A Relaxing assumptions

In this section we review the assumptions used throughout the paper and discuss if and how they could be relaxed.

A.1 Assumption 1

We consider how to remove Asm. 1. When $T_{\pi,t}(s) = 0$ we set $\widehat{\mu}_{\pi,t}(s)$ to an arbitrary default value μ_∞.[7] In this case, the prediction loss becomes

$$\mathcal{L}_n(\pi) = \frac{n}{S}\sum_{s\in\mathcal{S}} \mathbb{E}_\pi\big[\frac{\sigma^2(s)}{T_{\pi,n}(s)}|T_{\pi,n}(s) > 0\big] + E(\pi, n), \quad \text{with } E(\pi,n) := \frac{n}{S}\sum_{s\in\mathcal{S}}(\mu_\infty - \mu(s))^2\mathbb{P}(T_{\pi,n}(s) = 0).$$

Asm. 1 makes the simplification that $E(\pi, n) = 0$. In order to deal with the general case, we need to take care of the event $\{\exists s \in \mathcal{S}, T_{\pi,n}(s) = 0\}$ in which at least one state does not have any sample from which we could estimate its mean. An alternative is to consider that we initially have a "fictitious" observation equal to a fixed value at each state, which would introduce a small bias that tends to zero quickly. Another alternative could be to start by running a policy π_0 over the states of the MDP and as soon as each state is visited at least once, we set the time step equal to 1 and begin our analysis. In the framework of the learning algorithm FW-AME, Asm. 1 can be easily replaced in practice by considering an adaptive length τ_1 such that at least one sample of each state is collected at the end of the first episode (which is what we do in the experiments in Sect. 5). The length of this phase would be small as the following result applies.

Proposition 4. *For any policy $\pi \in \Pi^{SR}$, under Asm. 2, the term $E(\pi, n)$ decreases exponentially in n.*

Proof. Let $n > 1/\eta_{\pi,\min}$. Then setting $\epsilon = \eta(s) - 1/n > 0$ yields[8]

$$\begin{aligned}\mathbb{P}\big(T_{\pi,n}(s) = 0\big) &= \mathbb{P}\Big(T_{\pi,n}(s) < (\eta(s) - \epsilon)n\Big) \\ &\le \sqrt{\frac{2}{\eta_{\pi,\min}}}\exp\left(\frac{-n\gamma_{\mathrm{ps}}^\pi\Big(\eta(s) - \frac{1}{n}\Big)^2}{16\eta(s)(1-\eta(s))\Big(1 + \frac{1}{n\gamma_{\mathrm{ps}}^\pi}\Big) + 40\Big(\eta(s) - \frac{1}{n}\Big)}\right).\end{aligned}$$

We thus obtain for any stationary policy π and any budget $n > 1/\eta_{\pi,\min}$

$$E(\pi,n) \le \frac{n}{S}\sum_{s\in\mathcal{S}}(\mu_\infty - \mu(s))^2\sqrt{\frac{2}{\eta_{\pi,\min}}}\exp\left(\frac{-n\gamma_{\mathrm{ps}}^\pi\Big(\eta_{\pi,\min} - \frac{1}{n}\Big)^2}{8\Big(1 + \frac{1}{n\gamma_{\mathrm{ps}}^\pi}\Big) + 40\Big(\eta_{\pi,\max} - \frac{1}{n}\Big)}\right),$$

which proves the result. □

A.2 Assumption 3

The reversibility assumption (Asm. 3) can be removed and Prop. 1, Lem. 1 as well as the proof of Thm. 1 could be easily adjusted to handle the case of non-reversible policies. As a result, the reversibility condition does not need to hold for the algorithm FW-AME and its vanishing regret guarantees. This can be achieved by replacing Prop. 1 with a concentration result adapted from Paulin et al. (2015).

Proposition 5 (Thm. 3.10 and Prop. 3.14 from Paulin et al. (2015))**.** *Let us fix a stationary policy π which induces a time-homogeneous, ergodic Markov chain. We denote by P its transition matrix and by $\hat{P}$ the time-reversal matrix of P. We denote by $\eta_{\pi,\min} = \min_{s\in\mathcal{S}}\eta(s) > 0$ and $\eta_{\pi,\max} = \max_{s\in\mathcal{S}}\eta(s)$ where η is the chain's*

[7] Formally $\mu_\infty = +\infty$ yet we can also set it equal to a suitable finite value depending on the distributions. For example, if the state distributions are Gaussian and the means belong to an interval $[-\mu_{\max}, +\mu_{\max}]$, we can fix $\mu_\infty = 3\sigma_{\max} + \mu_{\max}$, which ensures that the mean estimate computed from one single sample has an overwhelming probability of being more accurate than the default value μ_∞ when there are no samples.

[8] Here we use the more general result for non-reversible chains reported in Prop. 5.

stationary distribution. We consider the pseudo-spectral gap $\gamma_{ps}^{\pi} = \max_{k \geq 1} \gamma(\hat{P}^k P^k)/k > 0$. *For a given state* s *and for every* $\epsilon > 0$

$$\mathbb{P}\Big(|\frac{T_{\pi,n}(s)}{n} - \eta(s)| > \epsilon\Big) \leq \sqrt{\frac{2}{\eta_{\pi,\min}}} \exp\left(\frac{-n\gamma_{ps}^{\pi}\epsilon^2}{16\eta(s)(1-\eta(s))\big(1+1/(n\gamma_{ps}^{\pi})\big) + 40\epsilon}\right).$$

In Sect. 4, the reversibility condition is intrinsically needed to relate the spectral gap with its spectral norm formulation, which is not possible for the pseudo-spectral gap. Nonetheless, rather than assuming that all policies are in the set of reversible stationary randomized policies Π^{SRR}, we could focus on computing a policy $\pi^{\star}_{\mathrm{FMH}}$ belonging to the restricted set Π^{SRR}, thus replacing the assumption with an additional constrained in the optimization problem.

A.3 Assumption 4

We can deal with the case when the MDP transition model p is unknown by following an optimistic approach similar to UCRL (Jaksch et al., 2010). We recall that the optimization problem solved by FW-AME at each episode is indeed equivalent to solving an MDP with known p and reward function set to $\nabla \widehat{\mathcal{L}}_{t_k-1}^{+}(\widetilde{\lambda}_k)$, which is already an optimistic evaluation of the true gradient. Whenever p is unknown, but an estimate and a confidence set are available, we can include the uncertainty of the estimate of p into the optimistic optimization of the MDP. Let us fix an episode k and $t = t_k - 1$ the time step at the end of the previous episode. We introduce the following set that is p-dependent and thus unknown

$$\Lambda_{\underline{\eta}}^{(p)} = \Big\{\lambda \in \Delta(\mathcal{S} \times \mathcal{A}) \ : \ \forall s \in \mathcal{S}, \ \sum_{b \in \mathcal{A}} \lambda(s,b) \geq 2\underline{\eta} \quad \text{and} \quad \sum_{b \in \mathcal{A}} \lambda(s,b) = \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} p(s|s',a)\lambda(s',a)\Big\}.$$

The aim is the solve the following problem

$$\min_{\lambda \in \Lambda_{\underline{\eta}}^{(p)}} \min_{p \in \mathcal{C}_t} \langle c, \lambda \rangle = \sum_{s,a} \nabla \widehat{\mathcal{L}}_t^{+}(\widetilde{\lambda}_k)(s,a)\lambda(s,a).$$

If we define over $\mathcal{S} \times \mathcal{A}$ the (bounded) reward function $r = -\nabla \widehat{\mathcal{L}}_t^{+}(\widetilde{\lambda}_k)$, we notice that the above problem can be reduced to the dual formulation of finding the policy that maximizes the average reward (Sect. 8, Puterman, 1994). As such, it becomes equivalent to solving the following problem

$$\max_{\pi \in \Lambda_{\underline{\eta}}^{(p)}} \max_{p \in \mathcal{C}_t} \rho_{\pi}(p), \tag{18}$$

where $\rho_{\pi}(p)$ is the gain of stationary policy π in the MDP with transition probability function p. The confidence set $\mathcal{C}_t$ defines a set of plausible transition probability functions at time t. Since the reward function is known, this corresponds to a set of plausible MDPs. Problem (18) thus returns an optimal policy in the plausible MDP with the largest gain. Lattimore and Szepesvári (Sect. 38, 2019) explicit the construction of $\mathcal{C}_t$ and explain that the solutions of (18) are guaranteed to exist and can be found efficiently.[9]

While a complete derivation of the regret bound for this algorithm is left for future work, we expect the final result of Thm. 1 to remain unchanged. In fact, the optimal p returned by problem (18) belongs to $\mathcal{C}_t$ so it is close to the real p up to a factor scaling in $1/T_t$ by construction of $\mathcal{C}_t$. Hence, if the number of visits of any state-action pair (and not just the number of any state visit as in the case of known p) is enforced to be proportional to the time step with high probability, then the derivation of the $\widetilde{O}(t^{-1/3})$ rate in the proof of Prop. 1 (cf. App. D) is unchanged.

[9] In a nutshell, the justification comes from introducing the extended Markov decision process $\widetilde{M}$ from Jaksch et al. (2010) and solving the average reward problem on that specific MDP using Extended Value Iteration. The fact that the extended action-sets of $\widetilde{M}$ are infinite is not problematic since $\mathcal{C}_t$ is a convex polytope and has finitely many extremal points; as a result restricting the confidence sets to these points makes the extended MDP finite without changing the optimal policy.

Algorithm 2 FMH

Require: $\eta^\star$ is the optimal stationary distribution of the convex problem (7).
Require: 3 parameters ρ_n, δ_n and $\underline{\eta}$ (typically set respectively to $S/n, 1/n$ and $\min_s \eta^\star(s)/2$).
Compute X_1 the optimal solution of the convex problem $(\mathcal{P}_1)$ with parameters ρ_n, δ_n and $\underline{\eta}$.
Deduce the corresponding state distribution η_1: $\eta_1(s) = \sum_{s' \in \mathcal{S}} X(s, s')$.
Compute the optimal stationary policy of the convex problem $(\mathcal{P}_2)$ with η_1 as target state distribution.

B Faster-Mixing Heuristic FMH

B.1 Derivation of the two-step method

FMH$(\pi^\star, n)$ receives as input a budget n and $\pi^\star$, the optimal solution of (7), and returns a stationary policy $\pi^\star_{\text{FMH}}$ by solving two convex optimization problems. An outline of FMH is provided in Alg. 2.

Step 1. In this step we first remove the stationarity constraint on η w.r.t. the MDP dynamics and replace it by a weaker but easier constraint involving the adjacency matrix of the MDP. Instead of using P as the kernel of the Markov chain associated to a policy, we consider it as a generic *transition matrix* that respects the possible transitions in the MDP, i.e., $P_{ij} = 0$ if $Q_{ij} = 0$. In this case, problem (11) becomes convex in P for a fixed η and convex in η for a fixed P, yet it is non-convex in both P and η. When P is fixed, η has no more degree of freedom (i.e., it can be directly derived from P), thus any framework of alternate minimization cannot be applied here. We notice that the constraint of reversibility $D_\eta P = P^T D_\eta$ is the toughest one to handle, since it involves both P and η and is not convex in P and η. This leads us to introduce the matrix variable $X = D_\eta P \in \mathbb{R}^{S \times S}$ (i.e., $X_{ij} = \eta_i P_{ij}$). The reversibility constraint on P thus simply translates to a symmetry constraint on X. More discussion on the characteristics of the matrix X is for example provided in Hsu et al. (2015). We then obtain the following optimization problem with variable X (and its corresponding η)

$$\begin{aligned} \underset{X,\ \eta}{\text{minimize}} \quad & \mathcal{L}_0(X) \ := \ \sum_{s \in \mathcal{S}} \frac{\sigma^2(s)}{\eta(s)} + \rho_n \frac{1}{1 - \|D_\eta^{-1/2} X D_\eta^{-1/2} - \sqrt{\eta}\sqrt{\eta}^T\|_2} \\ \text{subject to} \quad & X \geq 0, \quad X = X^T, \quad \sum_{j \in \mathcal{S}} X_{ij} = \eta_i \ \forall i \in \mathcal{S}, \quad X_{ij} = 0 \text{ if } Q_{ij} = 0, \quad \eta \geq \underline{\eta}, \quad \eta^T \mathbb{1} = 1. \end{aligned} \tag{19}$$

This problem is still non-convex in X and η. An idea could be to fix η and solve the convex problem in X (or equivalently P). The most straightforward choice for η is to use $\eta^\star$, the optimal stationary distribution of problem (7), and solve the convex problem of finding the fastest mixing Markov chain with stationary distribution $\eta^\star$ (Boyd et al., 2004). However the Markov chains whose stationary distributions are $\eta^\star$ might all mix poorly. Leveraging the intuition behind the regularized problem (11), we give more slack to η in order to find faster mixing Markov chains, at the cost of having $\mathcal{L}(\eta)$ slightly larger than $\mathcal{L}(\eta^\star)$, i.e., at the cost of a slightly worse asymptotic performance. We formalize this trade-off with the a parameter δ_n, which represents how close we allow the stationary distribution η to be to $\eta^\star$ with respect to the ℓ_2-norm (we pick the ℓ_2-norm in order to ensure the convexity of the resulting constraint). We thus focus on solving the following surrogate optimization problem $(\mathcal{P}_1)$

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \mathcal{L}_1(X) \ := \ \sum_{s \in \mathcal{S}} \frac{\sigma^2(s)}{\sum_{j \in \mathcal{S}} X_{sj}} + \rho_n \frac{1}{1 - \|D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2} \\ \text{subject to} \quad & X \geq 0, \quad X = X^T, \quad X_{ij} = 0 \text{ if } Q_{ij} = 0, \\ & \|D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2 \leq 1, \\ & \sum_{(i,j) \in \mathcal{S}^2} X_{ij} = 1, \quad \sum_{j \in \mathcal{S}} X_{ij} \geq \underline{\eta}, \quad \sum_{i \in \mathcal{S}} \big(\sum_{j \in \mathcal{S}} X_{ij} - \eta^\star_i \big)^2 \leq \delta_n^2, \end{aligned} \tag{$\mathcal{P}_1$}$$

where the small positive constant $\underline{\eta}$ should satisfy $\underline{\eta} \leq \min_s \eta^\star(s)$. Prop. 6 guarantees the convexity and feasibility of the optimization problem $(\mathcal{P}_1)$.

Proposition 6. *$(\mathcal{P}_1)$ is convex in X and well-defined for any δ_n.*

Proof. The convexity of $(\mathcal{P}_1)$ is easily obtained from the convexity of the non-regularized problem, the convexity of the function $X \mapsto 1/(1 - \|X\|_2)$ and the convexity of all the constraints. There can however exist some matrices X such that $\|D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2 \geq 1$, thus making $(\mathcal{P}_1)$ either undefined in its objective function (if the norm is equal to 1) or not satisfying one of the constraints. We thus need to ensure that for any fixed δ_n there exists at least one matrix X such that $\|D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2 < 1$ with the remaining constraints satisfied. To do so, we introduce the transition matrix $M^\star = (P^\star + \widehat{P^\star})/2$ with $\widehat{P^\star}$ the time-reversed transition matrix of $P^\star$ which is the transition matrix of the optimal policy for problem (7). Whereas $P^\star$ is not necessarily reversible w.r.t. $\eta^\star$, it is the case for $M^\star$, thus yielding $\mathrm{SLEM}(M^\star) < 1$. We also define $X = D_{\eta^\star} M^\star$. By construction of X, we have $X \geq 0$, $X = X^T$, $\sum_{i,j} X_{ij} = 1$, $X_{ij} = 0$ if $Q_{ij} = 0$ and $\sum_j X_{ij} \geq \underline{\eta}$. Furthermore, we have $\sum_i \big(\sum_j X_{ij} - \eta_i^\star\big)^2 = \sum_i \big(\sum_j \eta_i^\star M_{ij}^\star - \eta_i^\star\big)^2 = \sum_i (\eta_i^\star)^2 \big(\sum_j M_{ij}^\star - 1\big)^2 = 0$ which means that all the constraints are verified. In addition, since $M^\star$ is reversible w.r.t. $\eta^\star$, we have $\|D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2 = \|D_{\eta^\star}^{1/2} M^\star D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2 = \mathrm{SLEM}(M^\star) < 1$. This proves that $(\mathcal{P}_1)$ is well-defined. □

Solving the convex optimization problem $(\mathcal{P}_1)$ yields an optimal matrix $X_1 \in \mathbb{R}^{S \times S}$, from which we easily obtain the associated stationary distribution η_1 as well as the transition matrix of the associated Markov chain P_1

$$\eta_1(s) = \sum_{s' \in \mathcal{S}} X_1(s, s') \quad \text{and} \quad P_1(s, s') = \frac{X_1(s, s')}{\sum_{s'} X_1(s, s')}.$$

Step 2. The distribution η_1 is stationary w.r.t. the Markov chain P_1 (which is expected to have better mixing properties than $P^\star$), but it may not be feasible w.r.t. the MDP dynamics. As a result, we must now find a stationary policy π whose stationary distribution is closest to η_1. This is closely linked to the steady-state control problem from Akshay et al. (2013), where it is proved that for an ergodic MDP the problem of finding a stationary policy given a target stationary state distribution is effectively decidable in polynomial time. If the steady-state control problem admits a solution, such a policy can be computed by simply solving a polynomial-size linear program. More precisely, we seek a policy π in the set of non-negative reals $\{\pi_{s,a} | s \in \mathcal{S}, a \in \mathcal{A}_s\}$ such that

$$\forall s \in \mathcal{S}, \quad \sum_{s' \in \mathcal{S}, a \in \mathcal{A}_{s'}} \eta_1(s') p(s|s', a) \pi_{s',a} = \eta_1(s) \quad \text{and} \quad \sum_{a \in \mathcal{A}_s} \pi_{s,a} = 1.$$

If the steady-state control problem does not admit a solution, we seek a stationary policy whose stationary state distribution is closest to η_1 w.r.t. the ℓ^2-norm by solving the following convex optimization problem $(\mathcal{P}_2)$ in π

$$\begin{aligned} \underset{\pi}{\text{minimize}} \quad & \sum_s \Big(\eta_1(s) - \sum_{s' \in \mathcal{S}, a \in \mathcal{A}_{s'}} \eta_1(s') p(s|s', a) \pi_{s',a}\Big)^2 \\ \text{subject to} \quad & \forall s \in \mathcal{S}, \quad \pi_{s,a} \geq 0 \quad \forall a \in \mathcal{A}_s \quad \text{and} \quad \sum_{a \in \mathcal{A}_s} \pi_{s,a} = 1. \end{aligned} \tag{$\mathcal{P}_2$}$$

Since we do not know in advance if the steady-state control problem admits a solution or not, we directly solve problem $(\mathcal{P}_2)$ which encompasses both cases (its optimal value is 0 if the steady-state control problem admits a solution). This yields a policy denoted $\pi^\star_{\mathrm{FMH}}$.

B.2 A SDP formulation of FMH (FMH-SDP)

We notice that step 1 of FMH is by far the most computationally demanding, due to the complexity of the objective function and constrained set of problem $(\mathcal{P}_1)$. Fortunately, the symmetry constraint on X leads to the symmetry of the matrix $D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T$, which is a very useful property because it becomes easy to compute a subgradient of its spectral norm w.r.t. X (see e.g., Boyd et al. (Sect. 5.1, 2004)). We can thus apply subgradient descent to solve $(\mathcal{P}_1)$. However a projection on the constrained set is required at each step. We thus propose an alternative method to solve problem $(\mathcal{P}_1)$ that is projection-free and hence more computationally efficient. Since this approach uses semidefinite programming, the resulting heuristic is called FMH-SDP.

The key observation is that the regularizing term in $(\mathcal{P}_1)$ partially "takes into account" the non-regularized one through the last constraint $\|\eta - \eta^\star\| \leq \delta_n$. Furthermore, the regularizing term corresponds (up to composition

of a non-decreasing function) to minimizing the spectral norm of a symmetric matrix. Drawing inspiration from Boyd et al. (Sect. 2.3, 2004), we can express it as a semidefinite program (SDP) which can be solved efficiently using standard SDP solvers. Introducing a scalar variable s to bound the spectral norm, step 1 of FMH is replaced by the following SDP problem whose variables are the matrix X and the scalar s

$$\begin{aligned} \underset{X,s}{\text{minimize}} \quad & s \\ \text{subject to} \quad & -sI \preceq D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T \preceq sI \\ & X \geq 0, \quad X = X^T, \quad X_{ij} = 0 \text{ if } J_{ij} = 0 \\ & \sum_{(i,j)\in\mathcal{S}^2} X_{ij} = 1, \quad \sum_{j\in\mathcal{S}} X_{ij} \geq \underline{\eta}, \quad |\big(\sum_{j\in\mathcal{S}} X_{ij}\big)_i - \eta_i^\star| \leq (\delta_n)_i. \end{aligned} \tag{20}$$

FMH-SDP is not only more computationally efficient due to its SDP formulation but it also loses the dependency on the hyper-parameter ρ_n as only δ_n remains.

C Proofs

We first recall the performance loss suffered by the continuous relaxation in the bandit case, where the frequency $T_{\pi,n}/n$ is replaced by an allocation λ in the simplex. In order to keep the notation as consistent as possible, consider a stochastic bandit problem with S arms, let $\Delta_n = \{\eta_n \in [0,1]^S : \eta_n(s) = \frac{T_n(s)}{n}\}$ and Δ be the discrete and continuous simplex over S arms, where $\eta_n(s)$ is the frequency associated to T_n pulls. Since in this case a policy directly selects arms rather than actions, the objective functions $\mathcal{L}_n$ and $\mathcal{L}$ coincide and we can write

$$\mathcal{L}(\eta) = \frac{1}{S}\sum_s \frac{\sigma^2(s)}{\eta(s)},$$

where η may be either a discrete or a continuous allocation. We have the following.

Proposition 7. *Let $\eta_n^\star = \arg\min_{\eta_n\in\Delta_n} \mathcal{L}(\eta_n)$ be the optimal discrete allocation. As computing η_n^* is NP hard, a standard solution is to first compute $\eta^\star = \arg\min_{\eta\in\Delta} \mathcal{L}(\eta)$ and then round it to obtain $\widetilde{\eta}_n$. If $\widetilde{\eta}_n$ is computed using efficient apportionment techniques (Chapter 12, Pukelsheim, 2006), then for any budget $n > 2S$ we have*

$$\mathcal{L}(\widetilde{\eta}_n) - \mathcal{L}(\eta_n^\star) \leq \frac{2}{n}\sum_s \frac{\sigma^2(s)}{\eta^\star(s)} = \frac{2S}{n}\mathcal{L}(\eta^\star).$$

Furthermore, for any $n \geq 4/(S\eta_{\min}^2)$, where $\eta_{\min} = \min_s \eta^\star(s)$ we have

$$\mathcal{L}(\widetilde{\eta}_n) - \mathcal{L}(\eta_n^\star) \leq \frac{8\sigma_{\max}^2}{\eta_{\min}^3 n^2}.$$

Proof. Using efficient apportionment techniques for rounding we have (Lem. 12.8, Pukelsheim, 2006)

$$\min_s \frac{\widetilde{\eta}_n(s)}{\eta^\star(s)} \geq 1 - \frac{S}{n}, \qquad \text{i.e.,} \quad \forall s,\ \widetilde{\eta}_n(s) \geq \eta^\star(s)\Big(1 - \frac{S}{n}\Big),$$

which also implies the other direction as

$$\widetilde{\eta}_n(s) = 1 - \sum_{s'\neq s} \widetilde{\eta}_n(s') \leq 1 - \sum_{s'\neq s} \eta^\star(s') + \sum_{s'\neq s} \eta^\star(s')\frac{S}{n} \leq \eta^\star(s) + \frac{S}{n}.$$

Then we can bound the performance loss of $\widetilde{\eta}_n$ as

$$\mathcal{L}(\widetilde{\eta}_n) - \mathcal{L}(\eta_n^\star) = \mathcal{L}(\widetilde{\eta}_n) - \mathcal{L}(\eta^\star) + \underbrace{\mathcal{L}(\eta^\star) - \mathcal{L}(\eta_n^\star)}_{\leq 0} \leq \frac{1}{S}\sum_s \sigma^2(s)\Big(\frac{1}{\widetilde{\eta}_n(s)} - \frac{1}{\eta^\star(s)}\Big).$$

Under the assumption that $n > 2S$, we can bound each of the summands as

$$\frac{1}{\widetilde{\eta}_n(s)} - \frac{1}{\eta^\star(s)} = \frac{\eta^\star(s) - \widetilde{\eta}_n(s)}{\widetilde{\eta}_n(s)\eta^\star(s)} \leq \frac{\eta^\star(s)S/n}{(\eta^\star(s))^2(1 - S/n)} \leq \frac{2S}{\eta^\star(s)n},$$

which proves the $O(1/n)$ upper bound. Recalling the definition of $\mathcal{L}(\eta^\star)$ we obtain the final statement

$$\mathcal{L}(\widetilde{\eta}_n) - \mathcal{L}(\eta_n^\star) \leq 2\sum_s \frac{\sigma^2(s)}{\eta^\star(s)n} = \frac{2S}{n}\mathcal{L}(\eta^\star).$$

An even faster rate can be obtained exploiting the smoothness of $\mathcal{L}$. Let $\overline{\Delta} = \{\eta \in \Delta : \forall s, \eta(s) \geq \eta_{\min}/2\}$. Since $\widetilde{\eta}(s) \geq \eta^\star(s)(1 - S/n)$, for any any $n > 2S$ we have $\widetilde{\eta}_n, \eta^\star \in \overline{\Delta}$, hence using the $\overline{C}$-smoothness of $\mathcal{L}$ on $\overline{\Delta}$ with $\overline{C} = \frac{2\sigma_{\max}^2}{S(\eta_{\min}/2)^3}$, we can write (Thm. 12.10 Pukelsheim, 2006)

$$\mathcal{L}(\widetilde{\eta}_n) - \mathcal{L}(\eta_n^\star) \leq \frac{\overline{C}}{2}||\widetilde{\eta}_n - \eta^\star||_2^2 \leq \frac{8\sigma_{\max}^2}{\eta_{\min}^3 n^2},$$

which corresponds to an asymptotic rate of $O(1/n^2)$. Since the multiplicative constants are larger than those of the $O(1/n)$ rate, the rate $O(1/n^2)$ effectively starts when n is big enough. A rough bound on n for the second bound to be effectively smaller than the first is obtained by upper-bounding $\mathcal{L}(\eta^\star) \leq \sigma_{\max}^2/\eta_{\min}$ as

$$\frac{8\sigma_{\max}^2}{\eta_{\min}^3 n^2} \leq \frac{2\sigma_{\max}^2 S}{\eta_{\min} n} \iff n \geq \frac{4}{S\eta_{\min}^2},$$

which concludes the proof.

□

Proof of Proposition 1. The first statement is a direct application of the relationship between mixing and spectral gap. For any policy reversible and ergodic policy π, any starting state s' and any state s, we have from Diaconis et al. (Prop. 3, 1991)

$$\left|\mathbb{P}_\pi(s_t = s|s_1 = s') - \eta_\pi(s)\right| \leq \frac{1}{2}\sqrt{\frac{1 - \eta_\pi(s')}{\eta_\pi(s')}}(1 - \gamma_\pi)^t$$

Then the difference between the expected frequency and the stationary distribution is bounded as

$$\begin{aligned}\left|\frac{\mathbb{E}\left[T_{\pi,n}(s)\right]}{n} - \eta_\pi(s)\right| &\leq \frac{1}{n}\sum_{t=2}^{n}\left|\mathbb{P}_\pi(s_t = s|s_1 = \overline{s}) - \eta_\pi(s)\right| \\ &\leq \frac{1}{2\sqrt{\eta_{\min}}n}\sum_{t=1}^{n}(1 - \gamma_\pi)^t \leq \frac{1}{2\sqrt{\eta_{\min}}n\gamma_\pi}.\end{aligned}$$

□

Proof of Proposition 3. The state-action polytope Λ is closed, bounded and convex according to Puterman (Thm. 8.9.4, 1994). The problem (7) is thus convex in λ due to the convexity of the objective function and constraints. It is straightforward that $\pi_{\lambda^\star} \in \Pi^{\text{SR}}$. From Puterman (Thm. 8.8.1, 1994), the stationary distribution $\eta_{\pi_{\lambda^\star}}$ of $\pi_{\lambda^\star}$ is the unique solution of the system of equations $\sum_{s'} P_{\pi_{\lambda^\star}}(s|s')\eta_{\pi_{\lambda^\star}}(s') = \eta_{\pi_{\lambda^\star}}(s)$ (for each state s) subject to $\sum_s \eta_{\pi_{\lambda^\star}}(s) = 1$. Given that $\left(\sum_a \lambda^\star(s,a)\right)_s$ is a solution, it corresponds to the stationary distribution $\eta_{\pi_{\lambda^\star}}$. By contradiction, assume that there exists a policy $\overline{\pi} \in \Pi^{\text{SR}}$ such that $\mathcal{L}(\overline{\pi}, \eta_{\overline{\pi}}) < \mathcal{L}(\pi_{\lambda^\star}, \eta_{\pi_{\lambda^\star}})$. Then define for every state-action pair (s,a) the quantity $\overline{\lambda}(s,a) = \eta_{\overline{\pi}}(s)\overline{\pi}(a|s)$. It is evident that $\overline{\lambda} \in \Delta(\mathcal{S} \times \mathcal{A})$, furthermore for every state s, we have

$$\begin{aligned}\sum_{s',a} p(s|s',a)\overline{\lambda}(s',a) &= \sum_{s',a} p(s|s',a)\eta_{\overline{\pi}}(s)\overline{\pi}(a|s) = \sum_{s'} \eta_{\overline{\pi}}(s')\sum_a p(s|s',a)\overline{\pi}(a|s') \\ &= \sum_{s'} \eta_{\overline{\pi}}(s')P_{\overline{\pi}}(s|s') = \eta_{\overline{\pi}}(s) = \sum_a \overline{\lambda}(s,a),\end{aligned}$$

since by stationarity of the policy $\overline{\pi}$, the Markov chain transition matrix $P_{\overline{\pi}}$ is stationary w.r.t. $\eta_{\overline{\pi}}$. So $\overline{\lambda}$ satisfies the constraint of stationarity of (7), and

$$\mathcal{L}(\overline{\lambda}) = \sum_s \frac{\sigma^2(s)}{\eta_{\overline{\pi}}(s)} = \mathcal{L}(\overline{\pi}, \eta_{\overline{\pi}}) < \mathcal{L}(\pi_{\lambda^\star}, \eta_{\pi_{\lambda^\star}}) = \sum_s \frac{\sigma^2(s)}{\eta_{\pi_{\lambda^\star}}(s)} = \sum_s \frac{\sigma^2(s)}{\sum_a \lambda^\star(s,a)} = \mathcal{L}(\lambda^\star),$$

which contradicts the optimality of $\lambda^\star$ for problem (7) and thus proves that $\pi_{\lambda^\star}$ is the optimal solution of the problem (8). Finally, the upper bound on the smoothness parameter C_η on the restricted set Λ_η is derived using that the maximal eigenvalue of a symmetric block matrix with positive eigenvalues is bounded by above by the sum of maximal eigenvalues of its diagonal blocks. □

Proof of Lemma 1. The proof is a rather direct application of Prop. 1. We first recall the exact formulation of the term $\epsilon_\pi(s,n,\delta)$ in Prop. 1 (see e.g., Hsu et al. (Thm. 3, 2015), Paulin et al. (Thm. 3.8, 2015)):

$$\epsilon_\pi(s,n,\delta) := \sqrt{8\eta_\pi(s)(1-\eta_\pi(s))\frac{\ln(\frac{1}{\delta}\sqrt{\frac{2}{\eta_{\pi,\min}}})}{\gamma_\pi n}} + 20\frac{\ln(\frac{1}{\delta}\sqrt{\frac{2}{\eta_{\pi,\min}}})}{\gamma_\pi n}.$$

Let $\eta_{\pi,n}(s) = \frac{T_{\pi,n}(s)}{n}$ be the empirical frequency of visits to state s. Since we need all following statements to hold simultaneously for all states $s \in \mathcal{S}$ and all stationary policies $\pi \in \Pi^{\mathrm{SR}}$, we need to take a union bound over states and a cover over the action simplex at each state, which leads to tuning $\delta = \delta'/(SA^S)$ in the high-probability guarantees of Prop. 1, which then hold with probability $1-\delta'$. Furthermore, we have the following deterministic bound

$$\Big|\frac{1}{\eta_{\pi,n}(s)} - \frac{1}{\eta_\pi(s)}\Big| \le \max\{n, \frac{1}{\eta_\pi(s)}\},$$

where we used Asm. 1 to ensure that $1/\eta_{\pi,n}(s) \le n$. We introduce the event

$$\mathcal{E}_1(s,n,\delta) = \{\eta_{\pi,n}(s) \ge \eta_\pi(s) - \epsilon_\pi(s,n,\delta)\}.$$

Then we have

$$\begin{aligned}
\Big|\mathbb{E}\Big[\frac{1}{\eta_{\pi,n}(s)} - \frac{1}{\eta_\pi(s)}\Big]\Big| &\le \Big|\mathbb{E}\Big[\Big(\frac{1}{\eta_{\pi,n}(s)} - \frac{1}{\eta_\pi(s)}\Big)\mathbb{I}\{\mathcal{E}_1(s,n,\delta)\}\Big]\Big| + \Big|\mathbb{E}\Big[\Big(\frac{1}{\eta_{\pi,n}(s)} - \frac{1}{\eta_\pi(s)}\Big)\mathbb{I}\{\mathcal{E}_1^{\mathsf{c}}(s,n,\delta)\}\Big]\Big| \\
&\le \Big|\mathbb{E}\Big[\Big(\frac{1}{\eta_{\pi,n}(s)} - \frac{1}{\eta_\pi(s)}\Big)\mathbb{I}\{\mathcal{E}_1(s,n,\delta)\}\Big]\Big| + \max\{n, \frac{1}{\eta_\pi(s)}\}\mathbb{P}\{\mathcal{E}_1^{\mathsf{c}}(s,n,\delta)\} \\
&\le \Big|\mathbb{E}\Big[\frac{\eta_\pi(s) - \eta_{\pi,n}(s)}{\eta_{\pi,n}(s)\eta_\pi(s)}\mathbb{I}\{\mathcal{E}_1(s,n,\delta)\}\Big]\Big| + \max\{n, \frac{1}{\eta_\pi(s)}\}\delta' \\
&\le \frac{\big|\mathbb{E}\big[\eta_\pi(s) - \eta_{\pi,n}(s)\big]\big|}{\eta_\pi(s)\big(\eta_\pi(s) - \epsilon_\pi(s,n,\delta)\big)} + \max\{n, \frac{1}{\eta_\pi(s)}\}\delta' \\
&\le \frac{1}{2\sqrt{\eta_{\min}}n\gamma_\pi\eta_\pi^2(s)}\Big(1 + 2\frac{\epsilon_\pi(s,n,\delta)}{\eta_\pi(s)}\Big) + \max\{n, \frac{1}{\eta_\pi(s)}\}\delta',
\end{aligned}$$

where the last inequality follows from $1/(1-x) \le 1 + 2x$ for $0 < x \le 1/2$ which can be applied due to the condition that n is big enough so that $\epsilon_\pi(s,n,\delta) \le \eta_\pi(s)/2$. Since this condition requires $n \ge O(1/\eta_{\min}^2)$, we can resolve the maximum in the previous expression as $\max\{n, \frac{1}{\eta_\pi(s)}\} \le n$. Finally, setting $\delta' = 1/n^2$ translates to the inequality on the objective function

$$\big|\mathcal{L}_n(\pi) - \mathcal{L}(\pi,\eta_\pi)\big| \le \ell_n(\pi) := \frac{1}{S\sqrt{\eta_{\min}}n\gamma_\pi}\sum_{s\in\mathcal{S}}\frac{\sigma^2(s)}{\eta_\pi^2(s)}\Big(1 + 2\frac{\epsilon_\pi(s,n,\delta)}{\eta_\pi(s)}\Big),$$

from which we obtain the final statement as

$$\begin{aligned}
\mathcal{L}_n(\pi_{\lambda^\star}) - \mathcal{L}_n(\pi_n^\star) &\le \mathcal{L}(\pi_{\lambda^\star},\eta_{\pi_{\lambda^\star}}) + \ell_n(\pi_{\lambda^\star}) - \mathcal{L}(\pi_n^\star,\eta_{\pi_n^\star}) + \ell_n(\pi_n^\star) \\
&\le \ell_n(\pi_{\lambda^\star}) + \ell_n(\pi_n^\star).
\end{aligned}$$

□

Proof of Lemma 2. The proof relies on the concentration inequality in Eq. 9. We proceed through the following inequalities

$$\mathcal{L}_n(\pi_{\mathrm{reg}}^\star) \le \mathcal{L}(\pi_{\mathrm{reg}}^\star,\eta_{\pi_{\mathrm{reg}}^\star}) + \ell_n(\pi_{\mathrm{reg}}^\star) \le \mathcal{L}(\pi_n^\star,\eta_{\pi_n^\star}) + \ell_n(\pi_n^\star) \le \mathcal{L}_n(\pi_n^\star) + 2\ell_n(\pi_n^\star),$$

where in the first and last inequality we used Eq. 9, and where the second inequality follows from the definition of $\pi_{\mathrm{reg}}^\star$ as the optimal solution to the regularized problem. □

Proof of Lemma 3. Introducing the term $H := \mathcal{L}^{\mathrm{reg}}(\pi^\star_{\mathrm{FMH}}) - \mathcal{L}^{\mathrm{reg}}(\pi^\star_{\mathrm{reg}})$ where $\mathcal{L}^{\mathrm{reg}}$ is defined in Eq. 11, we have

$$\mathcal{L}_n(\pi^\star_{\mathrm{FMH}}) \leq \mathcal{L}^{\mathrm{reg}}(\pi^\star_{\mathrm{FMH}}) = H + \mathcal{L}^{\mathrm{reg}}(\pi^\star_{\mathrm{reg}}) \leq H + \mathcal{L}_n(\pi^\star_n) + 2\ell_n(\pi^\star_n).$$

Given the expression of $\ell_n(\pi)$ provided in Lem. 1, we can write

$$\mathcal{L}^{\mathrm{reg}}(\pi, \eta) = \mathcal{L}(\pi, \eta) + \frac{\rho_n}{1 - \|D_\eta^{1/2} P_\pi D_\eta^{-1/2} - \sqrt{\eta}\sqrt{\eta}^\mathsf{T}\|_2} + O(n^{-3/2}).$$

For notational simplicity we denote $\eta_{\mathrm{fmh}} = \eta_{\pi^\star_{\mathrm{FMH}}}$, $P_{\mathrm{fmh}} = P_{\pi^\star_{\mathrm{FMH}}}$, $\eta_{\mathrm{reg}} = \eta_{\pi^\star_{\mathrm{reg}}}$ and $P_{\mathrm{reg}} = P_{\pi^\star_{\mathrm{reg}}}$. We thus have

$$H = \sum_{s\in\mathcal{S}} \Big(\frac{\sigma^2(s)}{\eta_{\mathrm{fmh}}(s)} - \frac{\sigma^2(s)}{\eta_{\mathrm{reg}}(s)}\Big) + \rho_n \Big(\frac{1}{\gamma(P_{\mathrm{fmh}})} - \frac{1}{\gamma(P_{\mathrm{reg}})}\Big) + O(n^{-3/2}).$$

Given that η_{reg} is a stationary state distribution w.r.t. the MDP dynamics, we can write by optimality of $\eta^\star$ for the problem (7)

$$\sum_{s\in\mathcal{S}} \Big(\frac{\sigma^2(s)}{\eta_{\mathrm{fmh}}(s)} - \frac{\sigma^2(s)}{\eta_{\mathrm{reg}}(s)}\Big) \leq \sum_{s\in\mathcal{S}} \Big(\frac{\sigma^2(s)}{\eta_{\mathrm{fmh}}(s)} - \frac{\sigma^2(s)}{\eta^\star(s)}\Big) \leq \frac{\sigma^2_{\max}}{\underline{\eta}^2}\|\eta_{\mathrm{fmh}} - \eta^\star\|_1 \leq \frac{\sigma^2_{\max}\sqrt{S}}{\underline{\eta}^2}\|\eta_{\mathrm{fmh}} - \eta^\star\|_2.$$

Using successively the triangular inequality, the property guaranteed in $(\mathcal{P}_2)$ that η_{fmh} minimizes the distance $\|\cdot - \eta_1\|_2$ among all the stationary state distributions w.r.t. the MDP dynamics, and finally the property guaranteed in $(\mathcal{P}_1)$ of δ_n-proximity of η_1 to $\eta^\star$, we get

$$\|\eta_{\mathrm{fmh}} - \eta^\star\|_2 \leq \|\eta_{\mathrm{fmh}} - \eta_1\|_2 + \|\eta_1 - \eta^\star\|_2 \leq 2\|\eta_1 - \eta^\star\|_2 \leq 2\delta_n.$$

We conclude the proof using the fact that $\gamma(P_{\mathrm{fmh}})$ and $\gamma(P_{\mathrm{reg}})$ are larger than $\gamma_{\min}$. □

D Proof of Thm. 1

D.1 Preliminaries

We recall that the notation $u_n = \widetilde{O}(v_n)$ means that there exist $c > 0$ and $d > 0$ such that $u_n \leq c(\log n)^d v_n$ for sufficiently large n. By abuse of language we say that a stationary policy π belongs to $\Lambda_{\underline{\eta}}$ if $\forall s \in \mathcal{S}, \eta_\pi \geq 2\underline{\eta}$. For notational convenience we consider throughout the proof that we relax Asm. 1 (cf. App. A.1) and that the initial state s_1 is drawn from an arbitrary initial distribution over states and we collect its observation x_1. This leads to the configuration where at every time t exactly t state samples have been collected. We start our analysis with the two following technical lemmas.

Lemma 4. *Let $\delta \in (0,1)$. For any length $\tau > 0$, the following bound holds simultaneously for any state s and any policy $\pi \in \Lambda_{\underline{\eta}}$ with probability at least $1 - \delta$*

$$\Big|\frac{\sum_{t=1}^\tau \mathbb{I}\{\pi_t = s\}}{\tau} - \eta_\pi(s)\Big| \leq M(\tau,\delta) := \sqrt{\frac{2B}{\gamma_{\min}\tau}} + \frac{20B}{\gamma_{\min}\tau} \qquad \textit{with} \quad B = \log\Big(\frac{SA^S}{\delta}\sqrt{\frac{1}{\underline{\eta}}}\Big).$$

Proof. Pick any $\delta \in (0,1)$. Let π be a fixed policy whose stationary distribution is lower-bounded by $\eta_{\pi,\min}$ and whose associated Markov chain admits γ_π as spectral gap. For any length $\tau > 0$ and state s, we define $\nu_{\pi,\tau}(s) = \sum_{t=1}^\tau \mathbb{I}\{\pi_t = s\}$. From Prop. 1, for a fixed state $s \in \mathcal{S}$, the following bound holds with probability at least $1 - \delta$

$$\Big|\frac{\nu_{\pi,\tau}(s)}{\tau} - \eta_\pi(s)\Big| \leq \sqrt{8\eta_\pi(s)(1-\eta_\pi(s))\tilde{\epsilon}} + 20\tilde{\epsilon} \quad \text{where } \tilde{\epsilon} = \frac{\log(\frac{1}{\delta}\sqrt{\frac{2}{\eta_{\pi,\min}}})}{\gamma_\pi \tau}.$$

Since we need this statement to hold simultaneously for all states $s \in \mathcal{S}$ and all stationary policies $\pi \in \Lambda_{\underline{\eta}}$, we need to take a union bound over states and a cover over the action simplex at each state, which leads to tuning $\delta = \delta'/SA^S$ and thus yields with probability at least $1 - \delta$

$$|\frac{\nu_{\pi,\tau}(s)}{\tau} - \eta_\pi(s)| \leq \sqrt{8\eta_\pi(s)(1-\eta_\pi(s))\tilde{\epsilon}} + 20\tilde{\epsilon} \quad \text{where } \tilde{\epsilon} = \frac{\log(\frac{SA^S}{\delta}\sqrt{\frac{1}{\underline{\eta}}})}{\gamma_{\min}\tau}.$$

Using the fact that the function $x \mapsto x(1-x)$ is upper bounded by $1/4$ and setting $B = \log(\frac{SA^S}{\delta}\sqrt{\frac{1}{\underline{\eta}}})$ yields the desired high-probability result. □

Lemma 5. *Let $\delta \in (0,1)$. There exists a length $\tau_\delta > 0$ such that for any $T \geq \tau_\delta$, the following inequality holds simultaneously for any state s and any policy $\pi \in \Lambda_{\underline{\eta}}$ with probability at least $1-\delta$*

$$\sum_{t=1}^{T} \mathbb{I}\{\pi_t = s\} \geq \underline{\eta} T$$

Proof. Pick any $\delta \in (0,1)$. $M(\tau, \delta)$ is a decreasing function of τ, hence there exists a length τ_δ such that for any $T \geq \tau_\delta$, $M(T,\delta) \leq \underline{\eta}$. As a result, Lem. 4 guarantees that we have with probability at least $1-\delta$ simultaneously for any state s and any stationary policy $\pi \in \Lambda_{\underline{\eta}}$

$$|\frac{\sum_{t=1}^{T} \mathbb{I}\{\pi_t = s\}}{T} - \eta_\pi(s)| \leq \underline{\eta},$$

which yields in particular

$$\frac{\sum_{t=1}^{T} \mathbb{I}\{\pi_t = s\}}{T} \geq \eta_\pi(s) - \underline{\eta} \geq \underline{\eta}.$$

□

Restricting our attention to increasing episode lengths in FW-AME and using Lem. 5, we deduce the important property that for any $\delta \in (0,1)$, there exists an episode k_δ such that for all episodes k succeeding it (and including it), we have with probability at least $1-\delta$

$$\sum_{a \in \mathcal{A}} \widetilde{\lambda}_k(s,a) \geq \underline{\eta}, \quad \forall s \in \mathcal{S}, \quad \forall k \geq k_\delta. \tag{21}$$

More specifically, k_δ is the first episode whose length τ_{k_δ} verifies

$$M(\tau_{k_\delta}, \delta) = \sqrt{\frac{2B}{\gamma_{\min}\tau_{k_\delta}}} + \frac{20B}{\gamma_{\min}\tau_{k_\delta}} \leq \underline{\eta} \qquad \text{with} \quad B = \log\Big(\frac{SA^S}{\delta}\sqrt{\frac{1}{\underline{\eta}}}\Big). \tag{22}$$

We proceed by providing time-dependent lower and upper bounds on the true gradient $\nabla\mathcal{L}$, which is unknown. We denote by $\widehat{\mathcal{L}}_t^+$ the empirical optimistic approximation of $\mathcal{L}$ at any time t, i.e.,

$$\widehat{\mathcal{L}}_t^+(\lambda) = \sum_{s \in \mathcal{S}} \frac{1}{\sum_a \lambda(s,a)} \Big[\widehat{\sigma}_t^2(s) + 5R^2\sqrt{\frac{\log(\frac{4St}{\delta})}{T_t(s)}}\Big] = \widehat{\mathcal{L}}_t(\lambda) + \sum_{s \in \mathcal{S}} \frac{\alpha(t,s,\delta)}{\sum_a \lambda(s,a)}.$$

Here we used that $\nu(s)$ is an observation distribution supported in $[0,R]$. We note that this assumption can be easily extended to the general case of sub-Gaussian distributions as done in Carpentier et al. (2011). From Prop. 2, the following inequalities hold with probability at least $1-\delta$ for any λ, time t and state-action pair (s,a)

$$\nabla\widehat{\mathcal{L}}_t^+(\lambda)(s,a) = \nabla\widehat{\mathcal{L}}_t(\lambda)(s,a) - \frac{\alpha(t,s,\delta)}{(\sum_b \lambda(s,b))^2} \leq \nabla\mathcal{L}(\lambda)(s,a) \leq \nabla\widehat{\mathcal{L}}_t(\lambda)(s,a) + \frac{\alpha(t,s,\delta)}{(\sum_b \lambda(s,b))^2}. \tag{23}$$

Finally, let $T = t_K - 1$ be the final budget (i.e., the time at the end of the final episode $K-1$). For the sake of clarity and readability, we make the simplification that the logarithmic term $\log(T)$ behaves as a constant.

D.2 Core of the proof

We denote by ρ_{k+1} the approximation error at the end of each episode k (i.e., at time $t_{k+1}-1$). Recalling that $\beta_k = \tau_k/(t_{k+1}-1)$, we have

$$\rho_{k+1} = \mathcal{L}(\widetilde{\lambda}_{k+1}) - \mathcal{L}(\lambda^\star) = \mathcal{L}\big((1-\beta_k)\widetilde{\lambda}_k + \beta_k \widetilde{\psi}_{k+1}\big) - \mathcal{L}(\lambda^\star).$$

Let $\psi^\star_{k+1} = \operatorname{argmin}_{\lambda\in\Lambda_{\underline{\eta}}} \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \lambda\rangle$ be the state-action stationary distribution that "exact" FW would return at episode k. We have the following series of inequality

$$\begin{aligned}
\rho_{k+1} &\le \mathcal{L}(\widetilde{\lambda}_k) - \mathcal{L}(\lambda^\star) + \beta_k \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widetilde{\psi}_{k+1} - \widetilde{\lambda}_k\rangle + C_{\underline{\eta}}\beta_k^2 \\
&= \mathcal{L}(\widetilde{\lambda}_k) - \mathcal{L}(\lambda^\star) + \beta_k \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \psi^\star_{k+1} - \widetilde{\lambda}_k\rangle + C_{\underline{\eta}}\beta_k^2 + \beta_k \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widetilde{\psi}_{k+1} - \psi^\star_{k+1}\rangle \\
&\le \mathcal{L}(\widetilde{\lambda}_k) - \mathcal{L}(\lambda^\star) + \beta_k \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \lambda^\star - \widetilde{\lambda}_k\rangle + C_{\underline{\eta}}\beta_k^2 + \beta_k \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widetilde{\psi}_{k+1} - \psi^\star_{k+1}\rangle \\
&\le (1-\beta_k)\rho_k + C_{\underline{\eta}}\beta_k^2 + \beta_k \underbrace{\langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widehat{\psi}^+_{k+1} - \psi^\star_{k+1}\rangle}_{\epsilon_{k+1}} + \beta_k \underbrace{\langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widetilde{\psi}_{k+1} - \widehat{\psi}^+_{k+1}\rangle}_{\Delta_{k+1}},
\end{aligned} \tag{24}$$

where the first step follows from the $C_{\underline{\eta}}$-smoothness of $\mathcal{L}$, the second inequality comes from the FW optimization step and the definition of $\psi^\star_{k+1}$, which gives $\langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \psi^\star_{k+1} - \widetilde{\lambda}_k\rangle \le \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \lambda^\star - \widetilde{\lambda}_k\rangle$, the final step follows from the convexity of $\mathcal{L}$. The term ϵ_{k+1} measures the error due to an inaccurate estimate of the gradient and the term Δ_{k+1} refers to the discrepancy between the stationary state-action distribution $\widehat{\psi}^+_{k+1}$ and the empirical frequency $\widetilde{\psi}_{k+1}$ of its realization for τ_k steps.

Step 1 (Bound on error Δ_{k+1}). For any $k \ge k_\delta$, inequality (21) is verified and we can write

$$\langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widetilde{\psi}_{k+1} - \widehat{\psi}^+_{k+1}\rangle = \sum_s \frac{-\sigma^2(s)}{\big(\sum_b \widetilde{\lambda}_k(s,b)\big)^2} \sum_a \big(\widetilde{\psi}_{k+1}(s,a) - \widehat{\psi}^+_{k+1}(s,a)\big) \le \frac{S\sigma^2_{\max}}{\underline{\eta}^2} \Big\| \frac{\nu_{k+1}}{\tau_k} - \eta_{\widehat{\pi}^+_{k+1}} \Big\|_\infty.$$

Let $B = \log(\frac{SA^S}{\delta}\sqrt{\frac{1}{\underline{\eta}}})$. From Lem. 4, we have with probability at least $1-\delta$ simultaneously for every state s and every policy followed during the episode

$$\Big| \frac{\nu_{k+1}(s)}{\tau_k} - \eta_{\widehat{\pi}^+_{k+1}}(s) \Big| \le \sqrt{\frac{2B}{\gamma_{\min}\tau_k}} + \frac{20B}{\gamma_{\min}\tau_k}.$$

Hence we obtain the following bound on Δ_{k+1} with probability at least $1-\delta$

$$\Delta_{k+1} \le \frac{S\sigma^2_{\max}}{\underline{\eta}^2}\Big[\sqrt{\frac{2B}{\gamma_{\min}\tau_k}} + \frac{20B}{\gamma_{\min}\tau_k}\Big].$$

Step 2 (Bound on error ϵ_{k+1}). Using inequality (23), we get with probability at least $1-\delta$

$$\begin{aligned}
\langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \widehat{\psi}^+_{k+1}\rangle &= \sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \nabla\mathcal{L}(\widetilde{\lambda}_k)(s,a) \\
&\le \sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \nabla\widehat{\mathcal{L}}_{t_k-1}(\widetilde{\lambda}_k)(s,a) + \sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \frac{\alpha(t_k-1,s,\delta)}{(\sum_b \widetilde{\lambda}_k(s,b))^2} \\
&\le \sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \nabla\widehat{\mathcal{L}}^+_{t_k-1}(\widetilde{\lambda}_k)(s,a) + 2\sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \frac{\alpha(t_k-1,s,\delta)}{(\sum_b \widetilde{\lambda}_k(s,b))^2} \\
&\le \langle \nabla\widehat{\mathcal{L}}^+_{t_k-1}(\widetilde{\lambda}_k), \psi^\star_{k+1}\rangle + 2\sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \frac{\alpha(t_k-1,s,\delta)}{(\sum_b \widetilde{\lambda}_k(s,b))^2} \\
&\le \langle \nabla\mathcal{L}(\widetilde{\lambda}_k), \psi^\star_{k+1}\rangle + 2\sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \frac{\alpha(t_k-1,s,\delta)}{(\sum_b \widetilde{\lambda}_k(s,b))^2}.
\end{aligned}$$

For notational simplicity we denote by $T_k(s) = T_{t_k-1}(s)$ the number of visits of state s until the end of episode $k-1$ (i.e., at time $t_k - 1$). Using inequality (21) and an intersection bound over two high-probability events, we get with probability at least $1-2\delta$ for any episode $k \geq k_\delta$

$$\begin{aligned}\epsilon_{k+1} &\leq \sum_{s,a} \widehat{\psi}^+_{k+1}(s,a) \frac{10R^2}{\underline{\eta}^2} \sqrt{\log(\frac{4S(t_k-1)}{\delta})} \frac{1}{\sqrt{T_k(s)}} \\ &\leq \underbrace{c_0 \sum_{s,a} \widetilde{\psi}_{k+1}(s,a) \frac{1}{\sqrt{T_k(s)}}}_{v_k} + \underbrace{c_0 \sum_{s,a} \big(\widehat{\psi}^+_{k+1}(s,a) - \widetilde{\psi}_{k+1}(s,a)\big) \frac{1}{\sqrt{T_k(s)}}}_{\xi_{k+1}},\end{aligned}$$

where we define $c_0 = \dfrac{10R^2}{\underline{\eta}^2}\sqrt{\log(\frac{4ST}{\delta})}$. ξ_{k+1} can be bounded in the same vein as Δ_{k+1} using Lem. 4. The error ξ_{k+1} is of a higher order than Δ_{k+1} and for proof simplicity we consider the following loose bound which is satisfied with probability at least $1-\delta$

$$\xi_{k+1} \leq c_0 \Big[\sqrt{\frac{2B}{\gamma_{\min}\tau_k}} + \frac{20B}{\gamma_{\min}\tau_k} \Big].$$

Step 3 (putting everything together in (24)). For $k \geq k_\delta$, we get with probability at least $1-2\delta$

$$\Delta_{k+1} + \xi_{k+1} \leq \frac{c_1}{\sqrt{\tau_k}} + \frac{c_2}{\tau_k} \quad \text{with} \quad \begin{cases} c_1 = \Big(c_0 + \dfrac{S\sigma^2_{\max}}{\underline{\eta}^2}\Big)\sqrt{\dfrac{2B}{\gamma_{\min}}} \\ c_2 = \Big(c_0 + \dfrac{S\sigma^2_{\max}}{\underline{\eta}^2}\Big)\dfrac{20B}{\gamma_{\min}} \end{cases},$$

which provides the bound for $k \geq k_\delta$

$$\rho_{k+1} \leq (1-\beta_k)\rho_k + \beta_k\big(\frac{c_1}{\sqrt{\tau_k}} + \frac{c_2}{\tau_k}\big) + C_{\underline{\eta}}\beta_k^2 + \beta_k c_0 v_k. \tag{25}$$

Choosing episode lengths satisfying $t_k = \tau_1(k-1)^3 + 1$ yields

$$\tau_k = t_{k+1} - t_k = \tau_1(3k^2 - 3k + 1) \geq 3\tau_1 k^2 \quad \text{and} \quad \beta_k = \frac{\tau_k}{t_{k+1}-1} = \frac{3k^2-3k+1}{k^3} \in \Big[\frac{1}{k}, \frac{3}{k}\Big].$$

Consequently we get

$$\beta_k\big(\frac{c_1}{\sqrt{\tau_k}} + \frac{c_2}{\tau_k}\big) + C_{\underline{\eta}}\beta_k^2 \leq \frac{b_\delta}{k^2} \quad \text{with} \quad b_\delta = \frac{\sqrt{3}c_1}{\sqrt{\tau_1}} + \frac{c_2}{\tau_1 k_\delta} + 9C_{\underline{\eta}}. \tag{26}$$

Hence the recurrence inequality (25) becomes

$$\rho_{k+1} \leq (1 - \frac{1}{k})\rho_k + \frac{b_\delta}{k^2} + \beta_k c_0 v_k. \tag{27}$$

We pick an integer $q \geq (S/\tau_1)^{1/3} + 1$ such that $\rho_q \geq 0$ is satisfied.[10] We define the sequence $(u_n)_{n \geq q}$ as $u_q = \rho_q$ and

$$u_{n+1} = \big(1 - \frac{1}{n}\big)u_n + \frac{b_\delta}{n^2} + \beta_n c_0 v_n,$$

with b_δ the fixed positive constant defined in (26). From inequality (27), we have $\rho_k \leq u_k$ for $k \geq k_\delta$ and an immediate induction guarantees the positivity of the sequence (u_n). By rearranging we get

$$(n+1)u_{n+1} - nu_n = \frac{-u_n}{n} + \frac{b_\delta(n+1)}{n^2} + (n+1)\beta_n c_0 v_n \leq \frac{b_\delta(n+1)}{n^2} + (n+1)\beta_n c_0 v_n.$$

[10] Assuming this last condition is sensible since as the number of samples increases, $\widetilde{\lambda}$ gets closer to the stationary set $\Lambda_{\underline{\eta}}$ whose minimizer of $\mathcal{L}$ is $\lambda^\star$. The introduction of the term $(S/\tau_1)^{1/3}+1$ is motivated by the subsequent analysis of the series $\sum v_k$ in Lem. 6.

By telescoping and using the fact that $\beta_n \leq 3/n \leq 6/(n+1)$, we obtain

$$nu_n - qu_q \leq 2b_\delta \sum_{i=q}^{n-1} \frac{1}{i} + 6c_0 \sum_{i=q}^{n-1} v_i \leq 2b_\delta \log(\frac{n-1}{q-1}) + 6c_0 \sum_{i=q}^{n-1} v_i.$$

Let $K \geq k_\delta$. We thus have with probability at least $1-2\delta$

$$\rho_K \leq \frac{q\rho_q + 2b_\delta \log K}{K} + \frac{6c_0}{K} \sum_{k=q}^{K-1} v_k = \frac{\tau_1^{1/3}}{(t_K-1)^{1/3} + \tau_1^{1/3}} \Big(q\rho_q + 2b_\delta \log K + 6c_0 \sum_{k=q}^{K-1} v_k \Big). \tag{28}$$

We conclude the proof by plugging the result of Lem. 6 into inequality (28) which yields the desired high-probability bound $\rho_K = \widetilde{O}(1/t_K^{1/3})$.

Lemma 6. *Recalling that* $v_k = \sum_{s,a} \widetilde{\psi}_{k+1}(s,a) \frac{1}{\sqrt{T_k(s)}}$, *we have* $\sum v_k = \widetilde{O}(1)$.

Proof. Denoting $\mathcal{S} = \{1, 2, ..., S\}$ and recalling that $q \geq (S/\tau_1)^{1/3} + 1$, we have

$$\begin{aligned}
\sum_{k=q}^{K-1} v_k &= \sum_{k=q}^{K-1} \sum_{s,a} \widetilde{\psi}_{k+1}(s,a) \frac{1}{\sqrt{T_k(s)}} = \sum_{k=q}^{K-1} \sum_{s=1}^{S} \frac{\sqrt{\nu_{k+1}(s)}}{\tau_k} \frac{\sqrt{\nu_{k+1}(s)}}{\sqrt{T_k(s)}} \\
&\leq \sqrt{\sum_{k=q}^{K-1} \sum_{s=1}^{S} \frac{\nu_{k+1}(s)}{\tau_k^2}} \sqrt{\sum_{k=q}^{K-1} \sum_{s=1}^{S} \frac{\nu_{k+1}(s)}{T_k(s)}} = \sqrt{\underbrace{\sum_{k=q}^{K-1} \frac{1}{\tau_k}}_{\Sigma_1}} \sqrt{\underbrace{\sum_{k=q}^{K-1} \sum_{s=1}^{S} \Big(\frac{T_{k+1}(s)}{T_k(s)} - 1 \Big)}_{\Sigma_2}},
\end{aligned}$$

where the inequality uses the Cauchy-Schwarz inequality on the sum indexed doubly by the episodes and the states. Since the Riemann zeta function of 2 is upper bounded by 3, we have

$$\Sigma_1 \leq \frac{1}{3\tau_1} \sum_{k=q}^{K-1} \frac{1}{k^2} \leq \frac{1}{\tau_1}.$$

There remains to show that $\Sigma_2 = \widetilde{O}(1)$. We introduce the following related optimization problem. For any $K \geq q$, we have $t_K - 1 \geq S$ since we chose $q \geq (S/\tau_1)^{1/3} + 1$. Let $V^\star(K)$ be defined by

$$V^\star(K) \;=\; \max \sum_{k=q}^{K-1} \sum_{s=1}^{S} \big(h_{s,k} - 1 \big), \tag{29}$$

$$\text{s.t.} \quad h_{s,k} \geq 1 \quad \text{and} \quad \sum_{s=1}^{S} \prod_{k=q}^{K-1} h_{s,k} \leq t_K - 1. \tag{30}$$

We have for any episode k and state s, $T_{k+1}(s) \geq T_k(s)$ and

$$\sum_{s=1}^{S} \prod_{k=q}^{K-1} \frac{T_{k+1}(s)}{T_k(s)} = \sum_{s=1}^{S} \frac{T_K(s)}{T_q(s)} \leq t_K - 1.$$

Hence the sequence $\Big(\frac{T_{k+1}(s)}{T_k(s)} \Big)_{s,k}$ satisfies the constraints (30), thus $\Sigma_2 \leq V^\star(K)$. There remains to solve the optimization problem (29). Since the variables $h_{s,k}$ play interchangeable roles, there exists $h^\star = h_{s,k}$ for all s and k. From the second constraint in (30), we know that $h^\star \leq \big((t_K - 1)/S \big)^{1/(K-q)}$. Given that (29) is a maximization problem that increases proportionally with $h^\star$, when $t_K - 1 \geq S$ (so as to satisfy the first constraint), we finally

have $h^\star = \big((t_K - 1)/S\big)^{1/(K-q)}$. Consequently we have

$$\Sigma_2 \leq \sum_{k=q}^{K-1} \sum_{s=1}^{S} \Big(\Big(\frac{t_K - 1}{S} \Big)^{1/(K-q)} - 1 \Big) = \underbrace{\frac{\exp\Big(\frac{1}{K-q} \log\big(\frac{\tau_1 (K-1)^3}{S} \big) \Big) - 1}{\frac{1}{K-q} \log\big(\frac{\tau_1 (K-1)^3}{S} \big)}}_{\longrightarrow 1 \quad \text{when} \quad K \longrightarrow +\infty} \underbrace{S \log\big(\frac{\tau_1 (K-1)^3}{S} \big)}_{= \widetilde{\mathcal{O}}(1)},$$

which proves that $\Sigma_2 = \widetilde{O}(1)$. We conclude the proof using that $\sum_{k=q}^{K-1} v_{k+1} \leq \frac{1}{\sqrt{\tau_1}} \sqrt{\Sigma_2}$.

□

Algorithm 3 FW-AME w/ FMH-SDP

Input: $\widetilde{\lambda}_1 = 1/SA, \eta$
for $k = 1, 2, ..., K-1$ **do**
$\widehat{\psi}^+_{k+1} = \operatorname{argmin}_{\lambda \in \Lambda_\eta} \langle \nabla \widehat{\mathcal{L}}^+_{t_k - 1}(\widetilde{\lambda}_k), \lambda \rangle$
$\widehat{\pi}^+_{k+1}(a|s) = \dfrac{\widehat{\psi}^+_{k+1}(s,a)}{\sum_{b \in \mathcal{A}} \widehat{\psi}^+_{k+1}(s,b)}$
Compute $\pi^\star_{\text{FMH}} = \text{FMH-SDP}(\widehat{\pi}^+_{k+1}, \tau_k)$ with δ_{τ_k} defined in Eq. (31)
Execute $\pi^\star_{\text{FMH}}$ for τ_k steps, collect the samples and update $\widetilde{\lambda}_{k+1}$ as in Alg. 1
end for

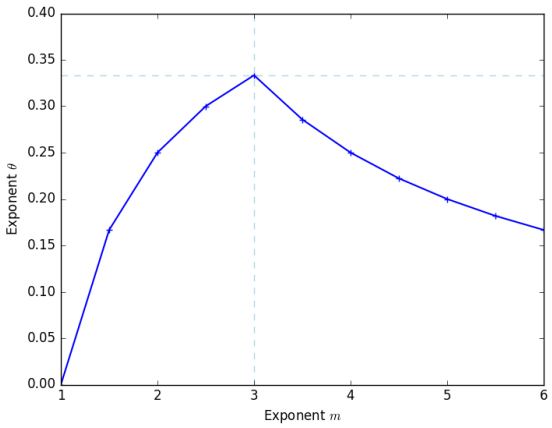

Figure 5: Exponent θ as a function of m (cf. App. D.3).

D.3 Optimality of the Episode Length

As explained in Sect. 3.2, an interesting open question is whether the regret bound obtained in Thm. 1 is optimal. Our analysis however yields the following optimality result: among all the episode lengths such that the time t is polynomial in the number of episodes k, i.e., among all the integers $m \geq 1$ such that t behaves as k^m, the value of m that optimizes convergence is $m = 3$. Indeed, we can apply the Euler method on inequality (25) which results in solving the differential equation $y' = \frac{-y}{x} + \frac{1}{x^2} + \frac{1}{x^{(m+1)/2}}$ and finding the largest θ such that $x^\theta y(x) = \widetilde{O}(1)$. θ is thus the largest value such that

$$\mathcal{L}(\widetilde{\lambda}_{k+1}) - \mathcal{L}(\lambda^\star) = \widetilde{O}\Big(\frac{1}{t^\theta}\Big) = \widetilde{O}\Big(\frac{1}{k^{\theta/m}}\Big).$$

Fig. 5 plots the exponent θ as a function of m and shows that θ reaches its maximal value of $1/3$ for $m = 3$, consequently yielding the regret $\widetilde{O}(1/t^{1/3})$.

D.4 FW-AME w/ FMH-SDP

The variant incorporating the framework of FMH-SDP is presented in Alg. 3 and its difference with Alg. 1 is highlighted in blue. The regret analysis is the same as in App. D.2 except that the error Δ_{k+1} in the recurrence inequality (24) goes from $\widetilde{\psi}_{k+1} - \widehat{\psi}^+_{k+1}$ to

$$\widetilde{\psi}^{\text{FMH}}_{k+1} - \widehat{\psi}^+_{k+1} = \widetilde{\psi}^{\text{FMH}}_{k+1} - \widehat{\psi}^{\text{FMH}}_{k+1} + \widehat{\psi}^{\text{FMH}}_{k+1} - \widehat{\psi}^+_{k+1},$$

where $\widetilde{\psi}_{k+1}^{\mathrm{FMH}} = \text{FMH-SDP}(\widehat{\psi}_{k+1}^{+}, \tau_k)$ and $\widetilde{\psi}_{k+1}^{\mathrm{FMH}}$ is its empirical realization for the τ_k steps of the episode. The new error Δ_{k+1} can thus be decomposed as follows

$$\underline{\eta}^2 \Delta_{k+1} \leq \sum_s \sigma^2(s) \Big| \frac{\nu_{k+1}^{\mathrm{FMH}}(s)}{\tau_k} - \eta_{\widetilde{\pi}_{k+1}^{\mathrm{FMH}}}(s) \Big| + \sum_s \sigma^2(s) \big| \eta_{\widetilde{\pi}_{k+1}^{\mathrm{FMH}}}(s) - \eta_{\widetilde{\pi}_{k+1}^{+}}(s) \big|,$$

where the first term is $O\Big(1/\sqrt{\gamma(\widehat{\psi}_{k+1}^{\mathrm{FMH}})\tau_k}\Big)$ and the second term is upper bounded by $\sum_s \sigma^2(s)\delta_{\tau_k}$ where δ_{τ_k} is the FMH-SDP parameter from problem (20).

Since w/o FMH-SDP we have $\gamma_{k+1} = O\Big(1/\sqrt{\gamma(\widehat{\psi}_{k+1}^{\mathrm{FMH}})\tau_k}\Big)$, this suggests that the slack variable δ_{τ_k} can decrease at least as $O(1/\sqrt{\tau_k})$ so as to guarantee that the order of the error Δ_{k+1} is unchanged. Furthermore, the component $\delta_{\tau_k}(s)$ is weighted by $\sigma^2(s)$ (which is unknown), hence we are encouraged to set

$$\delta_{\tau_k}(s) = \frac{\widehat{\Sigma} - \widehat{\sigma}_{t_k-1}^2(s)}{(S-1)\widehat{\Sigma}} \frac{1}{\sqrt{\tau_k}} \quad \text{where } \widehat{\Sigma} = \sum_{s \in \mathcal{S}} \widehat{\sigma}_{t_k-1}^2(s). \tag{31}$$

The regret analysis of FW-AME w/ FMH-SDP is thus unchanged and we recover the final rate in $O(t^{-1/3})$. In addition, if the heuristic is able to obtain an improvement in the mixing properties of the episodic policy (i.e., $\gamma(\widehat{\psi}_{k+1}^{\mathrm{FMH}})$ bigger than $\gamma(\widehat{\psi}_{k+1}^{+})$) that outweighs the error introduced by $\delta_{\tau_k}(s)$, then the regret performance at episode k of FW-AME w/ FMH-SDP is improved.

E Garnet MDPs

We detail here the process for generating Garnet[11] MDPs which we use in Sect. 5. A Garnet instance $\mathcal{G}(S, A, b, \sigma_{\min}^2, \sigma_{\max}^2)$ is characterized by 5 parameters. S and A are the number of states and actions respectively, and b is a branching factor specifying the number of possible next states for each state-action pair, i.e., the number of uniformly distributed non-zero entries in each line of the MDP transition matrix. We ensure the aperiodicity of the MDP by adding a non-zero probability (equal to 0.001) of self-loop for all state-action pairs. Since the state means are arbitrarily fixed, there remains to uniformly sample the state variances $\sigma^2(s)$ between $\sigma_{\min}^2$ and $\sigma_{\max}^2$ and randomly select two states whose variances are set respectively to $\sigma_{\min}^2$ and $\sigma_{\max}^2$. We likewise introduce reversible Garnet MDPs denoted by $\mathcal{G}_{\mathcal{R}}$. The generation process of $\mathcal{G}_{\mathcal{R}}$ is identical to $\mathcal{G}$ except that we set the branching factor to $b-1$ and ensure the reversibility of the MDP by randomly picking $a \in \mathcal{A}$ and $q \in (0,1)$ such that $p(s|s',a) = q$ for every pair (s,s') such that $Q(s,s') = 1$ (and finally normalize to obtain an admissible p).

We note that the Garnet procedure allows some control over the mixing properties of the MDP. Indeed, when A and b are small, only a few transitions are assigned significant probabilities so the speed of mixing is generally slower. For higher values of A and b, all the positive transition probabilities are of similar magnitude so the speed of mixing is generally faster.

[11] In full, Generalized Average Reward Non-stationary Environment Test-bench (Bhatnagar et al., 2009).