

A Failure of the negative entropy

In this appendix, we prove our lower bound result for OMD with the negative entropy from Section 3. We first restate the result.

Theorem 3.1. *For any $S \geq 6$, there exists an SSP instance with a fixed horizon of 3, sparsity level $M = 3$, an action space of size $A = 2$ and state space of size S such that the regret of OMD (1) with negative-entropy regularization and any step-size $\eta > 0$ after K episodes is $\mathbb{E}[R_K] = \Omega(\min\{\sqrt{K \log S}, K\})$.*

Proof. Fix K even, $S \geq 6$, $A = 2$ and $N = S - 5$. We first describe the SSP instance. Consider the following MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, s_0, g)$, where $\mathcal{S} = \{s_0, s_0^L, s_0^R, s_1^R, \dots, s_N^R, s_1\}$ and $\mathcal{A} = \{a_1, a_2\}$. The transitions and costs (in each episode k) are defined as:

- s_0 : $p(s_0^L | s_0, a) = p(s_0^R | s_0, a) = 1/2$ and $c_k(s_0, a) = 0$ for all $a \in \mathcal{A}$.
- s_0^L : $p(s_g^L | s_0^L, a) = 1$ for all $a \in \mathcal{A}$ and $c_k(s_0^L, a_1) = \frac{1+(-1)^k}{2}$, $c_k(s_0^L, a_2) = 1/2$.
- s_0^R : $p(s_g^R | s_0^R, a_1) = 1$, $p(s_i^R | s_0^R, a_2) = 1/N$ and $c_k(s_0^R, a_1) = 0$, $c_k(s_0^R, a_2) = 1$.
- s_i^R : $p(g | s_i^R, a) = 1$ and $c_k(s_i^R, a) = 0$ for all $a \in \mathcal{A}$.
- s_g^L : $p(g | s_g^L, a) = 1$ and $c_k(s_g^L, a) = 0$ for all $a \in \mathcal{A}$.
- s_g^R : $p(g | s_g^R, a) = 1$ and $c_k(s_g^R, a) = 0$ for all $a \in \mathcal{A}$.

An illustration is given in Figure 3. This SSP instance has a fixed horizon of 3 in the sense that all policies have a hitting time of exactly 3 (the states s_g^L and s_g^R are added to guarantee this). As a result we have that $T_\star = D = 3$. Also note that there are at most 3 state-action pairs that have non-zero cost, therefore the sparsity level $M = 3$.

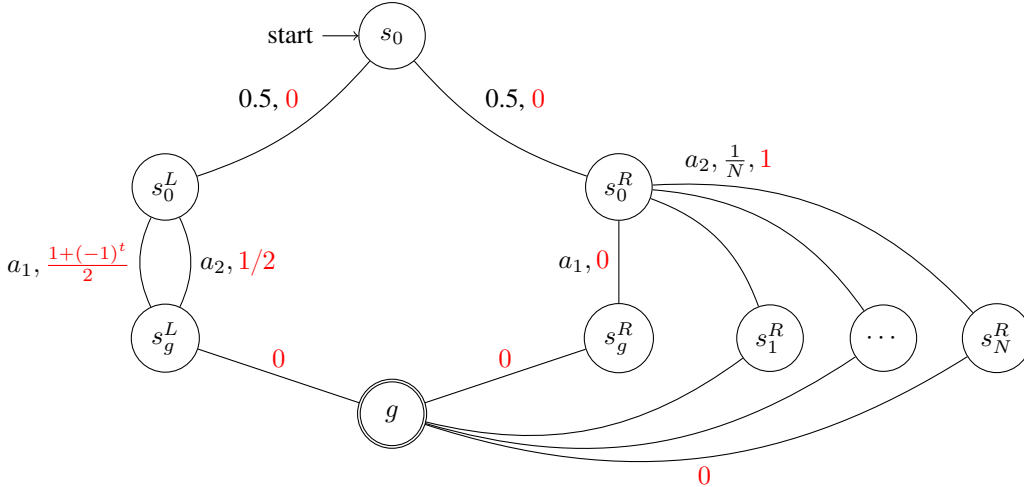


Figure 3: Diagram illustrating MDP construction for the proof of Theorem 3.1. When an action is not specified for an edge, then both actions give the same transition and cost. If an edge has a number in black, it is a transition probability; if it does not then the transition is deterministic. The costs are given in red. The formal description of the MDP is given above.

From Appendix B.1 in [24] (we can ignore the optimization over λ because we are in a fixed horizon setting), the update of OMD with negative entropy for any $k \geq 0$ can be computed by solving a convex optimization problem:

$$q_{k+1}(s, a) = q_k(s, a) e^{B_k^{v_{k+1}}(s, a)}, \quad \text{where} \quad B_k^v(s, a) = v(s) - \eta c_k(s, a) - \sum_{s' \in \mathcal{S}} p(s' | s, a) v(s'),$$

$$v_{k+1} = \arg \min_v \mathcal{D}_k(v),$$

$$\mathcal{D}_k(v) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_k(s, a) e^{B_k^v(s, a)} - v(s_0),$$

with $q_0(s, a) = 1$ and $c_0(s, a) = 0$. This allows us to compute exactly the points played by the algorithm on the SSP instance described above, and in turn compute the regret, from which the result will follow.

In the following few pages, we compute the occupancy measures played by OMD with the negative entropy on the SSP instance described earlier for all episodes, using the convex optimization problem above. We begin by computing expressions for $B_k^v(s, a)$ in each state:

- $B_k^v(s_0, a) = v(s_0) - \frac{1}{2}v(s_0^L) - \frac{1}{2}v(s_0^R)$ for all $a \in \mathcal{A}$
- $B_k^v(s_0^L, a) = v(s_0^L) - \eta c_k(s_0^L, a) - v(s_g^L)$ for all $a \in \mathcal{A}$
- $B_k^v(s_g^L) = v(s_g^L)$
- $B_k^v(s_0^R, a_1) = v(s_0^R) - v(s_g^R)$
- $B_k^v(s_0^R, a_2) = v(s_0^R) - \eta c_k(s_0^R, a_2) - \frac{1}{N} \sum_{i=1}^N v(s_i^R) = v(s_0^R) - \eta c_k(s_0^R, a_2) - v(s_1^R)$ since by symmetry $v(s_i^R) = v(s_1^R)$ for all $i \geq 1$ and any v solving the convex optimization problem specified in the OMD update.
- $B_k^v(s_i^R, a) = v(s_i^R) = v(s_1^R)$ for all $a \in \mathcal{A}$
- $B_k^v(s_g^R, a) = v(s_g^R)$ for all $a \in \mathcal{A}$

Plugging these into the optimization problem, we obtain (recall the notation $q(s) = \sum_{a \in \mathcal{A}} q(s, a)$):

$$\begin{aligned}
v_{k+1} &= \arg \min_v \mathcal{D}_k(v) = \arg \min_v \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_k(s, a) e^{B_k^v(s, a)} - v(s_0) \\
&= \arg \min_v q_k(s_0, a_1) e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} + q_k(s_0, a_2) e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} \\
&\quad + q_k(s_0^L, a_1) e^{v(s_0^L) - \eta c_k(s_0^L, a_1) - v(s_g^L)} + q_k(s_0^L, a_2) e^{v(s_0^L) - \eta c_k(s_0^L, a_2) - v(s_g^L)} \\
&\quad + q_k(s_g^L, a_1) e^{v(s_g^L)} + q_k(s_g^L, a_2) e^{v(s_g^L)} \\
&\quad + q_k(s_0^R, a_1) e^{v(s_0^R) - v(s_g^R)} + q_k(s_0^R, a_2) e^{v(s_0^R) - \eta c_k(s_0^R, a_2) - v(s_1^R)} \\
&\quad + \sum_{i=1}^N \left\{ q_k(s_i^R, a_1) e^{v(s_1^R)} + q_k(s_i^R, a_2) e^{v(s_1^R)} \right\} \\
&\quad + q_k(s_g^R, a_1) e^{v(s_g^R)} + q_k(s_g^R, a_2) e^{v(s_g^R)} \\
&\quad - v(s_0) \\
&= \arg \min_v q_k(s_0) e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} \\
&\quad + q_k(s_0^L, a_1) e^{v(s_0^L) - \eta c_k(s_0^L, a_1) - v(s_g^L)} + q_k(s_0^L, a_2) e^{v(s_0^L) - \eta c_k(s_0^L, a_2) - v(s_g^L)} \\
&\quad + q_k(s_g^L) e^{v(s_g^L)} \\
&\quad + q_k(s_0^R, a_1) e^{v(s_0^R) - v(s_g^R)} + q_k(s_0^R, a_2) e^{v(s_0^R) - \eta c_k(s_0^R, a_2) - v(s_1^R)} \\
&\quad + \sum_{i=1}^N \left\{ q_k(s_i^R) e^{v(s_1^R)} \right\} \\
&\quad + q_k(s_g^R) e^{v(s_g^R)} \\
&\quad - v(s_0).
\end{aligned}$$

This being a convex optimization problem, it can be solved by differentiating and setting to 0:

$$\begin{aligned}
\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0)} &= q_k(s_0) e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} - 1 = 0 \\
\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^L)} &= -0.5 q_k(s_0) e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} + q_k(s_0^L, a_1) e^{v(s_0^L) - \eta c_k(s_0^L, a_1) - v(s_g^L)} + q_k(s_0^L, a_2) e^{v(s_0^L) - \eta c_k(s_0^L, a_2) - v(s_g^L)} \\
&= -0.5 + q_k(s_0^L, a_1) e^{v(s_0^L) - \eta c_k(s_0^L, a_1) - v(s_g^L)} + q_k(s_0^L, a_2) e^{v(s_0^L) - \eta c_k(s_0^L, a_2) - v(s_g^L)} = 0
\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^L)} &= -q_k(s_0^L, a_1)e^{v(s_0^L) - \eta c_k(s_0^L, a_1) - v(s_g^L)} - q_k(s_0^L, a_2)e^{v(s_0^L) - \eta c_k(s_0^L, a_2) - v(s_g^L)} + q_k(s_g^L)e^{v(s_g^L)} \\ &= q_k(s_g^L)e^{v(s_g^L)} - 0.5 = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^R)} &= -0.5q_k(s_0)e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} + q_k(s_0^R, a_1)e^{v(s_0^R) - v(s_g^R)} + q_k(s_0^R, a_2)e^{v(s_0^R) - \eta c_k(s_0^R, a_2) - v(s_1^R)} \\ &= -0.5 + q_k(s_0^R, a_1)e^{v(s_0^R) - v(s_g^R)} + q_k(s_0^R, a_2)e^{v(s_0^R) - \eta c_k(s_0^R, a_2) - v(s_1^R)} = 0\end{aligned}$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_1^R)} = -q_k(s_0^R, a_2)e^{v(s_0^R) - \eta c_k(s_0^R, a_2) - v(s_1^R)} + e^{v(s_1^R)} \sum_{i=1}^N q_k(s_i^R) = 0$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^R)} = -q_k(s_0^R, a_1)e^{v(s_0^R) - v(s_g^R)} + q_k(s_g^R)e^{v(s_g^R)} = 0.$$

811 **Let's look specifically at the case $k = 0$** ($q_0(s, a) = 1, q_0(s) = 2, c_0(s, a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$).

812 For the left part of the MDP we have:

$$\begin{aligned}\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^L)} = 0 &\implies 2e^{v(s_0^L) - v(s_g^L)} = 0.5 \implies e^{v(s_0^L)} = 0.25e^{v(s_g^L)} \\ \frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^L)} = 0 &\implies e^{v(s_g^L)} = 0.25 \implies e^{v(s_0^L)} = 0.25^2 \\ &\implies q_1(s_0^L, a) = e^{B_0^v(s_0^L, a)} = \frac{0.25^2}{0.25} = 0.25, \text{ for all } a \in \mathcal{A}.\end{aligned}$$

813 For the right part of the MDP, we have:

$$\begin{aligned}\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^R)} = 0 &\implies e^{v(s_0^R) - v(s_g^R)} + e^{v(s_0^R) - v(s_1^R)} = 0.5 \implies e^{v(s_0^R)} = \frac{0.5}{e^{-v(s_g^R)} + e^{-v(s_1^R)}} \\ \frac{\partial \mathcal{D}_k(v)}{\partial v(s_1^R)} = 0 &\implies e^{v(s_0^R) - v(s_1^R)} = 2Ne^{v(s_1^R)} \implies e^{v(s_1^R)} = \frac{1}{\sqrt{2N}}e^{0.5v(s_0^R)} \\ \frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^R)} = 0 &\implies e^{v(s_0^R) - v(s_g^R)} = 2e^{v(s_g^R)} \implies e^{v(s_g^R)} = \frac{1}{\sqrt{2}}e^{0.5v(s_0^R)} \\ &\implies e^{v(s_0^R)} = \frac{0.5}{\sqrt{2}e^{-0.5v(s_0^R)} + \sqrt{2N}e^{-0.5v(s_0^R)}} \\ &\implies e^{0.5v(s_0^R)} = \frac{0.5}{\sqrt{2} + \sqrt{2N}} \\ &\implies e^{v(s_0^R)} = \frac{0.25}{(\sqrt{2} + \sqrt{2N})^2} \\ &\implies q_1(s_0^R, a_1) = e^{B_0^v(s_0^R, a_1)} = e^{v(s_0^R) - v(s_g^R)} = \sqrt{2}e^{0.5v(s_0^R)} = \frac{0.5}{1 + \sqrt{N}} \\ q_1(s_0^R, a_2) &= e^{B_0^v(s_0^R, a_2)} = e^{v(s_0^R) - v(s_1^R)} = \sqrt{2N}e^{0.5v(s_0^R)} = \frac{0.5\sqrt{N}}{1 + \sqrt{N}} \\ q_1(s_1^R, a) &= e^{B_0^v(s_1^R, a)} = e^{v(s_1^R)} = \frac{1}{\sqrt{2N}}e^{0.5v(s_0^R)} = \frac{0.25}{\sqrt{N}(1 + \sqrt{N})} \\ q_1(s_g^R, a) &= e^{B_0^v(s_g^R, a)} = e^{v(s_g^R)} = \frac{1}{\sqrt{2}}e^{0.5v(s_0^R)} = \frac{0.25}{1 + \sqrt{N}}.\end{aligned}$$

814 **Let's now look at general $k \geq 1$:** Since q_k is an occupancy measure, it satisfies the properties of
815 the dynamics of the MDP (see the definition of $\Delta(T)$ in Section 2) and we have that for any $s \in \mathcal{S}$:

816 $\sum_{a \in \mathcal{A}} q_k(s, a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p(s|s', a')q_k(s', a') + \mathbb{I}\{s = s_0\}$. In particular, this gives

817 $\bullet s = s_0: q_k(s_0) = 1$

818 $\bullet s = s_0^L: q_k(s_0^L) = 0.5q_k(s_0) = 0.5$

$$819 \quad \bullet \quad s = s_g^L: q_k(s_g^L) = q_k(s_0^L) = 0.5$$

$$820 \quad \bullet \quad s = s_0^R: q_k(s_0^R) = 0.5q_k(s_0) = 0.5$$

$$821 \quad \bullet \quad s = s_g^R: q_k(s_g^R) = q_k(s_0^R, a_1)$$

$$822 \quad \bullet \quad s = s_1^R: q_k(s_1^R) = \frac{1}{N}q_k(s_0^R, a_2)$$

823 This leads to the following simplifications in the derivatives of $\mathcal{D}_k(v)$:

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0)} = e^{v(s_0) - 0.5v(s_0^L) - 0.5v(s_0^R)} - 1 = 0$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^L)} = -0.5 + q_k(s_0^L, a_1)e^{v(s_0^L) - \eta c_k(s_0^L, a_1) - v(s_g^L)} + (1/2 - q_k(s_0^L, a_1))e^{v(s_0^L) - \eta c_k(s_0^L, a_2) - v(s_g^L)} = 0$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^L)} = 0.5e^{v(s_g^L)} - 0.5 = 0 \implies e^{v(s_g^L)} = 1$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^R)} = -0.5 + q_k(s_0^R, a_1)e^{v(s_0^R) - v(s_g^R)} + (1/2 - q_k(s_0^R, a_1))e^{v(s_0^R) - \eta - v(s_1^R)} = 0$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_1^R)} = -q_k(s_0^R, a_2)e^{v(s_0^R) - \eta - v(s_1^R)} + q_k(s_0^R, a_2)e^{v(s_1^R)} = 0 \implies e^{v(s_0^R)} = e^{\eta + 2v(s_1^R)}$$

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^R)} = -q_k(s_0^R, a_1)e^{v(s_0^R) - v(s_g^R)} + q_k(s_0^R, a_1)e^{v(s_g^R)} = 0 \implies e^{v(s_0^R)} = e^{2v(s_g^R)}.$$

824 **Left part of the MDP:**

$$\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^L)} = 0 \implies q_k(s_0^L, a_1)e^{v(s_0^L) - \eta \frac{1+(-1)^k}{2}} + (0.5 - q_k(s_0^L, a_1))e^{v(s_0^L) - 0.5\eta} = 0.5$$

$$\implies e^{v(s_0^L)} = \frac{0.5}{q_k(s_0^L, a_1)e^{-\eta \frac{1+(-1)^k}{2}} + (0.5 - q_k(s_0^L, a_1))e^{-0.5\eta}}$$

$$k+1=2 \implies e^{v(s_0^L)} = \frac{0.5}{0.25 + 0.25e^{-0.5\eta}}$$

$$\implies q_2(s_0^L, a_1) = q_1(s_0^L, a_1)e^{B_1^v(s_0^L, a_1)} = 0.25e^{v(s_0^L) - v(s_g^L)} = \frac{0.5}{1 + e^{-0.5\eta}}$$

$$k+1=3 \implies e^{v(s_0^L)} = \frac{0.5}{q_1(s_0^L, a_1)e^{-\eta} + (0.5 - q_1(s_0^L, a_1))e^{-0.5\eta}}$$

$$\implies e^{v(s_0^L)} = \frac{0.5}{\frac{0.5}{1+e^{-0.5\eta}}e^{-\eta} + \frac{0.5e^{-0.5\eta}}{1+e^{-0.5\eta}}e^{-0.5\eta}} = 0.25 \frac{e^{\eta}}{q_2(s_0^L, a_1)}$$

$$\implies q_3(s_0^L, a_1) = q_2(s_0^L, a_1)e^{v(s_0^L) - \eta - v(s_g^L)} = 0.25$$

$$k \text{ even} \implies q_k(s_0^L, a_1) = \frac{0.5}{1 + e^{-0.5\eta}}$$

$$k \text{ odd} \implies q_k(s_0^L, a_1) = 0.25,$$

825 where the last two lines follow by a straightforward induction. Hence the losses suffered by OMD on
826 the left part of the MDP are:

$$\begin{aligned} \sum_{k=1}^K \left\{ q_k(s_0^L, a_1)c_k(s_0^L, a_1) + q_k(s_0^L, a_2)c_k(s_0^L, a_2) \right\} &= \sum_{k=1}^K \left\{ q_k(s_0^L, a_1) \cdot \frac{1+(-1)^k}{2} + 0.5 \cdot (0.5 - q_k(s_0^L, a_1)) \right\} \\ &= \sum_{k=1}^K \left\{ q_k(s_0^L, a_1) \cdot \frac{(-1)^k}{2} + 0.25 \right\} \\ &= 0.25K + 0.5 \sum_{t=1}^{K/2} \left\{ q_{2t}(s_0^L, a_1) - q_{2t-1}(s_0^L, a_1) \right\} \end{aligned}$$

$$\begin{aligned}
&= 0.25K + 0.5 \sum_{t=1}^{K/2} \left\{ \frac{0.5}{1 + e^{-0.5\eta}} - 0.25 \right\} \\
&= 0.25K + 0.5 \frac{K}{2} \frac{0.5 - 0.25 - 0.25e^{-0.5\eta}}{1 + e^{-0.5\eta}} \\
&= 0.25K + \frac{K}{16} \cdot \frac{1 - e^{-0.5\eta}}{1 + e^{-0.5\eta}} \\
&= 0.25K + \frac{K}{16} \cdot \min\left\{\frac{\eta}{5}, \frac{1}{2}\right\}. \tag{6}
\end{aligned}$$

827 **Right part of the MDP:**

$$\begin{aligned}
\frac{\partial \mathcal{D}_k(v)}{\partial v(s_0^R)} = 0 &\implies e^{v(s_0^R)} = \frac{0.5}{q_k(s_0^R, a_1)e^{-v(s_g^R)} + (1/2 - q_k(s_0^R, a_1))e^{-\eta - v(s_1^R)}} \\
\frac{\partial \mathcal{D}_k(v)}{\partial v(s_1^R)} = 0 &\implies v(s_0^R) = \eta + 2v(s_1^R) \\
\frac{\partial \mathcal{D}_k(v)}{\partial v(s_g^R)} = 0 &\implies v(s_0^R) = v(s_g^R) \\
&\implies e^{v(s_0^R)} = \frac{0.5}{q_k(s_0^R, a_1)e^{-0.5v(s_0^R)} + (1/2 - q_k(s_0^R, a_1))e^{-\eta - 0.5v(s_0^R) + 0.5\eta}} \\
&\implies e^{0.5v(s_0^R)} = \frac{0.5}{q_k(s_0^R, a_1) + (1/2 - q_k(s_0^R, a_1))e^{-0.5\eta}} = \frac{0.5}{q_k(s_0^R, a_1) + q_k(s_0^R, a_2)e^{-0.5\eta}} \\
&\implies q_{k+1}(s_0^R, a_1) = q_k(s_0^R, a_1)e^{B_k^v(s_0^R, a_1)} = q_k(s_0^R, a_1)e^{v(s_0^R) - v(s_g^R)} = q_k(s_0^R, a_1)e^{0.5v(s_0^R)} \\
&\implies q_{k+1}(s_0^R, a_1) = 0.5 \frac{q_k(s_0^R, a_1)}{q_k(s_0^R, a_1) + (1/2 - q_k(s_0^R, a_1))e^{-0.5\eta}} \\
&\implies \frac{q_{k+1}(s_0^R, a_1)}{q_{k+1}(s_0^R, a_2)} = \frac{q_{k+1}(s_0^R, a_1)}{(1/2 - q_{k+1}(s_0^R, a_1))} = \frac{q_k(s_0^R, a_1)}{(1/2 - q_k(s_0^R, a_1))e^{-0.5\eta}} = e^{0.5\eta} \frac{q_k(s_0^R, a_1)}{q_k(s_0^R, a_2)} \\
&\implies \frac{q_{k+1}(s_0^R, a_1)}{q_{k+1}(s_0^R, a_2)} = e^{0.5k\eta} \frac{q_1(s_0^R, a_1)}{q_1(s_0^R, a_2)} = e^{0.5k\eta} \frac{0.5}{0.5\sqrt{N}} = \frac{1}{\sqrt{N}} e^{0.5k\eta} \\
&\implies q_{k+1}(s_0^R, a_2) = \frac{0.5}{1 + \frac{1}{\sqrt{N}}e^{0.5k\eta}} = \frac{0.5\sqrt{N}}{\sqrt{N} + e^{0.5\eta k}}.
\end{aligned}$$

828 This also holds for $k = 0$ (as shown above). Hence, the losses suffered by OMD on the right part of
829 the MDP are

$$\begin{aligned}
\sum_{k=1}^K q_k(s_0^R, a_2)c_k(s_0^R, a_2) &= \sum_{k=1}^K \frac{0.5\sqrt{N}}{\sqrt{N} + e^{0.5k\eta}} \\
&\geq \int_1^{K+1} \frac{0.5\sqrt{N}}{\sqrt{N} + e^{0.5\eta x}} dx = 0.5K - \int_1^{K+1} \frac{0.5e^{0.5\eta x}}{\sqrt{N} + e^{0.5\eta x}} dx \\
&= 0.5K - \left[\frac{1}{\eta} \log(\sqrt{N} + e^{0.5\eta x}) \right]_1^{K+1} \\
&= 0.5K - \frac{1}{\eta} \log(\sqrt{N} + e^{0.5\eta(K+1)}) + \frac{1}{\eta} \log(\sqrt{N} + e^{0.5\eta}) \\
&\geq 0.5K - \frac{1}{\eta} \log(2e^{0.5\eta(K+1)}) + \frac{1}{\eta} \log \sqrt{N} \quad \text{assuming } \sqrt{N} \leq e^{0.5\eta(K+1)} \\
&= 0.5K - 0.5(K+1) - \frac{1}{\eta} \log 2 + \frac{1}{2\eta} \log N \\
&= -0.5 + \frac{1}{2\eta} \log \frac{N}{4}.
\end{aligned}$$

830 If $\sqrt{N} > e^{0.5\eta(K+1)}$, then we have

$$\begin{aligned}
0.5K - \frac{1}{\eta} \log(\sqrt{N} + e^{0.5\eta(K+1)}) + \frac{1}{\eta} \log(\sqrt{N} + e^{0.5\eta}) &= 0.5K + \frac{1}{\eta} \log\left(\frac{\sqrt{N} + e^{0.5\eta}}{\sqrt{N} + e^{0.5\eta(K+1)}}\right) \\
&\geq 0.5K + \frac{1}{\eta} \log\left(\frac{e^{0.5\eta(K+1)} + e^{0.5\eta}}{2e^{0.5\eta(K+1)}}\right) \\
&\geq 0.5K + \frac{1}{\eta} \log\left(\frac{1 + e^{-0.5\eta K}}{2}\right) \\
&\geq 0.25K,
\end{aligned}$$

831 using that $\frac{1+e^{-0.5Kx}}{2} \geq e^{-0.25Kx}$ since $\cosh(x) \geq 1$. So we have

$$\sum_{k=1}^K q_k(s_0^R, a_2) c_k(s_0^R, a_2) \geq \min\left\{-0.5 + \frac{1}{2\eta} \log \frac{N}{4}, 0.25K\right\}. \quad (7)$$

832 Combining the losses from the left part in (6) and from the right part in (7), we have:

$$\begin{aligned}
\sum_{k=1}^K \langle q_k, c_k \rangle &= \sum_{k=1}^K \left\{ q_k(s_0^L, a_1) c_k(s_0^L, a_1) + q_k(s_0^L, a_2) c_k(s_0^L, a_2) \right\} + \sum_{k=1}^K \left\{ q_k(s_0^R, a_2) c_k(s_0^R, a_2) \right\} \\
&\geq 0.25K + \frac{K}{16} \cdot \min\left\{\frac{\eta}{5}, \frac{1}{2}\right\} + \min\left\{-0.5 + \frac{1}{2\eta} \log \frac{N}{4}, 0.25K\right\}.
\end{aligned}$$

833 **Regret lower-bound:** consider q_\star defined as follows:

- 834 • $q_\star(s_0, a) = 1/2$ for all $a \in \mathcal{A}$
- 835 • $q_\star(s_0^L, a_2) = 1/2, q_\star(s_0^L, a_1) = 0$
- 836 • $q_\star(s_g^L, a) = 1/4$ for all $a \in \mathcal{A}$
- 837 • $q_\star(s_0^R, a_1) = 1/2, q_\star(s_0^R, a_2) = 0, q_\star(s_i^R, a) = 0$
- 838 • $q_\star(s_g^R, a) = 1/4$ for all $a \in \mathcal{A}$

839 It is straightforward to check that q_\star satisfies the flow constraints and is an occupancy measure. We
840 obtain

$$\begin{aligned}
\sum_{k=1}^K \langle q_\star, c_k \rangle &= \sum_{k=1}^K \left\{ q_\star(s_0^L, a_2) \cdot 0.5 \right\} = 0.25K \\
\implies R_K &\geq \sum_{k=1}^K \langle q_k - q_\star, c_k \rangle \geq \frac{K}{16} \cdot \min\left\{\frac{\eta}{5}, \frac{1}{2}\right\} + \min\left\{-0.5 + \frac{1}{2\eta} \log \frac{N}{4}, 0.25K\right\} \\
&\geq \min\left\{\frac{1}{2} \sqrt{\frac{1}{10} K \log \frac{N}{4}}, \frac{K}{32}\right\} - 0.5.
\end{aligned}$$

841 Recalling that $N = S - 5$, we have $R_K = \Omega(\min\{\sqrt{K \log S}, K\})$ for an MDP where the sparsity
842 level is $M = 3$, concluding the proof. \square

843 B Efficient implementation of OMD using our regularizer

844 In this section, we describe how the OMD update with our regularizer from Section 4 defined in
 845 (4) can be computed efficiently. This closely follows Appendix B.1 of [24], who provide a similar
 846 description for the negative entropy.

847 Recall the regularizer $\psi_p(q) = p \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(s, a)^{1+1/p} - p$ for $q \in \mathbb{R}_{\geq 0}^\Gamma$. We have

$$\nabla \psi_p(q) = (p+1) \cdot q(s, a)^{1/p}.$$

848 The Bregman divergence is defined as:

$$\begin{aligned} D_{\psi_p}(q, q') &= \psi_p(q) - \psi_p(q') - \langle \nabla \psi_p(q'), q - q' \rangle \\ &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ p \cdot q(s, a)^{1+1/p} - p \cdot q'(s, a)^{1+1/p} \right\} \\ &\quad - (p+1) \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q'(s, a)^{1/p} q(s, a) - q'(s, a)^{1+1/p} \right\} \\ &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q'(s, a)^{1+1/p} + q(s, a) \cdot [p \cdot q(s, a)^{1/p} - (p+1) \cdot q'(s, a)^{1/p}] \right\}. \end{aligned}$$

849 Recall that OMD with the above regularizer computes the occupancy measures as follows - see (1):

$$q_1 = \arg \min_{q \in \Delta(T)} \psi_p(q), \quad q_{k+1} = \arg \min_{q \in \Delta(T)} \left\{ \eta \cdot \langle q, c_k \rangle + D_{\psi_p}(q, q_k) \right\}.$$

850 As shown in [20] (Theorem 6.15), each of these steps can be split into an unconstrained minimization
 851 step, and a projection step. Thus, q_1 can be computed as follows

$$\begin{aligned} q'_1 &= \arg \min_{q \in \mathbb{R}_{\geq 0}^\Gamma} \psi_p(q) \\ q_1 &= \arg \min_{q \in \Delta(T)} D_{\psi_p}(q, q'_1), \end{aligned}$$

852 where q'_1 has a closed-form solution $q'_1(s, a) = 1$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Similarly, q_{k+1} is
 853 computed as follows for every $k = 1, \dots, K-1$:

$$\begin{aligned} q'_{k+1} &= \arg \min_{q \in \mathbb{R}_{\geq 0}^\Gamma} \left\{ \eta \cdot \langle q, c_k \rangle + D_{\psi_p}(q, q_k) \right\} \\ q_{k+1} &= \arg \min_{q \in \Delta(T)} D_{\psi_p}(q, q'_{k+1}), \end{aligned}$$

854 where again q'_{k+1} has a closed-form solution $q'_{k+1}(s, a) = \left[q_k(s, a)^{1/p} - \frac{\eta}{p+1} c_k(s, a) \right]_+$ for every
 855 $s \in \mathcal{S}$ and $a \in \mathcal{A}$ (follows from straightforwardly differentiating above objective and setting to 0 and
 856 accounting for the non-negativity of occupancy measures) - we use notation $a_+ = \max\{0, a\}$.

857 For the projection step, we start by formulating it as a constrained convex optimization problem:

$$\begin{aligned} \min_{q \in \mathbb{R}^\Gamma} D_{\psi_p}(q, q'_{k+1}) \quad \text{s.t.} \quad & \sum_{a \in \mathcal{A}} q(s, a) - \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s'|s, a) q(s', a') = \mathbb{I}\{s = s_0\} \quad \forall s \in \mathcal{S} \\ & \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(s, a) \leq T \\ & q(s, a) \geq 0 \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \end{aligned}$$

858 The problem can be solved by considering the Lagrangian with Lagrange multipliers λ and $\{v(s)\}_{s \in \mathcal{S}}$:

$$\begin{aligned} \mathcal{L}(q, \lambda, v) &= D_{\psi_p}(q, q'_{k+1}) + \lambda \left(\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(s, a) - T \right) + \sum_s v(s) \left(\sum_{s', a'} P(s'|s, a) q(s', a') + \mathbb{I}\{s = s_0\} - \sum_a q(s, a) \right) \\ &= D_{\psi_p}(q, q'_{k+1}) + \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(s, a) \left(\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a) v(s') - v(s) \right) + v(s_0) - \lambda T, \end{aligned}$$

859 Differentiating the Lagrangian with respect to any $q(s, a)$ and setting to 0, we get

$$\begin{aligned} \frac{\partial \mathcal{L}(q, \lambda, v)}{\partial q(s, a)} &= \nabla \psi_p(q)(s, a) - \nabla \psi_p(q'_{k+1})(s, a) + \lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s) \\ &= (p+1)q(s, a)^{1/p} - (p+1)q'_{k+1}(s, a)^{1/p} + \lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s) = 0. \\ \implies q_{k+1}(s, a) &= \left[q'_{k+1}(s, a)^{1/p} - \frac{\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s)}{p+1} \right]_+^p. \end{aligned}$$

860 This formula is also valid for $k = 0$ by setting $c_0(s, a) = 0$ and $q_0(s, a) = 1$ for every $s \in \mathcal{S}$ and
861 $a \in \mathcal{A}$.

862 To compute the value of λ and v at the optimum, we write the dual problem $\mathcal{D}(\lambda, v) = \min_q \mathcal{L}(q, \lambda, v)$
863 by substituting q_{k+1} back into \mathcal{L} :

$$\mathcal{D}(\lambda, v) = D_{\psi_p}(q_{k+1}, q'_{k+1}) + \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_{k+1}(s, a) \left(\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s) \right) + v(s_0) - \lambda T.$$

864 Recall that $q'_{k+1}(s, a) = \left[q_k(s, a)^{1/p} - \frac{\eta}{p+1} c_k(s, a) \right]_+^p$, so (ignoring terms independent of λ, v , e.g.
865 $q'_{k+1}(s, a)$):

$$\begin{aligned} D_{\psi_p}(q_{k+1}, q'_{k+1}) &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q'_{k+1}(s, a)^{1+1/p} + q_{k+1}(s, a) \cdot [pq_{k+1}(s, a)^{1/p} - (p+1)q'_{k+1}(s, a)^{1/p}] \right\} \\ &\propto \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q_{k+1}(s, a) \cdot \left[p \left(q'_{k+1}(s, a)^{1/p} - \frac{\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s)}{p+1} \right)_+ \right. \right. \\ &\quad \left. \left. - (p+1)q'_{k+1}(s, a)^{1/p} \right] \right\} \\ &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q_{k+1}(s, a) \cdot \left[p \left(q'_{k+1}(s, a)^{1/p} - \frac{\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s)}{p+1} \right) \right. \right. \\ &\quad \left. \left. - (p+1)q'_{k+1}(s, a)^{1/p} \right] \right\} \quad \text{since if } q_{k+1}(s, a) = 0, \text{ then the whole term is 0} \\ &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q_{k+1}(s, a) \cdot \left[-q'_{k+1}(s, a)^{1/p} - \frac{p}{p+1} \left(\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s) \right) \right] \right\} \\ \implies \mathcal{D}(\lambda, v) &\propto \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q_{k+1}(s, a) \cdot \left[-q'_{k+1}(s, a)^{1/p} - \frac{p}{p+1} \left(\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s) \right) \right. \right. \\ &\quad \left. \left. + \left(\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s) \right) \right] \right\} + v(s_0) - \lambda T \\ &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left\{ q_{k+1}(s, a) \cdot \left[-q'_{k+1}(s, a)^{1/p} + \frac{\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s)}{p+1} \right] \right\} + v(s_0) - \lambda T \\ &= - \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_{k+1}(s, a)^{1+1/p} + v(s_0) - \lambda T \\ &= - \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left[q'_{k+1}(s, a)^{1/p} - \frac{\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s)}{p+1} \right]_+^{1+p} + v(s_0) - \lambda T. \end{aligned}$$

866 Maximizing the dual gives λ and v or equivalently, we can minimize the negation of the dual:

$$\lambda_{k+1}, v_{k+1} = \arg \min_{\lambda \geq 0, v} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left[q'_{k+1}(s, a)^{1/p} - \frac{\lambda + \sum_{s' \in \mathcal{S}} P(s'|s, a)v(s') - v(s)}{p+1} \right]_+^{1+p} - v(s_0) + \lambda T.$$

867 This is a convex optimization problem subject only to non-negativity constraints, and can be efficiently
868 solved using iterative methods (e.g. gradient descent).

Algorithm 1 Sparse-Agnostic Mirror Descent

Input: T, K, D
Initialize: $p \leftarrow \log 2^2 T$, $\eta \leftarrow \sqrt{pT^{1+1/p}/(pDK2^{2/p})}$, $m \leftarrow 1$, $q_1 \leftarrow \arg \min_{q \in \Delta(T)} \psi_p(q)$
for $k = 1, \dots, K$ **do**
 Play q_k and Observe c_k
 if $\|c_k\|_0 \leq m$ **then**
 $q_{k+1} = \arg \min_{q \in \Delta_T} \langle q, c_k \rangle + D_{\psi_p}(q, c_k)$
 else
 $m \leftarrow \lceil \log_2 \log_2 \|c_k\|_0 \rceil$
 $p \leftarrow \log 2^{2^m} T$
 $\eta \leftarrow \sqrt{pT^{1+1/p}/(pDK2^{2/p})}$
 $q_{k+1} \leftarrow \arg \min_{q \in \Delta(T)} \psi_p(q)$
 end if
end for

C The benefits of sparsity - upper bounds

We restate the main theorem proved in Section 4.

Theorem 4.1. Consider OMD with ψ_p as regularizer. If T is such that $q_{\pi^*} \in \Delta(T)$, $\eta = \sqrt{\frac{2pT^{1+1/p}}{KDM^{1/p}}}$, $p = 1 + \log(TM)$, then $\mathbb{E}[R_K] \leq O\left(\sqrt{DKT \log(MT)}\right)$.

We include the missing details from the proof given in Section 4. First, recall from (4) that

$$\begin{aligned}
 \psi_p(q) &= p \cdot \left(-1 + \|q\|_{1+1/p}^{1+1/p}\right) = p \cdot \left(-1 + \sum_{s \in S} \sum_{a \in A} |q(s, a)|^{1+1/p}\right) \\
 \implies \frac{\partial \psi_p(q)}{\partial q(s, a)} &= (p+1)q(s, a)^{1/p} \\
 \implies \frac{\partial^2 \psi_p(q)}{\partial q(s, a)^2} &= \left(1 + \frac{1}{p}\right)q(s, a)^{1/p-1} \\
 \implies \nabla \psi_p(q) &= (p+1)q^{1/p}, \quad \nabla^2 \psi_p(q) = \text{diag}\left(\frac{p+1}{p}q^{1/p-1}\right)
 \end{aligned}$$

We implicitly assumed here that ψ_p is defined on $\mathbb{R}_{\geq 0}^\Gamma$. The missing details are:

- ψ_p satisfies the condition (2) with $\alpha = 1$:

$$\begin{aligned}
 \nabla \psi_p(q) \in \left[\nabla \psi_p(q_k), \nabla \psi_p(q_k) - \eta c_k\right] &\implies q^{1/p}(s, a) \leq q_k^{1/p}(s, a) \\
 &\implies \frac{1}{q(s, a)} \geq \frac{1}{q_k(s, a)} \\
 &\implies \frac{1}{q^{1-1/p}(s, a)} \geq \frac{1}{q_k^{1-1/p}(s, a)} \\
 &\implies \nabla^2 \psi_p(q) \geq \nabla^2 \psi_p(q_k).
 \end{aligned}$$

- $\nabla^2 \psi_p(q)^{-1} = \text{diag}\left(\frac{p}{p+1}q^{1-1/p}\right)$: follows directly from the expression for $\nabla^2 \psi_p(q)$ above.

We now turn to the description of the parameter-free algorithm and the proof of its corresponding regret bound (Theorem 4.3).

For the unknown sparsity level, we use the same approach as in [16], which divides the episode horizon into segments, where each segment will run OMD from scratch with an increasing sparsity level guess. By definition $M = \max_{k \in [K]} \|c_k\|_0$ is the higher sparsity level of the cost vectors. Now let $B = \lceil \log_2 \log_2 M \rceil$, $\tau(0) = 0$ then for $1 \leq b \leq B$:

$$\tau(b) = \min \left\{ 1 \leq k \leq K \mid \|c_k\|_0 > 2^{2^b} \right\} \quad \text{and} \quad \tau(B) = K.$$

Algorithm 2 Fully Parameter-Free Online Mirror Descent for Sparse SSPs

Define $j_0 = \lceil \log_2 T^{\pi_f} \rceil (s_0) - 1$, $b(j) = 2^{j_0+j}$, $N = \lceil \log_2 K \rceil - j_0$, $\eta_j = \sqrt{\frac{\log N \sqrt{b(j)}}{b(j)DK}}$
Define $\psi_p(p) = \sum_{j=1}^N \frac{1}{\eta_j} p(j) \log p(j)$
Initialize: $p_1 \in \Delta_N$, such that $p_1(j) = \frac{\eta_j}{\eta_1 N}$, $\forall j \neq 1$
Initialize: N instances of Algorithm 1, with j -th instance $T = b(j)$
for $k = 1, \dots, K$ **do**
 Obtain occupancy measures q_k^j for $j \in [N]$
 Sample $j_k \sim p_k$, execute policy induced by $q_k^{j_k}$
 Receive c_k and send it to all instances.
 Compute $\ell_k(j) = \langle q_k^j, c_k \rangle$, $a_k(j) = 4\eta_j \ell_k^2(j)$
 Update $p_{k+1} = \arg \min_{p \in \Delta_N} \langle p, \ell_k + a_k \rangle + D_{\psi_p}(p, p_k)$
end for

883 This essentially says that $\tau(b)$ is the first episode in which the sparsity level of the loss vector exceeds
 884 2^b . Then we denote $b_k = \min \{b \leq 1 \mid \tau(b) \leq k\}$, which is the index of the only episode k belongs
 885 to. Therefore we partition the horizon $[K]$ as intervals $(I(b))_{b \in [B]}$:

$$I(b) = \begin{cases} [\tau(b) + 1, \tau(b)] & \text{if } \tau(b-1) < \tau(b) \\ \emptyset & \text{if } \tau(b-1) = \tau(b) \end{cases}$$

886 Now we define:

$$\begin{aligned}
 887 \quad p(b) &= \log 2^{2^b} T \\
 \eta(b) &= \sqrt{\frac{p(b)T^{1+1/p(b)}}{DK2^{2^b/p(b)}}}.
 \end{aligned}$$

888 We consider mirror descent updates with our regularizer:

$$\psi_{p(b)}(q) = p(b) \left(-1 + \|q\|_{1+1/p(b)}^{1+1/p(b)} \right),$$

889 then each mirror descent instance is defined by:

$$q_t = \nabla \psi_{p(b_t)}^* \left(\eta(b_t) \sum_{k' < k, k' \in I(b_t)} c_k \right), \quad k = 1, \dots, K.$$

890 **Lemma C.1.** *Running Algorithm Algorithm 1 guarantees:*

$$R_K \leq \mathcal{O} \left(\sqrt{DTK \log(TM)} + BM \sqrt{\frac{DT \log(TM)}{K}} \right)$$

891 *Proof.* On the time interval $I(b)$, we run OMD with regularizer $\psi_{p(b)}$, learning rate $\eta(b)$ and we
 892 consider the interval regret $R(b)$. Therefore, by the previous result, we see that the Penalty term is:

$$\psi_{p(b)}(q_\star) - \psi_{p(b)}(q_1) \leq p(b) \|q_\star\|^{1+1/p(b)} \leq p(b) T^{1+1/p(b)}$$

893 While for the Stability term:

$$\begin{aligned}
 \|c_k\|_{\nabla^2 \psi_p(q_k)^{-1}}^2 &= \eta(b) \sum_{k \in I(b)} \sum_{s,a} c_k(s,a)^{1-1/p(b)} q_k(s,a)^{1-1/p(b)} \\
 &= \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} \sum_{s,a} c_k(s,a)^{1-1/p(b)} q_k(s,a)^{1-1/p(b)} + \eta(b) \sum_{s,a} c_{\tau(b)}(s,a)^{1-1/p(b)} q_{\tau(b)}(s,a)^{1-1/p(b)}
 \end{aligned}$$

$$\begin{aligned}
&\leq \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} 2^{2^b} \left(\frac{1}{2^{2^b}} \sum_{s,a} c_k(s,a) q_k(s,a) \right)^{1-1/p(b)} + \eta(b) \tau(b) \left(\frac{1}{\tau(b)} \sum_{s,a} c_{\tau(b)}(s,a) q_{\tau(b)}(s,a) \right)^{1-1/p(b)} \\
&\leq \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} 2^{2^b/p(b)} \langle c_k, q_k \rangle^{1-1/p(b)} + \eta(b) M^{1/p(b)} \langle c_{\tau(b)}, q_{\tau(b)} \rangle^{1-1/p(b)} \\
&\leq \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} 2^{2^b/p(b)} \max \{ \langle c_k, q_k \rangle, 1 \} + \eta(b) M^{1/p(b)} \max \{ \langle c_{\tau(b)}, q_{\tau(b)} \rangle, 1 \} \\
&\leq \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} 2^{2^b/p(b)} (\langle c_k, q_k \rangle + 1) + \eta(b) M^{1/p(b)} (\langle c_{\tau(b)}, q_{\tau(b)} \rangle + 1) \\
&= \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} 2^{2^b/p(b)} \langle c_k, q_k \rangle + \eta(b) (I_{\tau(b)} - 1) 2^{2^b/p(b)} + \eta(b) M^{1/p(b)} + \eta(b) M^{1/p(b)} \langle c_{\tau(b)}, q_{\tau(b)} \rangle
\end{aligned}$$

894 Therefore

$$\begin{aligned}
&\sum_{k \in I(b)} \langle q_k - q_{\pi^*}, c_k \rangle \leq p(b) \frac{T^{1+1/p(b)}}{\eta(b)} + \eta(b) \sum_{\substack{k \in I(b) \\ k < \tau(b)}} 2^{2^b/p(b)} \langle c_k, q_k \rangle + \eta(b) (I_{\tau(b)} - 1) 2^{2^b/p(b)} \\
&\quad + \eta(b) M^{1/p(b)} + \eta(b) M^{1/p(b)} \langle c_{\tau(b)}, q_{\tau(b)} \rangle \\
\Rightarrow &\sum_{\substack{k \in I(b) \\ k < \tau(b)}} \langle q_k - q_{\pi^*}, c_k \rangle + \langle c_{\tau(b)}, q_{\tau(b)} \rangle \leq p(b) \frac{T^{1+1/p(b)}}{\eta(b)} + \eta(b) (I_{\tau(b)} - 1) 2^{2^b/p(b)} + \eta(b) M^{1/p(b)} \\
&\quad + \eta(b) 2^{2^b/p(b)} \sum_{\substack{k \in I(b) \\ k < \tau(b)}} \langle c_k, q_k \rangle + \eta(b) M^{1/p(b)} \langle c_{\tau(b)}, q_{\tau(b)} \rangle \\
\Rightarrow &R_T(b) \leq 4p(b) \frac{T^{1+1/p(b)}}{\eta(b)} + 4\eta(b) (I_{\tau(b)} - 1) 2^{2^b/p(b)} + 4\eta(b) M^{1/p(b)} \\
&\quad + 4\eta(b) 2^{2^b/p(b)} \sum_{\substack{k \in I(b) \\ k < \tau(b)}} \langle c_k, q_{\pi^*} \rangle + 4\eta(b) M^{1/p(b)} \langle c_{\tau(b)}, q_{\pi^*} \rangle \\
\Rightarrow &R_T(b) \leq 4p(b) \frac{T^{1+1/p(b)}}{\eta(b)} + 4\eta(b) I_{\tau(b)} 2^{2^b/p(b)} + 4\eta(b) M^{1/p(b)} \\
&\quad + 4\eta(b) 2^{2^b/p(b)} I_{\tau(b)} D + 4\eta(b) M^{1/p(b)} D
\end{aligned}$$

895 Tuning $\eta(b) = \sqrt{\frac{p(b)T^{1+1/p(b)}}{DK2^{2^b/p(b)}}}$:

$$\begin{aligned}
R_T(b) &\leq 4\sqrt{KDT^{1+1/p(b)}p(b)2^{2^b/p(b)}} + 4I_{\tau(b)}\sqrt{\frac{DT^{1+1/p(b)}2^{2^b/p(b)}p(b)}{K}} + 4M^{1/p(b)}\sqrt{\frac{Dp(b)T^{1+1/p(b)}}{K2^{2^b/p(b)}}} \\
&\leq 4\sqrt{KDT\log(T2^{2^b})} + 4I_{\tau(b)}\sqrt{\frac{DT\log(T2^{2^b})}{K}} + 4M\sqrt{\frac{DT\log(T2^{2^b})}{K}}
\end{aligned}$$

896 Note that:

$$\begin{aligned}
\log(2^{2^b}) &\leq 2^{\log_2 \log_2 M + 1} \log 2 \\
&= \log 2 \exp(\log(2^{\log_2 \log_2 M + 1})) \\
&= \log 2 \exp((\log_2 \log_2 M + 1) \log 2) \\
&= 2 \log 2 \exp((\log_2 \log_2 M)) \\
&= 2 \log 2 \log M \leq 2 \log M
\end{aligned}$$

897 And therefore $\log(T2^{2^b}) \leq 2 \log(TM)$.

$$\begin{aligned}
R_K &= \sum_{b=1}^B R(b) \leq \sum_{b=1}^B 4\sqrt{KDT \log(T2^{2^b})} + \sum_{b=1}^B 4I_{\tau(b)} \sqrt{\frac{DT \log(T2^{2^b})}{K}} + \sum_{b=1}^B 4M \sqrt{\frac{DT \log(T2^{2^b})}{K}} \\
&\leq 4\sqrt{KDT} \sum_{b=1}^B \sqrt{\log(T2^{2^b})} + 4\sqrt{\frac{DT}{K}} \sum_{b=1}^B I_{\tau(b)} \sqrt{\log(T2^{2^b})} + 4M \sqrt{\frac{DT}{K}} \sum_{b=1}^B \sqrt{\log(T2^{2^b})}
\end{aligned}$$

898 Now ignoring log log terms and knowing that $\sum_{b=1}^B I_{\tau(b)} = K$:

$$\begin{aligned}
R_K &\leq 4B\sqrt{KDT \log(TM)} + 4\sqrt{DTK \log(T2^{2^b})} + 4M\sqrt{\frac{DT}{K}} B\sqrt{\log(TM)} \\
&\leq \mathcal{O}\left(\sqrt{DKT \log(TM)} + BM\sqrt{\frac{DT \log(TM)}{K}}\right)
\end{aligned}$$

899

□

900 We now turn our attention to the unknown hitting time of the optimal policy T_\star , where we can exploit
901 the same technique presented in [8].

902 We will therefore run $N \approx \log K$ instances of Algorithm 1 where the j -th instance will set its
903 parameter T as $b(j)$ which is roughly 2^j , so that there always exists an instance j_\star such that $b(j_\star)$ is
904 very close to the unknown T_\star .

905 We will therefore run a scale invariant meta algorithm with a correction term as in [8] to obtain the
906 desired bound (details in Algorithm 2).

907 **Theorem 4.3.** For any $T_\star \geq B^2$, $K \geq MT$, Algorithm 2 guarantees:

$$\mathbb{E}[R_K] \leq \mathcal{O}\left(\sqrt{DT_\star K \log(T_\star M)} + BM\sqrt{\frac{DT_\star \log(T_\star M)}{K}} + \sqrt{DT_\star K \log N \sqrt{T_\star}}\right),$$

908 where $B = \lceil \log_2 \log_2 M \rceil$ and $N = \mathcal{O}(\log K)$.

909 *Proof.* We start with a regret decomposition into the regret of the meta algorithm w.r.t. finding j_\star and
910 the regret of the j_\star instance w.r.t. the best policy:

$$\begin{aligned}
\mathbb{E}[R_K] &= \mathbb{E}\left[\sum_{k=1}^K \sum_{j=1}^N p_k(j) \langle q_k^j, c_k \rangle - \sum_{k=1}^K \langle q_k^\star, c_k \rangle\right] \\
&= \mathbb{E}\left[\sum_{k=1}^K \sum_{j=1}^N p_k(j) \langle q_k^j, c_k \rangle - \langle q_k^{j_\star}, c_k \rangle\right] + \mathbb{E}\left[\sum_{k=1}^K \langle q_k^\star - q_k^{j_\star}, c_k \rangle\right]
\end{aligned}$$

$$= \underbrace{\mathbb{E} \left[\sum_{k=1}^K \langle p_k(j) - e_{j_\star}, c_k \rangle \right]}_{\text{Meta-Regret}} + \underbrace{\mathbb{E} \left[\sum_{k=1}^K \langle q_k^\star - q_k^{j_\star}, c_k \rangle \right]}_{j_\star - \text{Regret}}$$

911 where e_{j_\star} is the basis vector with the j_\star coordinate equal to 1. By Lemma C.1 the j_\star -Regret is
 912 bounded by a $\mathcal{O} \left(\sqrt{DTK \log(b(j_\star)M)} + BM \sqrt{\frac{Db(j_\star) \log(b(j_\star)M)}{K}} \right)$, where $b(j_\star) = \Theta(T_\star)$.

913 This also allows to say that:

$$\begin{aligned} \mathbb{E} \left[\sum_{k=1}^K \langle q_k^{j_\star}, c_k \rangle \right] &\leq \mathbb{E} \left[\sum_{k=1}^K \langle q_{\pi^\star}, c_k \rangle \right] + \mathcal{O} \left(\sqrt{DTK \log(b(j_\star)M)} + BM \sqrt{\frac{Db(j_\star) \log(b(j_\star)M)}{K}} \right) \\ &\leq \mathbb{E} \left[\sum_{k=1}^K \langle q_{\pi_f}, c_k \rangle \right] + \mathcal{O} \left(\sqrt{DTK \log(b(j_\star)M)} + BM \sqrt{\frac{Db(j_\star) \log(b(j_\star)M)}{K}} \right) \\ &\leq DK + \mathcal{O} \left(\sqrt{DTK \log(b(j_\star)M)} + BM \sqrt{\frac{Db(j_\star) \log(b(j_\star)M)}{K}} \right). \end{aligned}$$

914 For the meta algorithm regret, we can use Lemma 12 of [8] which guarantees that:

$$\sum_{k=1}^K \langle p_k - e_{j_\star}, c_k \rangle \leq \frac{2 + \log \left(N \sqrt{\frac{b(j_\star)}{b(1)}} \right)}{\eta_{j_\star}} + 4\eta_{j_\star} b(j_\star) \sum_{k=1}^K \langle q_k^{j_\star}, c_k \rangle.$$

915 Therefore, when $K \geq MT$:

$$\begin{aligned} \langle p_k - e_{j_\star}, c_k \rangle &\leq \mathcal{O} \left(\frac{\log N \sqrt{T_\star}}{\eta_{j_\star}} + \eta_{j_\star} T_\star \left(DK + \sqrt{DTK \log T_\star M} + BM \sqrt{\frac{DT_\star \log T_\star M}{K}} \right) \right) \\ &= \mathcal{O} \left(\frac{\log N \sqrt{T_\star}}{\eta_{j_\star}} + \eta_{j_\star} T_\star DK \right) = \mathcal{O} \left(\sqrt{DT_\star K \log N \sqrt{T_\star}} \right) \end{aligned}$$

916 with $\eta_j = \sqrt{\frac{\log N \sqrt{b(j)}}{b(j)DK}}$, giving the desired result.

917

□

D The benefits of sparsity - lower bound under sparsity

In this appendix, we prove our sparse lower bound result from Section 4.2. We first restate the result.

Theorem 4.4. *For any D, T_*, K, S, A with $T_* \geq D \geq 3 \log S$, $S(A-1) \geq 400$, $K \geq \frac{800T_*}{D} \log M$ and $M \geq 101$, there exists an SSP instance with stochastic M -sparse costs, S states and A actions such that its diameter is D , the expected hitting time of the optimal policy is T_* , and the expected regret with respect to the randomness of the losses for any learner after K episodes is $\mathbb{E}[R_K] \geq \Omega(\sqrt{KT_*D \log M})$.*

Proof. Fix $B = \lceil \frac{\log S/2}{\log 2} \rceil - 2$. Fix $N = 2^{B+1} \geq 2^{\frac{\log S/2}{\log 2} - 1} = \frac{S}{4}$. Fix $L = \min\{M-1, N \cdot (A-1)\} \geq \frac{M}{8}$. Fix $D' = D - B - 2$, $T' = T_* - B - 1$, with $T_* \geq D$.

We first describe the SSP instance with stochastic costs. Consider the following MDP $\mathcal{M} = (S, \mathcal{A}, p, s_0, g)$ illustrated in Figure 4 and that we formally define below.

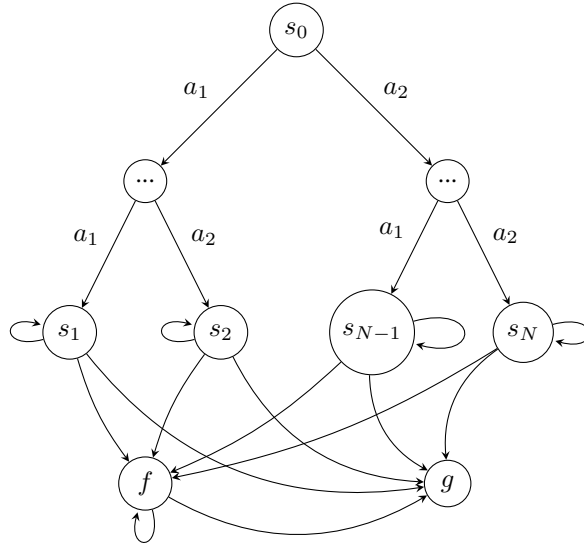


Figure 4: Diagram illustrating MDP construction for the proof of Theorem 4.4. Details are given below.

The first part of the states are represented by a binary tree of depth $B+2$ and allow us to formerly consider the N states at the bottom of the tree that matter, while avoiding an assumption on the existence of a state with $A \approx S$ actions as was done in prior work [8]. Each non-leaf node corresponds to a state with two actions transitioning (deterministically) to the left or right child respectively. The total number of nodes in the tree is

$$\sum_{i=0}^{B+1} 2^i = 2^{B+2} - 1 \leq 2^{\frac{\log S/2}{\log 2} + 1} - 1 = 2 \frac{S}{2} - 1 = S - 1.$$

The total number of leaf nodes is $N = 2^{B+1} \geq \frac{S}{4}$. Denote the set of states corresponding to the leaf nodes by $\mathcal{S}_\ell = \{s_1, \dots, s_N\}$. The root node is s_0 . There is also one additional state denoted by f (recall that the number of states in the tree is $\leq S-1$).

We consider the same action set across each state: $\mathcal{A} = \{a_1, \dots, a_{A-1}, a_f\}$. In the states of the binary tree where we have only described two actions, we can consider the other actions to remain in the same state deterministically with 0 cost.

The transitions and costs are defined as follows:

- For all states and actions in the tree that are not leaves, the transitions are specified above. The costs are all 0.
- For $s_i \in \mathcal{S}_\ell$, and $a_j \in \mathcal{A}$

- 944 – if $j = f$, then $p(f|s_i, a_f) = 1$ and $c_k(s_i, a_f) = 0$.
- 945 – if $j \in \{1, \dots, A-1\}$ and $j + (A-1) \cdot (i-1) \leq L$, then $p(g|s_i, a_j) = \frac{1}{T'}$, $p(s_i|s_i, a_j) =$
 946 $1 - \frac{1}{T'}$ and the cost is an independent sample from a Bernoulli distribution at each
 947 episode k : $c_k(s_i, a_j) \sim \text{Ber}\left(\frac{D'}{2T'}\right)$.
- 948 – if $j \in \{1, \dots, A-1\}$ and $j + (A-1) \cdot (i-1) > L$, then a_j is the same as a_f , i.e.
 949 $p(f|s_i, a_j) = 1$ and $c_k(s_i, a_j) = 0$.
- 950 • For f ,
 - 951 – $p(g|f, a_f) = \frac{1}{D'}$, $p(f|f, a_f) = 1 - \frac{1}{D'}$ and $c_k(f, a_f) = 1$.
 - 952 – for all $a_j \in \mathcal{A} \setminus \{a_f\}$, $p(f|f, a_j) = 1$ and $c_k(f, a_j) = 0$.

953 Denote the above distribution for c_k by \mathcal{D} . In each episode there are at most $L+1 \leq M$ non-zero
 954 costs, ensuring the condition on sparsity is respected.

955 For $i \in \{1, \dots, N\}$, let \mathcal{A}_i correspond to the actions in state $s_i \in \mathcal{S}_L$ which can transition directly to g
 956 and $\mathcal{A} \setminus \mathcal{A}_i$ corresponds to the actions which deterministically transition to f (e.g. if $(A-1) \cdot i \leq L$,
 957 then $\mathcal{A}_i = \{a_1, \dots, a_{A-1}\}$). For any proper policy π independent of the stochastically generated costs
 958 in episode k , we have

$$\begin{aligned}
 \mathbb{E}_{c_k \sim \mathcal{D}}[J_k^\pi(s_i)] &= \mathbb{E}_{c_k \sim \mathcal{D}}\left[\sum_{a \in \mathcal{A}_i} \pi(a|s_i) \left(c_k(s_i, a) + \left(1 - \frac{1}{T'}\right) J_k^\pi(s_i)\right) + J_k^\pi(f) \sum_{a \notin \mathcal{A}_i} \pi(a|s_i)\right] \\
 &= \sum_{a \in \mathcal{A}_i} \pi(a|s_i) \left(\mathbb{E}_{c_k \sim \mathcal{D}}[c_k(s_i, a)] + \left(1 - \frac{1}{T'}\right) \mathbb{E}_{c_k \sim \mathcal{D}}[J_k^\pi(s_i)]\right) + D' \cdot \sum_{a \notin \mathcal{A}_i} \pi(a|s_i) \\
 &= \sum_{a \in \mathcal{A}_i} \pi(a|s_i) \left(\frac{D'}{2T'} + \left(1 - \frac{1}{T'}\right) \mathbb{E}_{c_k \sim \mathcal{D}}[J_k^\pi(s_i)]\right) + D' \cdot \left(1 - \sum_{a \in \mathcal{A}_i} \pi(a|s_i)\right) \\
 \implies \mathbb{E}_{c_k \sim \mathcal{D}}[J_k^\pi(s_i)] &\left(1 - \left(1 - \frac{1}{T'}\right) \sum_{a \in \mathcal{A}_i} \pi(a|s_i)\right) = \frac{D'}{2T'} \sum_{a \in \mathcal{A}_i} \pi(a|s_i) + D' \cdot \left(1 - \sum_{a \in \mathcal{A}_i} \pi(a|s_i)\right) \\
 \implies \mathbb{E}_{c_k \sim \mathcal{D}}[J_k^\pi(s_i)] &= \frac{\frac{D'}{2T'} \sum_{a \in \mathcal{A}_i} \pi(a|s_i) + D' \cdot \left(1 - \sum_{a \in \mathcal{A}_i} \pi(a|s_i)\right)}{1 - \left(1 - \frac{1}{T'}\right) \sum_{a \in \mathcal{A}_i} \pi(a|s_i)} \\
 &= \frac{D'}{2} \cdot \frac{\frac{1}{T'} \sum_{a \in \mathcal{A}_i} \pi(a|s_i) + 2\left(1 - \sum_{a \in \mathcal{A}_i} \pi(a|s_i)\right)}{\frac{1}{T'} \sum_{a \in \mathcal{A}_i} \pi(a|s_i) + \left(1 - \sum_{a \in \mathcal{A}_i} \pi(a|s_i)\right)} \\
 &\geq \frac{D'}{2}.
 \end{aligned}$$

959 The optimal policy π^* is the policy that takes actions in the binary tree to reach state s_{i^*} and then
 960 $\pi^*(a_{j^*}|s_{i^*}) = 1$ for $i^*, j^* = \arg \min_{i, j: j + (A-1) \cdot i \leq L} \sum_{k=1}^K c_k(s_i, a_j)$. We have $J_k^{\pi^*}(s_0) = J_k^{\pi^*}(s_{i^*})$
 961 and for any $k \geq 1$

$$\begin{aligned}
 J_k^{\pi^*}(s_{i^*}) &= c_k(s_{i^*}, a_{j^*}) + \left(1 - \frac{1}{T'}\right) J_k^{\pi^*}(s_{i^*}) \\
 \implies J_k^{\pi^*}(s_0) &= T' c_k(s_{i^*}, a_{j^*}) \\
 \implies \sum_{k=1}^K J_k^{\pi^*}(s_0) &= T' \sum_{k=1}^K c_k(s_{i^*}, a_{j^*}) = T' \min_{i, j: j + (A-1) \cdot i \leq L} \sum_{k=1}^K c_k(s_i, a_j).
 \end{aligned}$$

962 Hence,

$$\mathbb{E}_{c_1, \dots, c_K \sim \mathcal{D}}[R_K] \geq \frac{D'}{2} \cdot K - T' \cdot \mathbb{E}_{c_1, \dots, c_K \sim \mathcal{D}}\left[\min_{i, j: j + (A-1) \cdot i \leq L} \sum_{k=1}^K c_k(s_i, a_j)\right]$$

$$\begin{aligned}
&= T' \cdot \left(\frac{D'}{2T'} \cdot K - \mathbb{E}_{c_1, \dots, c_K \sim \mathcal{D}} \left[\min_{i, j: j + (A-1) \cdot i \leq L} \sum_{k=1}^K c_k(s_i, a_j) \right] \right) \\
&= T' \cdot \mathbb{E}_{c_1, \dots, c_K \sim \mathcal{D}} \left[\max_{i, j: j + (A-1) \cdot i \leq L} \sum_{k=1}^K \left(\frac{D'}{2T'} - c_k(s_i, a_j) \right) \right]
\end{aligned}$$

963 We now apply Theorem F.1 with $p = 1 - \frac{D'}{2T'} \geq \frac{1}{2}$, $d = L \geq 100$ (since $S(A-1) \geq 400$ and
964 $M \geq 101$) and $n = K \geq \frac{800T_\star}{D} \log M \geq \frac{400T'}{D'} \log M \geq 200 \frac{p}{1-p} \log d$. We obtain:

$$\begin{aligned}
\sup_{c_1, \dots, c_K} \mathbb{E}[R_K] &\geq \mathbb{E}_{c_1, \dots, c_K \sim \mathcal{D}}[R_K] \geq 0.02T' \sqrt{K \left(1 - \frac{D'}{2T'}\right) \cdot \frac{D'}{2T'} \cdot \log L - 1.5T'} \\
&= \Omega\left(\sqrt{KT_\star D \log M}\right),
\end{aligned}$$

965 since $L \geq M/8$. Note that since $T_\star \geq D$, the hitting-time of the fast-policy is $D' + B + 2 = D$ and
966 the hitting time of the optimal is $T' + B + 1 = T_\star$, as required. This concludes the proof. \square

E Lower bound under unknown transitions

Theorem 5.1. *For any D, K, S, A with $S \geq 2, A \geq 16, D \geq 2$ and $K \geq SA$, there exists an SSP instance with $M = 1$, S states and A actions such that its diameter is D and the expected regret for any learner without knowledge of the transitions after K episodes is $\mathbb{E}[R_K] \geq \Omega(D\sqrt{SAK})$.*

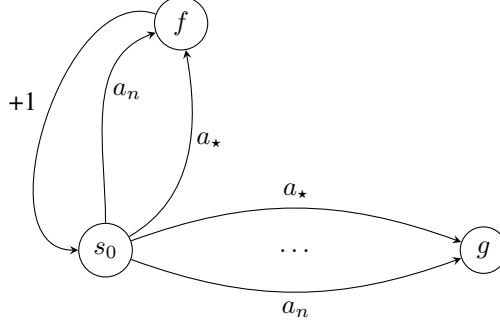


Figure 5: base case

Proof. The idea is to inject sparsity into the lower bound construction of [25] and to see if sparsity helps. Imagine the simple SSP in Figure 5, where at state s_0 there are A available actions, all with zero cost, while in the state f there is only one deterministic action with unit cost going back to s_0 . Among them, there exists an action a_* such that the transition probabilities are given by: $P(g \mid s_0, a) = \frac{1}{D} - \epsilon \mathbb{I}(a \neq a_*)$, and consequently, $P(f \mid s_0, a) = 1 - \frac{1}{D} + \epsilon \mathbb{I}(a \neq a_*)$. The cost is therefore only suffered when the selected action transitions to the f state. This will therefore not increase the hitting time of any proper deterministic policy while still inducing the desired sparsity.

Clearly, the optimal policy plays a_* at every time step to reach the goal as fast as possible and therefore $J^{\pi^*}(s_0) = D - 1$.

Now, denote with N_k the number of steps that the learner spends in s_0 in episode k and N_k^* the number of steps that the learner picks action a_* in episode k . Note that N_k is also the total cost that the learner suffers during episode k minus one (since the last transition will not be paid). Thanks to our construction we can still prove Lemma C.1 in [25] as follows:

Lemma E.1.

$$\mathbb{E}[N_k] - 1 - J^{\pi^*} = \epsilon \mathbb{E}[N_k - N_k^*]$$

Proof.

$$\begin{aligned} \mathbb{E}[N_k] &= \sum_{t=1}^{\infty} P[s_t = s_0] \\ &= 1 + \sum_{t=2}^{\infty} P[s_t = s_0] \\ &= 1 + \sum_{t=2}^{\infty} P[s_t = s_0 \mid s_{t+1} = s_0, a_{t-1} = a_*] P[s_{t+1} = s_0, a_{t-1} = a_*] \\ &\quad + \sum_{t=2}^{\infty} P[s_t = s_0 \mid s_{t+1} = s_0, a_{t-1} \neq a_*] P[s_{t+1} = s_0, a_{t-1} \neq a_*] \\ &= 1 + \left(1 - \frac{1}{D}\right) \sum_{t=2}^{\infty} P[s_{t+1} = s_0, a_{t-1} = a_*] + \left(1 - \frac{1-\epsilon}{D}\right) \sum_{t=2}^{\infty} P[s_{t+1} = s_0, a_{t-1} \neq a_*] \end{aligned}$$

Rearranging gives:

$$\mathbb{E}[N_k] - D = \epsilon \mathbb{E}[N_k - N_k^*]$$

Adding and subtracting 1 gives the desired result. \square

986 Hence:

$$\mathbb{E}[R_K] = \sum_{k=1}^K \sum_{i=1}^{I_k} c_k(s_k^i, a_k^i) - \sum_{k=1}^K J^{\pi^*}(s_0) = \mathbb{E}[N_K] - 1 - J^{\pi^*} = \epsilon[N - N^*]$$

987 where $N = \sum_{k=1}^K N_k$ and $N^* = \sum_{k=1}^K N_k^*$. Since we recovered Lemma C.1 in [25] as the starting
 988 block of the proof, following the derivation we can lower bound N in expectation and upper bound
 989 the expected value of N^* to retrieve

990 **Lemma E.2** (Theorem C.4 in [25]). *Suppose that $D \geq 2$, $\epsilon \in (0, 1/8)$ and $A > 16$, for the problem*
 991 *described above we have:*

$$\mathbb{E}[R_K] \geq \epsilon K D \left(\frac{1}{8} - 2\epsilon \sqrt{\frac{2K}{A}} \right)$$

992 Now consider the following MDP. Let \mathcal{S} be the set of states disregarding g and s_0 . The initial
 993 state s_0 has only one action which leads uniformly at random into one of the states $s \in \mathcal{S}$, where
 994 each one has its own optimal action a_s^* . Then the transition distributions are defined $P(g|a_s^*, s) =$
 995 $1/D$, $P(s|a_s^*, s) = 1 - 1/D$, and $P(g|a, s) = (1 - \epsilon)/D$, $P(s|a, s) = 1 - (1 - \epsilon)/D$ for any
 996 other action $a \in \mathcal{A} \setminus \{a_s^*\}$. Note that for each state, the learner is faced with a simple problem as
 997 the one described above. Therefore, we can apply Lemma E.2 for each state separately and lower
 998 bound the learner's expected regret the sum of the regrets suffered at each state, which would depend
 999 on the number of times each state $s \in \mathcal{S}$ is visited from the initial state. Since reaching each state
 1000 has uniform probability, there are many states (constant fraction) that are chosen $\Theta(K/S)$ times.
 1001 Summing the regret bounds and choosing ϵ , gives the desired bound.

1002 Denote by K_s the number of episodes the state $s \in \mathcal{S}$ is visited:

$$\mathbb{E}[R_K] \geq \sum_{s \in \mathcal{S}} \mathbb{E} \left[\epsilon K_s D \left(\frac{1}{8} - 2\epsilon \sqrt{\frac{2K_s}{A}} \right) \right] = \frac{\epsilon K D}{8} - 2\epsilon^2 D \sqrt{\frac{2}{A}} \sum_{s \in \mathcal{S}} \mathbb{E} [K_s^{3/2}]$$

1003 Then:

$$\sum_{s \in \mathcal{S}} \mathbb{E} [K_s^{3/2}] \leq \sum_{s \in \mathcal{S}} \sqrt{\mathbb{E} [K_s]} \sqrt{\mathbb{E} [K_s^2]} = \sum_{s \in \mathcal{S}} \sqrt{\mathbb{E} [K_s]} \sqrt{\mathbb{E} [K_s^2] + \mathbb{V} [K_s]} = \sum_{s \in \mathcal{S}} \sqrt{\frac{K}{S}} \sqrt{\frac{K^2}{S^2} + \frac{K(S-1)}{S^2}} \leq K \sqrt{\frac{2K}{S}}$$

1004 Leading to:

$$\mathbb{E}[R_K] \geq \frac{\epsilon K D}{8} - 2\epsilon^2 D K \sqrt{\frac{2K}{SA}} \geq \frac{1}{1024} D \sqrt{SAK}$$

1005 for $\epsilon = 1/64 \sqrt{SA/K}$ $K \geq SA$, concluding the proof. \square

F Lower bound on the maximum of asymmetric zero-mean random walks

We extend the lower bound of [21] to asymmetric zero-mean random-walks. We consider $p \geq 1/2$ because it simplifies the proof below (lower-bounding ψ by 1 and upper-bounding C in proof below) and is what we need in the proof of Section 4.4 in Appendix D (we use $p = 1 - D/2T^*$).

Theorem F.1. Fix $p \in [\frac{1}{2}, 1 - \frac{1}{n}]$. Consider random walks $Z_i^{(n)} = \sum_{t=1}^n X_t^i$, where

$$X_t^i = \begin{cases} -p, & \text{w.p. } 1-p \\ 1-p, & \text{w.p. } p. \end{cases}$$

If $n \geq 200 \frac{p}{1-p} \log d$ (also ensures that $p \leq 1 - \frac{1}{n}$) and $d \geq 100$. Then,

$$\mathbb{E}[\max_{1 \leq i \leq d} Z_i^n] \geq 0.02 \sqrt{np(1-p) \log d} - 1.5.$$

Proof. We follow the same lines as [21] who show a special case of the result for $p = 1/2$. We generalize it to $p > 1/2$.

Consider $Z^{(n)} = \sum_{t=1}^n X_t$, a random-walk of length n , then $B_n = Z^{(n)} + pn \sim B(n, p)$, Binomial distribution with parameters n and p .

F.1 1st part of the proof:

The 1st part of the proof is all about providing a lower bound on $\mathbb{P}(B_n \geq pn + t - 1)$ in (9) for any $t \in [1, np + 1]$.

Lemma F.2 (Generalized version of Lemma 4 of [21], Theorem 2 of [18]). Let n, k be integers satisfying $n \geq 1$ and $pn \leq k \leq n$. Define $x = \frac{k-pn}{\sqrt{p(1-p)n}}$. Then $B_n \sim B(n, p)$ satisfies

$$\mathbb{P}(B_n \geq k) \geq \sqrt{n} \binom{n-1}{k-1} p^{k-1/2} (1-p)^{n-k+1/2} \cdot \frac{1 - \Phi(x)}{\phi(x)},$$

where $\phi(x)$ is the PDF of a standard Normal and $\Phi(x)$ is the CDF. The proof can be found in [18].

Denote $D(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ as the KL-divergence between two Bernoullis.

Lemma F.3 (Generalized version of Theorem 5 of [21]). Let n, k be integers satisfying $n \geq 1$, and $np \leq k \leq n$. Define $x = \frac{k-pn}{\sqrt{p(1-p)n}}$. Then $B_n \sim B(n, p)$ satisfies

$$\mathbb{P}(B_n \geq k) \geq \frac{\exp\left(-nD\left(\frac{k}{n}, p\right)\right)}{e^{1/6} \sqrt{2\pi}} \cdot \frac{1 - \Phi(x)}{\phi(x)}.$$

Proof. For $k = n$, we verify the statement of the theorem directly. The left hand side is $\mathbb{P}(B_n \geq n) = p^n$. The right hand side is smaller because $\exp\{-nD(1, p)\} = p^n$ and for $x = \sqrt{n \frac{1-p}{p}} > 0$, we have $\frac{1-\Phi(x)}{\phi(x)} \leq \sqrt{2}$ (see e.g. Section 3.3 in [14]).

For $np \leq k < n$, we first bound the binomial coefficient $\binom{n}{k}$. Stirling's formula for the factorial [23] gives for any $n \geq 1$,

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n < n! < e^{1/12} \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

Since $0 < np \leq k \leq n - 1$, we can use this approximation for k, n and $n - k$ and obtain

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ &> \frac{n^n e^{-n} \sqrt{2\pi n}}{(e^{1/12} k^k e^{-k} \sqrt{2\pi k}) \cdot (e^{1/12} (n-k)^{n-k} e^{-(n-k)} \sqrt{2\pi(n-k)})} \\ &= \frac{1}{e^{1/6} \sqrt{2\pi}} \left(\frac{n}{k}\right)^k \left(\frac{n}{n-k}\right)^{n-k} \sqrt{\frac{n}{k(n-k)}} \end{aligned}$$

$$= \frac{1}{e^{1/6}\sqrt{2\pi}} \frac{1}{p^k(1-p)^{n-k}} \exp\left\{-nD\left(\frac{k}{n}, p\right)\right\} \sqrt{\frac{n}{k(n-k)}},$$

1031 since

$$\begin{aligned} D\left(\frac{k}{n}, p\right) &= \frac{k}{n} \log\left(\frac{k}{np}\right) + \left(1 - \frac{k}{n}\right) \log\left(\frac{1-k/n}{1-p}\right) = -\frac{k}{n} \log\left(\frac{np}{k}\right) - \frac{n-k}{n} \log\left(\frac{n(1-p)}{n-k}\right) \\ \implies \exp\left\{-nD\left(\frac{k}{n}, p\right)\right\} &= \left(\frac{np}{k}\right)^k \cdot \left(\frac{n(1-p)}{n-k}\right)^{n-k} = p^k(1-p)^{n-k} \left(\frac{n}{k}\right)^k \left(\frac{n}{n-k}\right)^{n-k}. \end{aligned}$$

1032 Since $k \geq 1$, we can write the binomial coefficient as $\binom{n-1}{k-1} = \frac{k}{n} \binom{n}{k}$. By Lemma F.2, we have

$$\begin{aligned} \mathbb{P}(B_n \geq k) &\geq \sqrt{n} \binom{n-1}{k-1} p^{k-1/2} (1-p)^{n-k+1/2} \cdot \frac{1-\Phi(x)}{\phi(x)} \\ &= \sqrt{n} \frac{k}{n} \binom{n}{k} p^{k-1/2} (1-p)^{n-k+1/2} \cdot \frac{1-\Phi(x)}{\phi(x)} \\ &\geq \frac{1}{e^{1/6}\sqrt{2\pi}} \frac{k}{\sqrt{n}} \sqrt{\frac{n}{k(n-k)}} \frac{p^{k-1/2} (1-p)^{n-k+1/2}}{p^k(1-p)^{n-k}} \exp\left\{-nD\left(\frac{k}{n}, p\right)\right\} \cdot \frac{1-\Phi(x)}{\phi(x)} \\ &= \frac{1}{e^{1/6}\sqrt{2\pi}} \sqrt{\frac{k}{n-k}} \cdot \sqrt{\frac{1-p}{p}} \exp\left\{-nD\left(\frac{k}{n}, p\right)\right\} \cdot \frac{1-\Phi(x)}{\phi(x)}. \end{aligned}$$

1033 The result follows from $\sqrt{\frac{k}{n-k}} \geq \sqrt{\frac{np}{n-np}} = \sqrt{\frac{p}{1-p}}$ for $np \leq k \leq n-1$. \square

1034 For $k = pn + xn$, the 2nd-order Taylor approximation of $u(x) = D\left(\frac{k}{n}, p\right) = D(p+x, p)$ around 0
1035 is $\frac{x^2}{2p(1-p)}$. We define $\psi : [-p, 1-p] \rightarrow \mathbb{R}$ as the ratio of the divergence and the approximation:

$$\psi(x) = D(p+x, p) \cdot \frac{2p(1-p)}{x^2}.$$

1036 In particular, we have that $1 \leq \psi(x) \leq \frac{p(1-p)}{(x+p)(1-p-x)}$ for $x \in [0, 1-p]$. This can be shown using
1037 Taylor's theorem on $u(x)$: for some $z \in [0, x]$,

$$\begin{aligned} D(p+x, p) &= \frac{u^{(2)}(z)}{2} x^2 = \left(\frac{1}{z+p} + \frac{1}{1-p-z}\right) \frac{x^2}{2} = \frac{x^2}{2(z+p)(1-p-z)} \\ \implies \frac{x^2}{2p(1-p)} &\leq D(p+x, p) \leq \frac{x^2}{2(x+p)(1-p-x)}, \end{aligned} \tag{8}$$

1038 since $\frac{1}{(x+p)(1-p-x)}$ is increasing on $[0, 1-p]$.

1039 Let $t \in [1, np+1]$ be a real number. By Lemma F.3 and Lemma 1 in [21] (also Mill's ratio for
1040 standard Gaussian [2]), we have

$$\begin{aligned} \mathbb{P}(B_n \geq pn+t-1) &= \mathbb{P}(B_n \geq [pn+t-1]) \\ &\geq \frac{\exp\left(-nD\left(\frac{[pn+t-1]}{n}, p\right)\right)}{e^{1/6}\sqrt{2\pi}} \cdot \frac{\pi}{\pi \frac{[pn+t-1]-np}{\sqrt{p(1-p)n}} + \sqrt{2\pi}} \\ &\geq \frac{\exp\left(-nD\left(\frac{pn+t}{n}, p\right)\right)}{e^{1/6}\sqrt{2\pi}} \cdot \frac{\pi}{\pi \frac{pn+t-np}{\sqrt{p(1-p)n}} + \sqrt{2\pi}} \\ &= \frac{\exp\left(-nD\left(p + \frac{t}{n}, p\right)\right)}{e^{1/6}\sqrt{2\pi}} \cdot \frac{\pi}{\pi \frac{t}{\sqrt{p(1-p)n}} + \sqrt{2\pi}} \\ &= e^{-1/6} \exp\left(-\frac{1}{2p(1-p)} \psi\left(\frac{t}{n}\right) \cdot \frac{t^2}{n}\right) \cdot \frac{1}{\frac{\sqrt{2\pi}t}{\sqrt{np(1-p)}} + 2}. \end{aligned} \tag{9}$$

1041 **F.2 2nd part of the proof:**

1042 We can now turn to the actual proof of the result. Define the event A equal to the case that at least
 1043 one of the $Z_i^{(n)}$ is greater or equal to $C\sqrt{np(1-p)\log d} - 1$. We will show this event / threshold
 1044 controls the expectation of the maximum. First, we define C and provide some upper and lower
 1045 bounds for it. Denote by $f(d) = \sqrt{2 - \frac{2\log\log d}{\log d}}$, then

$$C = C(d, n) = \frac{1}{\sqrt{\psi\left(\sqrt{\frac{2p(1-p)\log d}{n}}\right)}} \sqrt{2 - \frac{2\log\log d}{\log d}} = \frac{1}{\sqrt{\psi\left(\sqrt{\frac{2p(1-p)\log d}{n}}\right)}} f(d). \quad (10)$$

1046 We bound the two factors separately:

1047 • $z = \sqrt{\frac{2p(1-p)\log d}{n}} \in [0, \frac{1}{10}(1-p)]$ for $n \geq 200\frac{p}{1-p}\log d$ and so

$$1 \leq \psi(z) \leq \frac{p(1-p)}{(z+p)(1-p-z)} \leq \frac{1-p}{(1-p-\frac{1-p}{10})} = \frac{10}{9}. \quad (11)$$

1048 • The function $f(d)$ is as in [21]: decreasing on $(1, e^e]$, increasing on $[e^e, +\infty)$, and
 1049 $\lim_{d \rightarrow \infty} f(d) = \sqrt{2}$. Therefore for all $d \in [5, \infty)$,

$$1.12 \leq f(e^e) \leq f(d) \leq \max\{f(5), \sqrt{2}\} = \sqrt{2}$$

1050 This gives for $n \geq 200\frac{p}{1-p}\log d$,

$$1 \leq \frac{1.12}{\sqrt{10/9}} \leq C(d, n) \leq \sqrt{2} \quad (12)$$

1051 Since $p \geq 1/2$, if $n \geq 200\frac{p}{1-p}\log d$, then $n > \frac{200}{p(1-p)\log d}$ (if $d \geq 8$) and $n \geq 200\frac{1-p}{p}\log d$. The
 1052 above implies:

$$1 < C\sqrt{np(1-p)\log d} \leq np \leq np + 1. \quad (13)$$

1053 Finally, we bound the quantity of interest:

$$\begin{aligned} \mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n\right] &= \mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n | A\right] \cdot \mathbb{P}(A) + \mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n | A^C\right] \cdot (1 - \mathbb{P}(A)) \\ &\geq \mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n | A\right] \cdot \mathbb{P}(A) + \mathbb{E}\left[Z_1^{(n)} | A^C\right] \cdot (1 - \mathbb{P}(A)) \\ &= \mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n | A\right] \cdot \mathbb{P}(A) + \mathbb{E}\left[Z_1^{(n)} | Z_1^{(n)} < C\sqrt{np(1-p)\log d} - 1\right] \cdot (1 - \mathbb{P}(A)) \\ &\geq \mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n | A\right] \cdot \mathbb{P}(A) + \mathbb{E}\left[Z_1^{(n)} | Z_1^{(n)} \leq 0\right] \cdot (1 - \mathbb{P}(A)) \quad \text{by (13)} \\ &\geq (C\sqrt{np(1-p)\log d} - 1) \cdot \mathbb{P}(A) + \mathbb{E}\left[Z_1^{(n)} | Z_1^{(n)} \leq 0\right] \cdot (1 - \mathbb{P}(A)). \end{aligned} \quad (14)$$

1054 **First, we lower bound $\mathbb{E}\left[Z_1^{(n)} | Z_1^{(n)} \leq 0\right]$.** Let $\beta = \frac{1}{1 - \sqrt{\frac{2n}{\pi[n(1-p)](n-[n(1-p)])}}}$. For $n \geq \frac{200}{p(1-p)}\log d$,

1055 we have $n \geq \frac{205}{\pi p(1-p)} \geq \frac{200+\pi p}{\pi(1-p)p}$ and $\beta \leq \frac{10}{9}$.

1056 Then Lemma 2.2 in [22] combined with Lemma 8 in [12] give that for $Y_n \sim B(n, 1-p)$:

$$\mathbb{E}[Y_n | Y_n \geq n(1-p)] < n(1-p) + \beta\sqrt{np(1-p)} < n(1-p) + \frac{10}{9}\sqrt{np(1-p)}.$$

1057 Since $B_n = Z^{(n)} + pn \sim B(n, p)$ can be written as $n - Y_n$, we have:

$$\mathbb{E}\left[Z_1^{(n)} | Z_1^{(n)} \leq 0\right] = \mathbb{E}\left[B_n | B_n \leq np\right] - np$$

$$\begin{aligned}
&= \mathbb{E}[n - Y_n | n - Y_n \leq np] - np \\
&= n - \mathbb{E}[Y_n | Y_n \geq n(1-p)] - np \\
&\geq n - n(1-p) - \frac{10}{9}\sqrt{np(1-p)} - np \\
&= -\frac{10}{9}\sqrt{np(1-p)}. \tag{15}
\end{aligned}$$

1058 **Next, we lower-bound $\mathbb{P}(A)$:**

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^C) \\
&= 1 - \left(\mathbb{P}\left[Z_1^{(n)} < C\sqrt{np(1-p)\log d} - 1\right] \right)^d \\
&= 1 - \left(\mathbb{P}\left[B_n < np + C\sqrt{np(1-p)\log d} - 1\right] \right)^d \\
&= 1 - \left(1 - \mathbb{P}\left[B_n \geq np + C\sqrt{np(1-p)\log d} - 1\right] \right)^d \\
&\geq 1 - \exp\left(-d \cdot \mathbb{P}\left[B_n \geq np + C\sqrt{np(1-p)\log d} - 1\right]\right) \quad \text{since } 1 - x \leq e^{-x} \\
&\geq 1 - \exp\left(-d \cdot \frac{e^{-1/6} \exp\left(-\frac{1}{2p(1-p)}\psi\left(\frac{C\sqrt{np(1-p)\log d}}{n}\right) \cdot \frac{C^2 np(1-p)\log d}{n}\right)}{\frac{\sqrt{2\pi}C\sqrt{np(1-p)\log d}}{\sqrt{np(1-p)}} + 2}\right) \quad \text{using (9) and (13)} \\
&= 1 - \exp\left(-d \cdot \frac{e^{-1/6} \exp\left(-\frac{1}{2}\psi\left(C\sqrt{p(1-p)\log d/n}\right) \cdot C^2 \log d\right)}{\sqrt{2\pi}C\sqrt{\log d} + 2}\right) \\
&= 1 - \exp\left(-\frac{e^{-1/6} d^{1-\frac{C^2}{2}} \psi\left(C\sqrt{p(1-p)\log d/n}\right)}{\sqrt{2\pi}C\sqrt{\log d} + 2}\right) \\
&\geq 1 - \exp\left(-\frac{e^{-1/6} d^{1-\frac{C^2}{2}} \psi\left(\sqrt{\frac{2p(1-p)\log d}{n}}\right)}{2\sqrt{\pi}\log d + 2}\right) \quad \text{by (12).}
\end{aligned}$$

1059 We now use that $d^{1-\frac{C^2}{2}}\psi\left(\sqrt{2p(1-p)\log d/n}\right) = \log d$ by the definition of C in (10). Hence, we obtain:

$$\mathbb{P}(A) \geq 1 - \exp\left(-\frac{e^{-1/6} \log d}{2\sqrt{\pi}\log d + 2}\right) = 1 - g(d), \quad \text{for } g(d) = \exp\left(-\frac{e^{-1/6} \log d}{2\sqrt{\pi}\log d + 2}\right). \tag{16}$$

1060 **Putting everything together:** we plug (15) and (16) into (14) to get

$$\begin{aligned}
\mathbb{E}\left[\max_{1 \leq i \leq d} Z_i^n\right] &\geq (C\sqrt{np(1-p)\log d} - 1) \cdot (1 - g(d)) - \frac{10}{9}g(d)\sqrt{np(1-p)} \\
&= \frac{f(d) \cdot (1 - g(d))}{\sqrt{\psi\left(\sqrt{\frac{2p(1-p)\log d}{n}}\right)}} (\sqrt{np(1-p)\log d} - 2) - \frac{10}{9}g(d)\sqrt{np(1-p)} \quad \text{using (10)} \\
&\geq \frac{f(d)(1 - g(d))}{\sqrt{10/9}} (\sqrt{np(1-p)\log d} - 2) - \frac{10}{9}g(d)\sqrt{np(1-p)} \quad \text{using (11)} \\
&= \sqrt{np(1-p)\log d} \cdot \left(\frac{f(d)(1 - g(d))}{\sqrt{10/9}} - \frac{10}{9} \cdot \frac{g(d)}{\sqrt{\log(d)}} \right) - 2 \frac{f(d)(1 - g(d))}{\sqrt{10/9}} \\
&\geq 0.02\sqrt{np(1-p)\log d} - 1.5,
\end{aligned}$$

1061 for $d \geq 100$ (we also used that $\sqrt{np(1-p)\log d} > 2$ in the 2nd inequality). This gives the
1062 result. \square