R}^{|\Sigma|}_{\ge 0}$. By definition of w_j (Equation (2)), we have for all $ja \in \Sigma$

$$w_j + w_{\rho(j,a)} \le w_j + \max_{a'\in A_j} w_{\rho(j,a')} = 2 + 3 \max_{a'\in A_j} w_{\rho(j,a')} \le 3 + 3 \max_{a'\in A_j} w_{\rho(j,a')} = \frac{3}{2} w_j. \tag{18}$$

Plugging (18) into (17) yields the statement. □

Lemma 3. *Let $\bar{\boldsymbol{x}} \in \operatorname{co}\mathcal{T}$ be a sequence-form strategy. The inverse Hessian $(\nabla^2\varphi)^{-1}(\bar{\boldsymbol{x}})$ at $\bar{\boldsymbol{x}}$ can be expressed as:*

$$(\nabla^2\varphi)^{-1}(\bar{\boldsymbol{x}}) = \sum_{j\in\mathcal{J}} \sum_{a\in A_j} \frac{(\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})(\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})^\top}{w_j \bar{x}_{ja}}, \tag{3}$$

where $\circ$ denotes componentwise product of vectors.

Proof. Let

$$\boldsymbol{H} := \sum_{j\in\mathcal{J}} \sum_{a\in A_j} \frac{(\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})(\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})^\top}{w_j \bar{x}_{ja}}$$

be the proposed inverse Hessian matrix. We will prove that $\boldsymbol{H} = (\nabla^2\varphi)^{-1}(\bar{\boldsymbol{x}})$ by showing that $\nabla^2\varphi(\bar{\boldsymbol{x}}) \cdot \boldsymbol{H} = \boldsymbol{I}$ is the identity matrix. We break the proof into two steps:

- **Step one.** First, we show that for all sequences $ja \in \Sigma$ and $j'a' \in \Sigma$,

$$\left[\nabla^2\varphi(\bar{\boldsymbol{x}}) \cdot (\bar{\boldsymbol{x}} \circ \boldsymbol{u}_{ja})\right]_{j'a'} = \begin{cases} w_j & \text{if } j'a' = ja \\ -w_j \dfrac{\bar{x}_{ja}}{\bar{x}_{p_j}} & \text{if } p_j = j'a' \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

In order to prove (19), we start from Lemma 5:

$$\left[\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})\right]_{j'a'} = \sum_{j''\in\mathcal{J}}\sum_{a''\in A_{j''}}\frac{\partial^2\varphi(\bar{\boldsymbol{x}})}{\partial\bar{x}_{j'a'}\partial\bar{x}_{j''a''}}\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})_{j''a''}$$
$$= (w_{j'}+w_{\rho(j',a')})\cdot(\boldsymbol{u}_{ja})_{j'a'} - w_{j'}\cdot(\boldsymbol{u}_{ja})_{p_{j'}} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}}\sum_{a''\in A_{j''}}\frac{w_{j''}}{\bar{x}_{j'a'}}\bar{x}_{j''a''}\cdot(\boldsymbol{u}_{ja})_{j''a''}.$$

We now distinguish four cases, based on how ja relates to $p_{j'}$, $j'a'$, and $j''a''$:

– *First case:* $p_{j'} \succeq ja$, that is $p_{j'}, j'a'$ and $j''a''$ are all descendants of ja. Consequently, $(\boldsymbol{u}_{ja})_{p_{j'}} = (\boldsymbol{u}_{ja})_{j'a'} = (\boldsymbol{u}_{ja})_{j''a''} = 1$ for all $j''\in\mathcal{C}_{\rho(j',a')}$ and $a''\in A_{j''}$. Hence,

$$\left[\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})\right]_{j'a'} = w_{j'}+w_{\rho(j',a')} - w_{j''} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}}\sum_{a''\in A_{j''}}\frac{w_{j''}}{\bar{x}_{j'a'}}\bar{x}_{j''a''}$$
$$= w_{\rho(j',a')} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}}\left(w_{j''}\sum_{a''\in A_{j''}}\frac{\bar{x}_{j''a''}}{\bar{x}_{j'a'}}\right) = w_{\rho(j',a')} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}} w_{j''} = 0.$$

– *Second case:* $ja = j'a'$. In this case, $(\boldsymbol{u}_{ja})_{p_{j'}} = 0$, while $(\boldsymbol{u}_{ja})_{j'a'} = (\boldsymbol{u}_{ja})_{j''a''} = 1$ for all $j''\in\mathcal{C}_{\rho(j',a')}$ and $a''\in A_{j''}$. Hence,

$$\left[\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})\right]_{j'a'} = w_{j'}+w_{\rho(j',a')} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}}\sum_{a''\in A_{j''}}\frac{w_{j''}}{\bar{x}_{j'a'}}\bar{x}_{j''a''}$$
$$= w_{j'}+w_{\rho(j',a')} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}}\left(w_{j''}\sum_{a''\in A_{j''}}\frac{\bar{x}_{j''a''}}{\bar{x}_{j'a'}}\right)$$
$$= w_{j'}+w_{\rho(j',a')} - \sum_{j''\in\mathcal{C}_{\rho(j',a')}} w_{j''} = w_{j'} = w_j.$$

– *Third case:* $p_j = j'a'$ (that is, ja immediately follows $j'a'$). Then, $\left[\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})\right]_{j'a'} = -w_j\frac{\bar{x}_{ja}}{\bar{x}_{p_j}}$

– *Otherwise,* $j''a''\not\succeq ja$ for all $j''\in\mathcal{C}_{\rho(j',a')}$ and $a''\in A_{j''}$, and therefore $\left[\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})\right]_{j'a'} = 0$.

- **Step two.** Given $\sigma\in\Sigma\cup\{\varnothing\}$, let $\mathbf{1}_\sigma\in\mathbb{R}^{|\Sigma|}$ denote the vector that has a 1 in the entry corresponding to sequence σ, and 0 everywhere else (in particular, $\mathbf{1}_\varnothing = \mathbf{0}$). Then, (19) can be rewritten as

$$\frac{\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})}{w_j\bar{x}_{ja}} = \frac{1}{\bar{x}_{ja}}\mathbf{1}_{ja} - \frac{1}{\bar{x}_{p_j}}\mathbf{1}_{p_j}.$$

Therefore,

$$\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot\boldsymbol{H} = \sum_{j\in J}\sum_{a\in A_j}\frac{\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})}{w_j\bar{x}_{ja}}\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})^\top = \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\left(\frac{1}{\bar{x}_{ja}}\mathbf{1}_{ja} - \frac{1}{\bar{x}_{p_j}}\mathbf{1}_{p_j}\right)\cdot(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})^\top.$$
$$= \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{1}{\bar{x}_{ja}}\mathbf{1}_{ja}\cdot\left(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja} - \sum_{j'\in\mathcal{C}_{\rho(j,a)}}\sum_{a'\in A_{j'}}\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{j'a'}\right)^\top$$
$$= \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{1}{\bar{x}_{ja}}\mathbf{1}_{ja}\cdot\left[\bar{\boldsymbol{x}}\circ\left(\boldsymbol{u}_{ja} - \sum_{j'\in\mathcal{C}_{\rho(j,a)}}\sum_{a'\in A_{j'}}\boldsymbol{u}_{j'a'}\right)\right]^\top.$$

Using the definition of $\boldsymbol{u}_{ja}$, we obtain

$$\begin{aligned}\nabla^2\varphi(\bar{\boldsymbol{x}})\cdot\boldsymbol{H} &= \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{1}{\bar{x}_{ja}}\mathbf{1}_{ja}\cdot(\bar{\boldsymbol{x}}\circ\mathbf{1}_{ja})^\top\\ &= \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\mathbf{1}_{ja}\mathbf{1}_{ja}^\top\\ &= \boldsymbol{I},\end{aligned}$$

as we wanted to show. □

Corollary 1. *Let $\bar{\boldsymbol{x}}\in\operatorname{co}\mathcal{T}$ be a sequence-form strategy, and let $\boldsymbol{z}\in\mathbb{R}^{|\Sigma|}$. The local dual norm of $\boldsymbol{z}$ satisfies*

$$\|\boldsymbol{z}\|_{*,\bar{\boldsymbol{x}}}^2 = \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{(\boldsymbol{u}_{ja}^\top(\boldsymbol{z}\circ\bar{\boldsymbol{x}}))^2}{w_j\bar{x}_{ja}}. \tag{4}$$

Proof. By definition of local dual norm, using Lemma 3), and applying simple algebraic manipulations:

$$\|\boldsymbol{z}\|_{*,\bar{\boldsymbol{x}}}^2 = \boldsymbol{z}^\top\left(\sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja})^\top}{w_j\bar{x}_{ja}}\right)\boldsymbol{z} = \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{(\boldsymbol{z}^\top(\bar{\boldsymbol{x}}\circ\boldsymbol{u}_{ja}))^2}{w_j\bar{x}_{ja}} = \sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{(\boldsymbol{u}_{ja}^\top(\boldsymbol{z}\circ\bar{\boldsymbol{x}}))^2}{w_j\bar{x}_{ja}}.$$

□

C Analysis of Mirror Descent using Dilated Entropy DGF

We study some properties of Algorithm 1. The central result, Theorem 1, gives a bound on the cumulative regret expressed in term of (dual) local norms centered at the iterates produced by online mirror descent. Our first step is to introduce the "intermediate" iterate

$$\tilde{\boldsymbol{x}}^{t+1} := \operatorname*{arg\,min}_{\hat{\boldsymbol{x}} \in \mathbb{R}^{|\Sigma|}_{>0}} \Big\{ (\eta \tilde{\boldsymbol{\ell}}^t - \nabla \varphi(\bar{\boldsymbol{x}}^t))^\top \hat{\boldsymbol{x}} + \varphi(\hat{\boldsymbol{x}}) \Big\}, \tag{20}$$

which differs from $\bar{\boldsymbol{x}}^{t+1}$ in (6) in that the minimization problem is unconstrained. This intermediate iterate is known to be convenient for analyzing the regret accumulated by online mirror descent Abernethy and Rakhlin (2009). In particular, the following is well-known

Lemma 6 (Rakhlin (2009), Lemma 13)**.** *Online mirror descent satisfies, at all times T and for all mixed strategies $\boldsymbol{z} \in \operatorname{co} \mathcal{T}$, the regret bound*

$$R^T(\boldsymbol{z}) \leq \frac{\varphi(\boldsymbol{z})}{\eta} + \sum_{t=1}^{T} (\tilde{\boldsymbol{\ell}}^t)^\top (\bar{\boldsymbol{x}}^t - \tilde{\boldsymbol{x}}^{t+1}).$$

Using the structure of the dilated entropy DGF together with that of the game tree, we prove the following property, which will be fundamental in the analysis of online mirror descent based on local norms.

Proposition 2. *Let the quantity ψ^t_{ja} be defined for all sequences $ja \in \Sigma$ as*

$$\psi^t_{ja} := \boldsymbol{u}_{ja}^\top (\tilde{\boldsymbol{\ell}}^t \circ \bar{\boldsymbol{x}}^t) = \sum_{j'a' \succeq ja} \tilde{\ell}^t_{j'a'} \cdot \bar{x}^t_{j'a'}.$$

If $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}^{|\Sigma|}_{\geq 0}$, the intermediate iterate $\tilde{\boldsymbol{x}}^{t+1}$ satisfies

$$\frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{ -\frac{\eta}{w_j \bar{x}^t_{ja}} \psi^t_{ja} \right\} \leq \frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} \leq \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}}.$$

In particular, Proposition 2 is a fundamental step for the following theorem, which bounds the length of the step (as measured according to the local norm $\|\cdot\|_{\bar{\boldsymbol{x}}^t}$) between the last decision $\bar{\boldsymbol{x}}^t$ and the next intermediate iterate $\tilde{\boldsymbol{x}}^{t+1}$ as a function of the stepsize parameter η and the dual local norm of the loss $\tilde{\boldsymbol{\ell}}^t$ that was last observed:

Proposition 3. *Let D be the maximum depth of any node in the SDP. If $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}^{|\Sigma|}_{\geq 0}$, then*

$$\|\bar{\boldsymbol{x}}^t - \tilde{\boldsymbol{x}}^{t+1}\|_{\bar{\boldsymbol{x}}^t} \leq \eta \sqrt{3D} \cdot \|\tilde{\boldsymbol{\ell}}^t\|_{*, \bar{\boldsymbol{x}}^t}. \tag{21}$$

Proof. By Corollary 2, $\bar{\boldsymbol{x}}^t - \tilde{\boldsymbol{x}}^{t+1} \in \mathbb{R}^{|\Sigma|}_{\geq 0}$. Hence, we can apply Lemma 2:

$$\|\bar{\boldsymbol{x}}^t - \tilde{\boldsymbol{x}}^{t+1}\|^2_{\bar{\boldsymbol{x}}^t} \leq \frac{3}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \frac{w_j}{\bar{x}^t_{ja}} (\bar{x}^t_{ja} - \tilde{x}^{t+1}_{ja})^2 = \frac{3}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} w_j \bar{x}^t_{ja} \left(1 - \frac{\tilde{x}^{t+1}_{ja}}{\bar{x}^t_{ja}} \right)^2.$$

Using Inequality (27),

$$\begin{aligned}
\|\bar{\boldsymbol{x}}^t - \tilde{\boldsymbol{x}}^{t+1}\|^2_{\bar{\boldsymbol{x}}^t} &\leq \frac{3\eta^2}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} w_j \bar{x}^t_{ja} \left(\sum_{j'a' \preceq ja} \frac{\psi^t_{j'a'}}{w_{j'} \bar{x}^t_{j'a'}} \right)^2 \\
&\leq \frac{3D\eta^2}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \sum_{j'a' \preceq ja} \frac{w_j \bar{x}^t_{ja}}{w^2_{j'} (\bar{x}^t_{j'a'})^2} (\psi^t_{j'a'})^2 \\
&= \frac{3D\eta^2}{2} \sum_{j \in \mathcal{J}} \sum_{a \in A_j} \sum_{j'a' \preceq ja} \frac{w_j \bar{x}^t_{ja}}{w_{j'} \bar{x}^t_{j'a'}} \frac{(\psi^t_{j'a'})^2}{w_{j'} \bar{x}^t_{j'a'}},
\end{aligned}$$

where the second inequality follows from applying Cauchy-Schwarz. Now using double counting we derive

$$\frac{3D\eta^2}{2}\sum_{j\in\mathcal{J}}\sum_{a\in A_j}\sum_{j'a'\preceq ja}\frac{w_j\bar{x}^t_{ja}}{w_{j'}\bar{x}^t_{j'a'}}\frac{(\psi^t_{j'a'})^2}{w_{j'}\bar{x}^t_{j'a'}}=\frac{3D\eta^2}{2}\sum_{j\in\mathcal{J}}\sum_{a\in A_j}\left(\frac{(\psi^t_{ja})^2}{w_j\bar{x}^t_{ja}}\sum_{j'a'\succeq ja}\frac{w_{j'}\bar{x}^t_{j'a'}}{w_j\bar{x}^t_{ja}}\right)$$
$$\leq 3D\eta^2\sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{(\psi^t_{ja})^2}{w_j\bar{x}^t_{ja}}.$$

where the second inequality follows from Lemma 8. Finally, plugging in definition of ψ^t_{ja} and using Corollary 1, we have

$$\|\bar{\boldsymbol{x}}^t-\tilde{\boldsymbol{x}}^{t+1}\|^2_{\bar{\boldsymbol{x}}^t}\leq 3D\eta^2\sum_{j\in\mathcal{J}}\sum_{a\in A_j}\frac{(\boldsymbol{u}^\top_{ja}(\tilde{\boldsymbol{\ell}}^t\circ\bar{\boldsymbol{x}}^t))^2}{w_j\bar{x}^t_{ja}}=3D\eta^2\|\tilde{\boldsymbol{\ell}}^t\|^2_{*,\bar{\boldsymbol{x}}^t}.$$

Taking the square root of both sides yields the statement. □

Lemma 6 can be used to derive a regret bound for $\tilde{\mathcal{R}}$ expressed in term of local norms. In particular, using the generalized Cauchy-Schwarz inequality together with Proposition 3, we obtain

$$(\tilde{\boldsymbol{\ell}}^t)^\top(\bar{\boldsymbol{x}}^t-\tilde{\boldsymbol{x}}^{t+1})\leq\|\tilde{\boldsymbol{\ell}}^t\|_{*,\bar{\boldsymbol{x}}^t}\cdot\|\bar{\boldsymbol{x}}^t-\tilde{\boldsymbol{x}}^{t+1}\|_{\bar{\boldsymbol{x}}^t}$$
$$\leq\eta\sqrt{3D}\cdot\|\tilde{\boldsymbol{\ell}}^t\|^2_{*,\bar{\boldsymbol{x}}^t}.$$

Substituting the last inequality into the bound of Lemma 6, we obtain the following:

Lemma 7. *At all times t, each intermediate iterate $\tilde{\boldsymbol{x}}^{t+1}$ satisfies, for all $ja\in\Sigma$:*

$$\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}}=\frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}}\exp\left\{-\eta\frac{\ell^t_{ja}}{w_j}-\frac{w_{\rho(j,a)}}{w_j}+\frac{\xi^{t+1}_{ja}}{w_j}\right\},\tag{22}$$

where

$$\xi^{t+1}_{ja}:=\sum_{j'\in\mathcal{C}_{\rho(j,a)}}w_{j'}\sum_{a'\in A_{j'}}\frac{\tilde{x}^{t+1}_{j'a'}}{\tilde{x}^{t+1}_{ja}}.$$

Proof. The first-order optimality condition for the minimization problem (20) yields

$$\eta\boldsymbol{\ell}^t-\nabla\varphi(\bar{\boldsymbol{x}}^t)+\nabla\varphi(\tilde{\boldsymbol{x}}^{t+1})=\mathbf{0}.$$

Substituting the expression for $\nabla\varphi$ (Equation 15) into the optimality condition yields

$$\eta\ell_{ja}-w_j\log\frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}}+\sum_{j'\in\mathcal{C}_{\rho(j,a)}}w_{j'}\sum_{a'\in A_{j'}}\frac{\bar{x}^t_{j'a'}}{\bar{x}^t_{ja}}+w_j\log\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}}-\sum_{j'\in\mathcal{C}_{\rho(j,a)}}w_{j'}\sum_{a'\in A_{j'}}\frac{\tilde{x}^{t+1}_{j'a'}}{\tilde{x}^{t+1}_{ja}}=0$$

for all $ja\in\Sigma$. Using the fact that $\bar{\boldsymbol{x}}^t\in\operatorname{co}\mathcal{X}$, we can write $\sum_{a'\in A_{j'}}\frac{\bar{x}^t_{j'a'}}{\bar{x}^t_{ja}}=1$ and simplify the above condition into

$$\eta\ell_{ja}-w_j\log\frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}}+w_j\log\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}}+w_{\rho(j,a)}-\sum_{j'\in\mathcal{C}_{\rho(j,a)}}w_{j'}\sum_{a'\in A_{j'}}\frac{\tilde{x}^{t+1}_{j'a'}}{\tilde{x}^{t+1}_{ja}}=0,$$

where we used the equality $w_{\rho(j,a)}=\sum_{j'\in\mathcal{C}_{\rho(j,a)}}w_{j'}$ (Equation 2). Rearranging the terms yields the statement. □

Proposition 2. *Let the quantity ψ^t_{ja} be defined for all sequences $ja\in\Sigma$ as*

$$\psi^t_{ja}:=\boldsymbol{u}^\top_{ja}(\tilde{\boldsymbol{\ell}}^t\circ\bar{\boldsymbol{x}}^t)=\sum_{j'a'\succeq ja}\tilde{\ell}^t_{j'a'}\cdot\bar{x}^t_{j'a'}.$$

If $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}^{|\Sigma|}_{\geq 0}$, the intermediate iterate $\tilde{\boldsymbol{x}}^{t+1}$ satisfies

$$\frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\frac{\eta}{w_j \bar{x}^t_{ja}} \psi^t_{ja}\right\} \leq \frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} \leq \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}}.$$

Proof. For ease of notation, in this proof we will make use of the symbol $\mathcal{C}_{ja}$ to mean $\mathcal{C}_{\rho(j,a)}$. We prove the proposition by induction:

- **Base case**. For any ja with $\mathcal{C}_{ja} = \emptyset$ (and thus $\psi^t_{ja} = \tilde{\ell}^t_{ja} \bar{x}^t_{ja}$) we have by Lemma 7

$$\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} = \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\eta \frac{\tilde{\ell}^t_{ja}}{w_j}\right\} = \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\frac{\eta}{w_j \bar{x}^t_{ja}} \psi^t_{ja}\right\},$$

  which proves the lower bound. In order to prove the upper bound, it is enough to note that the argument of the $\exp$ is non-positive. Hence,

$$\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} = \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\frac{\eta}{w_j \bar{x}^t_{ja}} \psi^t_{ja}\right\} \leq \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}}.$$

- **Inductive step**. Suppose that the inductive hypothesis holds for all sequences $j'a' \succ ja$. Then, we have

$$\xi^{t+1}_{ja} = \sum_{j' \in \mathcal{C}_{ja}} \left(w_{j'} \sum_{a' \in A_{j'}} \frac{\tilde{x}^{t+1}_{j'a'}}{\tilde{x}^{t+1}_{ja}} \right) \geq \sum_{j' \in \mathcal{C}_{ja}} \left(w_{j'} \sum_{a' \in A_{j'}} \frac{\bar{x}^t_{j'a'}}{\bar{x}^t_{ja}} \exp\left\{-\frac{\eta}{w_{j'} \bar{x}^t_{j'a'}} \psi^t_{j'a'}\right\} \right). \tag{23}$$

  Furthermore, for all $ja \in \Sigma$, using Equation (2) we have

$$w_{\rho(j,a)} = \sum_{j' \in \mathcal{C}_{ja}} w'_j = \sum_{j' \in \mathcal{C}_{ja}} w_{j'} \sum_{a' \in A_{j'}} \frac{\bar{x}^t_{j'a'}}{\bar{x}^t_{ja}},$$

  where the last equality follows from the fact that $\bar{\boldsymbol{x}}^t$ is a valid sequence-form strategy. Hence, we can rewrite (22) as

$$\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} = \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\eta \frac{\tilde{\ell}^t_{ja}}{w_j} - \frac{1}{w_j} \sum_{j' \in \mathcal{C}_{ja}} w_{j'} \sum_{a' \in A_{j'}} \frac{\bar{x}^t_{j'a'}}{\bar{x}^t_{ja}} + \frac{\xi^{t+1}_{ja}}{w_j}\right\}. \tag{24}$$

  Plugging in the inductive hypothesis (23) into (24) and using the monotonicity of $\exp$, we obtain

$$\begin{aligned}
\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} &\geq \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\eta \frac{\tilde{\ell}^t_{ja}}{w_j} - \frac{1}{w_j} \left(\sum_{j' \in \mathcal{C}_{ja}} w_{j'} \sum_{a' \in A_{j'}} \frac{\bar{x}^t_{j'a'}}{\bar{x}^t_{ja}} \left(1 - \exp\left\{-\frac{\eta}{w_{j'} \bar{x}^t_{j'a'}} \psi^t_{j'a'}\right\}\right)\right)\right\} \\
&\geq \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\eta \frac{\tilde{\ell}^t_{ja}}{w_j} - \frac{\eta}{w_j} \left(\sum_{j' \in \mathcal{C}_{ja}} \sum_{a' \in A_{j'}} \frac{1}{\bar{x}^t_{ja}} \psi^t_{j'a'} \right)\right\},
\end{aligned} \tag{25}$$

  where the second inequality follows from the fact that $1 - e^{-x} \leq x$ for all $x \in \mathbb{R}$. Finally, using the definition of ψ^t_{ja} we find

$$\sum_{j' \in \mathcal{C}_{ja}} \sum_{a' \in A_{j'}} \psi^t_{j'a'} = \psi^t_{ja} - \tilde{\ell}^t_{ja} \bar{x}^t_{ja}. \tag{26}$$

  Plugging (26) into (25) we obtain

$$\frac{\tilde{x}^{t+1}_{ja}}{\tilde{x}^{t+1}_{p_j}} \geq \frac{\bar{x}^t_{ja}}{\bar{x}^t_{p_j}} \exp\left\{-\frac{\eta}{w_j \bar{x}^t_{ja}} \psi^t_{ja}\right\}.$$

  This completes the proof for the lower bound.

In order to prove the upper bound, we start from (22).

$$\frac{\tilde{x}_{ja}^{t+1}}{\tilde{x}_{p_j}^{t+1}} = \frac{\bar{x}_{ja}^{t}}{\bar{x}_{p_j}^{t}} \exp\left\{ -\eta \frac{\tilde{\ell}_{ja}^{t}}{w_j} - \frac{1}{w_j} \left(\sum_{j' \in \mathcal{C}_{ja}} w_{j'} \sum_{a' \in A_{j'}} \left(\frac{\bar{x}_{j'a'}^{t}}{\bar{x}_{ja}^{t}} - \frac{\tilde{x}_{j'a'}^{t+1}}{\tilde{x}_{ja}^{t+1}} \right) \right) \right\} \leq \frac{\bar{x}_{ja}^{t}}{\bar{x}_{p_j}^{t}}.$$

Using the inductive hypothesis $\tilde{x}_{ja}^{t+1}/\tilde{x}_{p_j}^{t+1} \leq \bar{x}_{ja}^{t}/\bar{x}_{p_j}^{t}$, we obtain

$$\begin{aligned}\frac{\tilde{x}_{ja}^{t+1}}{\tilde{x}_{p_j}^{t+1}} &\leq \frac{\bar{x}_{ja}^{t}}{\bar{x}_{p_j}^{t}} \exp\left\{ -\eta \frac{\tilde{\ell}_{ja}^{t}}{w_j} - \frac{1}{w_j} \left(\sum_{j' \in \mathcal{C}_{ja}} w_{j'} \sum_{a' \in A_{j'}} \left(\frac{\bar{x}_{j'a'}^{t}}{\bar{x}_{ja}^{t}} - \frac{\bar{x}_{j'a'}^{t}}{\bar{x}_{ja}^{t}} \right) \right) \right\} \leq \frac{\bar{x}_{ja}^{t}}{\bar{x}_{p_j}^{t}} \\ &\leq \frac{\bar{x}_{ja}^{t}}{\bar{x}_{p_j}^{t}} \exp\left\{ -\eta \frac{\tilde{\ell}_{ja}^{t}}{w_j} \right\} \leq \frac{\bar{x}_{ja}^{t}}{\bar{x}_{p_j}^{t}}.\end{aligned}$$

□

An immediate corollary of Proposition 2 is the following:

Corollary 2. *For all $ja \in \Sigma$,*

$$0 < \exp\left\{ - \sum_{j'a' \preceq ja} \frac{\eta}{w_j \bar{x}_{j'a'}^{t}} \psi_{j'a'}^{t} \right\} \leq \frac{\tilde{x}_{ja}^{t+1}}{\bar{x}_{ja}^{t}} \leq \frac{\tilde{x}_{p_j}^{t+1}}{\bar{x}_{p_j}^{t}} \leq 1.$$

In particular,

$$0 \leq 1 - \frac{\tilde{x}_{ja}^{t+1}}{\bar{x}_{ja}^{t}} \leq \sum_{j'a' \preceq ja} \frac{\eta}{w_j \bar{x}_{j'a'}^{t}} \psi_{j'a'}^{t}. \tag{27}$$

Proof. The first statement follows from applying Proposition 2 repeatedly on the path from the root of the decision tree to decision point j. The second statement holds from the first statement by noting that

$$1 - \frac{\tilde{x}_{ja}^{t+1}}{\bar{x}_{ja}^{t}} \leq 1 - \exp\left\{ - \sum_{j'a' \preceq ja} \frac{\eta}{w_j \bar{x}_{j'a'}^{t}} \psi_{j'a'}^{t} \right\} \leq \sum_{j'a' \preceq ja} \frac{\eta}{w_j \bar{x}_{j'a'}^{t}} \psi_{j'a'}^{t},$$

where we used the fact that $1 - e^{-x} \leq x$ for all $x \in \mathbb{R}$. □

Lemma 8. *For all sequences ja,*

$$\sum_{j'a' \succeq ja} \frac{w_{j'} y_{j'a'}}{w_j y_{ja}} \leq 2.$$

Proof. By induction.

- **Base case**. For any terminal decision $ja \in \Sigma$ (that is, $\mathcal{C}_{\rho(j,a)} = \emptyset$), we have

$$\sum_{j'a' \succeq ja} \frac{w_{j'} y_{j'a'}}{w_j y_{ja}} = \frac{w_j \bar{x}_{ja}}{w_j \bar{x}_{ja}} = 1 \leq 2.$$

- **Inductive step**. Suppose that the inductive hypothesis holds for all sequences $j'a' \succ ja$. Then,

$$\begin{aligned}\sum_{j'a' \succeq ja} \frac{w_{j'} y_{j'a'}}{w_j y_{ja}} &= 1 + \sum_{j' \in \mathcal{C}_{\rho(j,a)}} \sum_{a' \in A_{j'}} \left(\frac{w_{j'} y_{j'a'}}{w_j y_{ja}} \sum_{j''a'' \succeq j'a'} \frac{w_{j''} y_{j''a''}}{w_{j'} y_{j'a'}} \right) \\ &\leq 1 + 2 \sum_{j' \in \mathcal{C}_{\rho(j,a)}} \sum_{a' \in A_{j'}} \frac{w_{j'} y_{j'a'}}{w_j y_{ja}} \\ &= 1 + 2 \sum_{j' \in \mathcal{C}_{\rho(j,a)}} \frac{w_{j'}}{w_j} = 1 + \frac{2 w_{\rho(j,a)}}{w_j} \leq 2,\end{aligned}$$

where the first inequality follows by the inductive hypothesis, and the second inequality holds by definition of the weights in the dilated DGF (Equation 2). □

D Sampling Scheme

Lemma 9. *Let π be a distribution with finite support, and let $\boldsymbol{y} \sim \pi$. Then* $\operatorname{Im} \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^\top\right] = \operatorname{span} \operatorname{supp} \pi$.

Proof. We prove the statement by showing that the nullspace of $\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$ is equal to the orthogonal complement of $\operatorname{span} \operatorname{supp} \pi$, in symbols:

$$\ker \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] = (\operatorname{span} \operatorname{supp} \pi)^\perp.$$

This will immediately imply the statement using the well-known relationship $\operatorname{Im} \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] = (\ker \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top])^\perp$.

We start by showing $\ker \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] \subseteq (\operatorname{span} \operatorname{supp} \pi)^\perp$. Take $\boldsymbol{z} \in \ker \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$. Then,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]\boldsymbol{z} = 0 &\implies \boldsymbol{z}^\top \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]\boldsymbol{z} = 0 \\
&\implies \mathbb{E}[(\boldsymbol{z}^\top \boldsymbol{y})^2] = 0 \\
&\implies \boldsymbol{z}^\top \boldsymbol{y} = 0 \quad \forall \boldsymbol{y} \in \operatorname{supp} \pi \\
&\implies \boldsymbol{z}^\top \boldsymbol{y} = 0 \quad \forall \boldsymbol{y} \in \operatorname{span} \operatorname{supp} \pi.
\end{aligned}$$

We now look at the other direction, that is $(\operatorname{span} \operatorname{supp} \pi)^\perp \subseteq \ker \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$. Take $\boldsymbol{z} \in (\operatorname{span} \operatorname{supp} \pi)^\perp$. Then,

$$\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]\boldsymbol{z} = \mathbb{E}[\boldsymbol{y}(\boldsymbol{y}^\top \boldsymbol{z})] = \mathbb{E}[\boldsymbol{y} \cdot 0] = \mathbf{0}.$$

This implies $\boldsymbol{z} \in \ker \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$, and concludes the proof. □

Lemma 10. *Suppose that a distribution π^t over $\mathcal{T}$ is known, such that the support of π^t is full-rank (that is,* $\operatorname{span} \operatorname{supp} \pi^t = \operatorname{span} \mathcal{T}$*), and let $\boldsymbol{y}^t \sim \pi^t$. Furthermore, let $\boldsymbol{C}^{t-}$ be any generalized inverse of the autocorrelation matrix $\boldsymbol{C}^t$. Then, for all $\boldsymbol{z} \in \operatorname{span} \mathcal{X}_j$,*

$$\boldsymbol{C}^t \boldsymbol{C}^{t-} \boldsymbol{z} = \boldsymbol{z}, \text{ and } \boldsymbol{z}^\top \boldsymbol{C}^{t-} \boldsymbol{C}^t = \boldsymbol{z}^\top.$$

Proof. Since $\operatorname{Im} \boldsymbol{C}^t = \operatorname{span} \operatorname{supp} \pi^t$ (see Lemma 9) and $\operatorname{span} \operatorname{supp} \pi^t = \operatorname{span} \mathcal{X}$ by hypothesis, it must be $\boldsymbol{z} \in \operatorname{Im} \boldsymbol{C}^t$. Hence, there exists $\boldsymbol{v} \in \mathbb{R}^{|\Sigma|}$ such that $\boldsymbol{z} = \boldsymbol{C}^t \boldsymbol{v}$, and therefore

$$\boldsymbol{C}^t \boldsymbol{C}^{t-} \boldsymbol{z} = \boldsymbol{C}^t \boldsymbol{C}^{t-} \boldsymbol{C}^t \boldsymbol{v} = \boldsymbol{C}^t \boldsymbol{v} = \boldsymbol{z},$$

where the second equality follows by definition of generalized inverse. The proof of the second equality in the statement is analogous. □

Lemma 11. *Suppose that a distribution π^t over $\mathcal{X}$ is known, such that the support of π^t is full-rank (that is,* $\operatorname{span} \operatorname{supp} \pi^t = \operatorname{span} \mathcal{X}$*), and let $\boldsymbol{y}^t \sim \pi^t$. Furthermore, let $\boldsymbol{C}^{t-}$ be any generalized inverse of the autocorrelation matrix $\boldsymbol{C}^t$. Then, for all $\boldsymbol{z} \in \operatorname{span} \mathcal{X}$,*

$$\boldsymbol{z}^\top \boldsymbol{C}^{t-} \bar{\boldsymbol{x}}^t = (\bar{\boldsymbol{x}}^t)^\top \boldsymbol{C}^{t-} \boldsymbol{z} = 1.$$

Proof. Since $\operatorname{Im} \boldsymbol{C}^t = \operatorname{span} \operatorname{supp} \pi^t$ (see Lemma 9) and $\operatorname{span} \operatorname{supp} \pi^t = \operatorname{span} \mathcal{X}$ by hypothesis, there exists $\boldsymbol{v} \in \mathbb{R}^{|\Sigma|}$ such that $\boldsymbol{z} = \boldsymbol{C}^t \boldsymbol{v}$. Furthermore, $\bar{\boldsymbol{x}}^t = \boldsymbol{C}^t \boldsymbol{\tau}$ where $\boldsymbol{\tau}$ is any vector such that $\boldsymbol{z}^\top \boldsymbol{\tau} = 1$ for all $\boldsymbol{z} \in \operatorname{co} \mathcal{T}$ (such vector must exist because $\mathbf{0}$ is not in the affine hull of $\mathcal{T}$). Hence,

$$\begin{aligned}
\boldsymbol{z}^\top \boldsymbol{C}^{t-} \bar{\boldsymbol{y}}^t &= \boldsymbol{v}^\top \boldsymbol{C}^t \boldsymbol{C}^{t-} \boldsymbol{C}^t \boldsymbol{\tau} \\
&= \boldsymbol{v}^\top \boldsymbol{C}^t \boldsymbol{\tau} \\
&= \boldsymbol{z}^\top \boldsymbol{\tau} = 1.
\end{aligned}$$

The proof that $(\bar{\boldsymbol{x}}^t)^\top \boldsymbol{C}^{t-} \boldsymbol{z} = 1$ is analogous. □

Proposition 1. *Let π^t be the conditional distribution over $\mathcal{T}$, given the previous decisions $\boldsymbol{y}^1, \dots, \boldsymbol{y}^{t-1}$, and suppose that the support of π^t is full-rank (that is,* $\operatorname{span} \operatorname{supp} \pi^t = \operatorname{span} \mathcal{T}$*). Let $\boldsymbol{C}^t := \mathbb{E}_t[\boldsymbol{y}^t (\boldsymbol{y}^t)^\top]$ be the autocorrelation matrix of $\boldsymbol{y}^t$, and let $\boldsymbol{C}^{t-}$ be any generalized inverse of $\boldsymbol{C}^t$, that is any matrix such that $\boldsymbol{C}^t \boldsymbol{C}^{t-} \boldsymbol{C}^t = \boldsymbol{C}^t$. Then, for any $\boldsymbol{b}^t \perp \operatorname{dir} \mathcal{T}$, the random variable*

$$\tilde{\boldsymbol{\ell}}^t := [(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t] \cdot (\boldsymbol{C}^{t-} \boldsymbol{y}^t + \boldsymbol{b}^t), \tag{8}$$

satisfies ($\star$).

Proof. For all $\boldsymbol{z} \in \operatorname{dir} \mathcal{X}$,

$$\boldsymbol{z}^\top \mathbb{E}_t[\tilde{\boldsymbol{\ell}}^t] = \boldsymbol{z}^\top \mathbb{E}_t\left[((\boldsymbol{y}^t)^\top \boldsymbol{\ell}^t)\, \boldsymbol{C}^{t-}\, \boldsymbol{y}^t\right] = \boldsymbol{z}^\top \mathbb{E}_t\left[\boldsymbol{C}^{t-}\, \boldsymbol{y}^t (\boldsymbol{y}^t)^\top\, \boldsymbol{\ell}^t\right] = \boldsymbol{z}^\top \boldsymbol{C}^{t-}\, \boldsymbol{C}^t \boldsymbol{\ell}^t.$$

Using the inclusion $\operatorname{dir} \mathcal{T} \subseteq \operatorname{span} \mathcal{T}$ together with Lemma 10 gives the statement. □

Unbiasedness of the Sampling Scheme

Lemma 4. *The sampling scheme given by Algorithm 2 is unbiased, that is,* $\mathbb{E}_t[\boldsymbol{y}^t] = \bar{\boldsymbol{x}}^t$.

Proof. We prove by induction over the structure of the sequential decision process that for all $v \in \mathcal{J} \cup \mathcal{K}$,

$$\mathbb{E}_t[\boldsymbol{y}_v^t] = \bar{\boldsymbol{x}}_v^t$$

- **First case:** $v \in \mathcal{J}$ is a terminal decision point. Let $A_v = \{a_1, \dots, a_n\}$. Then, $\bar{\boldsymbol{x}}_v^t = (\bar{x}_{va_1}^t, \dots, \bar{x}_{va_n}^t) \in \Delta^n$ and

$$\mathbb{E}_t[\boldsymbol{y}_v^t] = \sum_{i=1}^{n} \bar{x}_{va_i}^t \boldsymbol{e_i} = \bar{\boldsymbol{x}}_v^t.$$

- **Second case:** $v \in \mathcal{K}$ is an observation point. Let $\mathcal{C}_v = \{j_1, \dots, j_n\}$ be the set of decision points that are immediately reachable after v. From (13), $\bar{\boldsymbol{x}}_v^t$ is in the form $\bar{\boldsymbol{x}}_v^t = (\bar{\boldsymbol{x}}_{j_1}^t, \dots, \bar{\boldsymbol{x}}_{j_n}^t) \in \prod_{i=1}^n \operatorname{co} \mathcal{T}_{j_i}$. It follows that

$$\mathbb{E}_t[\boldsymbol{y}_v^t] = \mathbb{E}_t\left[\begin{pmatrix} \boldsymbol{y}_{j_1}^t \\ \vdots \\ \boldsymbol{y}_{j_n}^t \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}_t[\boldsymbol{y}_{j_1}^t] \\ \vdots \\ \mathbb{E}_t[\boldsymbol{y}_{j_n}^t] \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{x}}_{j_1}^t \\ \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t \end{pmatrix} = \bar{\boldsymbol{x}}_v^t,$$

  where the second equality follows from the independence of the sampling scheme and the third equality follows from the inductive hypothesis.

- **Third case:** $v \in \mathcal{J}$ is a non-terminal decision point. Let $\mathcal{C}_j = \{k_1, \dots, k_n\}$ be the set of observation points that are immediately reachable after v. From Equation (14), $\bar{\boldsymbol{x}}_j^t$ must be in the form $\bar{\boldsymbol{x}}_j^t = (\lambda_1^t, \dots, \lambda_n^t, \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t, \dots, \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t)$, where $\boldsymbol{\lambda}^t = (\lambda_1^t, \dots, \lambda_n^t) \in \Delta^n$. It follows that

$$\mathbb{E}_t[\boldsymbol{y}_v^t] = \mathbb{E}_t\left[\begin{pmatrix} \boldsymbol{y}_v^t \\ \boldsymbol{y}_{j_1}^t \\ \vdots \\ \boldsymbol{y}_{j_n}^t \end{pmatrix}\right] = \begin{pmatrix} \sum_{i=1}^n \lambda_i^t \boldsymbol{e_i} \\ \lambda_1^t \mathbb{E}_t[\boldsymbol{y}_{j_1}^t] \\ \vdots \\ \lambda_n^t \mathbb{E}_t[\boldsymbol{y}_{j_n}^t] \end{pmatrix} = \begin{pmatrix} \boldsymbol{\lambda}^t \\ \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t \\ \vdots \\ \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t \end{pmatrix} = \bar{\boldsymbol{x}}_v^t,$$

  where the second equality follows from the independence of the sampling scheme and the third equality follows from the inductive hypothesis. This concludes the proof. □

Autocorrelation Matrix of the Sampling Scheme

Terminal Decision Points Let $v \in \mathcal{J}$ be a terminal decision point and let $A_v = \{a_1, \dots, a_n\}$. The sequence-form strategy space $\operatorname{co} \mathcal{T}_v$ is the probability simplex Δ^n. Hence, at all times t, $\bar{\boldsymbol{x}}_v^t = (\lambda_1^t, \dots, \lambda_n^t) \in \Delta^n$. In this case When asked to sample a pure sequence-form strategy, we draw $a \in \{1, \dots, n\}$ according to the distribution specified by λ_a^t and return the standard basis vector $\boldsymbol{y}_v^t = \boldsymbol{e}_a \in \mathcal{T}_v$.

Lemma 12. *Let $v \in \mathcal{J}$ be a terminal decision point, and let $\bar{\boldsymbol{x}}_v^t = (\lambda_1^t, \dots, \lambda_n^t) \in \Delta^n = \operatorname{co} \mathcal{T}_v$ in accordance with* (12)*. The autocorrelation matrix of the sampling scheme that picks $\boldsymbol{y}_j^t \in \mathcal{T}_v$ is*

$$\boldsymbol{C}_j^t = \begin{pmatrix} \lambda_1^t & & \\ & \ddots & \\ & & \lambda_n^t \end{pmatrix}.$$

Non-Terminal Decision Points Let $v \in \mathcal{J}$ be a non-terminal decision point. In order to sample a pure sequence-form strategy we first sample $a \in \{1, \dots, |A_v|\}$ according to the distribution specified by $\{\bar{x}_{va}\}_{a \in A_v} \in \Delta^{|A_v|}$. Then, we set $y_{va}^t = 1$, and recursively sample $\boldsymbol{y}_{\rho(v,a)}^t$ by calling into the sampling scheme for $\rho(v, a)$.

Lemma 13. *Let $\mathcal{T}$ be a SDP rooted in non-terminal decision point j. Let $\mathcal{C}_j = \{k_1, \dots, k_n\}$ be the set of observation points that are immediately reachable after j. In accordance with* (14)*, $\bar{\boldsymbol{x}}_j^t$ is in the form $\bar{\boldsymbol{x}}_j^t = (\lambda_1^t, \dots, \lambda_n^t, \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t, \dots, \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t)$, where $(\lambda_1^t, \dots, \lambda_n^t) \in \Delta^n$. Let $\boldsymbol{C}_{k_i}^t$ for $i \in 1, \dots, n$ be the autocorrelation matrix of the unbiased sampling scheme picking $\boldsymbol{y}_{k_i}^t \in \mathcal{T}_{k_i}$ using $\bar{\boldsymbol{x}}_{k_i}/\lambda_i$. The autocorrelation matrix of the sampling scheme picking $\boldsymbol{y}_j^t$ is*

$$\boldsymbol{C}_j^t = \left(\begin{array}{ccc|ccc} \lambda_1^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_1}^t)^\top & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_n}^t)^\top \\ \hline \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t & & & \lambda_1^t \boldsymbol{C}_{k_1}^t & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t & & & \lambda_n^t \boldsymbol{C}_{k_n}^t \end{array}\right).$$

Observation Points Let $v \in \mathcal{K}$ be an observation point. In order to sample $\boldsymbol{y}_v^t$ given a $\bar{\boldsymbol{x}}_v^t = (\bar{\boldsymbol{x}}_{j_1}^t, \dots, \bar{\boldsymbol{x}}_{j_n}^t)$, we call into the sampling schemes for nodes $j_1, \dots, j_n$ by making n independent calls to $\textsc{Sample}(j_i, \bar{\boldsymbol{x}}_{j_i}^t)$ for $i = 1, \dots, n$.

Lemma 14. *Let $\mathcal{T}$ be a SDP rooted in observation point k. Let $\mathcal{C}_k = \{j_1, \dots, j_n\}$ be the set of decision points that are immediately reachable after k. In accordance with* (13)*, $\bar{\boldsymbol{x}}_k^t$ is in the form $\bar{\boldsymbol{x}}_k^t = (\bar{\boldsymbol{x}}_{j_1}^t, \dots, \bar{\boldsymbol{x}}_{j_n}^t)$. Let $\boldsymbol{C}_{j_i}^t$ for $i \in 1, \dots, n$ be the autocorrelation matrix of the unbiased sampling scheme picking $\boldsymbol{y}_{j_i}^t \in \mathcal{T}_{j_i}$ using $\bar{\boldsymbol{x}}_{j_i}$. The autocorrelation matrix of the sampling scheme picking $\boldsymbol{y}_k^t \in \mathcal{T}$ is*

$$\boldsymbol{C}_k^t = \begin{pmatrix} \boldsymbol{C}_{j_1}^t & \bar{\boldsymbol{x}}_{j_1}^t(\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \bar{\boldsymbol{x}}_{j_1}^t(\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \bar{\boldsymbol{x}}_{j_2}^t(\bar{\boldsymbol{x}}_{j_1}^t)^\top & \boldsymbol{C}_{j_2}^t & \cdots & \bar{\boldsymbol{x}}_{j_2}^t(\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t(\bar{\boldsymbol{x}}_{j_1}^t)^\top & \bar{\boldsymbol{x}}_{j_n}^t(\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t \end{pmatrix}.$$

Lemma 12. *Let $v \in \mathcal{J}$ be a terminal decision point, and let $\bar{\boldsymbol{x}}_v^t = (\lambda_1^t, \dots, \lambda_n^t) \in \Delta^n = \operatorname{co} \mathcal{T}_v$ in accordance with* (12)*. The autocorrelation matrix of the sampling scheme that picks $\boldsymbol{y}_j^t \in \mathcal{T}_v$ is*

$$\boldsymbol{C}_j^t = \begin{pmatrix} \lambda_1^t & & \\ & \ddots & \\ & & \lambda_n^t \end{pmatrix}.$$

Proof. It follows from the definition of the sampling scheme that

$$\boldsymbol{C}_{t,j} = \mathbb{E}_t[\boldsymbol{y}_j^t(\boldsymbol{y}_j^t)^\top] = \sum_{i=1}^n \lambda_i(\boldsymbol{e}_i \boldsymbol{e}_i^\top) = \begin{pmatrix} \lambda_1^t & & \\ & \ddots & \\ & & \lambda_n^t \end{pmatrix}.$$

□

Lemma 13. *Let $\mathcal{T}$ be a SDP rooted in non-terminal decision point j. Let $\mathcal{C}_j = \{k_1, \dots, k_n\}$ be the set of observation points that are immediately reachable after j. In accordance with* (14)*, $\bar{\boldsymbol{x}}_j^t$ is in the form $\bar{\boldsymbol{x}}_j^t = (\lambda_1^t, \dots, \lambda_n^t, \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t, \dots, \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t)$, where $(\lambda_1^t, \dots, \lambda_n^t) \in \Delta^n$. Let $\boldsymbol{C}_{k_i}^t$ for $i \in 1, \dots, n$ be the autocorrelation matrix of the unbiased sampling scheme picking $\boldsymbol{y}_{k_i}^t \in \mathcal{T}_{k_i}$ using $\bar{\boldsymbol{x}}_{k_i}/\lambda_i$. The autocorrelation matrix of the sampling scheme picking $\boldsymbol{y}_j^t$ is*

$$\boldsymbol{C}_j^t = \left(\begin{array}{ccc|ccc} \lambda_1^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_1}^t)^\top & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_n}^t)^\top \\ \hline \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t & & & \lambda_1^t \boldsymbol{C}_{k_1}^t & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t & & & \lambda_n^t \boldsymbol{C}_{k_n}^t \end{array}\right).$$

Proof. It follows from the definition of the sampling scheme that

$$\begin{aligned}
\boldsymbol{C}_{t,j} &= \mathbb{E}_t[\boldsymbol{y}_j^t(\boldsymbol{y}_j^t)^\top] \\
&= \sum_{i=1}^{n} \lambda_i \mathbb{E}_t\left[(\boldsymbol{e}_i^\top, \boldsymbol{0}, \ldots, \boldsymbol{0}, (\boldsymbol{y}_{k_i}^t)^\top, \boldsymbol{0}, \ldots, \boldsymbol{0})^\top (\boldsymbol{e}_i^\top, \boldsymbol{0}, \ldots, \boldsymbol{0}, (\boldsymbol{y}_{k_i}^t)^\top, \boldsymbol{0}, \ldots, \boldsymbol{0})\right] \\
&= \left(\begin{array}{ccc|ccc}
\lambda_1^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_1}^t)^\top & & \\
& \ddots & & & \ddots & \\
& & \lambda_n^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_n}^t)^\top \\
\hline
\lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t & & & \lambda_1^t \boldsymbol{C}_{k_1}^t & & \\
& \ddots & & & \ddots & \\
& & \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t & & & \lambda_n^t \boldsymbol{C}_{k_n}^t
\end{array}\right).
\end{aligned}$$

□

Lemma 14. *Let $\mathcal{T}$ be a SDP rooted in observation point k. Let $\mathcal{C}_k = \{j_1, \ldots, j_n\}$ be the set of decision points that are immediately reachable after k. In accordance with* (13)*, $\bar{\boldsymbol{x}}_k^t$ is in the form $\bar{\boldsymbol{x}}_k^t = (\bar{\boldsymbol{x}}_{j_1}^t, \ldots, \bar{\boldsymbol{x}}_{j_n}^t)$. Let $\boldsymbol{C}_{j_i}^t$ for $i \in 1, \ldots, n$ be the autocorrelation matrix of the unbiased sampling scheme picking $\boldsymbol{y}_{j_i}^t \in \mathcal{T}_{j_i}$ using $\bar{\boldsymbol{x}}_{j_i}$. The autocorrelation matrix of the sampling scheme picking $\boldsymbol{y}_k^t \in \mathcal{T}$ is*

$$\boldsymbol{C}_k^t = \begin{pmatrix}
\boldsymbol{C}_{j_1}^t & \bar{\boldsymbol{x}}_{j_1}^t(\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \bar{\boldsymbol{x}}_{j_1}^t(\bar{\boldsymbol{x}}_{j_n}^t)^\top \\
\bar{\boldsymbol{x}}_{j_2}^t(\bar{\boldsymbol{x}}_{j_1}^t)^\top & \boldsymbol{C}_{j_2}^t & \cdots & \bar{\boldsymbol{x}}_{j_2}^t(\bar{\boldsymbol{x}}_{j_n}^t)^\top \\
\vdots & \vdots & \ddots & \vdots \\
\bar{\boldsymbol{x}}_{j_n}^t(\bar{\boldsymbol{x}}_{j_1}^t)^\top & \bar{\boldsymbol{x}}_{j_n}^t(\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t
\end{pmatrix}.$$

Proof. It follows from the definition of the sampling scheme that

$$\begin{aligned}
\boldsymbol{C}_{t,k} &= \mathbb{E}_t[\boldsymbol{y}_k^t(\boldsymbol{y}_k^t)^\top] \\
&= \mathbb{E}_t\left[\begin{pmatrix}\boldsymbol{y}_{j_1}\\ \boldsymbol{y}_{j_2}\\ \vdots \\ \boldsymbol{y}_{j_n}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_{j_1}\\ \boldsymbol{y}_{j_2}\\ \vdots \\ \boldsymbol{y}_{j_n}\end{pmatrix}^\top\right] = \begin{pmatrix}
\mathbb{E}_t[\boldsymbol{y}_{j_1}^t(\boldsymbol{y}_{j_1}^t)^\top] & \mathbb{E}_t[\boldsymbol{y}_{j_1}^t]\mathbb{E}_t[\boldsymbol{y}_{j_2}^t]^\top & \cdots & \mathbb{E}_t[\boldsymbol{y}_{j_1}^t]\mathbb{E}_t[\boldsymbol{y}_{j_n}^t]^\top \\
\mathbb{E}_t[\boldsymbol{y}_{j_2}^t]\mathbb{E}_t[\boldsymbol{y}_{j_1}^t]^\top & \mathbb{E}_t[\boldsymbol{y}_{j_2}^t(\boldsymbol{y}_{j_2}^t)^\top] & \cdots & \mathbb{E}_t[\boldsymbol{y}_{j_2}^t]\mathbb{E}_t[\boldsymbol{y}_{j_n}^t]^\top \\
\vdots & \vdots & \ddots & \vdots \\
\mathbb{E}_t[\boldsymbol{y}_{j_n}^t]\mathbb{E}_t[\boldsymbol{y}_{j_1}^t]^\top & \mathbb{E}_t[\boldsymbol{y}_{j_n}^t]\mathbb{E}_t[\boldsymbol{y}_{j_2}^t]^\top & \cdots & \mathbb{E}_t[\boldsymbol{y}_{j_n}^t(\boldsymbol{y}_{j_n}^t)^\top]
\end{pmatrix} \\
&= \begin{pmatrix}
\boldsymbol{C}_{j_1}^t & \bar{\boldsymbol{x}}_{j_1}^t(\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \bar{\boldsymbol{x}}_{j_1}^t(\bar{\boldsymbol{x}}_{j_n}^t)^\top \\
\bar{\boldsymbol{x}}_{j_2}^t(\bar{\boldsymbol{x}}_{j_1}^t)^\top & \boldsymbol{C}_{j_2}^t & \cdots & \bar{\boldsymbol{x}}_{j_2}^t(\bar{\boldsymbol{x}}_{j_n}^t)^\top \\
\vdots & \vdots & \ddots & \vdots \\
\bar{\boldsymbol{x}}_{j_n}^t(\bar{\boldsymbol{x}}_{j_1}^t)^\top & \bar{\boldsymbol{x}}_{j_n}^t(\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t
\end{pmatrix}.
\end{aligned}$$

□

Generalized Inverse of the Autocorrelation Matrix

Proposition 4. *Let $j \in \mathcal{J}$ be a terminal decision point. The (generalized) inverse of the autocorrelation matrix $\boldsymbol{C}_j^t$ defined in Lemma 12 is*

$$\begin{pmatrix}
1/\lambda_1^t & & \\
& \ddots & \\
& & 1/\lambda_n^t
\end{pmatrix}.$$

The matrix is well defined in virtue of Observation 3.

Proposition 5. *Let $j \in \mathcal{J}$ be a non-terminal decision point, and let $\mathcal{C}_j = \{k_1, \ldots, k_n\}$ be the observation points immediately*

reachable after j. Finally, for all $i = 1, \dots, n$, let $\boldsymbol{C}_{k_i}^{t-}$ be any generalized inverse for $\boldsymbol{C}_{k_i}^t$. The matrix

$$\left(\begin{array}{ccc|c} 0 & & & \\ & \ddots & & \mathbf{0} \\ & & 0 & \\ \hline & & & \frac{1}{\lambda_1^t}\boldsymbol{C}_{k_1}^{t-} & & \\ \mathbf{0} & & & & \ddots & \\ & & & & & \frac{1}{\lambda_n^t}\boldsymbol{C}_{k_n}^{t-} \end{array}\right)$$

is a generalized inverse for the autocorrelation matrix $\boldsymbol{C}_j^t$ defined in Lemma 13. The matrix is well defined in virtue of Observation 3.

Proof. Let $\boldsymbol{C}_j^{t-}$ be the matrix proposed by the statement. Using Lemma 13,

$$\boldsymbol{C}_j^t\,\boldsymbol{C}_j^{t-}\,\boldsymbol{C}_j^t = \boldsymbol{C}_j^t \left(\begin{array}{ccc|c} 0 & & & \\ & \ddots & & \mathbf{0} \\ & & 0 & \\ \hline & & & \frac{1}{\lambda_1^t}\boldsymbol{C}_{k_1}^{t-} & & \\ \mathbf{0} & & & & \ddots & \\ & & & & & \frac{1}{\lambda_n^t}\boldsymbol{C}_{k_n}^{t-} \end{array}\right) \left(\begin{array}{ccc|ccc} \lambda_1^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_1}^t)^\top & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_n}^t)^\top \\ \hline \lambda_1^t\bar{\boldsymbol{x}}_{k_1}^t & & & \lambda_1^t\boldsymbol{C}_{k_1}^t & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t\bar{\boldsymbol{x}}_{k_n}^t & & & \lambda_n^t\boldsymbol{C}_{k_n}^t \end{array}\right)$$

$$= \left(\begin{array}{ccc|ccc} \lambda_1^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_1}^t)^\top & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_n}^t)^\top \\ \hline \lambda_1^t\bar{\boldsymbol{x}}_{k_1}^t & & & \lambda_1^t\boldsymbol{C}_{k_1}^t & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t\bar{\boldsymbol{x}}_{k_n}^t & & & \lambda_n^t\boldsymbol{C}_{k_n}^t \end{array}\right) \left(\begin{array}{ccc|ccc} & & & \mathbf{0} & & \\ & & & & & \\ & & & & & \\ \hline \boldsymbol{C}_{k_1}^{t-}\bar{\boldsymbol{x}}_{k_1}^t & & & \boldsymbol{C}_{k_1}^{t-}\boldsymbol{C}_{k_1}^t & & \\ & \ddots & & & \ddots & \\ & & \boldsymbol{C}_{k_n}^{t-}\bar{\boldsymbol{x}}_{k_n}^t & & & \boldsymbol{C}_{k_n}^{t-}\boldsymbol{C}_{k_n}^t \end{array}\right)$$

$$= \left(\begin{array}{ccc|ccc} \lambda_1^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_1}^t)^\top & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t & & & \lambda_1^t(\bar{\boldsymbol{x}}_{k_n}^t)^\top \\ \hline \lambda_1^t\bar{\boldsymbol{x}}_{k_1}^t & & & \lambda_1^t\boldsymbol{C}_{k_1}^t & & \\ & \ddots & & & \ddots & \\ & & \lambda_n^t\bar{\boldsymbol{x}}_{k_n}^t & & & \lambda_n^t\boldsymbol{C}_{k_n}^t \end{array}\right) = \boldsymbol{C}_j^t,$$

Where the third equality uses Lemmas 10 and 11. This concludes the proof. □

Proposition 6. *Let $k \in \mathcal{K}$ be an observation point, and let $\mathcal{C}_k = \{j_1, \dots, j_n\}$ be the decision points immediately reachable after k. Finally, for all $i = 1, \dots, n$, let $\boldsymbol{C}_{j_i}^{t-}$ be any generalized invers for $\boldsymbol{C}_{j_i}^t$, and let*

$$\boldsymbol{\mu}_k^t := \begin{pmatrix} \boldsymbol{C}_{j_1}^{t-}\,\bar{\boldsymbol{x}}_{j_1}^t \\ \vdots \\ \boldsymbol{C}_{j_n}^{t-}\,\bar{\boldsymbol{x}}_{j_n}^t \end{pmatrix}.$$

The matrix

$$\begin{pmatrix} \boldsymbol{C}_{j_1}^{t-} & & \\ & \ddots & \\ & & \boldsymbol{C}_{j_n}^{t-} \end{pmatrix} - \frac{n-1}{n^2} \cdot \boldsymbol{\mu}_k^t(\boldsymbol{\mu}_k^t)^\top.$$

is a generalized inverse for the autocorrelation matrix $\boldsymbol{C}_k^t$ defined in Lemma 14.

Proof. In order to reduce the notational burden, let

$$\boldsymbol{C}_k^{t\sim} \coloneqq \begin{pmatrix} \boldsymbol{C}_{j_1}^{t-} & & \\ & \ddots & \\ & & \boldsymbol{C}_{j_n}^{t-} \end{pmatrix}.$$

With that, we have

$$\begin{aligned}
\boldsymbol{C}_k^t \, \boldsymbol{C}_k^{t\sim} \, \boldsymbol{C}_k^t &= \begin{pmatrix} \boldsymbol{C}_{j_1}^t & \cdots & \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \ddots & \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{j_1}^{t-} & & \\ & \ddots & \\ & & \boldsymbol{C}_{j_n}^{t-} \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{j_1}^t & \cdots & \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \ddots & \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{C}_{j_1}^t & \cdots & \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \ddots & \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{j_1}^{t-} \boldsymbol{C}_{j_1}^t & \cdots & \boldsymbol{C}_{j_1}^{t-} \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \ddots & \vdots \\ \boldsymbol{C}_{j_n}^{t-} \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^{t-} \boldsymbol{C}_{j_n}^t \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{C}_{j_1}^t + (n-1)\, \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & n\, \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & n\, \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ n\, \bar{\boldsymbol{x}}_{j_2}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & \boldsymbol{C}_{j_2}^t + (n-1)\, \bar{\boldsymbol{x}}_{j_2}^t (\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & n\, \bar{\boldsymbol{x}}_{j_2}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \vdots & \ddots & \vdots \\ n\, \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & n\, \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_2}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t + (n-1)\, \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \end{pmatrix} \\
&= \boldsymbol{C}_k^t + (n-1)\, \bar{\boldsymbol{x}}_k^t (\bar{\boldsymbol{x}}_k^t)^\top, \qquad (28)
\end{aligned}$$

where we repeatedly used Lemmas 10 and 11. At the same time, we have

$$\boldsymbol{C}_k^t \, \boldsymbol{\mu}_k^t = \begin{pmatrix} \boldsymbol{C}_{j_1}^t & \cdots & \bar{\boldsymbol{x}}_{j_1}^t (\bar{\boldsymbol{x}}_{j_n}^t)^\top \\ \vdots & \ddots & \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t (\bar{\boldsymbol{x}}_{j_1}^t)^\top & \cdots & \boldsymbol{C}_{j_n}^t \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{j_1}^{t-} \bar{\boldsymbol{x}}_{j_1}^t \\ \vdots \\ \boldsymbol{C}_{j_n}^{t-} \bar{\boldsymbol{x}}_{j_n}^t \end{pmatrix} = \begin{pmatrix} n\, \bar{\boldsymbol{x}}_{j_1}^t \\ \vdots \\ n\, \bar{\boldsymbol{x}}_{j_n}^t \end{pmatrix} = n\, \bar{\boldsymbol{x}}_k^t, \qquad (29)$$

where again we used Lemmas 10 and 11. Putting (28) and (29) together, we obtain

$$\boldsymbol{C}_k^t \left(\boldsymbol{C}_k^{t\sim} - \frac{n-1}{n^2} \cdot \boldsymbol{\mu}_k^t (\boldsymbol{\mu}_k^t)^\top \right) \boldsymbol{C}_k^t = \boldsymbol{C}_k^t + (n-1)\, \bar{\boldsymbol{x}}_k^t (\bar{\boldsymbol{x}}_k^t)^\top - (n-1)\, \bar{\boldsymbol{x}}_k^t (\bar{\boldsymbol{x}}_k^t)^\top = \boldsymbol{C}_k^t,$$

as we wanted to show. □

Propositions 4 to 6 give a way to compute the generalized inverse $\boldsymbol{C}^{t-}$ needed in (8). It is immediate to see that the vector $\boldsymbol{\mu}_k^t$ in Proposition 6 is orthogonal to $\operatorname{dir} \mathcal{T}_k$ for all $k \in \mathcal{K}$. Hence, we construct the vector $\boldsymbol{b}^t \perp \operatorname{dir} \mathcal{T}$ inductively to cancel the effect of all $\boldsymbol{\mu}_k^t$, as follows (we use the same symbols as Lemmas 12 to 14):

- At all terminal decision points $j \in \mathcal{J}$, we let $\boldsymbol{b}_j^t = \boldsymbol{0}$.
- At all non-terminal decision points $j \in \mathcal{J}$, we let $\boldsymbol{b}_j^t$ be dependent on the action $a \in \{1, \dots, |A_j|\}$ that was selected at j by the pure sequence-form strategy $\boldsymbol{y}^t$. With that, we let

$$\boldsymbol{b}_j^t = \left(\boldsymbol{0}, \dots, \frac{1}{\lambda_a^t} \boldsymbol{b}_{k_a}^t, \boldsymbol{0}, \dots, \boldsymbol{0} \right).$$

- At all observation points $k \in \mathcal{K}$, we let

$$\boldsymbol{b}_k^t = (\boldsymbol{b}_{j_1}^t, \dots, \boldsymbol{b}_{j_n}^t) + \frac{n-1}{n} \boldsymbol{\mu}_k^t.$$

Given the particular choice of generalized inverse $\boldsymbol{C}^{t-}$ and shifting vector $\boldsymbol{b}^t$, we can compute the loss estimate $\tilde{\boldsymbol{\ell}}^t$ according

to Proposition 1 as in Algorithm 3 .

Algorithm 7: LOSSESTIMATE$(v, \bar{\boldsymbol{x}}_v^t, \boldsymbol{y}_v^t)$

Input: $v \in \mathcal{J} \cup \mathcal{K}$,
$\bar{\boldsymbol{x}}_v^t \in \operatorname{co} \mathcal{T}_v$ strategy output by $\tilde{\mathcal{R}}$
$\boldsymbol{y}_v^t \in \mathcal{T}_v$ *pure* strategy output by $\mathcal{R}$

Output: $\tilde{\boldsymbol{\ell}}_v^t$

if $v \in \mathcal{J}$ *is terminal* **then**
  [▷ Let $\bar{\boldsymbol{x}}_v^t = (\lambda_1, \ldots, \lambda_n) \in \Delta^n$ as per (12)
  [▷ Let $\boldsymbol{y}_v^t = \boldsymbol{e}_i$ as per (9)
  [▷ Let $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t$ be the bandit feedback received
  return $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \cdot \left(0, \ldots, 0, \frac{1}{\lambda_i}, 0, \ldots, 0\right)^\top$
else if $v \in \mathcal{J}$ *is non-terminal* **then**
  [▷ Let $\{k_1, ..., k_n\} = \mathcal{C}_v$
  [▷ Let $\bar{\boldsymbol{x}}_v^t = (\lambda_1^t, ..., \lambda_n^t, \lambda_1^t \bar{\boldsymbol{x}}_{k_1}^t, ..., \lambda_n^t \bar{\boldsymbol{x}}_{k_n}^t)$ as per (14)
  [▷ $\boldsymbol{y}_v^t = (\boldsymbol{e}_i, \mathbf{0}, \ldots, \boldsymbol{y}_i, \ldots, \mathbf{0})$ as per (11)
  return $\begin{pmatrix} \mathbf{0} \\ \vdots \\ \frac{1}{\lambda_i^t} \cdot \text{LOSSESTIMATE}(k_i, \bar{\boldsymbol{x}}_{k_i}^t, \boldsymbol{y}_{k_i}^t) \\ \vdots \\ \mathbf{0} \end{pmatrix}$
else if $v \in \mathcal{K}$ **then**
  [▷ Let $\{j_1, ..., j_n\} = \mathcal{C}_v$
  [▷ $\bar{\boldsymbol{x}}_v^t = (\bar{\boldsymbol{x}}_{j_1}^t, \ldots, \bar{\boldsymbol{x}}_{j_n}^t)$ as per (13)
  [▷ $\boldsymbol{y}_v^t = (\boldsymbol{y}_{j_1}^t, \ldots, \boldsymbol{y}_{j_n}^t)$ as per (10)
  return $\begin{pmatrix} \text{LOSSESTIMATE}(j_1, \bar{\boldsymbol{x}}_{j_1}^t, \boldsymbol{y}_{j_1}^t) \\ \vdots \\ \text{LOSSESTIMATE}(j_n, \bar{\boldsymbol{x}}_{j_n}^t, \boldsymbol{y}_{j_n}^t) \end{pmatrix}$

Expected Local Dual Norm of Loss Estimate

Lemma 15. *For each node $v \in \mathcal{J} \cup \mathcal{K}$, let N_v be the number of terminal sequences in the subtree rooted at v. Then, for all $v \in \mathcal{J} \cup \mathcal{K}$,*

$$0 \le (\boldsymbol{b}_v^t)^\top \bar{\boldsymbol{x}}_v^t \le N_v - 1.$$

Proof. By induction on the structure of the SDP.

- **First case:** $v \in \mathcal{J}$ is a terminal decision point. In this case the statement holds trivially since $\boldsymbol{b}_v^t = \mathbf{0}$.

- **Second case:** $v \in \mathcal{K}$ is an observation point. We have

$$\begin{aligned}(\boldsymbol{b}_v^t)^\top \bar{\boldsymbol{x}}_v^t &= \left(\begin{pmatrix} \boldsymbol{b}_{j_1}^t \\ \vdots \\ \boldsymbol{b}_{j_n}^t \end{pmatrix} + \frac{n-1}{n} \begin{pmatrix} \boldsymbol{C}_{j_1}^{t-} \bar{\boldsymbol{x}}_{j_1}^t \\ \vdots \\ \boldsymbol{C}_{j_n}^{t-} \bar{\boldsymbol{x}}_{j_n}^t \end{pmatrix}\right)^\top \begin{pmatrix} \bar{\boldsymbol{x}}_{j_1}^t \\ \vdots \\ \bar{\boldsymbol{x}}_{j_n}^t \end{pmatrix} \\ &= (n-1) + \sum_{i=1}^n (\boldsymbol{b}_{j_i}^t)^\top \bar{\boldsymbol{x}}_{j_i}^t,\end{aligned}$$

where the second equality is an application of Lemma 11. The lower bound is straightforward. For the upper bound, using the

inductive hypothesis we find

$$\begin{aligned}(\boldsymbol{b}_v^t)^\top \bar{\boldsymbol{x}}_v^t &\le (n-1) + \sum_{i=1}^{n}(N_{j_i}-1)\\ &= -1 + \sum_{i=1}^{n} N_{j_i} = N_v - 1.\end{aligned}$$

- **Third case:** $v \in \mathcal{J}$ is a non-terminal decision point. In this case, we have

$$(\boldsymbol{b}_v^t)^\top \bar{\boldsymbol{x}}_v^t = \left(\frac{1}{\lambda_a^t}\boldsymbol{b}_{k_a}^t\right)^\top (\lambda_a^t \bar{\boldsymbol{x}}_{k_a}^t) = (\boldsymbol{b}_{k_a}^t)^\top \bar{\boldsymbol{x}}_{k_a}^t.$$

Both the lower bound and the upper bound follow trivially from applying the inductive hypothesis. □

Theorem 4. *Assume that the bandit information $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \in [0,1]$ at all times t. Then, at all times t, the loss estimate $\tilde{\boldsymbol{\ell}}^t \in \mathbb{R}_{\ge 0}^{|\Sigma|}$ returned by Algorithm 3 satisfies*

$$\mathbb{E}_t\left[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t + \boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\right] \le 2\cdot|\Sigma|^2.$$

Proof. For each node $v \in \mathcal{J}\cup\mathcal{K}$, let N_v be the number of terminal sequences in the subtree rooted at v. We will prove by induction over the tree structure that, for any node $v \in \mathcal{J}\cup\mathcal{K}$,

$$\mathbb{E}_t\left[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t + \boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\right] \le \left(2 - \frac{1}{w_v}\right)N_v^2.$$

That will be enough to conclude, since by hypothesis $(\boldsymbol{\ell}^t)^\top \boldsymbol{y}^t \in [0,1]$, and therefore

$$\begin{aligned}\mathbb{E}_t[\|\tilde{\boldsymbol{\ell}}^t\|_{*,\bar{\boldsymbol{x}}^t}^2] &= \mathbb{E}_t\left[\left\|\left((\boldsymbol{\ell}^t)^\top\boldsymbol{y}^t\right)\cdot\left(\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t + \boldsymbol{b}_v^t\right)\right\|_{*,\bar{\boldsymbol{x}}_v^t}^2\right]\\ &= \mathbb{E}_t\left[((\boldsymbol{\ell}^t)^\top\boldsymbol{y}^t)^2\cdot\left\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t + \boldsymbol{b}_v^t\right\|_{*,\bar{\boldsymbol{x}}_v^t}^2\right]\\ &\le \mathbb{E}_t\left[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t + \boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\right].\end{aligned}$$

- **First case**: $v \in \mathcal{J}$ is a terminal decision point. Let $A_v = \{a_1,\dots,a_n\}$. Then $\bar{\boldsymbol{x}}_v^t = (\lambda_1,\dots,\lambda_n) \in \Delta^n$ (Equation (12)) and $\boldsymbol{y}_v^t$ is of the form $\boldsymbol{y}_v^t = \boldsymbol{e}_i$ with probability λ_i. Then, using Proposition 4

$$\mathbb{E}_t\left[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t + \boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\right] = \sum_{i=1}^{n}\lambda_i\cdot\frac{1}{\lambda_i} = n \le \frac{3}{2}N_v^2 = \left(2-\frac{1}{w_j}\right)N_v^2.$$

- **Second case**: $v \in \mathcal{J}$ is a non-terminal decision point. Let $\mathcal{C}_v = \{k_1,\dots,k_n\}$ be the set of observation points that are immediately reachable after v. From Equation (14), $\bar{\boldsymbol{x}}_j^t$ must be in the form $\bar{\boldsymbol{x}}_j^t = (\lambda_1^t,\dots,\lambda_n^t,\lambda_1^t\bar{\boldsymbol{x}}_{k_1}^t,\dots,\lambda_n^t\bar{\boldsymbol{x}}_{k_n}^t)$, where $\boldsymbol{\lambda}^t = (\lambda_1^t,\dots,\lambda_n^t) \in \Delta^n$. Also, from (11), $\boldsymbol{y}_v^t = (\boldsymbol{e}_i, \boldsymbol{0},\dots,\boldsymbol{y}_{k_i}^t,\dots,\boldsymbol{0})$ with probability $\lambda_i^t\cdot\pi(\boldsymbol{y}_{k_i}^t)$ where π is a distribution

over $\mathcal{T}_{k_i}$. Hence,

$$\begin{aligned}
\mathbb{E}_t\Big[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t+\boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\Big] &= \sum_{i=1}^n \lambda_i^t\,\pi(\boldsymbol{y}_{k_i}^t)\cdot\left\|\begin{pmatrix}\mathbf{0}\\ \vdots\\ \frac{1}{\lambda_i^t}\big(\boldsymbol{C}_{k_i}^{t-}\boldsymbol{y}_{k_i}^t+\boldsymbol{b}_{k_i}^t\big)\\ \vdots\\ \mathbf{0}\end{pmatrix}\right\|_{*,\bar{\boldsymbol{x}}_v^t}^2\\
&= \sum_{i=1}^n\sum_{\boldsymbol{y}_{k_i}^t\in\mathcal{T}_{k_i}}\lambda_i^t\,\pi(\boldsymbol{y}_{k_i}^t)\cdot\frac{1}{(\lambda_i^t)^2}\cdot\left(\lambda_i^t\,\|\boldsymbol{C}_{k_i}^{t-}\boldsymbol{y}_{k_i}^t+\boldsymbol{b}_{k_i}^t\|_{*,\bar{\boldsymbol{x}}_{k_i}^t}^2+\frac{\lambda_i^t}{w_j}[(\boldsymbol{C}_{k_i}^{t-}\boldsymbol{y}_{k_i}^t+\boldsymbol{b}_{k_i}^t)^\top\bar{\boldsymbol{x}}_{k_i}^t]^2\right)\\
&= \sum_{i=1}^n\left(\sum_{\boldsymbol{y}_{k_i}^t\in\mathcal{T}_{k_i}}\pi(\boldsymbol{y}_{k_i}^t)\cdot\|\boldsymbol{C}_{k_i}^{t-}\boldsymbol{y}_{k_i}^t+\boldsymbol{b}_{k_i}^t\|_{*,\bar{\boldsymbol{x}}_{k_i}^t}^2\right)+\frac{1}{w_j}\sum_{i=1}^n\big(1+(\boldsymbol{b}_{k_i}^t)^\top\bar{\boldsymbol{x}}_{k_i}^t\big)^2\\
&= \sum_{i=1}^n\mathbb{E}_t\Big[\|\boldsymbol{C}_{k_i}^{t-}\boldsymbol{y}_{k_i}^t+\boldsymbol{b}_{k_i}^t\|_{*,\bar{\boldsymbol{x}}_{k_i}^t}^2\Big]+\frac{1}{w_j}\sum_{i=1}^n\big(1+(\boldsymbol{b}_{k_i}^t)^\top\bar{\boldsymbol{x}}_{k_i}^t\big)^2,
\end{aligned}$$

where we used Corollary 1 in the second equality and Lemma 11 in the third equality. Using Lemma 15, we obtain

$$\begin{aligned}
\mathbb{E}_t\Big[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t+\boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\Big] &= \sum_{i=1}^n\mathbb{E}_t\Big[\|\boldsymbol{C}_{k_i}^{t-}\boldsymbol{y}_{k_i}^t+\boldsymbol{b}_{k_i}^t\|_{*,\bar{\boldsymbol{x}}_{k_i}^t}^2\Big]+\frac{1}{w_j}\sum_{i=1}^n N_{k_i}^2\\
&\le \sum_{i=1}^n\left(2-\frac{1}{w_{k_i}}+\frac{1}{w_j}\right)N_{k_i}^2,
\end{aligned}$$

where the inequality follows from applying the inductive hypothesis. By definition of the DGF weights (2), we have $w_j\ge 2w_{k_i}$ for all $i=1,\dots,n$. Hence, $-1/w_{k_i}\le -2/(w_j)$ and therefore:

$$\begin{aligned}
\mathbb{E}_t\Big[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t+\boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\Big] &\le \sum_{i=1}^n\left(2-\frac{2}{w_j}+\frac{1}{w_j}\right)N_{k_i}^2\\
&= \left(2-\frac{1}{w_j}\right)\sum_{i=1}^n N_{k_i}^2\\
&\le \left(2-\frac{1}{w_j}\right)\left(\sum_{i=1}^n N_{k_i}\right)^2\\
&\le \left(2-\frac{1}{w_j}\right)N_j^2.
\end{aligned}$$

- **Third case:** $v\in\mathcal{K}$ is an observation point. Let $\mathcal{C}_v=\{j_1,\dots,j_n\}$ be the set of decision points that are immediately reachable after v. We have

$$\begin{aligned}
\mathbb{E}_t\Big[\|\boldsymbol{C}_v^{t-}\boldsymbol{y}_v^t+\boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_v^t}^2\Big] &= \sum_{i=1}^n\mathbb{E}_t\Big[\|\boldsymbol{C}_{j_i}^{t-}\boldsymbol{y}_{j_i}^t+\boldsymbol{b}_v^t\|_{*,\bar{\boldsymbol{x}}_{j_i}^t}^2\Big]\\
&\le \sum_{i=1}^n\left(2-\frac{1}{w_{j_i}}\right)N_{j_i}^2\\
&\le \left(2-\frac{1}{w_v}\right)\sum_{i=1}^n N_{j_i}^2\le\left(2-\frac{1}{w_v}\right)N_v^2.
\end{aligned}$$

□