Appendix A DYNAMIC DATA STRUCTURE FOR POLICY EVALUATION

In this section, we describe a sample-efficient data structure that allows us to get an unbiased sample from the stationary distribution of any deterministic policy.

For motivation, consider estimating the average reward of a single policy π, for which we can use CFTP to get a sample s from the stationary distribution $\mu(\pi)$ (Theorem 1). Then we sample $R(s,a)$ and get an unbiased estimate of $\rho(\pi)$. To estimate $\rho(\pi)$ to an accuracy of ε with confidence δ we average $O\left(\frac{1}{\varepsilon^2}\log(1/\delta)\right)$ such samples.

By Theorem 6 it takes $O\left(T^{\pi}_{\text{mix}}|S|\right)$ to get one sample from $\mu(\pi)$, so in total we would need $O\left(\frac{1}{\varepsilon^2}\log(1/\delta)\,T^{\pi}_{\text{mix}}|S|\right)$ to estimate the average reward of each single policy π to an accuracy of ε with confidence δ. Naively, to estimate the average reward of each of the $|A|^{|S|}$ policies separately, we need a fresh set of samples for every policy for a total of $O\left(\frac{1}{\varepsilon^2}\log(1/\delta)\sum_{\pi}(T^{\pi}_{\text{mix}})|S|\right)$ samples.

Instead, we propose to allow estimates of different policies to share samples by maintaining a matrix D that we use to estimate the reward of any policy π. Each column of D corresponds to a state-action pair. Each row contains, a sample of $R(s,a)$ and a sample $s' \sim P^a(s,\cdot)$ obtained using the MDP's generative model, for each state-action pair (s,a). We get an unbiased estimate of $\rho(\pi)$ for some policy π as follows. We focus on the columns of D that represent pairs $(s,\pi(s))$. The restriction of each row to these columns gives a random mapping from states to next states in the Markov chain induced by π. We now use these samples to run CFTP on this Markov chain, where row t gives the random mapping f_{-t} of Algorithm 1. CFTP gives a sample s from $\mu(\pi)$ and then from the entry in D to which all simulations coalesce. The sample $R(s,\pi(s))$ is an unbiased sample of $\rho(\pi)$.

The matrix D is empty at the beginning and we add rows to it on demand when we estimate $\rho(\pi)$ for a policy π. To analyze the expected size of D, observe that $n = O(T^{\pi}_{\text{mix}}|S|)$ rows are needed to get an unbiased estimate of $\rho(\pi)$ (Theorem 6). This, in turn requires $O(n|S||A|)$ calls to the generative model (to fill these n rows of D).

To get unbiased samples from the Markov chain of a different policy π', we use the rows of D that were already generated for estimates of previous policies (restricted to a different set of columns). If there are not enough rows for Algorithm 1 to give a sample from $\mu(\pi')$, we add rows to D until coalescence occurs. To get ε-approximate estimates with confidence $1-\delta$ for a set of policies Π we need to maintain $O\left(\frac{1}{\varepsilon^2}\log\left(|\Pi|/\delta\right)\right)$ independent copies of D and average the unbiased estimates that they return.

In summary, the number of rows that we add to D depends on the largest mixing time of a policy, which we evaluate and on the approximation guarantee ε and confidence requirement δ. Theorem 10 states the overall sample complexity of D for evaluating the reward of a set of policies Π.[3]

Theorem 10. *Assume that we use D as described above to estimate $\rho(\pi)$ for every π in a set of policies Π such that with probability at least $1-\delta$, it holds simultaneously for all $\pi \in \Pi$ that $|\tilde{\rho}(\pi)-\rho(\pi)| \le \varepsilon$ where $\tilde{\rho}(\pi)$ is our estimate of $\rho(\pi)$. Then the expected number of calls made to the generative model is $O\left(n|S|^2|A|\bar{T}_{mix}\right)$, where $n = \frac{1}{\varepsilon^2}\log\left(|\Pi|/\delta\right)$ and $\bar{T}_{mix}$ is an upper bound on the mixing time of all policies in Π.*

Notice that when Π is the set of all deterministic policies, then $|\Pi| = |A|^{|S|}$ and the sample complexity is $O\left(\frac{1}{\varepsilon^2}|S|^3|A|\log\left(|A|/\delta\right)\bar{T}_{\text{mix}}\right)$. This reveals the advantage of using the dynamic data structure: We can estimate the reward of exponentially many policies with a polynomial number of samples.

Proof. Let $Z_i = \rho(\pi)_i - \rho(\pi)$. Then $\mathbb{E}(Z_i) = 0$, and $|Z_i| \le 1$. Chernhoff bound implies that given independent random variables $Z_1, ..., Z_n$ where $|Z_i| \le 1$, $\mathbb{E}Z_i = 0$, then $\text{Prob}(\sum_{i=1}^n Z_i > a) < e^{-\frac{a^2}{2n}}$. Hence, Chernoff bound implies that (for any n) $\text{Prob}(\sum_{i=1}^n Z_i > \frac{n\varepsilon}{2}) < e^{-\frac{\varepsilon^2 n}{8}}$. This implies that $\text{Prob}(\sum_{i=1}^n \left(\rho(\pi)_i - \rho(\pi)\right) > \frac{n\varepsilon}{2}) = \text{Prob}(\widetilde{\rho}(\pi)-\rho(\pi) > \frac{\varepsilon}{2}) < e^{-\frac{\varepsilon^2 n}{8}}$.

Similarly, we can define $Z_i = \rho(\pi)-\widetilde{\rho}(\pi)$ and get that $\text{Prob}(\rho(\pi)-\widetilde{\rho}(\pi) > \frac{\varepsilon}{2}) < e^{-\frac{\varepsilon^2 n}{8}}$. Hence, we get that $\text{Prob}(|-\rho(\pi)| > \frac{\varepsilon}{2}) < 2e^{-\frac{\varepsilon^2 n}{8}}$. So far we have restricted our attention to a fixed policy π. Using the so-called union bound , we have that the probability that some $\pi \in \Pi$ deviates by more than $\frac{\varepsilon}{2}$ is bounded by $2me^{-\frac{\varepsilon^2 n}{8}}$. Plugging $n = -\frac{8}{\varepsilon^2}\ln\left(\frac{\delta}{2m}\right)$ concludes our proof. □

[3] Notice that while the sample complexity depends on the maximum mixing time of a policy in Π our algorithm does not need to know it.

Appendix B A simple proof of the Propp-Wilson theorem

Theorem (Restatement of Theorem 6)**.** Let μ be the stationary distribution of an ergodic Markov chain with $|S|$ states. We run $|S|$ simulations of the chain, each starting at a different state. When two or more simulations coalesce, we merge them into a single simulation. With probability at least $1-\delta$, all $|S|$ chains are merged after at most $512|S|T_{\text{mix}}\log(1/\delta)$ iterations.

The proof of this Theorem is as follows.

We split time into blocks of size T_{mix}. By the end of the first block, each chain is distributed with some distribution P for which $\text{TV}[P,\mu] \leq 1/8$. We check which of the chains arrive at the same state; chains that do—coalesce. Next, we utilize the Markov property and condition on the states arrived by the chains. On this event, we continue simulating the chains until the end of the next block and continue in this manner.

This is analogous to the following balls-and-bins process. We have $|S|$ balls and $|S|$ bins where the balls simulate the chains, and the bins simulate the states. Each ball j has a distribution P_j over the bins where $\text{TV}[P_j,\mu] \leq 1/8$. We throw the balls into the bins. After that, take one ball out of each nonempty bin and discard the remaining balls. We throw the balls taken out again, and repeat this process until we are left with a single ball.

The following Lemma shows a bound on the expected number of balls removed at each iteration.

Lemma 11. *Assume $2 \leq m \leq n$. Suppose each ball $j = 1,\ldots,m$ is distributed by P_j, and that there is a distribution μ such that $TV[P_j,\mu] \leq 1/8$ for all $j = 1,\ldots,m$. Then the expected number of nonempty bins is at most $m - m^2/256n$.*

Proof. Suppose we throw the balls one by one into the bins. We say that a ball *coalesces* if it is thrown into a nonempty bin. Thus, the number of nonempty bins by the end of the process is exactly m minus the total number of coalescences. Hence we proceed by lower bounding the expected number of coalescences.

We split the balls into two disjoint groups of (roughly) equal sizes: M of size $\lceil m/2 \rceil$ and M^c of size $\lfloor m/2 \rfloor$. We first throw the balls in M and *thereafter* the balls in M^c. The total number of coalescences is, therefore, lower bounded by the number of coalescences that occur between the balls in M^c and those in M. We continue by showing that the probability of a ball in M^c to coalesce with any ball in M is at least $m/64n$. Then, the expected total number coalescences is at least

$$|M^c| \cdot \frac{m}{64n} = \left\lfloor \frac{m}{2} \right\rfloor \cdot \frac{m}{64n} \geq \frac{m^2}{256n},$$

since $m \geq 2$.

Indeed, let Q_k denote the probability distribution over the bins of some ball $k \in M^c$, and P_j denote the probability distribution of $j \in M$. Ball k coalesces with a ball in M if it was thrown into a bin that was not empty after the first phase. Thus, we split the bins into two groups: those who are likely to be empty after the first phase, and those that are not. Let $S = \{i \in [n] : \sum_{j\in M} P_j(i) \leq 1\}$. The proof continues differently for two cases: either k is likely to be thrown into into a bin in S or not. If $Q_k(S) \geq 1/2$, Lemma 14 below states that the probability of a coalescence is at least

$$\frac{|M|}{32n} \geq \frac{m}{64n}$$

since $|M| \geq m/2$. If $Q_k(S) < 1/2$, Lemma 15 found below implies that the probability of a coalescence is at least $1/4 \geq m/64n$. □

Having proven Lemma 11, it remains to use it to show that the expected number of iterations is $O(n)$, which we prove in the following Lemma.

Lemma 12. *Suppose we have $|S|$ balls distributed by P_j, $j = 1,\ldots,n$, where $TV[P_j,\mu] \leq 1/8$. Throw the balls into the bins. Thereafter, take one ball out of each nonempty bin and throw these balls again. Let m_t be the number of balls remaining at iteration t, where $m_0 = n$. Then, $\mathbb{E}m_t \leq 256n/t$.*

Proof. We show that $\mathbb{E}m_t \leq 256n/t$. Denote $\bar{m}_s = \mathbb{E}m_s$. By Jensen's inequality and Lemma 11,

$$\bar{m}_s \leq \mathbb{E}m_{s-1} - \frac{\mathbb{E}m_{s-1}^2}{256n} \leq \bar{m}_{s-1} - \frac{\bar{m}_{s-1}^2}{256n} \leq \bar{m}_{s-1} - \frac{\bar{m}_{s-1}\bar{m}_s}{256n}$$

as $\bar{m}_s \leq \bar{m}_{s-1}$ in particular. Dividing both sides of the inequality by $\bar{m}_s \bar{m}_{s-1}$ gives

$$\frac{1}{\bar{m}_{s-1}} \leq \frac{1}{\bar{m}_s} - \frac{1}{256n} .$$

By summing over $s = 1, \ldots, t$ we obtain

$$\frac{1}{\bar{m}_0} \leq \frac{1}{\bar{m}_t} - \frac{t}{256n} .$$

Finally, we use $\bar{m}_0 \geq 0$ and rearrange the inequality above to gets the claim of the Lemma. □

With the Lemma at hand, the proof of Theorem 6 is as follows. After $512n$ iterations, the process is complete with probability at least $\frac{1}{2}$ by Markov's inequality. If it is not done, we condition on the remaining set of balls and run the process for another $512n$ iterations. Once again, the process is complete with probability at least $\frac{1}{2}$. Repeating this procedure for $\log_2(1/\delta)$ times, we conclude that the procedure is complete with probability at least $1 - \delta$. This finishes the proof of Theorem 6. □

We finish this Section by proving Lemmas 14 and 15. We begin with Lemma 13 that is needed for the proof of Lemma 14.

Lemma 13. *Let P and Q be two distribution on $\{1, \ldots, n\}$ such that $TV[P, Q] \leq 1/4$. Let $S \subseteq [n]$ be such that $Q(S) \geq \frac{1}{2}$. Let x be an element drawn from P and let y be an element draws from Q such that x and y are independent. Then $\Pr[\exists i \in S : x = y = i] \geq 1/16n$.*

Proof. Define $B = \{i : P(i) > Q(i)\}$. Then

$$\begin{aligned}
\Pr[\exists i \in S : x = y = i] &= \sum_{i \in S} P(i)Q(i) \\
&= \sum_{i \in S \cap B} P(i)Q(i) + \sum_{i \in S \cap B^c} P(i)Q(i) \\
&\geq \sum_{i \in S \cap B} Q^2(i) + \sum_{i \in S \cap B^c} P^2(i) \\
&\geq \frac{\big(Q(S \cap B) + P(S \cap B^c)\big)^2}{|S|} \\
&= \frac{\big(Q(S) - \big(Q(S \cap B^c) - P(S \cap B^c)\big)\big)^2}{|S|} \\
&\geq \frac{\big(Q(S) - \mathrm{TV}[P, Q]\big)^2}{n} \\
&\geq \frac{\big(1/2 - 1/4\big)^2}{n} \\
&= \frac{1}{16n} ,
\end{aligned}$$

where the fourth derivation follows from the Cauchy-Schwarz inequality. □

Lemma 14. *Suppose we first throw a set of balls $j \in M$ with probability distributions P_j. Thereafter, we throw an additional ball with probability distribution Q such that $TV[P_j, Q] \leq 1/4$ for every $j \in M$. Additionally, assume that $Q(S) \geq 1/2$ for $S = \{i \in [n] : \sum_{j \in M} P_j(i) \leq 1\}$. Then, the probability that Q is thrown into a nonempty bin is at least $|M|/32n$.*

Proof. The probability that bin $i \in S$ is nonempty is

$$1 - \prod_{j \in M} \big(1 - P_j(i)\big) \geq 1 - \exp\left(- \sum_{j \in M} P_j(i) \right) \geq \big(1 - e^{-1}\big) \sum_{j \in M} P_j(i) \geq \frac{1}{2} \sum_{j \in M} P_j(i) ,$$

using the inequality $1 - x \leq e^{-x}$ and $1 - e^{-x} \geq \left(1 - e^{-1}\right) x$ that holds for any $x \in [0, 1]$. The probability that Q is thrown into a nonempty bin is at least that of it being thrown into a nonempty bin $i \in S$. This is is at least

$$\sum_{i \in S} Q(i) \cdot \frac{1}{2} \sum_{j \in M} P_j(i) = \frac{1}{2} \sum_{j \in M} \sum_{i \in S} Q(i) P_j(i) \,,$$

where $\sum_{i \in S} Q(i) P_j(i)$ is the probability that both Q and P_j end up in to same bin in S. As $\mathrm{TV}[P_j, Q] \leq 1/4$ and $Q(S) \geq 1/2$, Lemma 13 implies that the latter probability is at least $1/16n$. Therefore, the probability of that the additional ball is thrown into a nonempty bin is at least

$$\frac{1}{2} |M| \cdot \frac{1}{16n} = \frac{|M|}{32n} \,.$$

□

Lemma 15. *Suppose with first throw a set of balls M with probability distributions P_j over the bins for every $j \in M$. Thereafter, we throw an additional ball with probability distribution Q such that $|Q - P_j| \leq 1/4$ for all $j \in M$. Additionally, denote*

$$S = \left\{ i \in [n] : \sum_{j \in M} P_j(i) \leq 1 \right\} \,,$$

and suppose that $Q(S) < 1/2$. Then, the probability that Q is thrown into a nonempty bin is at least $1/4$.

Proof. The probability of bin $i \notin S$ not being empty is

$$1 - \prod_{j \in M} \left(1 - P_j(i)\right) \geq 1 - \exp\left(- \sum_{j \in M} P_j(i) \right) \geq 1 - \exp(-1) \geq \frac{1}{2} \,.$$

The probability that Q is thrown into a nonempty bin is at least its probability of it being thrown into a nonempty bin in S^c which is exactly

$$\frac{1}{2} Q(S^c) \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \,.$$

□

Appendix C Proofs for Section 4

C.1 Multiplicative weights

We begin with a classic result on the Hedge algorithm.

Algorithm 4 Hedge

Input: number of experts k, number of iterations T.
Let $\beta = \sqrt{\frac{\log k}{T}}$
Initialize $W^{(1)}(i) = 1$, for $i = 1, \ldots, k$.
for $t = 1, \ldots, T$ **do**
  Set $w^{(t)}(i) = \frac{W^{(t)}(i)}{\sum_{i=1}^{k} W^{(t)}(i)}$, for $i = 1, \ldots, k$.
  Observe $c_t(i)$, for $i = 1, \ldots, k$.
  Incur loss $\sum_{i=1}^{k} w^{(t)}(i) c_t(i_t)$
  Update weights $W^{(t+1)}(i) = W^{(t)}(i) \cdot \exp\left(-\beta c_t(i)\right)$, $\forall i \in [1, .., k]$.
end for

Theorem 16 ((Freund & Schapire, 1997))**.** *Assume that $0 \leq c_t(i) \leq 1$ for all $t = 1, \ldots, T$. Hedge (Algorithm 4) satisfies that for any strategy $w \in \Delta_k$:*

$$\sum w_t \cdot c_t - \sum w \cdot c_t \leq 2\sqrt{T \log k}.$$

Note that in Algorithm 2 and in Algorithm 3, we actually run the Hedge algorithm with the estimates $\tilde{g}_t(i)$ as the costs $c_t(i)$. We obtain $\tilde{g}_t(i)$ by shifting and scaling $g_t(i)$, so that $\tilde{g}_t(i) \in [0, 1]$ and we can apply Theorem 16.

Corollary 17. *Let $-B \leq g_t(i) \leq B$, and $\tilde{g}_t(i) = (g_t(i) + B)/2B$. Assume that we run the Hedge algorithm with costs $c_t(i)$ equal to $\tilde{g}_t(i)$. We have that*

$$\frac{1}{T}\left(\sum w_t \cdot g_t - \min_{w \in \Delta_k} \sum w \cdot g_t\right) \leq 4B\sqrt{\frac{\log k}{T}},$$

Proof. The losses $\tilde{g}_t(i)$ satisfy the conditions of Theorem 16. Therefore,

$$\sum w_t \cdot \tilde{g}_t - \min_{w \in \Delta_k} \sum w \cdot \tilde{g}_t \leq 2\sqrt{T \log k}.$$

This implies that

$$\sum w_t \cdot (g_t + B\mathbf{1})/2B - \min_{w \in \Delta_k} \sum w \cdot (g_t + B\mathbf{1})/2B \leq 2\sqrt{T \log k},$$

where $\mathbf{1}$ denotes a vector of ones. Multiplying by $2B$ gives

$$\sum w_t \cdot (g_t + B\mathbf{1}) - \min_{w \in \Delta_k} \sum w \cdot (g_t + B\mathbf{1}) \leq 4B\sqrt{T \log k}.$$

Observing that $\forall w \in \Delta_k$, $w \cdot B\mathbf{1} = B$ we get that

$$\sum w_t \cdot g_t + B - \min_{w \in \Delta_k} \sum w \cdot g_t - B \leq 4B\sqrt{T \log k}$$

as stated. □

C.2 Estimating the feature expectations of the expert

We begin this subsection with Lemma 18 that bounds the number of samples needed from the expert in order to get a good approximation of the expectations of its features.

Lemma 18. *For any ε, δ, given $m \geq \frac{2\ln(2k/\delta)}{\varepsilon^2}$ samples from the stationary distribution π^E, with probability at least $1 - \delta$, the approximate feature expectations $\hat{\Phi}_E$ satisfy that $\|\hat{\Phi}_E - \Phi_E\|_\infty \leq \varepsilon$.*

Proof. By Hoeffding's inequality we get that

$$\forall i \in [1,..,k] \;\; \Pr(|\hat{\Phi}_E(i) - \Phi_E(i)| \geq \varepsilon) \leq 2\exp(-m\varepsilon^2/2).$$

Applying the union bound over the features we get that

$$\Pr(\exists i \in [1,..,k], s.t., |\hat{\Phi}_E(i) - \Phi_E(i)| \geq \varepsilon) \leq 2k\exp(-m\varepsilon^2/2).$$

This is equivalent to

$$\Pr(\forall i \in [1,..,k] \;\; |\hat{\Phi}_E(i) - \Phi_E(i)| \leq \varepsilon) \geq 1 - 2k\exp(-m\varepsilon^2/2).$$

and to

$$\Pr(\|\hat{\Phi}_E - \Phi_E\|_\infty \leq \varepsilon) \geq 1 - 2k\exp(-m\varepsilon^2/2).$$

The Lemma now follows by substituting the value of m. □

Theorem (8)**.** Assume we run Algorithm 2 for $T = \frac{144\log k}{\varepsilon^2}$ iterations, using $m = \frac{18\log(2k/\delta)}{\varepsilon^2}$ samples from $\mu(\pi^E)$. Let $\bar{\psi}$ be the mixed policy returned by the algorithm. Let $v^\star$ be the game value as in Eq. (3). Then, we have that $\rho(\bar{\psi}) - \rho(\pi^E) \geq v^\star - \varepsilon$ with probability at least $1 - \delta$, where ρ is the average of any reward of the form $r(s) = w \cdot \phi(s)$ where $w \in \Delta_k$.

Proof. Corollary 17 with $B = 1$ and $T = \frac{144\log k}{\varepsilon^2}$ gives that

$$\frac{1}{T}\left(\sum w_t \cdot g_t - \min_{w\in\Delta_k}\sum w \cdot g_t\right) \leq \frac{\varepsilon}{3}, \tag{4}$$

where $g_t(i) = \Phi(\pi^{(t)})[i] - \widetilde{\Phi}^E[i]$. Note also that Lemma 18 with $m = \frac{18\log(2k/\delta)}{\varepsilon^2}$ gives that $\|\hat{\Phi}_E - \Phi_E\|_\infty \leq \frac{\varepsilon}{3}$, which implies that, for any $w \in \Delta_k$:

$$w \cdot \hat{\Phi}_E \leq w \cdot \Phi_E + \varepsilon/3, \tag{5}$$

and

$$w \cdot \Phi_E \leq w \cdot \hat{\Phi}_E + \varepsilon/3, \tag{6}$$

Now, let $\bar{w} = \frac{1}{T}\sum_{t=1}^T w^{(t)}$, and recall that $\bar{\psi}$ is the mixed policy that assigns probability $\frac{1}{T}$ to $\pi^{(t)}$ for all $t \in \{1,\ldots,T\}$. Thus,

$$\begin{aligned}
v^\star &= \max_{\psi\in\Psi}\min_{w\in\Delta_k}\left[w\cdot\Phi(\psi) - w\cdot\Phi_E\right] \\
&= \min_{w\in\Delta_k}\max_{\psi\in\Psi}\left[w\cdot\Phi(\psi) - w\cdot\Phi_E\right] && \text{(von Neumann's minimax theorem)} \\
&\leq \min_{w\in\Delta_k}\max_{\psi\in\Psi}\left[w\cdot\Phi(\psi) - w\cdot\hat{\Phi}_E\right] + \varepsilon/3 && \text{(Eq. (5))} \\
&\leq \max_{\psi\in\Psi}\left[\bar{w}\cdot\Phi(\psi) - \bar{w}\cdot\hat{\Phi}_E\right] + \varepsilon/3 \\
&= \max_{\psi\in\Psi}\frac{1}{T}\sum_{t=1}^T\left[w^{(t)}\cdot\Phi(\psi) - w^{(t)}\cdot\hat{\Phi}_E\right] + \varepsilon/3 && \text{(Definition of } \bar{w}) \\
&\leq \frac{1}{T}\sum_{t=1}^T\max_{\psi\in\Psi}\left[w^{(t)}\cdot\Phi(\psi) - w^{(t)}\cdot\hat{\Phi}_E\right] + \varepsilon/3 \\
&= \frac{1}{T}\sum_{t=1}^T\left[w^{(t)}\cdot\Phi(\pi^{(t)}) - w^{(t)}\cdot\hat{\Phi}_E\right] + \varepsilon/3 && (\pi^{(t)} \text{ is optimal w.r.t the reward } w^{(t)}) \\
&\leq \frac{1}{T}\min_{w\in\Delta_k}\sum_{t=1}^T\left[w\cdot\Phi(\pi^{(t)}) - w\cdot\hat{\Phi}_E\right] + 2\varepsilon/3 && \text{(Eq. (4))} \\
&= \min_{w\in\Delta_k}\left[w\cdot\Phi(\bar{\psi}) - w\cdot\hat{\Phi}_E\right] + 2\varepsilon/3 && \text{(Definition of } \bar{\psi}) \\
&\leq \min_{w\in\Delta_k}\left[w\cdot\Phi(\bar{\psi}) - w\cdot\Phi_E\right] + \varepsilon && \text{(Eq. (6))} \\
&\leq w^\star\cdot\Phi(\bar{\psi}) - w^\star\cdot\Phi_E + \varepsilon && \text{(For any } w^* \in \Delta_k) \\
&= \rho(\bar{\psi}) - \rho(\pi^E) + \varepsilon.
\end{aligned}$$

□

C.3 Estimating the game matrix directly

In this section we prove Theorem 9. Our proof uses the following version of Azuma's concentration bound.

Lemma 19 (Azuma inequality)**.** *Let $\{y_t\}_{t=1}^T$ be a sequence of random variables such that $-b \le y_t \le b$, for $1 \le t < T$. Let $E_t = y_t - \mathbb{E}[y_t \mid y_1, ..., y_{t-1}]$ be the martingale difference sequence defined over the sequence $\{y_t\}_{t=1}^T$. Then*

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} E_t\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{T\varepsilon^2}{8b^2}\right)$$

Theorem (Restatement of Theorem 9)**.** Assume we run Algorithm 3 for T iterations, and there exists a parameter b, such that for any ℓ, $\Pr(\|g_t\|_\infty \ge \ell \cdot b) \le e^{-\ell}$. Let $\bar{\psi}$ be the mixed policy returned by the algorithm. Let $v^\star$ be the game value as in Eq. (3). Then, there exists a constant c such that for $T \ge cB\log^2 B$ where $B = \frac{b^2 \log^3 k \log^2(1/\delta)}{\varepsilon^2}$, we have that $\rho(\bar{\psi}) - \rho(\pi^E) \ge v^\star - \varepsilon$ with probability at least $1-\delta$, where ρ is the average of any reward of the form $r(s) = w \cdot \phi(s)$ where $w \in \Delta_k$.

Proof. Let $\ell = \max\{\log\left(\frac{T}{\delta}\right), \log\left(\frac{1}{\varepsilon}\right)\}$. Then for any t we have that $\Pr(\|g_t\|_\infty \ge \ell \cdot b) \le \frac{\delta}{T}$. By the union bound it follows that with probability $1-\delta$ for all times $t = 1, \ldots, T$, we have that $\|g_t\|_\infty \le \ell b$. We denote by $\mathcal{F}$ the subspace of our probability space that includes all runs of the algorithm in which $\|g_t\|_\infty \le \ell b$ for all $t = 1, \ldots, T$. We have that at least $1-\delta$ fraction of the runs of the algorithm are in $\mathcal{F}$.

By the definition of g_t we have that $\mathbb{E}\left[g_t | g_1, ..., g_{t-1}\right] = \Phi(\pi^{(t)}) - \Phi_E$. Furthermore $w^{(t)}$ depends only on $g_1, ..., g_{t-1}$ and not on g_t. It follows that the random variables $E_t = w^{(t)} \cdot g_t - E[w^{(t)} \cdot g_t \mid g_1, ..., g_{t-1}] = w^{(t)} \cdot \left(g_t - (\Phi(\pi^{(t)}) - \Phi_E)\right)$ is a martingale difference sequence. We would like to apply Azuma's inequality to this sequence, but the difficulty is that the variables E_t are unbounded.

To deal with this problem we define new variables $\bar{g}_t$ as follows

$$\bar{g}_t = \begin{cases} g_t & \|g_t(i)\|_\infty \le \ell b, \\ 0 & \text{otherwise}, \end{cases}$$

and we define the martingale difference sequence $\bar{E}_t = w^{(t)} \cdot \bar{g}_t - E[w^{(t)} \cdot \bar{g}_t \mid g_1, ..., g_{t-1}]$. Unfortunately, $E[w^{(t)} \cdot \bar{g}_t \mid g_1, ..., g_{t-1}]$ does not equal to $\Phi(\pi^{(t)}) - \Phi_E$. But we can bound the difference as follows.

$$\begin{aligned} &|E[w^{(t)} \cdot g_t \mid g_1, ..., g_{t-1}] - E[w^{(t)} \cdot \bar{g}_t \mid g_1, ..., g_{t-1}]| \\ &\qquad \le \int_{x=\ell b}^{\infty} \Pr\left(w^{(t)} \cdot g_t > x\right) \mathrm{dx} - \int_{x=-\ell b}^{\infty} \Pr\left(w^{(t)} \cdot g_t < x\right) \mathrm{dx} \\ &\qquad \le \int_{x=\ell b}^{\infty} \Pr\left(\|g_t\|_\infty \ge x\right) \mathrm{dx} \\ &\qquad = \int_{x=\ell}^{\infty} \Pr\left(\|g_t\|_\infty \ge xb\right) \mathrm{dx} \\ &\qquad \le \int_{x=\ell}^{\infty} e^{-x} \mathrm{dx} = e^{-\ell} \le \varepsilon\,, \end{aligned} \tag{7}$$

where the first inequality follows from the formula $E(Y) = \int_{x=0}^{\infty} \Pr(Y > x) - \int_{x=0}^{-\infty} \Pr(Y < x)$ (which is derived from the more familiar formula $E(Y) = \int_{x=0}^{\infty} \Pr(Y > x)$ for a nonnegative variable Y). The second inequality follows since $w \in \Delta_k$ and the last equality follows by the definition of ℓ. By applying Azuma's inequality to the sequence $\bar{E}_t$ we get that

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} \bar{E}_t\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{T\varepsilon^2}{8(\ell b)^2}\right)\,.$$

Our choice of T guarantees that

$$2\exp\left(-\frac{T\varepsilon^2}{8(\ell b)^2}\right) \le \delta\,.$$

So we also have that within the subspace $\mathcal{F}$

$$\Pr_{\mathcal{F}}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\bar{E}_t\right| \geq \varepsilon\right) \leq \frac{\delta}{1-\delta}\,. \tag{8}$$

But in $\mathcal{F}$, $\bar{g}_t = g_t$ and therefore $\bar{E}_t = E_t - E[w^{(t)} \cdot g_t \mid g_1, ..., g_{t-1}] + E[w^{(t)} \cdot \bar{g}_t \mid g_1, ..., g_{t-1}]$. So by Eq. (7)),

$$|E_t - \bar{E}_t| \leq \varepsilon\,. \tag{9}$$

It follows from Equations (8) and (9) that within $\mathcal{F}$:

$$\Pr_{\mathcal{F}}\left(\left|\frac{1}{T}\sum_{t=1}^{T}E_t\right| \geq 2\varepsilon\right) \leq \frac{\delta}{1-\delta}\,. \tag{10}$$

Let $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$, and recall that $\bar{\psi}$ is the mixed policy that assigns probability $\frac{1}{T}$ to $\pi^{(t)}$ for all $t \in \{1, \ldots, T\}$. We have that

$$\begin{aligned}
v^\star &= \max_{\psi\in\Psi}\min_{w\in\Delta_k}\left[w\cdot\Phi(\psi) - w\cdot\Phi_E\right] \\
&= \min_{w\in\Delta_k}\max_{\psi\in\Psi}\left[w\cdot\Phi(\psi) - w\cdot\Phi_E\right] && \text{von Neumann's minimax theorem} \\
&\leq \max_{\psi\in\Psi}\left[\bar{w}\cdot\Phi(\psi) - \bar{w}\cdot\Phi_E\right] \\
&= \max_{\psi\in\Psi}\frac{1}{T}\sum_{t=1}^{T}\left[w^{(t)}\cdot\Phi(\psi) - w^{(t)}\cdot\Phi_E\right] && \text{Definition of } \bar{w} \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\max_{\psi\in\Psi}\left[w^{(t)}\cdot\Phi(\psi) - w^{(t)}\cdot\Phi_E\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[w^{(t)}\cdot\Phi(\pi^{(t)}) - w^{(t)}\cdot\Phi_E\right] && \pi^{(t)} \text{ is optimal w.r.t the reward } w^{(t)}
\end{aligned} \tag{11}$$

Now we continue our derivation assuming that the run of the algorithm is in $\mathcal{F}$. We use Equation (10) and say that with probability $1 - \frac{\delta}{1-\delta}$ the expression in (11) is bounded by

$$\frac{1}{T}\sum_{t=1}^{T} w^{(t)}\cdot g_t + 2\varepsilon\,. \tag{12}$$

Our choice of T also guarantees that

$$4\ell b\sqrt{\frac{\log k}{T}} \leq \varepsilon\,.$$

and therefore for a run in $\mathcal{F}$, the bound on the regret of Hedge in Corollary 17 implies that expression in (12) is bounded by

$$\frac{1}{T}\min_{w\in\Delta_k}\sum_{t=1}^{T} w\cdot g_t + 3\varepsilon \tag{13}$$

Let $w_{\min} \in \Delta_k$ be the vector achieving the minimum in Equation (13). To finish the proof we need to bound Equation (13) with

$$\frac{1}{T}\sum_{t=1}^{T} w_{\min}\cdot\left(g_t - (\Phi(\pi^{(t)}) - \Phi_E)\right)\,. \tag{14}$$

For this we would like to apply Azuma's inequality to each of the k martingale differences sequences $X_t(i) = g_t(i) - \mathbb{E}[g_t(i) \mid g_1, \ldots, g_{t-1}] = g_t(i) - (\Phi(\pi^{(t)}[i] - \Phi_E[i])$. As before, since the $g_t(i)$'s are unbounded we look instead at the martingale sequence $\bar{X}_t(i) = \bar{g}_t(i) - \mathbb{E}[\bar{g}_t(i) \mid g_1, \ldots, g_{t-1}]$.

Unfortunately, as before, $\mathbb{E}[\bar{g}_t(i) \mid g_1, \ldots, g_{t-1}]$ does not equal to $\mathbb{E}[g_t(i) \mid g_1, \ldots, g_{t-1}]$. But we can bound the difference as follows.

$$\begin{aligned}
|\mathbb{E}[g_t(i) - \bar{g}_t(i) \mid g_1, ..., g_{t-1}]| &\leq \int_{x=\ell b}^{\infty} \Pr\left(g_t(i) > x\right) \mathrm{dx} - \int_{x=-\ell b}^{\infty} \Pr\left(g_t(i) < x\right) \mathrm{dx} \\
&\leq \int_{x=\ell b}^{\infty} \Pr\left(\|g_t\|_\infty \geq x\right) \mathrm{dx} = \int_{x=\ell}^{\infty} \Pr\left(\|g_t\|_\infty \geq xb\right) \mathrm{dx} \\
&\leq \int_{x=\ell}^{\infty} e^{-x} \mathrm{dx} = e^{-\ell} \leq \varepsilon \,, \qquad (15)
\end{aligned}$$

where the inequalities follow from the same reasons as in Eq. (7).

By applying Azuma's inequality to the sequence $\bar{X}_t(i)$ we get that $\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} \bar{X}_t(i)\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{T\varepsilon^2}{8(\ell b)^2}\right)$, and our choice of T guarantees that $2\exp\left(-\frac{T\varepsilon^2}{8(\ell b)^2}\right) \leq \frac{\delta}{k}$. So we also have that within the subspace $\mathcal{F}$

$$\Pr_{\mathcal{F}}\left(\left|\frac{1}{T}\sum_{t=1}^{T} \bar{X}_t(i)\right| \geq \varepsilon\right) \leq \frac{\delta}{k(1-\delta)} \,. \qquad (16)$$

But in $\mathcal{F}$, $\bar{g}_t(i) = g_t(i)$ and therefore $\bar{X}_t(i) = X_t(i) - \mathbb{E}[g_t(i) \mid g_1, ..., g_{t-1}] + \mathbb{E}[\bar{g}_t(i) \mid g_1, ..., g_{t-1}]$. So by Eq. (15)),

$$|X_t(i) - \bar{X}_t(i)| \leq \varepsilon \,. \qquad (17)$$

It follows from Equations (16) and (17) that within $\mathcal{F}$:

$$\Pr_{\mathcal{F}}\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_t(i)\right| \geq 2\varepsilon\right) \leq \frac{\delta}{k(1-\delta)} \,. \qquad (18)$$

By applying the union bound over the features we get that

$$\Pr_{\mathcal{F}}\left(\exists i \in [1, .., k], s.t., \left|\frac{1}{T}\sum_{t=1}^{T} X_t(i)\right| \geq 2\varepsilon\right) \leq \frac{\delta}{1-\delta}.$$

This is equivalent to

$$\Pr_{\mathcal{F}}\left(\forall i \in [1, .., k] \;\; \left|\frac{1}{T}\sum_{t=1}^{T} X_t(i)\right| \leq 2\varepsilon\right) \geq 1 - \frac{\delta}{1-\delta}. \qquad (19)$$

Equation (19) implies that with probability $1 - \frac{\delta}{1-\delta}$ in $\mathcal{F}$, for any $w \in \Delta_k$ it holds that:

$$\frac{1}{T}\sum_{t=1}^{T} w \cdot \left(g_t - (\Phi(\pi^{(t)}) - \Phi_E)\right) \leq 2\varepsilon \,.$$

Since it is true for any w, we get that that we can upper bound Equation (13) by

$$\frac{1}{T}\min_{w \in \Delta_k} \sum_{t=1}^{T} \left[w \cdot \Phi(\pi^{(t)}) - w \cdot \Phi_E\right] + 5\varepsilon \,. \qquad (20)$$

The theorem now follows[4] since the expression in the last equation is smaller than $\rho(\bar{\psi}) - \rho(\pi^E) + 5\varepsilon$ where ρ is the average reward of the form $r(s) = w\phi(s)$ for any $w \in \Delta_k$. □

[4]We have to scale down ε by 5. We also have to scale down δ by 3 since our bound fails to hold with probability 3δ. Indeed, with probabilty $\leq \delta$ our run is not in $\mathcal{F}$, and with probability $1 - \delta$ it is in $\mathcal{F}$, and either of the bounds in Equation (12) and (20) fails – which happens with probability $\leq \frac{2\delta}{1-\delta}$.