Local and Adaptive Mirror Descents in Extensive-Form Games

Côme Fiegel
CREST - FairPlay, ENSAE Paris
Palaiseau, France
come.fiege@normalesup.org

Pierre Ménard
ENS Lyon
Lyon, France

Tadashi Kozuno
OMRON SINIC X
Tokyo, Japan

Rémi Munos
Google DeepMind
Paris, France

Vianney Perchet
CREST - FairPlay, ENSAE Paris, Criteo AI Lab
Paris, France

Michal Valko
INRIA

Abstract

We study how to learn ε-optimal strategies in zero-sum imperfect information games (IIG) with *trajectory feedback*. In this setting, players update their policies sequentially, based on their observations over a fixed number of episodes denoted by T. As noted by Steinberger et al. (2020) and McAleer et al. (2022), most existing procedures suffer from high variance due to the use of importance sampling over sequences of actions. To reduce this variance, we consider a *fixed sampling* approach, where players still update their policies over time, but with observations obtained through a given fixed sampling policy. Our approach is based on an adaptive Online Mirror Descent (OMD) algorithm that applies OMD locally to each information set, using individually decreasing learning rates and a *regularized loss*. We show that this approach guarantees a convergence rate of $\tilde{\mathcal{O}}(T^{-1/2})$ with high probability and has a near-optimal dependence on the game parameters when applied with the best theoretical choices of learning rates and sampling policies. To achieve these results, we generalize the notion of OMD stabilization, allowing for time-varying regularization with convex increments.

1 Introduction

The extensive-form representation of a game (Osborne & Rubinstein, 1994) can be depicted as a tree whose nodes correspond to the game states. At each state, the players choose some available actions and, based on these choices, the game transitions to the next state among the current state's children.

In imperfect information games (IIGs), players may only have access to partial information about the current game state upon taking action. Therefore, the state space is partitioned for each player into multiple information sets, which consist of indistinguishable states from the player's perspective. With perfect recall (Kuhn, 1950), when players remember their previous moves, each space of information sets also has a tree structure.

We focus more specifically on zero-sum IIGs represented in an extensive form under the perfect recall assumption, where the gains of one player, conventionally called the max-player, are equal to the losses of his opponent, the min-player. The primary goal is to design an algorithm learning ε-optimal strategies (von Neumann, 1928). To achieve this, one can use the self-play framework, where an agent controls both players for T episodes. At the beginning of each episode, the agent prescribes a strategy for each player. The agent then observes the play and updates the players' strategies for the next episode based on the outcome of the game. After T episodes, this protocol returns a guess of

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

https://doi.org/10.52202/079017-1806

strategies with a small exploitability gap (Ponsen et al., 2011). In this learning framework, the agent has very limited feedback, only observing the rewards along each sampled trajectory, as opposed to richer feedback that would for example include all possible rewards and all transition probabilities, (Zinkevich et al., 2007; Hoda et al., 2010; Tammelin, 2014; Kroer et al., 2015; Burch et al., 2019) unrealistic in large games.

To deal with this learning framework, a well-studied approach is to unilaterally minimize the regret of each player during the interactions with the game, i.e. the difference between the cumulative gain the player would have obtained had he played the best fixed a posteriori policy and the cumulative gain obtained by following the sequence of policies. The key observation is that by minimizing the regret of both players, the average policies over the sequence of policies generated during the process converge toward optimal strategies at the rate of order $\mathcal{O}(1/\sqrt{T})$ (Cesa-Bianchi & Lugosi, 2006; Kozuno et al., 2021). Regret minimizers such as CFR-based algorithm or online mirror descent (OMD) (Hoda et al., 2010; Kroer et al., 2015) can be used, leading to optimal rates (with respect to the game size) with the latter option (Bai et al., 2022; Fiegel et al., 2023).

Since the agent only observes trajectories of the game, an importance sampling estimate (Auer et al., 2003) of gain (or loss) is fed to the regret minimizer. However, the estimate of this loss usually suffers from high variance due to two reasons. First, the same sequence of policies is used to minimize the regret and to collect the trajectories, making the players strive to fulfill two competing goals: play a policy with small regret and play a policy leading to a small variance gain estimate. Second, importance sampling is applied to sequences of actions, that have in large games a very small probability of being played, leading to empirically large importance sampling weights and ultimately inflating the variance of the gain estimates.

To mitigate this issue, regularization and biasing the estimates can help (Kozuno et al., 2021; Bai et al., 2020). However, the high variance of the gain estimates remains problematic with large games, for which the algorithms are generally coupled with function approximation (Steinberger et al., 2020; McAleer et al., 2022). For instance, neural networks are particularly susceptible to noise (Zhang et al., 2021). A natural question is thus whether it is possible to learn optimal strategies without relying on importance-sampling over the sequence of actions.

To this aim, we consider a particular case of the self-play framework: the fixed policy sampling framework (Lanctot et al., 2009). In this setting, a fixed policy is used to collect the trajectories of the game. Precisely, at each round, one player, let's say the min-player, follows the fixed sampling policy to play against the current policy of the max-player. The collected trajectory is then used to update the current policy of the min-player. In the next episode, the max-player will follow a sampling policy against the current policy of the min-player, and so on. The outcome sampling MCCFR algorithm adopts this framework to update the two players' policy by regret minimization, feeding the CFR algorithm with gain estimated via importance sampling (Lanctot et al., 2009; Bai et al., 2020; Farina et al., 2021b).

Recently, McAleer et al. (2022) proposed the ESCHER algorithm that removes the need for importance sampling in this framework. In particular, as the CFR algorithm is invariant by re-scaling of the gains and the weights of the sampling policy are fixed, ESCHER can directly operate with the unweighted history cumulative gain (Bai et al., 2020). Unfortunately, it still requires access to an oracle that provides this history of cumulative gains at an arbitrary information set.

Nonetheless, the insight of McAleer et al. (2022) cannot be used directly for OMD-based algorithms as they are not scale-invariant. Furthermore, the OMD-based algorithms generally work at the global game level whereas CFR-based algorithms work at the local level of the information set (Bai et al., 2020), making local adaptation to the problem easier.

Contributions We make the following main contributions:

- We propose the LocalOMD algorithm, in the fixed policy sampling framework, that allows adaptive learning rates and does not require importance-sampling over the sequence of actions but only for the current action. We explain how it can simply be seen as a regret minimization procedure applied to a local loss on each information set, similarly to Farina et al. (2019b).

https://doi.org/10.52202/079017-1806

- We prove that LocalOMD achieves, in this fixed sampling framework, a $\widetilde{\mathcal{O}}\left(1/\varepsilon^2\right)$[1] sample complexity with *any* choice of non-degenerate sampling policy, ignoring the game and policy-dependent parameters.
- With an appropriate sampling policy and choice of learning rates, we prove that LocalOMD, recover the $\widetilde{\mathcal{O}}\left(H^3(A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2\right)$ near-optimal sample complexity for learning ε-optimal strategies in a tabular setting, where H is the height of the tree, $A_{\mathcal{X}}$ the total number of available actions for the min-player and $B_{\mathcal{Y}}$ the same quantity for the max-player. This sample complexity was also achieved in the fixed policy framework by BalancedCFR (Bai et al., 2022), but with a less generalizable procedure that updates the policy at one depth at a time.
- We generalize the dual-stabilization technique introduced by Fang et al. (2020) to analyze OMD with a time-varying regularization as long as the increments of the regularization are convex.
- Our tabular experiments reveal that our algorithm yields comparable results to existing baselines while demonstrating a reduced variance in loss estimation.

2 Settings and fixed sampling procedure

2.1 Extensive-form games and regret

Game definition We consider a finite zero-sum IIG game $(\mathcal{S},\mathcal{X},\mathcal{Y},\mathcal{A},\mathcal{B},p,\ell)$ with perfect recall. Given two behavioral policies $\mu=(\mu(\cdot|x))_{x\in\mathcal{X}}$ and $\nu=(\nu(\cdot|y))_{y\in\mathcal{Y}}$, one episode of such game proceeds as follows: An initial game state $s_1\sim p(\cdot|s_0)$ is first sampled in the set of states $\mathcal{S}$ according to the transition function p, starting from the root s_0 of the tree. At depth h, the min- and max-players respectively observe the information set x_h and y_h associated with the current state s_h in the spaces of information sets $\mathcal{X}$ and $\mathcal{Y}$ (these spaces being two partitions of $\mathcal{S}$), then simultaneously choose and execute actions $a_h\sim\mu(\cdot|x_h)$ and $b_h\sim\nu(\cdot|y_h)$ in the sets of legal actions $\mathcal{A}(x_h)$ and $\mathcal{B}(y_h)$. As a result, the state transitions to a new state $s_{h+1}\sim p(\cdot|s_h,a_h,b_h)$ in $\mathcal{S}$, with the min- and max- players getting respectively the losses $\ell_h\sim\ell(\cdot|s_h,a_h,b_h)$ in $[0,1]$ and $1-\ell_h$ according to the loss distribution ℓ. This is repeated until a final state s_H of a fixed depth H is reached, after which the episode finishes.

Policies and actions We will denote by Π_{min} and Π_{max} the set of behavioral policies of the min- and max- players. Because of the perfect recall assumption, such policies, with an independent stochastic choice of action for each information set, are enough to describe the entire set of strategies (Laraki et al., 2019). We will also denote by $A_{\mathcal{X}}$ and $B_{\mathcal{Y}}$ the total number of actions for respectively the min- and max- players, i.e. $A_{\mathcal{X}}:=\sum_{x\in\mathcal{X}}|\mathcal{A}(x)|$ and $B_{\mathcal{Y}}=\sum_{y\in\mathcal{Y}}|\mathcal{B}(y)|$.

Regret and ε-optimal strategies We are interested in learning ε-optimal policies through self-play over multiple episodes. A useful notion for this objective is the regret as explained in the introduction. We first define the value $V^{\mu,\nu}=\mathbb{E}^{\mu,\nu}[\sum_{h=1}^{H}\ell_h]$ as the expected sum of losses (for the min-player) with respect to a pair of policies $(\mu,\nu)\in\Pi_{\text{min}}\times\Pi_{\text{max}}$. Given a sequence $(\mu^t,\nu^t)_{t\in[T]}$ in $\Pi_{\text{min}}\times\Pi_{\text{max}}$, the regrets of the min- and max- players are then defined by

$$\mathfrak{R}^T_{\text{min}}:=\max_{\mu^\dagger\in\Pi_{\text{min}}}\sum_{t=1}^{T}(V^{\mu^t,\nu^t}-V^{\mu^\dagger,\nu^t})\quad\text{and}\quad\mathfrak{R}^T_{\text{max}}:=\max_{\nu^\dagger\in\Pi_{\text{max}}}\sum_{t=1}^{T}(V^{\mu^t,\nu^\dagger}-V^{\mu^t,\nu^t}).$$

Minimizing the regret of both players leads to the computation of an ε-optimal profile (equivalent to an ε-Nash equilibrium for two players zero-sum games) through the computation of an average of the policies. The following theorem quantifies this statement under the perfect recall assumption.

Theorem 2.1. *(Cesa-Bianchi & Lugosi, 2006; Kozuno et al., 2021) From a sequence $(\mu^t,\nu^t)_{t\in[T]}$ in $\Pi_{\text{min}}\times\Pi_{\text{max}}$ a certain time-averaged profile $(\overline{\mu},\overline{\nu})$ is ε-optimal with $\varepsilon=\left(\mathfrak{R}^T_{\text{min}}+\mathfrak{R}^T_{\text{max}}\right)/T$.*

It especially shows that both averaged strategies converge to the set of optimal strategies as long as the regret of both players is sub-linear.

We now focus on the min-player point of view because of the symmetry of the game. Indeed, the following ideas will apply exactly the same way to the max-player, using the losses $1-\ell_h$ instead.

[1] For algorithms with a probability at least $1-\delta$ of a correct output, the symbol $\widetilde{\mathcal{O}}$ hides dependencies logarithmic in $A_{\mathcal{X}},B_{\mathcal{Y}}$ and δ

https://doi.org/10.52202/079017-1806

Algorithm 1 Learning procedures with fixed sampling policies for two players

```
Input: Fixed sampling policies $\mu^s$ and $\nu^s$. Initial policies $\mu^1$ and $\nu^1$ and update procedure for each player
For $t = 1$ to $T$
    The min-player observes the full outcome of an episode with the policies $(\mu^s, \nu^t)$
    The max-player observes the full outcome of an episode with the policies $(\mu^t, \nu^s)$
    The min- and max-player respectively update $\mu^{t+1}$ and $\nu^{t+1}$ based on their past observations
Output: The time-averaged policies $\overline{\mu}, \overline{\nu}$ of Theorem 2.1
```

Perfect recall and realization plan Under the perfect recall assumption, players do not forget their past observations and actions. We can then assume, for any information set $x \in \mathcal{X}$ and action $a \in \mathcal{A}(x)$, the existence of a unique depth $h \in [H]$ and history $(x_1, a_1, ..., x_h, a_h)$ such that $x_h = x$ and $a_h = a$. Using this unique history, we define the realization plan $\mu_{1:} \in \mathbb{R}^{A_{\mathcal{X}}}$ (von Stengel, 1996) associated to a policy $\mu \in \Pi_{\min}$ with, for any $x \in \mathcal{X}$ and $a \in \mathcal{A}(x)$ by $\mu_{1:}(x, a) := \Pi_{i=1}^h \mu(a_i|x_i)$. It denotes the combined probability of choosing actions that lead to (x, a). We will especially define $Q_{\max} := \{\mu_{1:}, \mu \in \Pi_{\min}\}$ the treeplex, i.e. the set of all possible realization plans.

Loss and regret linearization For ν a max-player policy, the unique history also allows for the definition of adversarial transitions $p^\nu_{1:} \in \mathbb{R}^{\mathcal{X}}$ and adversarial losses $\ell^\nu \in \mathbb{R}^{A_{\mathcal{X}}}$ with:

$$p^\nu_{1:}(x) := p(x_1|s_0) \prod_{i=2}^{h} p^\nu(x_i|x_{i-1}, a_{i-1}) \quad \text{and} \quad \ell^\nu(x, a) := p^\nu_{1:}(x)\ell^\nu_h(x, a)$$

where $p(x_1|s_0)$ is the probability that x_1 is initially observed by the min-player, and, assuming that the max-player policy is set to ν, $p^\nu(\cdot|(x_{i-1}, a_{i-1}))$ denotes the probability of transitioning to x_i when (x_{i-1}, a_{i-1}) is reached, and ℓ^ν_h the average loss ℓ_h associated to a when x is reached. Similarly to the realization plan, the adversarial transitions denote the combined probability of both Nature and max-player actions that lead to x, assuming that the min-player plays the actions $(a_1, ..., a_{h-1})$.

Using a chain-rule argument, we get the relation $V^{\mu,\nu} = \langle \ell^\nu, \mu_{1:} \rangle$, given a pair of policies $(\mu, \nu) \in \Pi_{\min} \times \Pi_{\max}$, where $\langle \cdot, \cdot \rangle$ is the standard inner product of $\mathbb{R}^{A_{\mathcal{X}}}$, defined by $\langle z_1, z_2 \rangle := \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}(x)} z_1(x, a) z_2(x, a)$. The regret can then be rewritten

$$\mathfrak{R}^T_{\min} = \max_{\mu^\dagger \in \Pi_{\min}} \sum_{t=1}^{T} \left\langle \ell^t, \mu^t_{1:} - \mu^\dagger_{1:} \right\rangle$$

where $\ell^t := \ell^{\nu^t}$, which effectively reduces the problem to a linear regret problem over the convex polytope $Q_{\min}$ of realization plans.

Several techniques exist to sequentially choose policies $(\mu^t)_{t \in [T]}$ minimizing $\mathfrak{R}^T_{\min}$, assuming that the losses ℓ^t are observed after each round t (Hoda et al., 2010). However, in the *trajectory feedback* setting, these losses are not observed, and can only be estimated from the observation of the trajectories $(x^t_1, a^t_1, ..., x^t_H, a^t_H)$ and partial losses $(\ell^t_1, ..., \ell^t_H)$ of each round.

2.2 Fixed sampling policy

In the *fixed sampling* framework (Lanctot et al., 2009), both players always use the same policy for the observations of the trajectory. However, the two observations can not be done simultaneously with such an approach, as the learning would then be quite naive. The solution, summarized in Algorithm 1, is for the two players to take turns between an observation phase, in which they play their fixed sampling policy μ^s or ν^s, and an interaction phase, in which they play their updated policy μ^t or ν^t. The underlying idea is that the observation phase lets each player observe how the game unfolds against the opponent in its interaction phase, playing its updated policy. Given upper-bounds of the regrets $\mathfrak{R}^T_{\min}$ and $\mathfrak{R}^T_{\max}$ associated to the sequence $(\mu^t, \nu^t)_{t \in [T]}$, the previous Theorem 2.1 then characterizes the optimality of the outputted time-averaged profile $(\overline{\mu}, \overline{\nu})$.

While theoretically optimal algorithms already exist using simultaneous regret minimization procedures (Bai et al., 2022; Fiegel et al., 2023), this framework allows for the removal of the global importance sampling term of the loss, which reduces the variance to make algorithms more suitable beyond the tabular setting (McAleer et al., 2022). Indeed, as the probability of choosing a sequence of action reaching a given information set is fixed, the average estimations of the losses do not need to be re-weighted based on the inverse of a changing probability. This re-weighting eventually leads to unstable function approximation, e.g. with neural networks, as this probability can be very small.

Furthermore, the fixed sampling framework also allows aggressive policies more focused on exploitation, as the observation side is handled by the sampling strategy. The downside is that this sampling policy must be fixed in advance, which requires defining a good sampling policy beforehand.

From now on, we again focus on the min-player for the same symmetry reasons.

Estimated regret Based on the min-player observations, we define $\hat{\mathfrak{R}}^T_{\text{min}}$ the estimated regret by

$$\hat{\mathfrak{R}}^T_{\text{min}} := \max_{\mu^\dagger \in \Pi_{\min}} \sum_{t=1}^{T} \left\langle \widehat{\ell}^t,\, \mu^t_{1:} - \mu^\dagger_{1:} \right\rangle$$

where the $\widehat{\ell}^t$ are the importance-sampling estimated loss vectors, defined for each information set x of depth h and action $a \in \mathcal{A}(x)$ by

$$\widehat{\ell}^t(x,a) := \frac{\mathbb{I}_{\left\{x = x^t_h, a = a^t_h\right\}}}{\mu^s_{1:}(x,a)} \ell^t_h$$

with x^t_h the visited information set, a^t_h the chosen action and ℓ^t_h the loss at depth h of episode t.

The following theorem states that upper-bounding this estimated regret is enough to upper-bound the actual regret, up to an additional additive term. Its proof is given in Appendix B and relies on Bernstein-type inequalities.

Theorem 2.2. *Assume that the estimated losses are obtained with a fixed positive sampling policy* μ^s *as above. Then, for any sequence* $(\mu^t)_{t\in[T]}$ *of* $\Pi_{\min}$ *and any* $\delta \in (0,1)$*, the following bound holds with a probability at least* $1-\delta$

$$\mathfrak{R}^T_{min} \leq \max\left\{\hat{\mathfrak{R}}^T_{min}, 0\right\} + 4\sqrt{\iota H \kappa(\mu^s) T}$$

where $\iota := \log\left(\frac{A_{\mathcal{X}}+1}{\delta}\right)$ *and* $\kappa(\mu^s) := \max_{\mu \in \Pi_{\min}} \sum_{x\in\mathcal{X}} \sum_{a \in \mathcal{A}_x} \frac{\mu_{1:}(x,a)}{\mu^s_{1:}(x,a)}$.

A similar proposition is proved by Farina et al. (2020). Our bound is specific to the importance-sampling loss estimator, but tighter by a factor $\sqrt{\kappa(\mu^s)/H}$.

Remark 2.3. The quantity $\kappa(\mu^s)$ can be efficiently computed recursively for each of the sub-trees induced by an information set $x \in \mathcal{X}$, and we will denote by $\kappa(\mu^s|x)$ the associated quantities. The same recursion shows that the *balanced policy* $\mu^\star$, which plays proportionally to the total number of actions of each sub-tree, minimizes all these local quantities and satisfies $\kappa(\mu^\star) = A_{\mathcal{X}}$. The related computations are provided in Appendix C.

3 Adaptive Mirror Descent

We shall now focus on the update procedure the min-player can use to minimize this estimated regret. Let us first define some important notions of convex optimization.

Definition 3.1. Let $\Omega \subset \mathbb{R}^n$ be a non-empty open convex, and $\overline{\Omega}$ be its closure. A function $\Psi : \overline{\Omega} \to \mathbb{R}$ is said to be Legendre if Ψ is strictly convex, continuously differentiable on Ω and $\forall y \in \overline{\Omega}\backslash\Omega,\ \lim_{x \to y} \|\nabla\Psi(x)\| = +\infty$. The Bregman divergence $\mathbf{D}_\Psi : \overline{\Omega} \times \Omega \to \mathbb{R}$ of a Legendre function Ψ is defined as $\mathbf{D}_\Psi(x,y) := \Psi(x) - \Psi(y) - \langle \nabla\Psi(y), x - y\rangle$. The Fenchel conjugate $\Psi^\star : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of Ψ is defined by $\Psi^\star(\xi) = \sup_{x\in\overline{\Omega}} \langle \xi, x\rangle - \Psi(x)$.

3.1 Online Mirror Descent and dilated entropy

In an extensive-form game with perfect recall, algorithms based on the Online Mirror Descent (OMD) typically compute at each time step t the update

$$\mu^{t+1} = \operatorname*{arg\,min}_{\mu \in \Pi_{\min}} \left\langle \widehat{\ell}^t, \mu_{1:} \right\rangle + \mathbf{D}_\Psi(\mu_{1:}, \mu^t_{1:}) \qquad \text{(OMD)}$$

https://doi.org/10.52202/079017-1806

where $\widehat{\ell}^t$ is the estimated loss and $\Psi : Q_{\min} \to \mathbb{R}$ a Legendre regularizer.

Dilated entropy A common choice of such regularizer is the dilated entropy (Hoda et al., 2010; Kroer et al., 2015). It requires for each $x \in \mathcal{X}$ a Legendre regularizer Ψ_x over a convex domain $\overline{\Omega_x} \subset \mathbb{R}_{\geq 0}^{|\mathcal{A}(x)|}$ that contains the simplex $\Delta_{\mathcal{A}(x)} := \left\{\mu,\ \sum_{a\in\mathcal{A}(x)} \mu(a) = 1\right\}$. For a given list of positive weights $\alpha = (\alpha(x))_{x\in\mathcal{X}}$, the dilated entropy $\Psi_\alpha^{\mathrm{dil}}$ satisfies for any $\mu \in \Pi_{\min}$:

$$\Psi_\alpha^{\mathrm{dil}}(\mu_{1:}) := \sum_{x\in\mathcal{X}} \alpha(x)\mu_{1:}(x)\Psi_x\left(\mu(\cdot|x)\right)$$

where $\mu_{1:}(x) := \sum_{a\in\mathcal{A}(x)} \mu_{1:}(x,a)$. Using this dilated entropy as the regularizer, the OMD updates become

$$\mu^{t+1} = \operatorname*{arg\,min}_{\mu\in\Pi_{\min}} \left\langle \widehat{\ell}^t, \mu_{1:} \right\rangle + \mathbf{D}_\alpha^{\mathrm{dil}}(\mu_{1:}, \mu_{1:}^t)$$

where $\mathbf{D}_\alpha^{\mathrm{dil}}(\mu_{1:}, \mu_{1:}^t) := \sum_{x\in\mathcal{X}} \alpha(x)\mu_{1:}(x)\mathbf{D}_x(\mu_{1:}(\cdot|x), \mu_{1:}^t(\cdot|x))$ and $(\mathbf{D}_x)_{x\in\mathcal{X}}$ are the individual Bregman divergences of the $(\Psi_x)_{x\in\mathcal{X}}$. The benefits of this regularization are that it efficiently suits the structure of the game and that the associated updates are easily computed recursively, starting from the final states.

3.2 Stabilized OMD algorithm

The regularizer Ψ sometimes needs to change over time. For example, when T is unknown, a regularizer of the form $\Psi^t = \Psi/\eta^t$ is usually considered, with $\eta^t = t^{-1/2}$ the learning rate. Fiegel et al. (2023) gives another example of time-varying regularization, adapting the regularization to the game structure that is assumed to be initially unknown. The previous updates (OMD) do not however allow adaptive regularization in general. In fact, even the simple learning rate decrease $\eta^{t+1} = t^{-1/2}$ can lead to a linear regret dependence with time (Orabona & Pál, 2018).

In this part, we shall consider more generally a sequence of Legendre regularizers $(\Psi^t)_{t\in[T]}$ defined on a convex domain $\overline{\Omega} \subset \mathbb{R}^n$, and that the player chooses a sequence of primal iterates $(w^t)_{t\in[T]}$ (respectively the updated realization plans $(\mu_{1:}^t)_{t\in[T]}$ of our settings) in a closed convex set $\mathcal{C}$ (respectively the treeplex $Q_{\min}$) included in $\overline{\Omega}$, according to a sequence of dual increments $(\xi^t)_{t\in[T]}$ in $\mathbb{R}^n$ (respectively the estimated losses $(\widehat{\ell}^t)_{t\in[T]}$) observed sequentially.

Fang et al. (2020) proposed in the presence of non-increasing learning rates, to use a technique called dual-stabilization to recover the classical OMD bounds. We noticed that their updates can be interpreted as

$$w^{t+1} = \operatorname*{arg\,min}_{w\in\mathcal{C}} \left\langle \xi^t, w \right\rangle + \mathbf{D}_{\Psi^t}\left(w, w^t\right) + \mathbf{D}_{\Psi^{t+1}-\Psi^t}\left(w, w^1\right) \qquad \text{(GDS-OMD)}$$

with $\Psi^{t+1} - \Psi^t$ incremental functions assumed to be convex, generalizing their special case $\Psi^{t+1} = \Psi/\eta^{t+1}$. The following theorem, proven in Appendix D shows that classical OMD guarantees can be recovered with these updates.

Theorem 3.2. *Let $(w^t)_{t\in[T]}$ be a sequence of primal iterates generated by the updates (GDS-OMD), with convex incremental functions. Then for any $w^\dagger \in \overline{\Omega}$,*

$$\sum_{t=1}^T \left\langle \xi^t, w^t - w^\dagger \right\rangle \leq \mathbf{D}_{\Psi^T}(w^\dagger, w^1) + \sum_{t=1}^T \mathbf{D}_{\Psi^{t,\star}}\left(\nabla\Psi^t(w^t) - \xi^t, \nabla\Psi^t(w^t)\right)$$

where the $(\Psi^{t,\star})_{t\in[T]}$ are the respective Fenchel conjugates of the $(\Psi^t)_{t\in[T]}$.

Compared to the guarantees obtained with previous adaptive procedures, such as `Ada-MD` (Joulani et al., 2017), the first term of the bound is stated with respect to w^1 instead of the sequence $(w^t)_t$, which is important for some $(\Psi^t)_t$ sequences (Orabona & Pál, 2018).

Remark 3.3. `AdaGrad` for stochastic gradient descent (Duchi et al., 2011) is an interesting example of regularizatiom with convex increments (and not only through a decreasing learning rate). It uses the adaptive regularization $\Psi^{t+1} = \|\cdot\|^2_{(G^t)^{1/2}}$, where G^t is a positive semi-definite matrix defined with the gradients g_k by either $G^t = \sum_{k=1}^t g_k g_k^T$ or, more efficiently, by $G^t = \mathrm{Diag}\left(\sum_{k=1}^t g_k g_k^T\right)$.

Algorithm 2 LocalOMD

```
Input:
    Sampling policy $\mu^s \in \Pi_{\min}$ and initial policy $\mu^1 \in \Pi_{\min}$
    Bregman divergences $\mathbf{D}_x$ for each information set $x \in \mathcal{X}$
    Sequences of (possibly adaptive) learning rates $(\eta^t(x))_{t,x}$ for each round $t$ and information set $x$.
For $t = 1$ to $T$
    Observes the outcome of an episode using the fixed strategy $\mu^s$
    $q^t_{H+1} \leftarrow 0$
    For $h = H$ to 1:
        $\widetilde{\ell}^t_h \leftarrow \mathbb{I}_{\{a=a^t_h\}} \left(\ell^t_h + q^t_{h+1}\right) \Big/ \mu^s(a^t_h|x^t_h)$
        $\mu^{t+1}(\cdot|x) \leftarrow \arg\min_{\mu\in\Delta_{\mathcal{A}(x)}} h^t_x(\mu)$
        $q^t_h \leftarrow \min_{\mu\in\Delta_{\mathcal{A}(x)}} h^t_x(\mu)$
    where $h^t_x(\mu) := \left\langle \widetilde{\ell}^t_h, \mu \right\rangle + \frac{1}{\eta^t(x^t_h)} \mathbf{D}_x\left(\mu, \mu^t(\cdot|x^t_h)\right) + \left(\frac{1}{\eta^{t+1}(x^t_h)} - \frac{1}{\eta^{t'+1}(x^t_h)}\right) \mathbf{D}_x\left(\mu, \mu^1(\cdot|x^t_h)\right)$
    and $t'$ is the last round in which $x^t_h$ was visited
    For all non-visited $x \in \mathcal{X}$:
        $\mu^{t+1}(\cdot|x) \leftarrow \mu^t(\cdot|x)$
Output: The time-averaged policy $\overline{\mu}$
```

Adaptive dilatation In the extensive-form game setting based on the dilated entropy $\Psi^{\text{dil}}_{\alpha}$, this stabilization can be applied to have weights $(\alpha^t(x))_{x\in\mathcal{X},t\in[T]}$ that vary with times. The convexity assumption of $\Psi^{\text{dil}}_{\alpha^{t+1}} - \Psi^{\text{dil}}_{\alpha^t}$ then rewrites to having locally non-decreasing weights for each $x \in \mathcal{X}$. In this particular case, the updates are obtained with the formula

$$\mu^{t+1} = \underset{\mu\in\Pi_{\min}}{\arg\min} \left\langle \widehat{\ell}^t, \mu_{1:} \right\rangle + \mathbf{D}^{\text{dil}}_{\alpha^t}(\mu, \mu^t) + \mathbf{D}^{\text{dil}}_{\alpha^{t+1}-\alpha^t}(\mu, \mu^1)\,. \tag{DDS-OMD}$$

4 LocalOMD algorithm

4.1 Algorithm

Let us now consider the fixed sampling framework introduced in Section 2.2. Given a sequence $(\eta^t(x))_{t\in[T]}$ of locally non-increasing learning rates for each $x \in \mathcal{X}$, we introduce the LocalOMD algorithm described in Algorithm 2, that uses the updates (DDS-OMD) above with the adaptive weights $\alpha^t(x) = 1/(\mu^s_{1:}(x)\eta^t(x))$. Dividing the loss by the importance sampling term $1/\mu^s_{1:}(x)$ through the learning rates lets it bypass the large variance that this rate can introduce.

Local loss This algorithm can be interpreted as one that locally applies the updates (GDS-OMD) using the local loss $\widetilde{\ell}^t_h$, a regularized version of the sum of subsequent losses. Even though this algorithm results from a global minimization procedure, the local loss only uses the probability $\mu^s(a|x)$ of choosing the last action $a \in \mathcal{A}(x)$ in the important sampling, instead of the combined probability $\mu^s_{1:}(x, a)$ of the realization plan. A similar decomposition was observed by Farina et al. (2019a) for the non-stochastic settings, in which both players directly observe the gradient associated with their policies.

For this reason, the local loss will consistently be at most of order $\mathcal{O}(HA)$. Meanwhile, the loss used by Fiegel et al. (2023) can be of order $\mathcal{O}(A_{\mathcal{X}})$ (approximately A^H in the worst case), even with IX exploration attempting to alleviate the importance sampling issue. This presents a challenge for potential applications involving function approximation, where $A_{\mathcal{X}}$ becomes very large (McAleer et al., 2022). For instance, such high-variance estimates could lead to highly unstable training dynamics of a policy parametrized with a neural network.

Complexity At each iteration, the algorithm only needs to update the policy along the observed trajectory, so the time complexity per iteration is only H times the cost of a local update. If $\mathbf{D}_x$ is the Kullback-Leibler divergence, the local updates then simplify to some Exponential Weights updates

https://doi.org/10.52202/079017-1806

and the total time complexity of an iteration becomes $\mathcal{O}(HA)$, where A is an upper-bound on the local number of actions.

4.2 Theoretical analysis

The analysis of `LocalOMD`, detailed in Appendix E is derived from Theorem 3.2 that bounds the estimated regret. The results on the real regret are then obtained with Theorem 2.2. We now present two choices of regularization and their associated guarantees.

Adaptive rates As `LocalOMD` treats each information set $x \in \mathcal{X}$ as a separate problem through the local losses $\widetilde{\ell}_h^t$, an interesting choice is to consider the same adaptive rates that would be used in the K-armed bandit problems. The following theorem provides an upper bound in this case.

Theorem 4.1. *(Informal, exact statement in Appendix E)*
For a large class of regularizers $(\Psi_x)_{x\in\mathcal{X}}$ and learning rates $(\eta^t(x)_{x\in\mathcal{X},t\in[T]})$, the regret has a $\mathcal{O}(\sqrt{T\log(1/\delta)})$ upper bound (hiding the game-dependent terms) with a probability at least $1-\delta$. Such learning rates include, for all $x \in \mathcal{X}$ of depth h,

$$\eta^t(x) = \eta \Big/ \sqrt{\sum_{k=1}^{t} \mathbb{I}_{\{x=x_h^k\}}}\ , \quad \textit{or the adaptive version } \eta^t(x) = \eta \Big/ \sqrt{\sum_{k=1}^{t} \mathbb{I}_{\{x=x_h^k\}} \left(\widetilde{\ell}_h^k\right)^2}\ .$$

The adaptive learning rates mentioned for this theorem generally enjoy better performances in practice. Furthermore, they require no initial computation and are easily updated.

Optimal rates The following theorem uses a constant learning rate that locally depends on the $\kappa(\mu^s|x)$ quantities of Remark 2.3, and on the $A := \max_{x\in\mathcal{X}}|\mathcal{A}(x)|$ quantity that upper bounds the local number of available actions on the whole tree.

Theorem 4.2. *Using `LocalOMD` with μ^1 as the uniform policy, with the learning rates $\eta^t(x) = \eta/\kappa(\mu^s|x)$ where $\eta = \sqrt{\log(A)\kappa(\mu^s)/(3HT)}$, and with Ψ_x the Shannon entropy $\Psi_x(\mu) = \sum_{a\in\mathcal{A}(x)} \mu(a)\log(\mu(a))$, we have with a probability at least $1-\delta$ and $\iota = \log(2(A_{\mathcal{X}}+1)/\delta)$,*

$$\mathfrak{R}^T_{\min} \leq \left(4 + 2\sqrt{3}\right) H^{3/2}\sqrt{\log(A)\iota\kappa(\mu^s)T}\,.$$

Note that these rates are not adaptive and thus do not require the stabilization introduced in Section 3.2. When using the balanced policy $\mu^\star$ as the sampling policy, for which $\kappa(\mu^\star) = A_{\mathcal{X}}$, we obtain with Theorem 2.1 the rate $\widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{A_{\mathcal{X}}T}\right)$, near-optimal up to the H dependency (Bai et al., 2022).

5 Experiments

We implemented `LocalOMD`, with the parameters of Theorem 4.1 and Theorem 4.2, then tested it against the theoretically optimal `BalancedCFR` (Bai et al., 2022) using the balanced policy as the sample policy, and `BalancedFTRL` (Fiegel et al., 2023). The algorithms were compared on three standard benchmark games: Kuhn poker (Kuhn, 1950), Leduc poker (Southey et al., 2005) and liars dice, using the version 1.4 of the OpenSpiel library (Lanctot et al., 2019) under the Apache 2.0 license. The learning rates (and the IX parameters for the relevant algorithms) were optimized independently for each algorithm using a grid search. The code is available at `https://github.com/anon5493/LocalOMD-experiments`.

The results are given with respect to the total number of episodes used for learning. This technically disadvantages the fixed sampling algorithms, as these require more than one episode at each round t while still performing a single update on the policy of each player. The exploitability gap, along with the variance across the different instances of the simulation, is plotted in Figure 1, top. Note that this variance across the instances is different from the variance of the estimated loss vector $\widehat{\ell}^t$ our method tries to reduce, which is plotted in the Figure 1, down.

Focusing on the exploitability gap, we observe that the two versions of `LocalOMD` behave similarly and constantly beat `BalancedCFR`, mainly because the latter needs to update each depth with independent samples, thus needing H times more episodes overall. The results of BalancedFTRL are more comparable, exhibiting for example better performances on liars dice but worse on Leduc poker.

In the second figure, we observe that the algorithms based on a fixed sampling procedure indeed get a smaller variance in their loss estimation as the sampling policy stays consistently balanced. BalancedCFR again gets worse results compared to LocalOMD as the losses of each depth are only estimated every H iteration, which increases its variance.

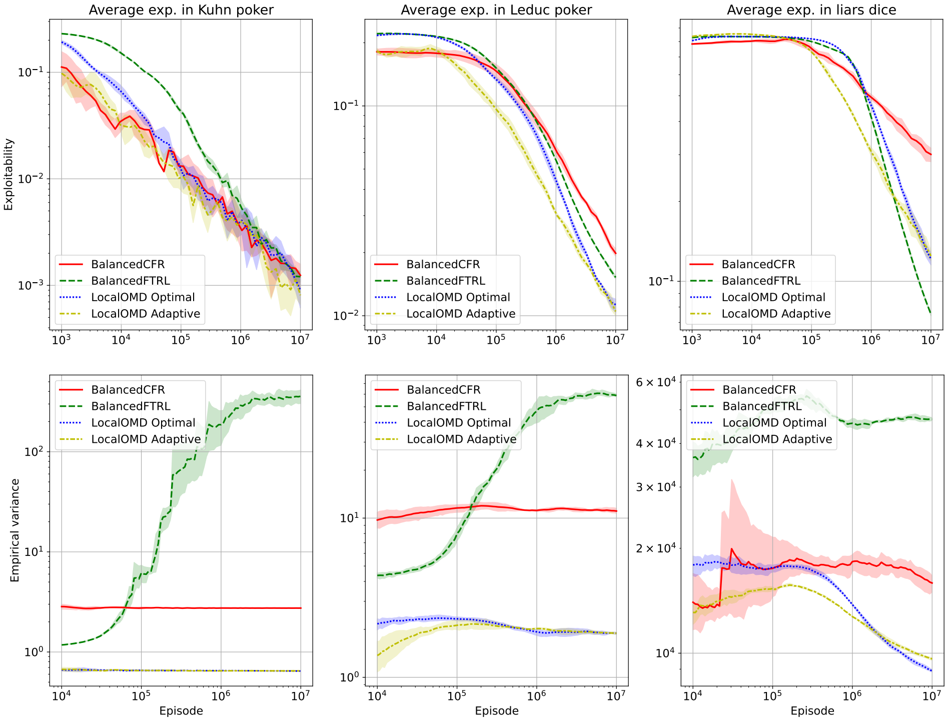

Figure 1: Performances over 5 simulations of various algorithms with respect to the total number of episodes. The vertical axis denotes the exploitability gap $\max_{(\mu,\nu)\in\Pi_{\min}\times\Pi_{\max}} V^{\overline{\mu},\nu} - V^{\mu,\overline{\nu}}$ (top) and the empirical variance of the $\widehat{\ell}^t$ vectors over time (bottom), with all rewards scaled between 0 and 1. The total numbers of actions are $A_{\mathcal{X}} = B_{\mathcal{Y}} = 12$ for Kuhn poker, $A_{\mathcal{X}} = B_{\mathcal{Y}} = 1092$ for Leduc poker, and $A_{\mathcal{X}} = B_{\mathcal{Y}} = 24570$ for Liars dice.

6 Conclusion

We studied the use of a fixed sampling OMD procedure for the computation of ε-optimal strategies. This approach relies, for each player, on an uncoupling between the observation policy and the interaction policy as described in Algorithm 1. This uncoupling is in direct contrast with the more restrictive semi-bandit setting usually considered for self-play, where these two policies must coincide by design. Notice that this is not the standard exploration/exploitation trade-off, as even in the expert setting (with full information), some kind of exploration is still required.

While the balanced observation policy gets the optimal rates in the worst case, it may not always be the best one for a given game. An alternate choice is to instead use for the observations the current average policy (Gibson et al., 2012). This choice can be adapted to the fixed sampling framework, by restarting the algorithm after a certain number of episodes and using the computed average as the new sampling policy.

The proposed algorithm LocalOMD also enjoys simultaneously two interpretations: one as a Mirror Descent type algorithm working at the global level, with a single update performed at each iteration over the whole tree; and one as regret minimizers working locally at each information set, which makes it very similar to a CFR algorithm despite a fundamentally different approach.

We would like to conclude by providing the following interesting research directions.

Problem-dependent optimality For a given game structure and fixed sampling policy μ^s, is there a policy-dependent lower bound $\mathcal{O}(\sqrt{\kappa(\mu^s)T})$ on the regret? We wonder if the $\kappa(\mu^s)$ quantity of Remark 2.3 denotes some sort of complexity related to the problem.

General sum game Using the same techniques as Bai et al. (2022), in a general sum game with potentially more than two players, `LocalOMD` can be shown to converge to an ε-approximate normal-form coarse correlated equilibrium. Are convergences to other forms of correlated equilibrium possible using this fixed sampling policy framework?

On-policy algorithms Is it also possible to remove the importance-sampling of the previous actions in the usual semi-bandit framework that observes with the current policy? The answer is not obvious since the current approach heavily relies on the fact that the sampling policy is fixed.

7 Acknowledgements

This research was supported in part by the French National Research Agency (ANR) in the framework of the PEPR IA FOUNDRY project (ANR-23-PEIA-0003) and through the grant DOOM ANR-23-CE23-0002. It was also funded by the European Union (ERC, Ocean, 101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2159–2170. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/172ef5a94b4dd0aa120c6878fc29f70c-Paper.pdf.

Bai, Y., Jin, C., Mei, S., and Yu, T. Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning*, 2022.

Burch, N., Moravčík, M., and Schmid, M. Revisiting CFR+ and Alternating Updates. *Journal of Artificial Intelligence Research*, 64:429–443, 2019.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006. ISBN 978-0-521-84108-5. doi: 10.1017/CBO9780511546921.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5527–5540. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3b2acfe2e38102074656ed938abf4ac3-Paper.pdf.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Fang, H., Harvey, N., Portella, V., and Friedlander, M. Online mirror descent and dual averaging: Keeping pace in the dynamic case. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3008–3017. PMLR, November 2020.

Farina, G., Kroer, C., and Sandholm, T. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*, 2019a.

https://doi.org/10.52202/079017-1806

Farina, G., Kroer, C., and Sandholm, T. Regret circuits: Composability of regret minimizers. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1863–1872. PMLR, 2019b. URL http://proceedings.mlr.press/v97/farina19b.html.

Farina, G., Kroer, C., and Sandholm, T. Stochastic Regret Minimization in Extensive-Form Games. In *International Conference on Machine Learning*, 2020.

Farina, G., Kroer, C., and Sandholm, T. Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent. In *AAAI Conference on Artificial Intelligence*, 2021a. URL https://arxiv.org/abs/2007.14358.

Farina, G., Kroer, C., and Sandholm, T. Bandit Linear Optimization for Sequential Decision Making and Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*, 2021b.

Fiegel, C., Ménard, P., Kozuno, T., Munos, R., Perchet, V., and Valko, M. Adapting to game trees in zero-sum imperfect information games, 2023.

Gibson, R., Burch, N., Lanctot, M., and Szafron, D. Efficient monte carlo counterfactual regret minimization in games with many player actions. volume 3, 12 2012.

Gordon, G. J. No-regret Algorithms for Online Convex Programs. In *Advances in Neural Information Processing Systems*, 2007.

Hart, S. and Mas-Colell, A. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Hoda, S., Gilpin, A., Peña, J., and Sandholm, T. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research*, 2010. URL https://kilthub.cmu.edu/ndownloader/files/12101699.

Johanson, M., Bard, N., Lanctot, M., Gibson, R., and Bowling, M. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, pp. 837–846, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0981738125.

Joulani, P., György, A., and Szepesvari, C. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In *International Conference on Algorithmic Learning Theory*, 2017. URL https://api.semanticscholar.org/CorpusID:28926102.

Koller, D., Megiddo, N., and Von Stengel, B. Efficient Computation of Equilibria for Extensive Two-Person Games. *Games and Economic Behavior*, 14(2):247–259, 1996.

Kozuno, T., Ménard, P., Munos, R., and Valko, M. Learning in two-player zero-sum partially observable Markov games with perfect recall. In *Neural Information Processing Systems*, 2021.

Kroer, C., Waugh, K., Kilinç-Karzan, F., and Sandholm, T. Faster first-order methods for extensive-form game solving. In *Economics and Computation*, 2015. ISBN 978-1-4503-3410-5. doi: 10.1145/2764468.2764476.

Kroer, C., Farina, G., and Sandholm, T. Solving large sequential games with the excessive gap technique. In *Neural Information Processing Systems*, 2018.

Kroer, C., Waugh, K., Kılınç-Karzan, F., and Sandholm, T. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179(1):385–417, 2020.

Kuhn, H. W. Extensive games. *Proceedings of the National Academy of Sciences*, 36(10):570–576, 1950.

Kuhn, H. W. Extensive Games and the Problem of Information. *Annals of Mathematics Studies*, 28: 193–216, 1953.

Lanctot, M., Waugh, K., Zinkevich, M., and Bowling, M. Monte-Carlo sampling for regret minimization in extensive games. In *Neural Information Processing Systems*, 2009.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., De Vylder, B., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., and Ryan-Davis, J. Openspiel: A framework for reinforcement learning in games, 2019. URL https://arxiv.org/abs/1908.09453.

Laraki, R., Renault, J., and Sorin, S. *Mathematical Foundations of Game Theory*. Springer, October 2019. doi: 10.1007/978-3-030-26646-2. URL https://hal.science/hal-03070434.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Lee, C.-W., Kroer, C., and Luo, H. Last-iterate convergence in extensive-form games. In *Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/file/77bb14f6132ea06dea456584b7d5581e-Paper.pdf.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7001–7010. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21z.html.

McAleer, S., Farina, G., Lanctot, M., and Sandholm, T. ESCHER: Eschewing importance sampling in games by computing a history value function to estimate regret. *CoRR*, abs/2206.04122, 2022. doi: 10.48550/arXiv.2206.04122. URL https://doi.org/10.48550/arXiv.2206.04122.

McMahan, H. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18:1–50, 08 2017.

Munos, R., Pérolat, J., Lespiau, J.-B., Rowland, M., De Vylder, B., Lanctot, M., Timbers, F., Hennes, D., Omidshafiei, S., Gruslys, A., Azar, M. G., Lockhart, E., and Tuyls, K. Fast computation of nash equilibria in imperfect information games. In *International Conference on Machine Learning*, 2020.

Nemirovski, A. Prox-Method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1): 127–152, 2005.

Orabona, F. and Pál, D. Scale-free online learning. *Theor. Comput. Sci.*, 716:50–69, 2018. doi: 10.1016/j.tcs.2017.11.021. URL https://doi.org/10.1016/j.tcs.2017.11.021.

Osborne, M. J. and Rubinstein, A. *A Course in Game Theory*. The MIT Press, 1994. ISBN 0-262-65040-1.

Ponsen, M., De Jong, S., and Lanctot, M. Computing Approximate Nash Equilibria and Robust Best-Responses Using Sampling. *Journal of Artificial Intelligence Research*, 42:575–605, 2011.

Romanovsky, J. V. Reduction of a game with complete memory to a matricial game. *Dokl. Akad. Nauk SSSR*, 144:62–64, 1962.

Schmid, M., Burch, N., Lanctot, M., Moravčík, M., Kadlec, R., and Bowling, M. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. *CoRR*, abs/1809.03057, 2018. URL http://arxiv.org/abs/1809.03057.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2992–3002. PMLR, 2020. URL http://proceedings.mlr.press/v108/sidford20a.html.

Southey, F., Bowling, M., Larson, B., Piccione, C., Burch, N., Billings, D., and Rayner, C. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertaintyin Artificial Intelligence (UAI)*, pp. 550–558, 2005.

Steinberger, E., Lerer, A., and Brown, N. DREAM: deep regret minimization with advantage baselines and model-free learning. *CoRR*, abs/2006.10410, 2020. URL https://arxiv.org/abs/2006.10410.

Strens, M. A Bayesian Framework for Reinforcement Learning. In *International Conference on Machine Learning*, 2000.

Tammelin, O. Solving large imperfect information games using CFR+. *arXiv preprint arXiv:1407.5042*, 2014.

von Neumann, J. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. ISSN 0025-5831; 1432-1807/e.

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

Waugh, K. and Bagnell, J. A. A unified view of large-scale zero-sum equilibrium computation. *CoRR*, abs/1411.5007, 2014. URL http://arxiv.org/abs/1411.5007.

Wei, C., Lee, C., Zhang, M., and Luo, H. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4259–4299. PMLR, 2021. URL http://proceedings.mlr.press/v134/wei21a.html.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/36e729ec173b94133d8fa552e4029f8b-Paper.pdf.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3674–3682. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/xie20a.html.

Zhang, B. H. and Sandholm, T. Finding and Certifying (Near-) Optimal Strategies in Black-Box Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1166–1178. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0cc6ee01c82fc49c28706e0918f57e2d-Paper.pdf.

Zhou, Y., Li, J., and Zhu, J. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=Syg-ET4FPS.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. *Neural Information Processing Systems*, 2007.

https://doi.org/10.52202/079017-1806

Appendix and Checklist

A Related works

In this section, we review previous works on learning an ε-optimal strategy in IIGs.

Full feedback When the game is known, that is the information set structure space, transitions probability, and reward function are provided, a first line of work recasts the setting through the sequence-form representation of a game as a linear program which can be solved efficiently (Romanovsky, 1962; von Stengel, 1996; Koller et al., 1996). A second line of work relies on first-order optimization methods for saddle point computation (Hoda et al., 2010; Kroer et al., 2015, 2018, 2020; Munos et al., 2020; Lee et al., 2021). In particular Hoda et al. (2010); Kroer et al. (2018) relies on the Nesterov smoothing technique Nesterov (2005) whereas Kroer et al. (2015, 2020) use the `MirrorProx` algorithm (Nemirovski, 2004). These methods have a rate of convergence of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H, A_{\mathcal{X}}, B_{\mathcal{Y}})/\varepsilon)$.

A third approach, counterfactual regret minimization (Zinkevich et al., 2007), leverages local regret minimization, i.e. minimizing a type of regret at each information set. Popular algorithms are based on the regret-matching algorithm (Hart & Mas-Colell, 2000; Gordon, 2007) such as `CFR` algorithm (Zinkevich et al., 2007) or based on a close variant of regret-matching, e.g. `CFR+` (Tammelin, 2014; Burch et al., 2019; Farina et al., 2021a). Note that other local regret minimizers could be used, see for example Waugh & Bagnell (2014); Farina et al. (2019b). These algorithms enjoy a guarantee of convergence of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H, A_{\mathcal{X}}, B_{\mathcal{Y}})/\varepsilon^2)$.

Nevertheless, all the methods described above need to explore *the whole information set tree* (or the whole state space) in order to compute one update. The cost of one traversal is of order $\mathcal{O}(X + Y)$ if the transitions and the actions of the other player are sampled; see for example the external-sampling `MCCFR` algorithm (Lanctot et al., 2009).

Trajectory feedback A way to tackle the aforementioned issues is to consider the agnostic setting where the *agent has no prior knowledge of the game and only observes trajectories of the game*. Precisely, the rewards and the transition probabilities are unknown.

Model-based A first method to deal with this limited feedback is to build a *model* of the game and then run any full feedback algorithm in this model. For example, Zhou et al. (2020) use *posterior sampling* (PS, Strens, 2000) to learn a model and then use the `CFR` algorithm in games sampled from the posterior. They obtain a convergence rate of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H, S, A, B)/\varepsilon^2)$ but only when the games are actually sampled according to the known prior. Instead, Zhang & Sandholm (2021) relies on the principle of optimism in the presence of uncertainty to incrementally build a model of the game. Then, the `CFR` algorithm is fed with *optimistic estimates* of the local regrets. They prove a high-probability sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H, S, A, B)/\varepsilon^2)$.

Model-free Another line of work (Lanctot et al., 2009; Johanson et al., 2012; Schmid et al., 2018; Farina et al., 2020) directly estimates the local regret via importance sampling that is then fed to the `CFR` algorithm. In particular, the outcome-sampling `MCCFR` (Lanctot et al., 2009; Farina et al., 2020) builds an importance sampling estimate of the counterfactual regret by playing according to a well-chosen *balanced policy*. Intuitively, this policy should ensure to *explore all the information sets*. Note that, depending on the structure of the information set space, playing uniformly over the actions at each information set is not necessarily a good choice. Instead, Farina et al. (2020) propose as a balanced policy to play action with probability proportional to the number of leaves in the sub-tree of possible next information sets. In particular, the outcome-sampling `MCCFR` algorithm requires the knowledge of the information set space structure to build its balanced policy. Nonetheless, in order to obtain ε-optimal strategies with high probability, `MCCFR` needs at most $\widetilde{\mathcal{O}}(H^3(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ realizations of the game (Farina et al., 2020; Bai et al., 2022).

Later, Kozuno et al. (2021) proposed to combine *Online Mirror Descent (`OMD`)* with *dilated Shannon entropy as regularizer* and importance sampling estimate of the losses of a player, see also Farina et al. (2021b). They prove a sample complexity, for the proposed algorithm, `IXOMD`, of order

https://doi.org/10.52202/079017-1806

$\widetilde{\mathcal{O}}(H^2(XA_{\mathcal{X}}+YB_{\mathcal{Y}})/\varepsilon^2)$. Interestingly, they do not need to know in advance the structure of the information set space to obtain this bound. However, the sample complexity of `IXOMD` does not match the lower bound for this setting which is of order $\mathcal{O}((A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2)$. Recently, Bai et al. (2022) proposed the `Balanced OMD` algorithm that enjoys also relies on `OMD` but with a dilated entropy weighted by the realization plans of balanced policies as regularizers. For this algorithm, they prove a sample complexity of order $\widetilde{\mathcal{O}}(H^3(A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2)$.

Perfect information Markov game Another line of work considers Markov game Kuhn (1953) with *perfect* information and limited feedback. However, it does not assume perfect recall. Sidford et al. (2020); Zhang et al. (2020); Daskalakis et al. (2020); Wei et al. (2021) consider the case where a *generative model* is available whereas Wei et al. (2017); Bai et al. (2020); Xie et al. (2020); Liu et al. (2021) deal with the *trajectory feedback* case. Although this setting is related to ours there is no direct comparison between the two.

B Regret estimation

In this section, we aim to establish Theorem 2.2 of the main paper. We start by stating a Bernstein-type inequality that we will use multiple times. It can be found e.g. in Exercise 5.15 by Lattimore & Szepesvári (2020). We provide a short proof below as we did not find any for this precise statement.

Lemma B.1. *Let $(U^t)_{t\in[T]}$ be a sequence of random variables with respect to a filtration $\mathcal{F}$, and $\gamma>0$ be a fixed constant such that for all t, $\gamma U^t\le 1$. Then with a probability of at least $1-\delta'$:*

$$\sum_{t=1}^T\left(U^t-\mathbb{E}\left[U^t\middle|\mathcal{F}^{t-1}\right]\right)\le\gamma\sum_{t=1}^T\mathbb{E}\left[(U^t)^2\middle|\mathcal{F}^{t-1}\right]+\frac{1}{\gamma}\log(\frac{1}{\delta'})$$

Proof. For any $t\in[T]$, using the inequalities $\exp(x)\le 1+x+x^2$ for all $x\le 1$ and $1+x\le\exp(x)$ for all $x\in\mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\gamma U^t\right)\middle|\mathcal{F}^{t-1}\right]&\le\mathbb{E}\left[1+\gamma U^t+\gamma^2(U^t)^2\middle|\mathcal{F}^{t-1}\right]\\&=1+\gamma\mathbb{E}\left[U^t\middle|\mathcal{F}^{t-1}\right]+\gamma^2\mathbb{E}\left[(U^t)^2\middle|\mathcal{F}^{t-1}\right]\\&\le\exp\left(\gamma\mathbb{E}\left[U^t\middle|\mathcal{F}^{t-1}\right]+\gamma^2\mathbb{E}\left[(U^t)^2\middle|\mathcal{F}^{t-1}\right]\right).\end{aligned}$$

This implies that the random process $(S_t)_{t\in[T]}$ defined by

$$S_t:=\exp\left(\sum_{k=1}^t\gamma\left(U^k-\mathbb{E}\left[U^k\middle|\mathcal{F}^{k-1}\right]\right)-\sum_{k=1}^t\gamma^2\mathbb{E}\left[(U^k)^2\middle|\mathcal{F}^{k-1}\right]\right)$$

is a super-martingale, with $S_0=1$. Using the Markov inequality, we then get

$$\mathbb{P}\left(\frac{1}{\gamma}\log(S_T)>\frac{1}{\gamma}\log\left(\frac{1}{\delta'}\right)\right)=\mathbb{P}\left(S_T>\frac{1}{\delta'}\right)\le\delta'\,\mathbb{E}(S_T)\le\delta'$$

which immediately yields the stated inequality with probability at least $1-\delta'$. □

This lemma is then used for Theorem 2.2. The filtration $(\mathcal{F}^t)_{t\in[T]}$ will be used, such that $\mathcal{F}^t$ is the sigma-algebra of all variables of the self-play algorithm up to the execution of episode $t+1$.

Theorem B.2. *Assume that the estimated losses are obtained with a fixed positive sampling policy μ^s as above. Then, for any sequence $(\mu^t)_{t\in[T]}$ of $\Pi_{\min}$ and any $\delta\in(0,1)$, the following bound holds with a probability at least $1-\delta$*

$$\mathfrak{R}^T_{min}\le\max\left\{\hat{\mathfrak{R}}^T_{min},0\right\}+4\sqrt{\iota H\kappa(\mu^s)T}$$

where

$$\iota:=\log\left(\frac{A_{\mathcal{X}}+1}{\delta}\right)\quad\textit{and}\quad\kappa(\mu^s):=\max_{\mu\in\Pi_{\min}}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}_x}\frac{\mu_{1:}(x,a)}{\mu^s_{1:}(x,a)}.$$

Proof. We want to show that, with probability at least $1-\delta$, that

$$\sum_{t=1}^{t} \left\langle \ell^t - \widehat{\ell}^t, \mu_{1:}^t - \mu_{1:} \right\rangle \le 4\sqrt{\iota H \kappa(\mu^s) T}$$

holds for all $\mu \in \Pi_{\min}$. Then the property follows after re-organizing the inequality and maximizing over μ. In order to do so, we divide this term into two parts:

$$\sum_{t=1}^{T} \left\langle \ell^t - \widehat{\ell}^t, \mu_{1:}^t - \mu_{1:} \right\rangle = \underbrace{\sum_{t=1}^{T} \left\langle \widehat{\ell}^t - \ell^t, \mu_{1:} \right\rangle}_{\text{EST I}} + \underbrace{\sum_{t=1}^{T} \left\langle \ell^t - \widehat{\ell}^t, \mu_{1:}^t \right\rangle}_{\text{EST II}} .$$

We will furthermore assume that $HT \ge \iota\kappa(\mu^s)$, as otherwise, $4\sqrt{\iota H\kappa(\mu^s)T} \le 4HT$ and the property immediately follows from $\mathfrak{R}_{\text{min}}^T \le HT$.

Upper bound of EST I For all $x \in \mathcal{X}$ of depth h and $a \in \mathcal{A}(x)$, we apply Lemma B.1 to the random process

$$U_{x,a}^t = \ell_h^t \mathbb{I}_{\{x = x_h^t, a = a_h^t\}}$$

with $\delta' = \delta/(AX+1)$ and a fixed $\gamma_1 \in (0,1]$ we will specify later. This yields, with a probability at least $1-\delta'$, that

$$\begin{aligned}\sum_{t=1}^{T} \left(\ell_h^t \mathbb{I}_{\{x = x_h^t, a = a_h^t\}} - \mathbb{E}\left[\ell_h^t \mathbb{I}_{\{x = x_h^t, a = a_h^t\}} \middle| \mathcal{F}^{t-1} \right] \right) &\le \gamma_1 \sum_{t=1}^{T} \mathbb{E}\left[\left(\ell_h^t\right)^2 \mathbb{I}_{\{x = x_h^t, a = a_h^t\}} \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_1} \\ &\le \gamma_1 \sum_{t=1}^{T} \mathbb{E}\left[\ell_h^t \mathbb{I}_{\{x = x_h^t, a = a_h^t\}} \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_1} .\end{aligned}$$

By definition of the estimated loss, $\ell_h^t \mathbb{I}_{\{x = x_h^t, a = a_h^t\}} / \mu_{1:}^s(x,a) = \widehat{\ell}^t(x,a)$. We thus divide by $\mu_{1:}^s(x,a)$ both sides of the inequality, and the unbiasedness of the loss estimator yields

$$\sum_{t=1}^{T} \left[\widehat{\ell}^t(x,a) - \ell^t(x,a) \right] \le \gamma_1 \sum_{t=1}^{T} \ell^t(x,a) + \frac{\iota}{\gamma_1 \mu_1^s : (x,a)} .$$

This inequality holds for all (x,a) with a probability of at least $1 - \delta A_{\mathcal{X}}/(A_{\mathcal{X}}+1)$. Taking the scalar product with any $\mu \in \Pi_{\min}$ then gives

$$\begin{aligned}\sum_{t=1}^{T} \left\langle \widehat{\ell}^t - \ell^t, \mu_{1:} \right\rangle &\le \gamma_1 \sum_{t=1}^{T} \left\langle \ell^t, \mu_{1:} \right\rangle + \frac{1}{\gamma_1} \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}(x)} \frac{\mu_{1:}(x,a)}{\mu_{1:}^s(x,a)} \\ &\le \gamma_1 HT + \frac{\iota}{\gamma_1} \kappa(\mu^s) .\end{aligned}$$

Using $\gamma_1 = \sqrt{\iota\kappa(\mu^s)/(HT)} \le 1$ (by assumption), finally yields

$$\text{EST I} \le 2\sqrt{\iota H \kappa(\mu^s) T} .$$

Upper bound of EST II For this upper bound, we apply Lemma B.1 directly to the sequence $U^t = \left\langle -\widehat{\ell}^t, \mu_{1:}^t \right\rangle$. We now choose $\gamma_2 \in \mathbb{R}_+$ (no further assumption is needed on γ_2 as the sequence is negative) and apply the lemma to get with probability at least $1 - \delta/(A_{\mathcal{X}}+1)$

$$
\begin{aligned}
\sum_{t=1}^{T} \left\langle \ell^t - \widehat{\ell}^t, \mu_{1:}^t \right\rangle &\leq \gamma_2 \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \widehat{\ell}^t, \mu_{1:}^t \right\rangle^2 \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_2} \\
&= \gamma_2 \sum_{t=1}^{T} \mathbb{E}\left[\left(\sum_{h=1}^{H} \left(\ell_h^t\right) \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}(x)} \mathbb{I}_{\left\{x=x_h^t, a=a_h^t\right\}} \frac{\mu_{1:}^t(x,a)}{\mu_{1:}^s(x,a)} \right)^2 \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_2} \\
\text{(Cauchy-Schwarz)} \quad &\leq \gamma_2 H \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=1}^{H} \left(\ell_h^t\right)^2 \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}(x)} \mathbb{I}_{\left\{x=x_h^t, a=a_h^t\right\}} \frac{\mu_{1:}^t(x,a)^2}{\mu_{1:}^s(x,a)^2} \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_2} \\
&\leq \gamma_2 H \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=1}^{H} \ell_h^t \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}(x)} \mathbb{I}_{\left\{x=x_h^t, a=a_h^t\right\}} \frac{\mu_{1:}^t(x,a)}{\mu_{1:}^s(x,a)^2} \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_2} \\
&= \gamma_2 H \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=1}^{H} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}(x)} \widehat{\ell}^t(x,a) \frac{\mu_{1:}^t(x,a)}{\mu_{1:}^s(x,a)} \middle| \mathcal{F}^{t-1} \right] + \frac{\iota}{\gamma_2} \\
&= \gamma_2 H \sum_{t=1}^{T} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}(x)} \ell^t(x,a) \frac{\mu_{1:}^t(x,a)}{\mu_{1:}^s(x,a)} + \frac{\iota}{\gamma_2} \\
\text{(as } \ell^t(x,a) \leq 1\text{)} \quad &\leq \gamma_2 H \sum_{t=1}^{T} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}(x)} \frac{\mu_{1:}^t(x,a)}{\mu_{1:}^s(x,a)} + \frac{\iota}{\gamma_2} \\
&\leq \gamma_2 H \kappa(\mu^s) T + \frac{\iota}{\gamma_2}\,.
\end{aligned}
$$

Taking $\gamma_2 = \sqrt{\frac{\iota}{H\kappa(\mu^s)T}}$ then leads to

$$
\sum_{t=1}^{T} \left\langle \ell^t - \widehat{\ell}^t, \mu_{1:}^t \right\rangle \leq 2\sqrt{\iota H \kappa(\mu^s) T}\,.
$$

Summing the two inequalities yields the inequality of the theorem with a probability of at least $1-\delta$. □

C Balanced policy and κ

This section deals with the $\kappa(\mu^s)$ and local $\kappa(\mu^s|x)$ of the main paper, and links it to the balanced policy $\mu^\star$.

Recursive κ computation Let μ^s be the positive sample policy. For any $\mu \in \Pi_{\min}$ and $x \in \mathcal{X}$ of depth h, we define $\kappa_\mu(\mu^s|x)$ the local sum of ratios against μ in the subtree induced by x, i.e.

$$
\kappa_\mu(\mu^s|x) := \sum_{x'\in\mathcal{X}, x \text{ is in the history of } x'} \; \sum_{a'\in\mathcal{A}(x')} \frac{\mu_{h:}(x',a')}{\mu_{h:}^s(x',a')}
$$

where, if $(x_1', a_1' ..., x_{h'}', a')$ is the history of (x', a'),

$$
\mu_{h:}(x',a') := \Pi_{i=h}^{h'} \, \mu(a_i'|x_i')\,.
$$

We then formally define $\kappa(\mu^s|x)$ as $\kappa(\mu^s|x) := \max_{\mu\in\Pi_{\min}} \kappa_\mu(\mu^s|x)$. For any $\mu \in \Pi_{\min}$, the following recursive formula stands

$$
\kappa_\mu(\mu^s|x) = \sum_{a\in\mathcal{A}(x)} \frac{\mu(a|x)}{\mu^s(a|x)} \left(1 + \sum_{x'\in\mathcal{X}, x' \text{ directly follows } (x,a)} \kappa_\mu(\mu^s|x') \right)
$$

that follows from the definition of $\kappa_\mu(\mu^s|x)$. The same kind of recursion can then be obtained for $\kappa(\mu^s|x)$, because each appearance of μ in the previous equality can be maximized independently (depending on different information sets). This yields

$$\begin{aligned}\kappa(\mu^s|x) &= \max_{\mu\in\Delta_{\mathcal{A}(x)}} \sum_{a\in\mathcal{A}(x)} \frac{\mu(a)}{\mu^s(a|x)}\left(1+\sum_{x'\in\mathcal{X},x' \text{ directly follows } (x,a)} \kappa(\mu^s|x')\right)\\ &= \max_{a\in\mathcal{A}(x)} \frac{1}{\mu^s(a|x)}\left(1+\sum_{x'\in\mathcal{X},x' \text{ directly follows } (x,a)} \kappa(\mu^s|x')\right), \qquad (1)\end{aligned}$$

which allows for a simple recursive computation of $\kappa(\mu^s|x)$. Finally, once the whole recursive computation is done, $\kappa(\mu^s)$ itself can be computed by, defining $\mathcal{X}_1$ the information sets of depth 1,

$$\kappa(\mu^s) = \sum_{x_1\in\mathcal{X}_1} \kappa(\mu^s|x_1)\,.$$

Balanced policy $\kappa(\mu^s|x)$ can also be minimized over $\mu^s \in \Pi_{\min}$ recursively from the leaves using the tree structure. Indeed, for each $x \in \mathcal{X}$, assuming that the minimizers of $\kappa(\mu^s|x')$ are already known for subsequent x', the policy $\mu^s \in \Delta_{\mathcal{A}(x)}$ that minimizes the maximum along the actions $a \in \mathcal{A}(x)$ can be computed from (1). Furthermore, if we define $A^\tau(x,a)$ and $A^\tau(x)$ the total number of actions in the subtrees respectively induced by (x,a) and x, i.e.

$$A^\tau(x,a) := 1 + \sum_{x'\in\mathcal{X},(x,a) \text{ is in the history of } x'} |\mathcal{A}(x')| \quad \text{and} \quad A^\tau(x) := \sum_{a\in\mathcal{A}(x)} A^\tau(x,a)\,,$$

we can show that $\min_{\mu^s\in\Pi_{\min}} \kappa(\mu^s|x) = A^\tau(x)$, and that the minimum is attained by the balanced policy $\mu^\star$ defined by

$$\mu^\star(a|x) := \frac{A^\tau(x,a)}{A^\tau(x)}\,.$$

Indeed, if we assume in (1) that the previous property holds for the $\kappa(\mu^s|x')$, then

$$\kappa(\mu^s|x) = \max_{a\in\mathcal{A}(x)} \frac{1}{\mu^s(a|x)}\left(1+\sum_{x'\in\mathcal{X},x' \text{ directly follows } (x,a)} A^\tau(x')\right) = \max_{a\in\mathcal{A}(x)} \frac{A^\tau(x,a)}{\mu^s(a|x)}$$

and the previous equality is minimized when the $\mu^s(a|x)$ are proportional to the $A^\tau(x,a)$, achieved by the balanced policy $\mu^\star$. With this policy, the same equality gives $\kappa(\mu^\star|x) = A^\tau(x)$, which concludes the induction.

Finally, computing $\kappa(\mu^\star)$ yields

$$\kappa(\mu^\star) = \sum_{x_1\in\mathcal{X}_1} \kappa(\mu^\star|x_1) = \sum_{x_1\in\mathcal{X}_1} A^\tau(x_1) = A_{\mathcal{X}}\,.$$

D Generalized dual stabilized online mirror descent

This section will establish the bound related to the updates (GDS-OMD) obtained with any Legendre function.

D.1 General Bregman divergence properties

We start this section by stating multiple properties of the Bregman divergence $\mathbf{D}_\Psi$ for Ψ a convex function, continuously differentiable on an open Ω and defined on $\overline{\Omega}$, proved by Cesa-Bianchi & Lugosi (2006).

Law of cosines : For any $x \in \overline{\Omega}$ and $w, z \in \Omega$, the following equality holds

$$\mathbf{D}_\Psi(x,w) = \mathbf{D}_\Psi(x,z) + \mathbf{D}_\Psi(z,w) - \langle \nabla\Psi(w) - \nabla\Psi(z), x - z\rangle\,.$$

Algorithm 3 Generalized dual-stabilized online mirror descent

```
Input:
    A sequence of dual increments $\xi^t$
    An open subset $\Omega \in \mathbb{R}^n$ and a closed convex $\mathcal{C}$ of $\overline{\Omega}$
    A sequence of Legendre regularizers $(\Psi^t)_{t\in[T]}$ on $\overline{\Omega}$ such that for all $t \in [T]$, $\Psi^{t+1} - \Psi^t$ is convex
    An initial primal iterate $w^1 \in \mathcal{C}$
Output:
    A sequence $(w^t)_{t\in[T]}$ of primal iterates
Algorithm:
For $t = 1$ to $T$
    $z^t = \nabla\Psi^t(w^t)$
    $y^{t+1} = z^t - \xi^t + \nabla\Psi^{t+1}(w_1) - \nabla\Psi^t(w^1)$
    $\hat{w}^{t+1} = \nabla\Psi^{t+1,\star}(y^{t+1})$
    $w^{t+1} = \Pi_{\mathcal{C}}^{\Psi^{t+1}}(\hat{w}^{t+1})$
```

Bregman projection : For $\mathcal{C}$ a closed convex of $\overline{\Omega}$, and Ψ strictly convex, we can define the Bregman projection $\Pi_{\mathcal{C}}^{\Psi}$ over $\overline{\Omega}$ by

$$\Pi_{\mathcal{C}}^{\Psi}(w) = \arg\min_{z\in\mathcal{C}} \mathbf{D}_{\Psi}(z, w)\,.$$

This Bregman projection satisfies a generalized Pythagorean inequality, for $w \in \Omega$ and $z \in \mathcal{C}$

$$\mathbf{D}_{\Psi}(z, w) \geq \mathbf{D}_{\Psi}(z, \Pi_{\mathcal{C}}^{\Psi}(w)) + \mathbf{D}_{\Psi}(\Pi_{\mathcal{C}}^{\Psi}(w), w)$$

Fenchel dual : We defined the Fenchel dual $\Psi^\star$ of a Legendre function Ψ for any $\xi \in \mathbb{R}^n$ by

$$\Psi^\star(\xi) = \sup_{w\in\overline{\Omega}} \langle \xi, w\rangle - \Psi(w)\,.$$

If we consider $\Omega^\star := \nabla\Psi(\Omega)$, it can be shown that $\nabla\Psi^\star$ is the inverse function of $\nabla\Psi$ over $\Omega^\star$, i.e. for any $w \in \Omega$, $\nabla\Psi^\star(\nabla\Psi(w)) = w$. Furthermore, for $w, z \in \Omega$,

$$\mathbf{D}_{\Psi}(w, z) = \mathbf{D}_{\Psi^\star}(\nabla\Psi^\star(z), \nabla\Psi^\star(y))\,.$$

Strong convexity: Ψ is said to be 1-strongly convex with respect to a norm $\|\cdot\|$ if for all $w, z \in \Omega$

$$\Psi(z) \geq \Psi(w) + \langle \nabla\Psi(w), z - w\rangle + \frac{1}{2}\|w - z\|^2\,.$$

In this case, the Bregman divergence of the Fenchel dual $\Psi^\star$ satisfies for any $\xi_1, \xi_2 \in \Omega^\star$

$$\mathbf{D}_{\Psi^\star}(\xi_1, \xi_2) \leq \|\xi_1 - \xi_2\|_\star^2$$

where $\|\cdot\|_\star$ is the dual norm of $\|\cdot\|$.

D.2 GDS-OMD Analysis

We will assume in the following parts that the updates of the following algorithm are properly defined, which happens when all vectors y^{t+1} belong to the Fenchel dual space $\Omega^{t+1,\star} := \nabla\Psi^{t+1}(\Omega)$. We make the same assumption on the regular OMD iterates $z^t - \xi^t$.

We start by giving an equivalent formulation of the updates (GDS-OMD) through Algorithm 3.

Proposition D.1. *Algorithm 3 computes the updates* (GDS-OMD) *if they are properly defined, i.e. computes the sequence of primal iterates defined by*

$$w^{t+1} = \arg\min_{w\in\mathcal{C}} \left\langle \xi^t, w\right\rangle + \mathbf{D}_{\Psi^t}\left(w, w^t\right) + \mathbf{D}_{\Psi^{t+1}-\Psi^t}\left(w, w^1\right)\,.$$

Proof. By definition of $\hat{w}^{t+1}$ in Algorithm 3, we have for all iterations $t \in [T]$ and $w \in \mathcal{C}$

$$\begin{aligned}\mathbf{D}_{\Psi^{t+1}}(w, \hat{w}^{t+1}) &= \Psi^{t+1}(w) - \left\langle \nabla\Psi^{t+1}(\hat{w}^{t+1}), w \right\rangle + C_1 \\ &= \Psi^t(w) + \left(\Psi^{t+1}(w) - \Psi^t(w)\right) - \left\langle y^{t+1}, w \right\rangle + C_1 \\ &= \left\langle \xi^t, w \right\rangle + \left(\Psi^t(w) - \left\langle \nabla\Psi^t(w^t), w \right\rangle\right) + \\ &\qquad \left(\Psi^{t+1}(w) - \Psi^t(w) - \left\langle \nabla\Psi^{t+1}(w^1) - \nabla\Psi^t(w^1), w \right\rangle\right) + C_1 \\ &= \left\langle \xi^t, w \right\rangle + \mathbf{D}_{\Psi^t}\left(w, w^t\right) + \mathbf{D}_{\Psi^{t+1}-\Psi^t}\left(w, w^1\right) + C_2\end{aligned}$$

where C_1 and C_2 are constants independent of the choice of w (but not independent of the other variables). As $w^{t+1} = \arg\min_{w \in \mathcal{C}} \mathbf{D}_{\Psi^{t+1}}(w, \hat{w}^{t+1})$, the updates of Algorithm 3 coincide with the updates (GDS-OMD), as both minimize the same function at each iteration up to an additive constant. □

The updates of Algorithm 3 are then used to show Theorem 3.2 below. Compared to the ones of McMahan (2017) that also allow adaptive regularization, these updates do not suffer from the potential linear rates observed by Orabona & Pál (2018).

Theorem D.2. *Let $(w^t)_{t\in[T]}$ be a sequence of primal iterates generated by the updates (GDS-OMD), with convex incremental functions. Then for any $w^\dagger \in \overline{\Omega}$,*

$$\sum_{t=1}^{T} \left\langle \xi^t, w^t - w^\dagger \right\rangle \leq \mathbf{D}_{\Psi^T}(w^\dagger, w^1) + \sum_{t=1}^{T} \mathbf{D}_{\Psi^{t,\star}}\left(\nabla\Psi^t(w^t) - \xi^t, \nabla\Psi^t(w^t)\right)$$

Proof. We can assume, without any incidence on the $(w^t)_{t\in[T]}$ sequence, that $\Psi^{T+1} = \Psi^T$. We also define for all $t \in [T]$ the notations $\varphi^t = \Psi^{t+1} - \Psi^t$ and

$$\hat{q}^t = \left\langle \xi^t, \hat{w}^{t+1} \right\rangle + \mathbf{D}_{\Psi^t}(\hat{w}^{t+1}, w^t) + \mathbf{D}_{\varphi^t}(\hat{w}^{t+1}, w^1)\,.$$

We then divide the sum into a stability and a penalty terms:

$$\sum_{t=1}^{T} \left\langle \xi^t, w^t - w^\dagger \right\rangle = \underbrace{\sum_{t=1}^{T} \left(\hat{q}^t - \left\langle \xi^t, w^\dagger \right\rangle\right)}_{\text{penalty}} + \underbrace{\sum_{t=1}^{T} \left(\left\langle \xi^t, w^t \right\rangle - \hat{q}^t\right)}_{\text{stability}}$$

and we look at upper-bounding these two terms.

Penalty term: For all $t \in [T]$, using the law of cosines on the Bregman divergences of Ψ^t and φ^t, we have the two equalities:

$$\mathbf{D}_{\Psi^t}(w^\dagger, w^t) = \mathbf{D}_{\Psi^t}(w^\dagger, \hat{w}^{t+1}) + \mathbf{D}_{\Psi^t}(\hat{w}^{t+1}, w^t) - \left\langle \nabla\Psi^t(w^t) - \nabla\Psi^t(\hat{w}^{t+1}), w^\dagger - \hat{w}^{t+1} \right\rangle$$

and

$$\mathbf{D}_{\varphi^t}(w^\dagger, w^1) = \mathbf{D}_{\varphi^t}(w^\dagger, \hat{w}^{t+1}) + \mathbf{D}_{\varphi^t}(\hat{w}^{t+1}, w^1) - \left\langle \nabla\varphi^t(w^1) - \nabla\varphi^t(\hat{w}^{t+1}), w^\dagger - \hat{w}^{t+1} \right\rangle\,.$$

Summing these two equalities, we get

$$\begin{aligned}&\mathbf{D}_{\Psi^t}(w^\dagger, w^t) + \mathbf{D}_{\varphi^t}(w^\dagger, w^1) \\ &\quad = \mathbf{D}_{\Psi^{t+1}}(w^\dagger, \hat{w}^{t+1}) + \mathbf{D}_{\Psi^t}(\hat{w}^{t+1}, w^t) + \mathbf{D}_{\varphi^t}(\hat{w}^{t+1}, w^1) - \left\langle \xi^t, w^\dagger - \hat{w}^{t+1} \right\rangle \\ &\quad = \mathbf{D}_{\Psi^{t+1}}(w^\dagger, \hat{w}^{t+1}) + \hat{q}^t - \left\langle \xi^t, w^\dagger \right\rangle\end{aligned}$$

as by definition of $\hat{w}^{t+1}$ and y^{t+1},

$$\nabla\Psi^{t+1}(\hat{w}^{t+1}) = y^{t+1} = -\xi_t + \nabla\Psi^t(w^t) + \nabla\varphi^t(w^1)\,.$$

Furthermore, as $w^{t+1} = \Pi_{\mathcal{C}}^{t+1}(\hat{w}^{t+1})$, the Pythagorean inequality for the Bregman divergence yields that

$$\mathbf{D}_{\Psi^{t+1}}(w^\dagger, \hat{w}^{t+1}) \geq \mathbf{D}_{\Psi^{t+1}}(w^\dagger, w^{t+1}) + \mathbf{D}_{\Psi^{t+1}}(w^{t+1}, \hat{w}^{t+1}) \geq \mathbf{D}_{\Psi^{t+1}}(w^\dagger, w^{t+1})\,.$$

https://doi.org/10.52202/079017-1806

Injecting this in the previous equality and telescoping leads to

$$\begin{aligned}\sum_{t=1}^{T}\left(\hat{q}^t-\left\langle \xi^t, w^\dagger\right\rangle\right) &= \sum_{t=1}^{T}\left(\mathbf{D}_{\Psi^t}(w^\dagger, w^t)+\mathbf{D}_{\varphi^t}(w^\dagger, w^1)-\mathbf{D}_{\Psi^{t+1}}(w^\dagger, \hat{w}^{t+1})\right)\\ &\leq \sum_{t=1}^{T}\left(\mathbf{D}_{\Psi^t}(w^\dagger, w^t)+\mathbf{D}_{\varphi^t}(w^\dagger, w^1)-\mathbf{D}_{\Psi^{t+1}}(w^\dagger, w^{t+1})\right)\\ &= \mathbf{D}_{\Psi^{T+1}}(w^\dagger, w^1)-\mathbf{D}_{\Psi^{T+1}}(w^\dagger, w^{t+1})\\ &\leq \mathbf{D}_{\Psi^{T}}(w^\dagger, w^1)\end{aligned}$$

as $\Psi^T = \Psi^{T+1}$ by definition.

Stability term: We first notice, for all $t \in [T]$, that

$$\begin{aligned}\left\langle \xi^t, w^t\right\rangle-\hat{q}^t &= \left\langle \xi^t, w^t-\hat{w}^{t+1}\right\rangle-\mathbf{D}_{\Psi^t}(\hat{w}^{t+1}, w^t)-\mathbf{D}_{\varphi^t}(\hat{w}^{t+1}, w^1)\\ &\leq \left\langle \xi^t, w^t-\hat{w}^{t+1}\right\rangle-\mathbf{D}_{\Psi^t}(\hat{w}^{t+1}, w^t)\\ &\leq \left\langle \xi^t, w^t-\tilde{w}^{t+1}\right\rangle-\mathbf{D}_{\Psi^t}(\tilde{w}^{t+1}, w^t)\end{aligned}$$

where

$$\tilde{w}^{t+1} := \operatorname*{arg\,min}_{\tilde{w}\in\Omega}\left[\left\langle \xi^t, \tilde{w}\right\rangle+\mathbf{D}_{\Psi^t}(\tilde{w}, w^t)\right]$$

is the $\tilde{w}^{t+1}$ iterate that would be obtained using a classical OMD step with Ψ^t, without the stabilization. By optimality, it verifies

$$\nabla\Psi^t(\tilde{w}^{t+1}) = \nabla\Psi^t(w^t)-\xi^t$$

and the law of cosines then yields

$$\begin{aligned}\mathbf{D}_{\Psi^t}(w^t, w^t) &= \mathbf{D}_{\Psi^t}(w^t, \tilde{w}^{t+1})+\mathbf{D}_{\Psi^t}(\tilde{w}^{t+1}, w^t)-\left\langle \nabla\Psi^t(w^t)-\nabla\Psi^t(\tilde{w}^{t+1}), w^t-\tilde{w}^{t+1}\right\rangle\\ (0) &= \mathbf{D}_{\Psi^t}(w^t, \tilde{w}^{t+1})+\mathbf{D}_{\Psi^t}(\tilde{w}^{t+1}, w^t)-\left\langle \xi^t, w^t-\tilde{w}^{t+1}\right\rangle .\end{aligned}$$

Plugging this in the first inequality, we directly get

$$\left\langle \xi^t, w^t\right\rangle-\hat{q}^t \leq \mathbf{D}_{\Psi^t}(w^t, \tilde{w}^{t+1})$$

and we conclude using

$$\begin{aligned}\mathbf{D}_{\Psi^t}(w^t, \tilde{w}^{t+1}) &= \mathbf{D}_{\Psi^{t,\star}}(\nabla\Psi^t(\tilde{w}^{t+1}), \nabla\Psi^t(w^t))\\ &= \mathbf{D}_{\Psi^{t,\star}}(\nabla\Psi^t(w^t)-\xi^t, \nabla\Psi^t(w^t)) .\end{aligned}$$

□

E LocalOMD analysis

This section will focus on the dilated entropy approach to extensive-form games, and especially on the updates

$$\mu^{t+1} = \operatorname*{arg\,min}_{\mu\in\Pi_{\min}}\left\langle \widehat{\ell}^t, \mu_{1:}\right\rangle+\mathbf{D}^{\mathrm{dil}}_{\alpha^t}(\mu, \mu^t)+\mathbf{D}^{\mathrm{dil}}_{\alpha^{t+1}-\alpha^t}(\mu, \mu^1) \qquad \text{(GDS-OMD dilated)}$$

that are used by LocalOMD.

E.1 General analysis

The following proposition shows that each update of this form can be computed recursively starting from the leaves of the tree. It requires for any $t \in [T]$ the vector q^t that satisfies for any $x \in \mathcal{X}$ of depth h

$$q^t(x) = \min_{\mu\in\Pi_{\min}}\left\langle \widehat{\ell}^{t,\to x}, \mu_{h:}^{\to x}\right\rangle+\mathbf{D}^{\mathrm{dil},\to\mathrm{x}}_{\alpha^t}(\mu, \mu^t)+\mathbf{D}^{\mathrm{dil},\to\mathrm{x}}_{\alpha^{t+1}-\alpha^t}(\mu, \mu^1)$$

where $\to x$ means that the quantity is considered on the sub-tree induced by x rather than the full information set tree, and $\mu_{h:}$ is defined in Appendix C.

https://doi.org/10.52202/079017-1806

Proposition E.1. *Consider the previous updates (GDS-OMD dilated) and the vectors* $(q^t)_{t\in[T]}$ *above. Both* μ^{t+1} *and* q^t *can be computed recursively starting from the leaves of the tree through*

$$\mu^{t+1} = \operatorname*{arg\,min}_{\mu\in\Delta_{\mathcal{A}(x)}} h_x^t(\mu) \quad \textit{and} \quad q^t(x) = \min_{\mu\in\Delta_{\mathcal{A}(x)}} h_x^t(\mu)$$

where

$$h_x^t(\mu) = \left\langle \widetilde{\ell}^t(x,\cdot), \mu \right\rangle + (1/\alpha^t(x))\,\mathbf{D}_x(\mu, \mu^t(\cdot|x)) + \big(1/\alpha^{t+1}(x) - 1/\alpha^t(x)\big)\ \mathbf{D}_x(\mu, \mu^1(\cdot|x))$$

and the regularized loss $\widetilde{\ell}^t(x,a)$ *is defined by*

$$\widetilde{\ell}^t(x,a) := \widehat{\ell}^t(x,a) + \sum_{x'\in\mathcal{X}|x' \textit{ directly follows } (x,a)} q^t(x')\,.$$

Proof. First, note that μ^{t+1} is the unique minimizer associated to each $q^t(x_1)$ for x_1 the information set of depth 1. Indeed, each of the sub-tree induced by the x_1 can be considered as an independent problem. The idea will be to recursively minimize the $q^t(x)$, starting from the leaves (i.e. the final information sets x_H), and compute $\mu^{t+1}(|x)$ as the associated minimizer at each information set.

This minimization is done through, at each $x \in \mathcal{X}$ of depth h, with a decomposition of $q^t(x)$. Indeed, separating the induced tree by x between the root and the rest of the tree leads to

$$\begin{aligned}
q^t(x) = \operatorname*{arg\,min}_{\mu\in\Pi_{\min}} & \left\langle \widehat{\ell}^t(x,\cdot), \mu(\cdot|x) \right\rangle + (1/\alpha^t(x))\,\mathbf{D}_x(\mu(\cdot|x), \mu^t(\cdot|x)) \\
& + \big(1/\alpha^{t+1}(x) - 1/\alpha^t(x)\big)\ \mathbf{D}_x(\mu(\cdot|x), \mu^1(\cdot|x)) \\
& + \sum_{a\in\mathcal{A}(x)} \mu(a|x) \sum_{x'\in\mathcal{X}|\text{x' directly follows } (x,a)} \left[\left\langle \widehat{\ell}^{t,\to x'}, \mu_{h+1:}^{\to x'} \right\rangle + \mathbf{D}_{\alpha^t}^{\text{dil},\to\text{x}'}(\mu,\mu^t) + \mathbf{D}_{\alpha^{t+1}-\alpha^t}^{\text{dil},\to\text{x}'}(\mu,\mu^1)\right] \\
= \operatorname*{arg\,min}_{\mu\in\Delta_{\mathcal{A}(x)}} & \left\langle \widehat{\ell}^t(x,\cdot), \mu \right\rangle + (1/\alpha^t(x))\,\mathbf{D}_x(\mu, \mu^t(\cdot|x)) + \big(1/\alpha^{t+1}(x) - 1/\alpha^t(x)\big)\ \mathbf{D}_x(\mu, \mu^1(\cdot|x)) \\
& + \sum_{a\in\mathcal{A}(x)} \mu(a) \sum_{x'\in\mathcal{X}|\text{x' directly follows } (x,a)} q^t(x') \\
= \operatorname*{arg\,min}_{\mu\in\Delta_{\mathcal{A}(x)}} & \left\langle \widetilde{\ell}^t(x,\cdot), \mu \right\rangle + (1/\alpha^t(x))\,\mathbf{D}_x(\mu, \mu^t(\cdot|x)) + \big(1/\alpha^{t+1}(x) - 1/\alpha^t(x)\big)\ \mathbf{D}_x(\mu, \mu^1(\cdot|x)) \\
= \operatorname*{arg\,min}_{\mu\in\Delta_{\mathcal{A}(x)}} & h_x(\mu)
\end{aligned}$$

as each minimization on $\mu \in \Pi_{\min}$ is done on independent components. This justifies the recursive computation of both μ^{t+1} and q^t. □

This proposition directly provides the proof of correctness of `LocalOMD`, for which the regularized losses at time step t are non-null only on the trajectory with

$$\widetilde{\ell}^t(x,a) = \frac{\mathbb{I}_{\{x=x_h^t, a=a_h^t\}}}{\mu_{1:}^s(x)} \widetilde{\ell}_h^t\,.$$

We now want to upper(bound the regret associated with this sequence μ^t. The following lemma gives a valuable property that links the regularized loss and the estimated loss.

Lemma E.2. *For any policy* $\mu' \in \Pi_{\min}$*, we have*

$$\left\langle \widetilde{\ell}^t, \mu'_{1:} \right\rangle - \sum_{x\in\mathcal{X}} \mu'_{1:}(x) q^t(x) = \left\langle \widehat{\ell}^t, \mu'_{1:} \right\rangle - \hat{q}^t$$

where $\hat{q}^t = \min_{\mu\in\Pi_{\min}} \left\langle \widehat{\ell}^t, \mu_{1:} \right\rangle + \mathcal{D}_{\alpha^t}^{\text{dil}}(\mu,\mu^t) + \mathcal{D}_{\alpha^{t+1}-\alpha^t}^{\text{dil}}(\mu,\mu^1)$

Proof. By definition of $\widetilde{\ell}^t$ we have, for any $\mu \in \Pi_{\min}$

$$\begin{aligned}
\left\langle \widetilde{\ell}^t, \mu'_{1:} \right\rangle &= \left\langle \widehat{\ell}^t, \mu'_{1:} \right\rangle + \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}_x} \mu'_{1:}(x,a) \sum_{x' | (x,a) \longrightarrow x'} q^t(x') \\
&= \left\langle \widehat{\ell}^t, \mu'_{1:} \right\rangle + \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}_x} \sum_{x' | (x,a) \longrightarrow x'} \mu'_{1:}(x') q^t(x') \\
&= \left\langle \widehat{\ell}^t, \mu'_{1:} \right\rangle + \sum_{x' \in \mathcal{X} \setminus \mathcal{X}_1} \mu'_{1:}(x') q^t(x') \\
&= \left\langle \widehat{\ell}^t, \mu'_{1:} \right\rangle + \sum_{x' \in \mathcal{X}} \mu'_{1:}(x') q^t(x') - \sum_{x' \in \mathcal{X}_1} q^t(x')
\end{aligned}$$

in which we identified the components of the second sum as the set of non-initial information sets. We then conclude using $\sum_{x \in \mathcal{X}_1} q^t(x) = \hat{q}^t$ by definition of the q^t terms. □

This lemma is then used to upper bound the estimated regret of the sequence generated by the updates (GDS-OMD dilated). Indeed, while we could apply Theorem 3.2, the associated stability term, which depends on the Fenchel dual of the dilated entropy, is not easy to upper bound. Nonetheless, the proof of the following theorem is mostly the same but with a slightly different definition of the stability and penalty terms.

Theorem E.3. *Let* $(\mu^t)_{t \in [T]}$ *be the sequence of policies generated by the updates* (GDS-OMD dilated)*. The following bound holds*

$$\hat{R}^T \leq \underbrace{\sup_{\mu^\dagger \in \Pi_{\min}} \mathbf{D}^{\mathrm{dil}}_{\alpha^T}(\mu^\dagger_{1:}, \mu^1_{1:})}_{\text{penalty}} + \underbrace{\sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \alpha^t(x) \mu^t_{1:}(x) \mathbf{D}^\star_x \left(\nabla \Psi_x(\mu^t_{1:}(\cdot|x)) - \frac{1}{\alpha^t(x)} \widetilde{\ell}^t(x, \cdot), \nabla \Psi_x(\mu^t_{1:}(\cdot|x)) \right)}_{\text{stability}} .$$

Proof. The separation between the stability and the penalty terms is the same as in Theorem 3.2, but with $\hat{q}^t$ (of Lemma E.2) defined after the projection rather than before. This leads to the decomposition

$$\hat{\mathcal{R}}^T = \underbrace{\max_{\mu^\dagger \in \Pi_{\min}} \sum_{t=1}^{T} \left(\hat{q}^t - \left\langle \widehat{\ell}^t, \mu^\dagger_{1:} \right\rangle \right)}_{\text{penalty}} + \underbrace{\sum_{t=1}^{T} \left(\left\langle \widehat{\ell}^t, \mu^t_{1:} \right\rangle - \hat{q}^t \right)}_{\text{stability}} .$$

Penalty term: This part is similar to the general theorem. The optimality of μ^{t+1} leads to, for any $t \in [T]$,

$$\nabla \Psi^{t+1}(\mu^{t+1}_{1:}) = -\widehat{\ell}^t - g^t + \nabla \Psi^t(\mu^t_{1:}) + \nabla \varphi^t(\mu^1_{1:}) .$$

where $g^t \in Q^{\perp}_{\max}$ and $\varphi^t = \Psi^{t+1} - \Psi^t$. We use the same two law of cosines as in Theorem 3.2

$$\begin{aligned}
\mathbf{D}_{\Psi^t}(\mu^\dagger_{1:}, \mu^t_{1:}) &= \mathbf{D}_{\Psi^t}(\mu^\dagger_{1:}, \mu^{t+1}_{1:}) + \mathbf{D}_{\Psi^t}(\mu^{t+1}_{1:}, \mu^t_{1:}) - \left\langle \nabla \Psi^t(\mu^t_{1:}) - \nabla \Psi^t(\mu^{t+1}_{1:}), \mu^\dagger_{1:} - \mu^{t+1}_{1:} \right\rangle \\
\mathbf{D}_{\varphi^t}(\mu^\dagger_{1:}, \mu^1_{1:}) &= \mathbf{D}_{\varphi^t}(\mu^\dagger_{1:}, \mu^{t+1}_{1:}) + \mathbf{D}_{\varphi^t}(\mu^{t+1}_{1:}, \mu^1_{1:}) - \left\langle \nabla \varphi^t(\mu^1_{1:}) - \nabla \varphi^t(\mu^{t+1}_{1:}), \mu^\dagger_{1:} - \mu^{t+1}_{1:} \right\rangle
\end{aligned}$$

which yields by summing

$$\begin{aligned}
&\mathbf{D}_{\Psi^t}(\mu^\dagger_{1:}, \mu^t_{1:}) + \mathbf{D}_{\varphi^t}(\mu^\dagger_{1:}, \mu^1_{1:}) \\
&\quad = \mathbf{D}_{\Psi^{t+1}}(\mu^\dagger_{1:}, \mu^{t+1}_{1:}) + \mathbf{D}_{\Psi^t}(\mu^{t+1}_{1:}, \mu^t_{1:}) + \mathbf{D}_{\varphi}(\mu^{t+1}_{1:}, \mu^1_{1:}) - \left\langle \widehat{\ell}^t + g^t, \mu^\dagger_{1:} - \mu^{t+1}_{1:} \right\rangle \\
&\quad = \mathbf{D}_{\Psi^{t+1}}(\mu^\dagger_{1:}, \mu^{t+1}_{1:}) + \hat{q}^t - \left\langle \widehat{\ell}^t, \mu^\dagger_{1:} \right\rangle
\end{aligned}$$

https://doi.org/10.52202/079017-1806

where we used $\left\langle g^t, \mu^\dagger_{1:} - \mu^{t+1}_{1:} \right\rangle = 0$ from the orthogonality. Summing over $t \in [T]$ then gives, by telescoping similarly to the general theorem,

$$\begin{aligned}
\sum_{t=1}^{T} \left(\hat{q}^t - \left\langle \widehat{\ell}^t, \mu^\dagger_{1:} \right\rangle \right) &= \sum_{t=1}^{T} \left(\mathbf{D}_{\Psi^t}(\mu^\dagger_{1:}, \mu^t_{1:}) + \mathbf{D}_{\varphi^t}(\mu^\dagger_{1:}, \mu^1_{1:}) - \mathbf{D}_{\Psi^{t+1}}(\mu^\dagger_{1:}, \mu^{t+1}_{1:}) \right) \\
&= \mathbf{D}_{\Psi^{T+1}}(\mu^\dagger_{1:}, \mu^1_{1:}) - \mathbf{D}_{\Psi^{T+1}}(\mu^\dagger_{1:}, \mu^{t+1}_{1:}) \\
&\leq \mathbf{D}_{\Psi^T}(\mu^\dagger_{1:}, \mu^1_{1:})
\end{aligned}$$

Stability term: From Lemma E.2 used with $\mu' = \mu^t$, we get an alternative expression of the stability term

$$\left\langle \widehat{\ell}^t, \mu^t_{1:} \right\rangle - \hat{q}^t = \left\langle \widetilde{\ell}^t, \mu^t_{1:} \right\rangle - \sum_{x \in \mathcal{X}} \mu^t_{1:}(x) q^t(x)$$

This shows the stability term can be decomposed in a positive linear combination

$$\left\langle \widehat{\ell}^t, \mu^t_{1:} \right\rangle - \hat{q}^t = \sum_{x \in \mathcal{X}} \mu^t_{1:}(x) \left[\left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) \right\rangle - q^t(x) \right]$$

and we will individually upperbound each of the terms of the combination. The method is again similar to the general theorem, but locally with the regularized loss. Defining $\Psi^t_x := \alpha^t(x)\Psi_x$ and $\varphi^t_x := \Psi^{t+1}_x - \Psi^t_x$, we have

$$\begin{aligned}
&\left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) \right\rangle - q^t(x) \\
&\quad = \left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) - \mu^{t+1}(\cdot|x) \right\rangle - \mathbf{D}_{\Psi^t_x}(\mu^{t+1}(\cdot|x), \mu^t(\cdot|x)) - \mathbf{D}_{\varphi^t_x}(\mu^{t+1}(\cdot|x), \mu^1(\cdot|x)) \\
&\quad \leq \left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) - \mu^{t+1}(\cdot|x) \right\rangle - \mathbf{D}_{\Psi^t_x}(\mu^{t+1}(\cdot|x), \mu^t(\cdot|x)) \\
&\quad \leq \left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) - \tilde{\mu}^{t+1}(\cdot|x) \right\rangle - \mathbf{D}_{\Psi^t_x}(\tilde{\mu}^{t+1}(\cdot|x), \mu^t(\cdot|x))
\end{aligned}$$

where

$$\tilde{\mu}^{t+1}(\cdot|x) := \operatorname*{arg\,min}_{\tilde{\mu} \in \Omega_x} \left[\left\langle \widetilde{\ell}^t(x, \cdot), \tilde{\mu} \right\rangle + \mathbf{D}_{\Psi^t_x}(\tilde{\mu}, \mu^t(\cdot|x)) \right]$$

By optimality, $\tilde{\mu}^{t+1}(\cdot|x)$ verifies

$$\nabla\Psi^t_x(\tilde{\mu}^{t+1}(\cdot|x)) = \nabla\Psi^t_x(\mu^t(\cdot|x)) - \widetilde{\ell}^t(x, \cdot)$$

and the law of cosines yields

$$\begin{aligned}
0 &= \mathbf{D}_{\Psi^t_x}(\mu^t(\cdot|x), \mu^t(\cdot|x)) \\
&= \mathbf{D}_{\Psi^t_x}(\mu^t(\cdot|x), \tilde{\mu}^{t+1}(\cdot|x)) + \mathbf{D}_{\Psi^t_x}(\tilde{\mu}^{t+1}(\cdot|x), \mu^t(\cdot|x)) - \\
&\qquad \left\langle \nabla\Psi^t_x(\mu^t(\cdot|x)) - \nabla\Psi^t_x(\tilde{\mu}^{t+1}(\cdot|x)), \mu^t(\cdot|x) - \tilde{\mu}^{t+1}(\cdot|x) \right\rangle \\
&= \mathbf{D}_{\Psi^t_x}(\mu^t(\cdot|x), \tilde{\mu}^{t+1}(\cdot|x)) + \mathbf{D}_{\Psi^t_x}(\tilde{\mu}^{t+1}(\cdot|x), \mu^t(\cdot|x)) - \left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) - \tilde{\mu}^{t+1}(\cdot|x) \right\rangle
\end{aligned}$$

Plugging this in the first inequality, we directly get

$$\left\langle \widetilde{\ell}^t(x, \cdot), \mu^t(\cdot|x) \right\rangle - q^t(x) \leq \mathbf{D}_{\Psi^t_x}(\mu^t(\cdot|x), \tilde{\mu}^{t+1}(\cdot|x))$$

and we get the individual upper bounds with

$$\begin{aligned}
\mathbf{D}_{\Psi^t_x}(\mu^t(\cdot|x), \tilde{\mu}^{t+1}(\cdot[x)) &= \alpha^t(x)\mathbf{D}_{\Psi_x}(\mu^t(\cdot|x), \tilde{\mu}^{t+1}(\cdot[x)) \\
&= \alpha^t(x)\mathbf{D}_{\Psi^\star_x}(\nabla\Psi_x(\tilde{\mu}^{t+1}(\cdot[x)), \nabla\Psi_x(\mu^t(\cdot|x))) \\
&= \alpha^t(x)\mathbf{D}_{\Psi^\star_x}\left(\nabla\Psi_x(\mu^t(\cdot|x)) - \frac{1}{\alpha^t(x)}\widetilde{\ell}^t(x, \cdot), \nabla\Psi_x(\mu^t(\cdot|x)) \right)
\end{aligned}$$

□

This upper bound on the estimated regret is then used with the learning rates considered in the main article.

E.2 Optimal rates analysis

We first consider the optimal rates of the main paper.

Theorem E.4. *Using* LocalOMD *with μ^1 as the uniform policy, with the learning rates $\eta^t(x) = \eta/\kappa(\mu^s|x)$ where $\eta = \sqrt{\log(A)\kappa(\mu^s)/(3HT)}$, and with Ψ_x the Shannon entropy $\Psi_x(\mu) = \sum_{a\in\mathcal{A}(x)} \mu(a)\log(\mu(a))$, the regret is bounded with a probability at least $1-\delta$ by*

$$\mathfrak{R}^T_{\min} \leq \left(4+2\sqrt{3}\right) H^{3/2}\sqrt{\log(A)\iota\kappa(\mu^s)T} \quad \textit{where} \quad \iota = \log(2(A_{\mathcal{X}}+1)/\delta)\,.$$

Proof. We apply Theorem E.3, using the relations $\alpha^t(x) = 1/(\mu^s_{1:}(x)\eta^t(x))$ and $\mathbb{I}_{\{x=x^t_h\}}\widetilde{\ell}^t_h = \mu^s_{1:}(x)\widetilde{\ell}^t(x,\cdot)$. We again separately bound the penalty and stability terms.

Penalty term : We will denote by PEN this term defined by

$$\mathrm{PEN} := \sup_{\mu^\dagger\in\Pi_{\min}} \mathbf{D}^{\mathrm{dil}}_{\alpha^T}(\mu^\dagger_{1:},\mu^1_{1:})\,.$$

By definition of the dilated entropy, we have, using that μ^1 is the uniform policy and that the Bregman divergence of the Shannon entropy is the Kullback-Leibler divergence,

$$\begin{aligned}
\mathrm{PEN} &= \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}} \frac{\mu^\dagger_{1:}(x)\kappa(\mu^s|x)}{\eta\mu^s_{1:}(x)} \mathbf{D}_\Psi(\mu^\dagger(\cdot|x),\mu^1(\cdot|x)) \\
&= \frac{1}{\eta}\sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}} \frac{\mu^\dagger_{1:}(x)\kappa(\mu^s|x)}{\mu^s_{1:}(x)} \sum_{a\in\mathcal{A}(x)} \mu^\dagger(a|x)\log(\mu^\dagger(a|x)/\mu^1(a|x)) \\
&\leq \frac{1}{\eta}\sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}} \frac{\mu^\dagger_{1:}(x)\kappa(\mu^s|x)}{\mu^s_{1:}(x)} \sum_{a\in\mathcal{A}(x)} \mu^\dagger(a|x)\log(1/\mu^1(a|x)) \\
&\leq \frac{\log(A)}{\eta}\sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}} \frac{\mu^\dagger_{1:}(x)}{\mu^s_{1:}(x)}\kappa(\mu^s|x) \\
&= \frac{\log(A)}{\eta}\sup_{\mu^\dagger\in\Pi_{\min}} \sum_{h=1}^H \sum_{x\in\mathcal{X}_h} \frac{\mu^\dagger_{1:}(x)}{\mu^s_{1:}(x)} \sup_{\mu'\in\Pi_{\min}} \sum_{x'\in\mathcal{X}|x \text{ is in the history of } x'} \sum_{a'\in\mathcal{A}(x')} \frac{\mu'_{h:}(x',a')}{\mu^s_{h:}(x',a')} \\
&\leq \frac{\log(A)}{\eta}\sum_{h=1}^H \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}_h} \frac{\mu^\dagger_{1:}(x)}{\mu^s_{1:}(x)} \sup_{\mu'\in\Pi_{\min}} \sum_{x'\in\mathcal{X}|x \text{ is in the history of } x'} \sum_{a'\in\mathcal{A}(x')} \frac{\mu'_{h:}(x',a')}{\mu^s_{h:}(x',a')} \\
\text{(by independance)} &= \frac{\log(A)}{\eta}\sum_{h=1}^H \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}_h} \frac{\mu^\dagger_{1:}(x)}{\mu^s_{1:}(x)} \sum_{x'\in\mathcal{X}|x \text{ is in the history of } x'} \sum_{a'\in\mathcal{A}(x')} \frac{\mu^\dagger_{h:}(x',a')}{\mu^s_{h:}(x',a')} \\
&= \frac{\log(A)}{\eta}\sum_{h=1}^H \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}_h} \sum_{x'\in\mathcal{X}|x \text{ is in the history of } x'} \sum_{a'\in\mathcal{A}(x')} \frac{\mu^\dagger_{1:}(x',a')}{\mu^s_{1:}(x',a')} \\
&= \frac{\log(A)}{\eta}\sum_{h=1}^H \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x'\in\mathcal{X}} \sum_{a'\in\mathcal{A}(x')} \frac{\mu^\dagger_{1:}(x',a')}{\mu^s_{1:}(x',a')} \\
&= \frac{\log(A)}{\eta} H\kappa(\mu^s)
\end{aligned}$$

where $\mathcal{X}_h$ is the set of information sets of depth h, the two sums being later merged on the basis of perfect recall. We now look at the stability term.

Stability term : We will denote by STA this term defined by

$$\mathrm{STA} := \sum_{t=1}^T \sum_{x\in\mathcal{X}} \alpha^t(x)\mu^t_{1:}(x)\mathbf{D}^\star_x\left(\nabla\Psi_x(\mu^t_{1:}(\cdot|x)) - \frac{1}{\alpha^t(x)}\widetilde{\ell}^t(x,\cdot), \nabla\Psi_x(\mu^t_{1:}(\cdot|x))\right)$$

https://doi.org/10.52202/079017-1806

We first look at an upper-bound of $\mathbf{D}_x^\star\left(\nabla\Psi_x(\mu_{1:}^t(\cdot|x)) - \frac{1}{\alpha^t(x)}\widetilde{\ell}^t(x,\cdot), \nabla\Psi_x(\mu_{1:}^t(\cdot|x))\right)$. In order to do so, we upper-bound (in the symmetric matrix sense) the Hessian of $\Psi_x^\star$ on $I := \left\{\nabla\Psi_x(\mu_{1:}^t(\cdot|x)) - \frac{\gamma}{\alpha^t(x)}\widetilde{\ell}^t(x,\cdot)\middle|\gamma \in [0,1]\right\}$.

Because $\Psi_x(\mu) = \sum_{a\in\mathcal{A}(x)}\mu(a)\log(\mu(a))$ is the Shannon entropy,

$$\nabla\Psi_x(\mu)(a) = \log(\mu(a)) + 1 \quad \text{and thus} \quad \nabla^2\Psi_x(\mu) = \text{Diag}\{(1/\mu(a))\}_{a\in\mathcal{A}(x)}$$

and the Hessian of $\Psi_x^\star$ is given by

$$\nabla^2\Psi^\star(y) = \nabla^2\Psi_x(y)^{-1} = \text{Diag}\{y(a)\}_{a\in\mathcal{A}(x)}\,.$$

In particular, it is upper bounded on I by the matrix D_μ defined by

$$D_\mu := \text{Diag}\{\mu(a)\}_{a\in\mathcal{A}(x)}$$

This yields that

$$\begin{aligned}\mathbf{D}_x^\star\left(\nabla\Psi_x(\mu_{1:}^t(\cdot|x)) - \frac{1}{\alpha^t(x)}\widetilde{\ell}^t(x,\cdot), \nabla\Psi_x(\mu_{1:}^t(\cdot|x))\right) &\le \frac{1}{2}\|\frac{1}{\alpha^t(x)}\widetilde{\ell}^t(x,\cdot)\|^2_{D^t_\mu(\cdot|x)}\\ &= \frac{1}{2\alpha^t(x)^2}\sum_{a\in\mathcal{A}(x)}\mu^t(a|x)\widetilde{\ell}^t(x,a)^2\end{aligned}$$

which leads to

$$\begin{aligned}\text{STA} &\le \sum_{t=1}^T\sum_{x\in\mathcal{X}}\frac{\mu_{1:}^t(x)}{2\alpha^t(x)}\sum_{a\in\mathcal{A}(x)}\mu^t(a|x)\widetilde{\ell}^t(x,a)^2\\ &= \frac{\eta}{2}\sum_{t=1}^T\sum_{x\in\mathcal{X}}\mathbb{I}_{\{x=x_h^t\}}\frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)}\frac{1}{\kappa(\mu^s|x)}\sum_{a\in\mathcal{A}(x)}\mathbb{I}_{\{a=a_h^t\}}\mu^t(a|x)\widetilde{\ell}_h^t(a)^2\,.\end{aligned}$$

We can first notice from recursively comparing the minimizer $\mu^{t+1}(\cdot|x_h^t)$ with $\mu^t(\cdot|x_h^t)$ that the regularized loss $\widetilde{\ell}_h^t(a_h^t)$, satisfies

$$\widetilde{\ell}_h^t(a_h^t) \le \left\langle\widehat{\ell}^{t,\to x}, \mu_{h+1:}^{t,\to x}\right\rangle\,,$$

re-using the notation at the beginning of the section, because the regularization does not evolve with time. The difficulty is now to upper bound STA with high probability. In order to do so, we use the Lemma B.1 on the sequence $(U^t)_{t\in[T]}$ defined by

$$U^t := \sum_{x\in\mathcal{X}}\mathbb{I}_{\{x=x_h^t\}}\frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)}\frac{1}{\kappa(\mu^s|x)}\sum_{a\in\mathcal{A}(x)}\mathbb{I}_{\{a=a_h^t\}}\mu^t(a|x)\widetilde{\ell}_h^t(a)^2$$

with $\gamma' = \gamma \in (0, 1/(H^2\kappa(\mu^s))]$ and $\delta' = \delta/2$. This yields with probability at least $1-\delta/2$

$$\sum_{t=1}^T U^t \le \sum_{t=1}^T\mathbb{E}\left[U^t\middle|\mathcal{F}^{t-1}\right] + \gamma\sum_{t=1}^T\mathbb{E}\left[(U^t)^2\middle|\mathcal{F}^{t-1}\right] + \iota/\gamma\,.$$

On the one hand, we have, using $\widehat{\ell}_h^t(a_h^t) \le \kappa(\mu^s|x)$ and the previous inequality that

$$\begin{aligned}\mathbb{E}\left[U^t\middle|\mathcal{F}^{t-1}\right] &\le \sum_{x\in\mathcal{X}}p^t(x)\mu^t(x)\sum_{a\in\mathcal{A}(x)}\left\langle\ell^{t,\to x}, \mu_{h:}^{t,\to x}\right\rangle\\ &\le \sum_{x\in\mathcal{X}}p^t(x)\mu^t(x)H\\ &\le H^2\,.\end{aligned}$$

https://doi.org/10.52202/079017-1806

On the other hand, using the same inequality,

$$\begin{aligned}\mathbb{E}\left[(U^t)^2\middle|\mathcal{F}^{t-1}\right] &= \mathbb{E}\left[\left(\sum_{x\in\mathcal{X}}\mathbb{I}_{\{x=x_h^t\}}\frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)}\sum_{a\in\mathcal{A}(x)}\mathbb{I}_{\{a=a_h^t\}}\left\langle\widehat{\ell}_h^t(a),\mu^t(a|x)\right\rangle\right)^2\middle|\mathcal{F}^{t-1}\right]\\
&\leq H\mathbb{E}\left[\sum_{x\in\mathcal{X}}\mathbb{I}_{\{x=x_h^t\}}\frac{\mu_{1:}^t(x)^2}{\mu_{1:}^s(x)^2}\sum_{a\in\mathcal{A}(x)}\mathbb{I}_{\{a=a_h^t\}}\left\langle\widehat{\ell}_h^t(a)^2,\mu^t(a|x)^2\right\rangle\middle|\mathcal{F}^{t-1}\right]\\
&\leq H\kappa(\mu^s)\mathbb{E}\left[\sum_{x\in\mathcal{X}}\mathbb{I}_{\{x=x_h^t\}}\frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)}\sum_{a\in\mathcal{A}(x)}\mathbb{I}_{\{a=a_h^t\}}\left\langle\widehat{\ell}_h^t(a),\mu^t(a|x)\right\rangle\middle|\mathcal{F}^{t-1}\right]\\
&\leq H\kappa(\mu^s)\sum_{x\in\mathcal{X}}p^t(x)\mu^t(x)\sum_{a\in\mathcal{A}(x)}\left\langle\ell^{t,\to x},\mu_{h:}^{t,\to x}\right\rangle\\
&\leq H\kappa(\mu^s)\sum_{x\in\mathcal{X}}p^t(x)\mu^t(x)H\\
&\leq H^3\kappa(\mu^s)\,.\end{aligned}$$

The following upper bound on the stability term thus holds

$$\text{STA} \leq \eta\left(H^2T + \gamma H^3\kappa(\mu^s)T + \frac{\iota}{\gamma}\right)\,.$$

Taking $\gamma = 1/(H^2\kappa(\mu^s))$, we obtain

$$\text{STA} \leq \eta\left(2H^2T + H^2\iota\kappa(\mu^s)\right)$$

As the bound of the theorem trivially holds if $T < \iota\kappa(\mu^s)$ (the regret being bounded by T anyway), we even have assuming $T \geq \iota\kappa(\mu^s)$

$$\text{STA} \leq 3\eta H^2T\,.$$

Conclusion: Combining all the previous bounds, the estimated regret is bounded, with a probability of at least $1-\delta/2$ by

$$\hat{\mathfrak{R}}^T \leq \frac{\log(A)}{\eta}H\kappa(\mu^s) + 3\eta H^2T\,.$$

Taking $\eta = \sqrt{\log(A)\kappa(\mu^s)/(3HT)}$, we obtain

$$\hat{\mathfrak{R}}^T \leq 2\sqrt{3}\,H^{3/2}\sqrt{\log(A)\iota\kappa(\mu^s)T}\,.$$

We finally conclude by combining this bound with Theorem 2.2 for the true regret, using $\delta' = \delta/2$, such that the two inequalities hold with a probability at least $1-\delta$. □

E.3 Adaptive rates analysis

We end this appendix by considering the adaptive setting. We will assume that all regularizers Ψ_x are 1-strongly convex with respect to some norms $\|\cdot\|_x$, and we will define

$$\begin{aligned}C_\Psi &:= \sup_{x\in\mathcal{X},\mu\in\Delta_{A_x}}\mathbf{D}_x(\mu,\mu^1(\cdot|x))\\
C_\Psi^\star &:= \sup_{x\in\mathcal{X},a\in\mathcal{A}_x}\|\mathbb{I}_{\{x,a\}}\|_x^\star\end{aligned}$$

where μ^1 is the initial policy considered in the algorithm and $\mathbb{I}_{\{x,a\}}$ is the loss vector $\ell(x,\cdot)$ equal to 1 for $a\in\mathcal{A}(x)$ and 0 for $a'\in\mathcal{A}(x)\backslash\{a\}$. The following theorem is the formal statement of Theorem 4.1 in the main article. While being quite general, the upper bound is unsurprisingly not as tight as the previous one.

https://doi.org/10.52202/079017-1806

Theorem E.5. *With such regularizers, assume that the learning rates are locally decreasing and let $\lambda_1, \lambda_2 \in \mathbb{R}_{>0}$ be two constants such that for all information set $x \in \mathcal{X}$,*

$$\max_{t\in[T-1]} \left[\frac{1}{\eta^{t+1}(x)} - \frac{1}{\eta^t(x)}\right] \leq \lambda_1 \quad \text{and} \quad 1/\eta^T(x) + \sum_{t=1}^{T} \eta^t(x)\mathbb{I}_{\{x=x_h^t\}} \leq \lambda_2\sqrt{T}$$

Then with a probability at least $1-\delta$, the regret of Algorithm 2 is upper-bounded by

$$\mathfrak{R}_{\max}^T \leq \left[2\left[(1+\lambda_1)\, C_\Psi C_\Psi^\star \kappa(\mu^s)\right]^2 \lambda_2|\mathcal{X}| + 4\sqrt{H\kappa(\mu^s)\iota}\right]\sqrt{T}$$

where $\iota = \log((A_\mathcal{X}+1)/\delta)$.

The proof of this theorem will be based on the following lemma that bounds the regularized loss using the λ_1 constant above.

Lemma E.6. *For all $t \in [T]$ and $h \in [H]$,*

$$\widetilde{\ell}_h^t(a_h^t) \leq (1+\lambda_1 C_\Psi)\kappa(\mu^s|x_h^t)\,.$$

Proof. The proof is done recursively on h, starting from the leaves. Indeed, for $h = H$, the property is immediate as $\widetilde{\ell}_h^t(a_h^t) \leq 1/\mu^s(a_H^t|x_H^t) \leq \kappa(\mu^s|x_H^t)$. If we assume that the property holds for a depth $h > 1$, then

$$\begin{aligned} q_h^t &= \min_{\mu\in\Delta_{\mathcal{A}(x_h^t)}} \left\langle \widetilde{\ell}_h^t, \mu\right\rangle + \frac{1}{\eta^t(x_h^t)}\mathbf{D}_x\left(\mu, \mu^t(\cdot|x_h^t)\right) + \left(\frac{1}{\eta^{t+1}(x_h^t)} - \frac{1}{\eta^t(x_h^t)}\right)\mathbf{D}_x\left(\mu, \mu^1(\cdot|x_h^t)\right) \\ &\leq \left\langle \widetilde{\ell}_h^t, \mu^t(\cdot|x_h^t)\right\rangle + \left(\frac{1}{\eta^{t+1}(x_h^t)} - \frac{1}{\eta^t(x_h^t)}\right)\mathbf{D}_x\left(\mu^t(\cdot|x_h^t), \mu^1(\cdot|x_h^t)\right) \\ &\leq \widetilde{\ell}_h^t(a_h^t) + \lambda_1 C_\Psi\,. \end{aligned}$$

Then

$$\begin{aligned} \widetilde{\ell}_{h-1}^t(a_{h-1}^t) &= (\ell_{h-1}^t + q_h^t)/\mu^s(a_{h-1}^t|x_{h-1}^t) \\ &\leq (1+\lambda_1 C_\Psi + \widetilde{\ell}_h^t(a_h^t))/\mu^s(a_{h-1}^t|x_{h-1}^t) \\ &\leq (1+\lambda_1 C_\Psi)(1+\kappa(\mu^s|x_h^t))/\mu^s(a_{h-1}^t|x_{h-1}^t) \\ &\leq (1+\lambda_1 C_\Psi)\kappa(\mu^s|x_{h-1}^t) \end{aligned}$$

which concludes the induction.

□

Proof. We now prove the theorem. We start with the estimated regret, that we decompose between the penalty term and the stability term using theorem E.3.

Penalty term: The penalty term PEN is bounded by

$$\begin{aligned} \mathrm{PEN} &\leq \sup_{\mu^\dagger\in\Pi_{\min}} \mathbf{D}_{\alpha^T}^{\mathrm{dil}}(\mu_{1:}^\dagger, \mu_{1:}^1) \\ &\leq \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}} \frac{1}{\eta^T(x)} \frac{\mu_{1:}^\dagger(x)}{\mu_{1:}^s(x)} \mathbf{D}_x(\mu^\dagger(\cdot|x), \mu^1(\cdot|x)) \\ &\leq C_\Psi\lambda_2\sqrt{T} \sup_{\mu^\dagger\in\Pi_{\min}} \sum_{x\in\mathcal{X}} \frac{\mu_{1:}^\dagger(x)}{\mu_{1:}^s(x)} \\ &\leq C_\Psi\lambda_2\kappa(\mu^s)\sqrt{T}\,. \end{aligned}$$

Stability term: For the stability term STA, we rely on Lemma E.6 and the 1-strong convexity of Ψ_x with respect to $\|\cdot\|_x$ (see Appendix D.1) and get

https://doi.org/10.52202/079017-1806

$$
\begin{aligned}
\mathrm{STA} &= \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \alpha^t(x) \mu_{1:}^t(x) \mathbf{D}_x^{\star} \left(\nabla \Psi_x(\mu_{1:}^t(\cdot|x)) - \frac{1}{\alpha^t(x)} \widetilde{\ell}^t(x, \cdot), \nabla \Psi_x(\mu_{1:}^t(\cdot|x)) \right) \\
&\leq \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \frac{\mu_{1:}^t(x)}{\alpha^t(x)} \|\widetilde{\ell}^t(x, \cdot)\|_x^{\star^2} \\
&\leq \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \eta^t(x) \mathbb{I}_{\{x = x_h^t\}} \frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)} \|\widetilde{\ell}_h^t\|_x^{\star^2} \\
&\leq [C_\Psi^\star]^2 \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \eta^t(x) \mathbb{I}_{\{x = x_h^t\}} \frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)} \left(\widetilde{\ell}_h^t(a_h^t) \right)^2 \\
&\leq [(1 + \lambda_1)\, C_\Psi C_\Psi^\star]^2 \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \eta^t(x) \mathbb{I}_{\{x = x_h^t\}} \frac{\mu_{1:}^t(x)}{\mu_{1:}^s(x)} \kappa(\mu^s|x)^2 \\
&\leq [(1 + \lambda_1)\, C_\Psi C_\Psi^\star]^2\, \kappa(\mu^s) \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \eta^t(x) \mathbb{I}_{\{x = x_h^t\}} \frac{1}{\mu_{1:}^s(x)} \kappa(\mu^s|x) \\
&\leq [(1 + \lambda_1)\, C_\Psi C_\Psi^\star \kappa(\mu^s)]^2 \sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \eta^t(x) \mathbb{I}_{\{x = x_h^t\}} \\
&\leq [(1 + \lambda_1)\, C_\Psi C_\Psi^\star \kappa(\mu^s)]^2\, \lambda_2 |\mathcal{X}| \sqrt{T}\,.
\end{aligned}
$$

Conclusion: By summing these two upper bounds we get

$$
\begin{aligned}
\hat{\mathfrak{R}}^T &\leq \left[C_\Psi \lambda_2 \kappa(\mu^s) + [(1 + \lambda_1)\, C_\Psi C_\Psi^\star \kappa(\mu^s)]^2\, \lambda_2 |\mathcal{X}| \right] \sqrt{T} \\
&\leq 2\, [(1 + \lambda_1)\, C_\Psi C_\Psi^\star \kappa(\mu^s)]^2\, \lambda_2 |\mathcal{X}| \sqrt{T}\,.
\end{aligned}
$$

The bound is finally obtained using the Theorem 2.2 that holds with a probability of at least $1 - \delta$ and links the estimated regret to the true regret.

□

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims are proven in the theorems or observed in the experiments.

   Guidelines:

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

2. **Limitations**

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

https://doi.org/10.52202/079017-1806

Answer: [Yes]

Justification: Some limitations and possible improvements are mentioned in the conclusion.

Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

   Justification: The proofs of the theorem (and the exact statement of Theorem 4.1) are provided in the Appendix.

   Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: All implemented algorithms are fully specified. An open-access library (OpenSpiel) is used for an implementation of the games.

https://doi.org/10.52202/079017-1806

Guidelines:

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

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

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

Answer: [Yes]

Justification: A link to an anonymous repository with instructions to run the code is provided.

Guidelines:

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

https://doi.org/10.52202/079017-1806

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: The few details necessary to redo the experiments are provided.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: The variance over the multiple simulation is visible on the figures.

   Guidelines:

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

8. **Experiments Compute Resources**

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

   Answer: [NA]

   Justification: The experiments were run on a personal computer as they do not require heavy computations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-1806

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: the research conducted in the paper fully conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: This work is mostly theoretical and we are not aware of any societal impact.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [NA]

https://doi.org/10.52202/079017-1806

Justification: This work presents no risk of misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The original paper behind OpenSpiel is cited. The version used and the license are specified.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

https://doi.org/10.52202/079017-1806

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects

Guidelines:

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

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

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1806