A. List of Notations

Notation	Meaning
A	Action space
T	Time horizon
t	Time index $t = 0..T$
H	Number of trajectories in batch data
D	Number of factors in each state
$[D]$	The set $1, 2, .., D$.
Γ	Domain of each factor in a state and (dual notation) the number of possible values for the factor
M	Markov Decision Process
ρ	Distribution of first state in MDP
$\underline{X}$	Input variable (represents previous state)
$\underline{Y}$	Output variable (represents next state)
$\underline{Y}(i)$	The i'th variable in the output.
Ψ	A subset of indices
$\underline{X}(\Psi)$	The corresponding subset of variables to Ψ
Φ_i	Indices of the parents for variable i
m	$\max_i \lvert\Phi_i\rvert$
F_i	$\Gamma^{\lvert\Phi_i\rvert} \times A$, the set of all realization-action pairs for the parents of node i
$\hat{\Phi}_i$	Indices found by `G-SCOPE` for variable i
$n(instance)$	Number of observations in the data fitting the instance
Θ_i	The set of realization-action pairs observed more than $N(\epsilon, \delta)$ for each $\hat{\Phi}_i$
ψ_i	A value signifying policies mismatch (bigger means higher mismatch)

B. Proof of Main Theorem & Supporting Lemmas

The proof of Theorem 1 is broken down into parts.

B.1. The Simulation Lemma

In this subsection, we derive a simulation lemma for MDPs, which essentially says that for a fixed policy two MDPs with similar transition probability distributions will result in similar value functions. Our simulation lemma is the same as presented in (Abbeel & Ng, 2005). To formalize what we mean by "similar" MDPs, we introduce the following assumption.

Definition 3. *Let $M = \langle S, A, P, R, \rho \rangle$ be an MDP and $K \subseteq S \times A$. M and K define an* **induced MDP** *$M_K = \langle S, A, P_K, R_K, \rho \rangle$, where*

$$P_K(\underline{Y}|\underline{X}, a) = \begin{cases} P(\underline{Y}|\underline{X}, a) & \text{if } (\underline{X}, a) \in K \\ 1 & \text{if } (\underline{X}, a) \notin K \wedge \underline{Y} = \underline{X} \\ 0 & \text{otherwise} \end{cases}$$

and

$$R_K(\underline{X}, a) = \begin{cases} R(\underline{X}, a) & \text{if } (\underline{X}, a) \in K \\ 0 & \text{otherwise} \end{cases} .$$

Definition 4. *Let $\epsilon > 0$, $M = \langle S, A, P, R, \rho \rangle$ be an MDP, and $K \subseteq S \times A$. An* **ϵ-induced MDP** *$\widehat{M} = \langle S, A, \widehat{P}, \widehat{R}, \rho \rangle$ with respect to M and K, satisfies*

$$\begin{aligned} &\forall_{(\underline{X},a)\in K} \|P(\cdot|\underline{X}, a) - \widehat{P}(\cdot|\underline{X}, a)\|_1 \leq \epsilon \ , \\ &\forall_{(\underline{X},a)\notin K} \forall_{\underline{Y}\in S} \widehat{P}(\underline{Y}|\underline{X}, a) = P_K(\underline{Y}|\underline{X}, a) \ \text{, and} \\ &\forall_{(\underline{X},a)\in S\times A} \widehat{R}(\underline{X}, a) = R_K(\underline{X}, a) \ . \end{aligned}$$

Assumption 4. *A4(ϵ, δ, π) : Let $\epsilon > 0$, $\delta \in (0, 1]$, π be a policy, and $M = \langle S, A, P, R, \rho \rangle$. There exists an ϵ-induced MDP $\widehat{M}$ with respect to M and the subset of the state-action space $K \subseteq S \times A$, such that the probability of*

encountering a state-action pair that is not in K while following π in M is small:

$$\Pr\left[\exists_{t\in[T]}(\underline{X}_t, a_t) \notin K \mid M, \pi\right] \leq \delta \ . \tag{11}$$

Lemma 1. *(Simulation Lemma; Abbeel & Ng 2005) Suppose Assumption 4 holds with A4(ϵ, δ, π), then*

$$|\tilde{\nu} - \nu| \leq \delta T + \epsilon T^2 \ , \tag{12}$$

where $\tilde{\nu} = \rho^\top V^\pi_{\widehat{M}}$ and $\nu = \rho^\top V^\pi_M$.

Proof.

$$\begin{aligned}
|\nu - \tilde{\nu}| &= |\rho^\top V^\pi_M - \rho^\top V^\pi_{\widehat{M}}| \\
&= |\rho^\top V^\pi_M - (\rho^\top V^\pi_{M_K} - \rho^\top V^\pi_{M_K}) - \rho^\top V^\pi_{\widehat{M}}| && \text{Insert } 0 = (\rho^\top V^\pi_{M_K} - \rho^\top V^\pi_{M_K}) \\
&\leq |\rho^\top V^\pi_M - \rho^\top V^\pi_{M_K}| + |\rho^\top V^\pi_{M_K} - \rho^\top V^\pi_{\widehat{M}}| && \text{By the triangle inequality.} \\
&\leq \delta T + |\rho^\top V^\pi_{M_K} - \rho^\top V^\pi_{\widehat{M}}| && \text{By (11).}
\end{aligned}$$

We represent by $P^\pi_{M_K}, P^\pi_{\widehat{M}} \in \mathbb{R}^{S\times S}$ and $R \in \mathbb{R}^S$ the transition matrices and rewards induced by the policy π. For any matrix A, we denote by $\|A\|_p$ the p-induced matrix norm $\|\cdot\|$. Notice that:

$$\begin{aligned}
\left\|P^\pi_{M_K} - P^\pi_{\widehat{M}}\right\|_\infty &= \max_{1\leq i\leq S} \sum_{j=1}^n \left|P^\pi_{M_K}(s_j|s_i,\pi) - P^\pi_{\widehat{M}}(s_j|s_i,\pi)\right| && \text{Norm definition} \\
&= \max_{1\leq i\leq S} \sum_{j=1}^n \left|\sum_a \pi(a|s_i)(P_{M_K}(s_j|s_i,a) - P_{\widehat{M}}(s_j|s_i,a))\right| && \text{Policy decomposition} \\
&\leq \max_{1\leq i\leq S} \sum_a \pi(a|s_i) \sum_{j=1}^n \left|P_{M_K}(s_j|s_i,a) - P_{\widehat{M}}(s_j|s_i,a)\right| && \text{Triangle inequality} \\
&\leq \max_{1\leq i\leq S} \sum_a \pi(a|s_i)\epsilon = \epsilon && \text{By Definition 4}
\end{aligned}$$

In addition, we use the following result (page 254 in Bhatia 1997): For any two matrices X, Y and induced norm:

$$\|X^m - Y^m\| \leq mM^{m-1}\|X - Y\|, \tag{13}$$

where $M = \max(\|X\|, \|Y\|)$. Since $P^\pi_{M_K}, P^\pi_{\widehat{M}}$ are stochastic, this inequality holds for the ∞-induced norm with $M = 1$. Now:

$$\begin{aligned}
|\rho^\top V^\pi_{M_K} - \rho^\top V^\pi_{\widehat{M}}| &= \left|\rho^\top \sum_{t=0}^T (P^\pi_{M_K})^t R - \rho^\top \sum_{t=0}^T (P^\pi_{\widehat{M}})^t R\right| && \text{Sum of rewards over steps} \\
&= \left|\rho^\top \left(\sum_{t=0}^T (P^\pi_{M_K})^t - \sum_{t=0}^T (P^\pi_{\widehat{M}})^t\right) R\right| \\
&\leq \|\rho\|_1 \left\|\sum_{t=0}^T (P^\pi_{M_K})^t - \sum_{t=0}^T (P^\pi_{\widehat{M}})^t\right\|_\infty \|R\|_\infty && \text{Hölder inequality and submultiplicative norm} \\
&\leq \sum_{t=0}^T \left\|(P^\pi_{M_K})^t - (P^\pi_{\widehat{M}})^t\right\|_\infty && \text{Triangle inequality and bounded reward} \\
&\leq \left\|P^\pi_{M_K} - P^\pi_{\widehat{M}}\right\|_\infty \sum_{t=0}^T t && \text{Equation 13 for each summand with } m = t \\
&\leq \epsilon T^2 && \text{Definition 4 as seen above}
\end{aligned}$$

Therefore, we can combine the results to obtain:

$$|\nu - \tilde{\nu}| \leq \delta T + \epsilon T^2 \tag{14}$$

□

B.2. Bounding the L_1-error in Estimates of the Transition Probabilities

In this subsection, we consider the number of samples needed to estimate the transition probabilities of various realization-action pairs. The samples we receive are from a trajectory. Each trajectory is independent. Unfortunately, samples observed at timestep t may depend on samples observed at previous timesteps. So the samples within a trajectory may not be independent. Therefore, we cannot apply the Weissman inequality (Weissman et al., 2003), which requires the samples to be independent and identically distributed. Instead, we derive a bound based on a martingale argument.

Definition 5. *A sequence of random variables $X_0, X_1, \ldots$ is a* **martingale** *provided that for all $i \geq 0$, we have*

$$\mathbb{E}\left[|X_i|\right] < \infty \ \text{, and} \tag{15}$$

$$\mathbb{E}\left[X_{i+1} \mid X_0, X_1, X_2, \ldots, X_i\right] = X_i \ . \tag{16}$$

Theorem 2. *(Azuma's inequality) Let $\varepsilon > 0$ and $X_1, X_2, \ldots$ be a martingale such that $|X_{i+1} - X_i| < b_i$ for $i \geq 1$, then for all $m \geq 1$*

$$\Pr\left[|X_m - X_1| \geq \varepsilon\right] \leq 2\exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^{m} b_i}\right) \ . \tag{17}$$

Definition 6. *Let $X_1, X_2, \ldots, X_m$ be any set of random variables with support in Γ and $f : \Gamma^m \to \mathbb{R}$ is a function. A* **Doob martingale** *is the sequence*

$$B_0 = \mathbb{E}_{X_1, X_2, \ldots, X_m}\left[f(X_1, X_2, \ldots, X_m)\right] \ \text{, and}$$
$$B_i = \mathbb{E}_{X_{i+1}, X_{i+2}, \ldots, X_m}\left[f(X_1, X_2, \ldots, X_m) | X_1, X_2, \ldots, X_i\right] \ \text{, for } i = 1, 2, \ldots, m \ .$$

Lemma 2. *Let $\varepsilon > 0$, Γ be a finite set, $\vec{X} = \langle X_1, X_2, \ldots, X_m \rangle$ be a collection of $m \geq 1$ random variables with support in Γ generated by an unknown process, and $f_x(\vec{X}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{X_i = x\}$ for all $x \in \Gamma$. We denote by $\mu(x) = \mathbb{E}\left[f_x(\vec{X})\right]$ for all $x \in \Gamma$. Then*

$$\Pr\left[|f_x(\vec{X}) - \mu(x)| \geq \varepsilon\right] \leq 2\exp\left(\frac{-\varepsilon^2 m}{2}\right) \ , \tag{18}$$

for all $x \in \Gamma$ and

$$\Pr\left[\|\hat{\mu} - \mu\|_1 \geq \varepsilon\right] \leq 2|\Gamma|\exp\left(\frac{-\varepsilon^2 m}{2|\Gamma|^2}\right) \ , \tag{19}$$

where $\hat{\mu}(x) = f_x(\vec{X})$.

Proof. First, notice that $\vec{X}$ and $m \cdot f_x(\cdot)$ define a Doob martingale such that $|B_{i+1} - B_i| \leq 1$ for $i = 1, 2, \ldots m$. By applying Azuma's inequality, we obtain

$$\Pr\left[|B_m - B_0| \geq m\varepsilon\right] \leq 2\exp\left(\frac{-(\varepsilon m)^2}{2\sum_{i=1}^{m} 1}\right) \ ,$$
$$\Pr\left[|f_x(\vec{X}) - \mu(x)| \geq \varepsilon\right] \leq 2\exp\left(\frac{-\varepsilon^2 m}{2}\right) \ ,$$

which proves (18).

Now the union bound gives

$$\Pr\left[\|\hat{\mu}-\mu\| \geq \sum_{x\in\Gamma}\frac{\varepsilon}{|\Gamma|}\right] \leq \sum_{x\in\Gamma} 2\exp\left(\frac{-\varepsilon^2 m}{2|\Gamma|^2}\right) \ ,$$

which proves (19). □

Lemma 3. *Let $\epsilon, \delta > 0$, and $\Psi \subseteq [D]$, if there are*

$$N \geq \frac{2\Gamma^2}{\epsilon^2}\log\frac{2\Gamma}{\delta}$$

samples of the realization-action pair (v, a) obtained from independent trajectories of π_b, then

$$\|\Pr(Y(i)|X(\Psi)=v,a) - \widehat{\Pr}(Y(i)|X(\Psi)=v,a)\|_1 \leq \epsilon \ , \tag{20}$$

with probability at least $1-\delta$.

Proof. Since the samples are taken from the behavior distribution, $\widehat{\Pr}(Y(i)=y|X(\Psi)=v,a) = \frac{n(y,v,a)}{n(v,a)} = \frac{1}{n(v,a)}\sum_{k=1}^{N}\mathbb{I}\{Y_k(i)=y, X_k(\Psi)=v, a_k=a\}$. By Lemma 2:

$$\Pr(\|\Pr(Y(i)|X(\Psi)=v,a) - \widehat{\Pr}(Y(i)|X(\Psi)=v,a)\|_1 \geq \epsilon) \leq 2|\Gamma|\exp\left(\frac{-N\epsilon^2}{2|\Gamma|^2}\right) \tag{21}$$

Setting $\delta = 2|\Gamma|\exp(\frac{-N\epsilon^2}{2|\Gamma|^2})$ we obtain $N = \frac{2|\Gamma|^2}{\epsilon^2}\log\left(\frac{2|\Gamma|}{\delta}\right)$. □

B.3. Bounding the Number of Trajectories

In this subsection, we derive a bound on the number of trajectories needed to derive a model that evaluates the target policy accurately. Notice that the learned model does not need to be accurate everywhere – only the regions of the state space where the target policy is likely to visit (and in a FMDP only the parent realizations that the target policy is likely to visit). Our analysis takes advantage of this. When the behavior policy visits the parent realizations that the target policy is likely to visit, then the number of trajectories can be small. On the other hand, if the behavior policy never visits parent realizations that the target policy visits, then the number of trajectories may be infinite.

We will make use of the following Proposition proved in Li (2009).

Proposition 1. *(Li, 2009, Lemma 56) Let $k \in \mathbb{N}$, $\mu, \delta \geq (0,1)$, $B_1, B_2, \ldots B_m$ be a sequence of m independent Bernoulli random variables such that $\mathbb{E}[B_i] \geq \mu$ for $i = 1, 2, \ldots, m$, and*

$$m \geq \frac{2}{\mu}\left(k + \ln\frac{1}{\delta}\right) \ , \tag{22}$$

then

$$\Pr\left[\sum_{i=1}^{m} B_i \geq k\right] \geq 1-\delta \ . \tag{23}$$

Proposition 1 tells us the number of experiments we need to perform on a Bernoulli distribution to observe at least k successes with high probability. The following corollary modifies the statement of Proposition 1 to tell us the number of experiments we need to perform to see a high probability set of outcomes from a categorical distribution at least k times with high probability.

Corollary 1. *(to Proposition 1) Let $\delta \in (0,1]$, $k \geq 1$, Γ be a finite set, $\rho \in \mathcal{M}(\Gamma)$ be a probability distribution with outcomes from Γ, and $X_1, X_2, \ldots, X_m$ be independent random variables sampled from ρ. Let*

$S_m^k = \{x \in \Gamma \mid \sum_{i=1}^m \mathbb{I}\{X_i = x\} \geq k\}$ be the set of elements encountered k or more times and $\bar{S}_m^k = \Gamma \backslash S_m^k$ be its complement. If

$$m \geq \frac{2|\Gamma|}{\delta}\left(k + \ln\frac{|\Gamma|}{\delta}\right) \ , \tag{24}$$

then, with probability at least $1-\delta$,

$$\Pr_{x\sim\rho}\left[x \in \bar{S}_m^k\right] < \delta \ , \tag{25}$$

the set of outcomes visited less than k times has total probability mass less than δ.

Proof. Consider an infinite sequence of random variables $X_1, X_2, \ldots$ distributed according to ρ. Denote by $j[1] < j[2] < \cdots < j[k]$ the indices resulting in the event that $X_i \in S_i^k$ and $X_i \notin S_{i-1}^k$. Notice that we let an index $j[l]$ be infinite in the case that the event never occurs. However, Γ only contains $|\Gamma|$ elements, so an element can be added to $S_\cdot^k$ at most $|\Gamma|$ times. Notice that $S_{j[l]}^k = S_{j[l]+1}^k = ... = S_{j[l+1]-1}^k$ for $l = 1, 2, \ldots, |\Gamma| - 1$. We construct Bernoulli random variables

$$B_i = \begin{cases} 1 & \text{if } X_i \notin S_{i-1}^k \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

for $i \geq 1$. So $B_{j[l]+1}, B_{j[l]+2}, \ldots, B_{j[l+1]}$ are independent, identically distributed Bernoulli random variables for $l = 1, 2, \ldots, |\Gamma| - 1$ (but $B_{j[l]}$ and $B_{j[l+1]}$ are not independent). Suppose that $\Pr[B_{j[l]} = 1] \geq \delta$ for some $l \in \{1, 2, 3, \ldots, |\Gamma|\}$ (this is at least true for $\Pr[B_{j[1]} = 1] = 1 \geq \delta$), then by Proposition 1, (with $\mu \leftarrow \delta, \delta \leftarrow \frac{\delta}{|\Gamma|}$)

$$j[l+1] - (j[l]+1) \leq \frac{2}{\delta}\left(k + \ln\frac{|\Gamma|}{\delta}\right) \ ,$$

with probability at least $1 - \frac{\delta}{|\Gamma|}$. Since there are only $|\Gamma|$ elements in Γ, this can only happen at most $|\Gamma|$ times. Thus, by the union bound, after $m \geq \frac{2|\Gamma|}{\delta}\left(k + \ln\frac{|\Gamma|}{\delta}\right)$ samples, either all $|\Gamma|$ outcomes have been observed or $\Pr[B_m = 1] < \delta$, with probability at least $1 - |\Gamma|\frac{\delta}{|\Gamma|} = 1 - \delta$. If we have observed all $|\Gamma|$ elements then (25) holds trivially. On the other hand if $\Pr[B_m = 1] < \delta$, then

$$\begin{aligned} \delta &> \Pr_{x\sim\rho}[x \notin S_{m-1}^k] \ , && \text{By the definition of } B_m \text{ (26).} \\ &\geq \Pr_{x\sim\rho}[x \notin S_m^k], && \text{The probability of a success} \\ & && \text{decreases because } S_{m-1}^k \subseteq S_m^k. \\ &= \Pr_{x\sim\rho}[x \in \bar{S}_m^k] \ . \end{aligned}$$

□

Proposition 2. *Let $\delta \in (0, 1]$, $k \geq 1$, Γ be a finite set, $\rho, \mu \in \mathcal{M}(\Gamma)$ be probability distributions with outcomes from Γ, and $X_1, X_2, \ldots, X_n$ be independent random variables sampled from ρ. Let $S_n^k = \{x \in \Gamma \mid \sum_{i=1}^n \mathbb{I}\{X_i = x\} \geq k\}$ be the set of elements encountered k or more times and $\bar{S}_n^k = \Gamma \backslash S_n^k$ be its complement. If*

$$n \geq \frac{2|\Gamma|}{\delta}\left(k + \ln\frac{|\Gamma|}{\delta}\right) \ , \tag{27}$$

then, with probability at least $1-\delta$,

$$\Pr_{x\sim\mu}\left[x \in \bar{S}_n^k\right] < \psi\delta \ ,$$

where $\psi = \max_{x\in\Gamma} \frac{\mu(x)}{\rho(x)}$ (taking $\frac{0}{0} = 0$).

Proof. We want to show $\Pr_{x\sim\mu}\left[x \in \bar{S}_n^k\right] < \psi\delta$. By applying Corollary 1, we have that $\Pr_{x\sim\rho}\left[x \in \bar{S}_n^k\right] < \delta$ with probability at least $1 - \delta$. It suffices to show that $\Pr_{x\sim\mu}\left[x \in \bar{S}_n^k\right] \leq \psi \Pr_{x\sim\rho}\left[x \in \bar{S}_n^k\right] \leq \psi\delta$.

$$
\begin{aligned}
\Pr_{x\sim\mu}\left[x \in \bar{S}_n^k\right] &= \sum_{x\in\bar{S}_n^k} \mu(x) \\
&= \sum_{x\in\bar{S}_n^k} \mu(x)\frac{\rho(x)}{\rho(x)} \\
&= \sum_{x\in\bar{S}_n^k} \rho(x)\frac{\mu(x)}{\rho(x)} \\
&\leq \left(\max_{y\in\Gamma} \frac{\mu(y)}{\rho(y)}\right) \sum_{x\in\bar{S}_n^k} \rho(x) \\
&= \psi \Pr_{x\sim\rho}\left[x \notin S_n\right] \ .
\end{aligned}
$$

□

For completeness we introduce the following proposition that is used to prove our lemma.

Proposition 3. *(Osband & Van Roy, 2014) Let $\underline{Y}(i)$ be a set of variables indexed by $i \in [D]$, v_i a realization of $\underline{Y}(i)$, $v = (v_1, ..., v_D)$ and $\Pr_1, \Pr_2$ be two factorized probability distributions over $\underline{Y}$:*

$$
\Pr_j(\underline{Y}) = \prod_{i=1}^{D} \Pr_j(\underline{Y}(i)) \qquad j = 1, 2 \ . \tag{28}
$$

Then

$$
||\Pr_1(\underline{Y} = v) - \Pr_2(\underline{Y} = v)||_1 \leq \sum_{i=1}^{D} ||\Pr_1(\underline{Y}(i) = v_i) - \Pr_2(\underline{Y}(i) = v_i)||_1 \ . \tag{29}
$$

Lemma 4. *Let $\epsilon, \delta > 0$. If the number of trajectories*

$$
H \geq \frac{4AD\Gamma^m}{\delta}\left(\frac{2\Gamma^2}{\epsilon^2}\ln\left(\frac{4AD\Gamma^{m+1}}{\delta}\right) + \ln\left(\frac{2AD\Gamma^m}{\delta}\right)\right) \ ,
$$

then, with probability at least $1-\delta$, there is a subset of state-action pairs

$$
K = \left\{(\underline{X}, a) \in S \times A \mid \| \Pr(\underline{Y}|\underline{X}, a) - \widehat{\Pr}(\underline{Y}|\underline{X}, a)\|_1 \leq D\epsilon\right\} \ ,
$$

such that:

$$
\Pr\left[\exists_{t\in[T]}(\underline{X}_t, a_t) \notin K \mid M, \pi\right] < \frac{T\sum_{i=1}^{D}\psi_i\delta}{2D} \tag{30}
$$

where $\psi_i = \max_{(v,a)\in F_i} \frac{\sum_{t=1}^{T}\Pr(X_t(\Phi_i)=v, a_t=a|\pi)}{\sum_{t=1}^{T}\Pr(X_t(\Phi_i)=v, a_t=a|\pi_b)}$.

Proof. For every $i \in [D]$, we define the random variable $\underline{W}$:

For a given trajectory sample a time t uniformly and set $\underline{W} = (X_t(\Phi_i), a_t)$.

Notice that $\underline{W}$ is distributed according to the distribution induced by the behavior policy π_b and that $\underline{W}$ receives one of $A\Gamma^{|\Phi_i|}$ values. We denote the distribution of $\underline{W}$ by ρ and over the target policy by μ. Setting $k = \frac{2\Gamma^2}{\epsilon^2}\ln\left(\frac{2\Gamma}{\delta_1}\right)$ and using Proposition 2 we obtain that having:

$$
H \geq \frac{2A\Gamma^{|\Phi_i|}}{\delta_2}\left(\frac{2\Gamma^2}{\epsilon^2}\ln\left(\frac{2\Gamma}{\delta_1}\right) + \ln\frac{A\Gamma^{|\Phi_i|}}{\delta_2}\right) \ , \tag{31}
$$

samples from ρ, with probability at least $1 - \delta_2$,

$$\Pr_{(v,a)\sim\mu}\left[(v,a) : n(v,a) \leq \frac{2\Gamma^2}{\epsilon^2}\ln\left(\frac{2|\Gamma|}{\delta_1}\right)\right] < \psi_i\delta_2 \ ,$$

where $\psi_i = \max_{(v,a)\in F_i} \frac{\mu(v,a)}{\rho(v,a)} = \frac{\sum_{t=1}^{T}\Pr(X_t(\Phi_i)=v,a_t=a|\pi)}{\sum_{t=1}^{T}\Pr(X_t(\Phi_i)=v,a_t=a|\pi_b)}$ (taking $\frac{0}{0}=0$).

By Lemma 3 and given our choice for $k \equiv N(\epsilon,\delta_1)$, if we have observed $N(\epsilon,\delta_1)$ samples from $\Pr(\underline{Y}(i)|\underline{X}(\Phi_i) = v,a)$, then our estimate $\widehat{\Pr}$ satisfies

$$\left\|\Pr(\underline{Y}(i)|(\underline{X}(\Phi_i)=v,a)) - \widehat{\Pr}(\underline{Y}(i)|(\underline{X}(\Phi_i)=v,a))\right\|_1 \leq \epsilon \ ,$$

with probability at least $1-\delta_1$. Now denote by

$$K_i = \left\{(v,a)\in F_i \mid \left\|\Pr(\underline{Y}(i)|(\underline{X}(\Phi_i)=v,a)) - \widehat{\Pr}(\underline{Y}(i)|(\underline{X}(\Phi_i)=v,a))\right\|_1 \leq \epsilon\right\} \ ,$$

the set of realization-action pairs for predicting the i^{th} output variable where the empirical distribution estimated from trajectory data is ϵ-close to the true distribution.

By applying the union bound over at most $A\Gamma^{\Phi_i}$ realization-action pairs, after H trajectories, we have

$$\Pr\left[\exists_{t\in[T]}(\underline{X}_t(\Phi_i),a_t) \notin K_i \mid M,\pi\right] \leq T\psi_i\delta_2 \ ,$$

with probability at least $1-(\delta_2+\delta_1 A\Gamma^{|\Phi_i|})$. By applying the union bound again over all D output variables, we obtain

$$\sum_{i=1}^{D}\Pr\left[\exists_{t\in[T]}(\underline{X}_t(\Phi_i),a_t) \notin K_i \mid M,\pi\right] \leq T\sum_{i=1}^{D}\psi_i\delta_2$$

with probability at least $1 - D(\delta_2 + A\Gamma^{|\Phi_i|}\delta_1)$. Notice that this implies

$$\begin{aligned}\Pr\left[\exists_{t\in[T]}(\underline{X}_t,a_t) \notin K \mid M,\pi\right] &\leq \sum_{i=1}^{D}\Pr\left[\exists_{t\in[T]}(\underline{X}_t(\Phi_i),a_t) \notin K_i \mid M,\pi\right] \\ &\leq T\sum_{i=1}^{D}\psi_i\delta_2 \ ,\end{aligned}$$

holds with probability at least $1 - D(\delta_2 + A\Gamma^{|\Phi_i|}\delta_1) \geq 1 - D(\delta_2 + A\Gamma^m\delta_1)$.

The bound over $\|\Pr(\underline{Y}|\underline{X},a) - \widehat{\Pr}(\underline{Y}|\underline{X},a)\|_1$ directly results from Proposition 3.

Setting:

$$\begin{aligned}\frac{\delta}{2} &= D\delta_2 \ \Rightarrow \delta_2 = \frac{\delta}{2D} \\ \frac{\delta}{2} &= \delta_1 AD|\Gamma|^m \ \Rightarrow \delta_1 = \frac{\delta}{2AD\Gamma^m}\end{aligned} \tag{32}$$

We can rewrite H in terms of ϵ,δ:

$$H \geq \frac{4AD\Gamma^m}{\delta}\left(\frac{2\Gamma^2}{\epsilon^2}\ln\left(\frac{4AD\Gamma^{m+1}}{\delta}\right) + \ln\left(\frac{2AD\Gamma^m}{\delta}\right)\right) \tag{33}$$

□

B.4. Error due to Greedy Parent Selection

Lemma 5. *Suppose Assumptions 1, 2 and 3 hold. Let* $\epsilon > 0, \delta_1 > 0$*, and*

$$\frac{C_1}{4} > \epsilon + \frac{C_2}{4} \ .$$

After applying G-SCOPE*, for every* $i \in [D]$*,* $(v, a) \in \Theta_i$*, and every* $w \in \Gamma^{|\Phi_i|}$ *satisfying* $w(\hat{\Phi}_i) = v(\Phi_i), N(w, a) \geq N(\epsilon, \delta_1)$*:*

$$\| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i) = w, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \leq (4D + 1)\epsilon + D^2 C_3 \tag{34}$$

with probability at least $1 - 2D(m + 1)(D + 1 - m)A\Gamma^{m+1}\delta_1$.

Proof. This Lemma is only concerned about realization-action pairs for which there are enough samples. G-SCOPE will not consider the score of realization-action pairs that do not have enough sample. When constructing the structure, this automatically discard realization-action pairs containing non-parents that do not meet the number of samples required to have an estimation error bounded by ϵ with high probability. Hence, in what follows, we will always consider the worse case where there are always enough samples to estimate such probabilities.

To simplify notation, let

$$\widehat{\alpha}(k, v, v_k, a) = \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v, v_k), a)\|_1 \tag{35}$$

$$\alpha(k, v, v_k, a) = \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v, v_k), a)\|_1 \tag{36}$$

$$\alpha^*(k) = \max_{v, v_k, a} \alpha(k, v, v_k, a) \tag{37}$$

$$(v^*, v_k^*, a^*) = \arg\max_{v, v_k, a} \widehat{\alpha}(k, v, v_k, a) \tag{38}$$

$$\widehat{\alpha}^*(k) = \alpha(k, v^*, v_k^*, a^*) \tag{39}$$

$$. \tag{40}$$

We want to bound the probability that G-SCOPE adds any non-parent variable. The G-SCOPE algorithm can only select a variable k to add to the parent set only if the following sufficient condition holds:

$$\widehat{\alpha}^*(k) > \max_{j \in D \setminus \hat{\Phi}_i} \max_{v, v_j, a} \widehat{\alpha}(j, v, v_j, a) \ . \tag{41}$$

We break up this first part of the proof into two distinct, successive cases.

1. $\exists k \in \Phi_i^s$ that is not in $\hat{\Phi}_i$ (G-SCOPE has not added all of the strong parents yet), and
2. $\Phi_i^s \subseteq \hat{\Phi}_i$ (G-SCOPE has added all strong parents).

Case 1 (G-SCOPE has not added all of the strong parents):
Let $k \in \Phi_i^s$ that has not been added yet ($k \notin \hat{\Phi}_i$) such that k verifies Assumption 1, and j be a non-parent variable. We know such a k and corresponding realization-action pair which had been exhibited $N(w, a)$ times exist, since we assume there is at least one realization-action pair of the full parents with enough samples (since otherwise the requested bound holds trivially). We want to bound the probability that

$$\widehat{\alpha}^*(k) - \max_{v, v_j, a} \widehat{\alpha}(j, v, v_j, a) > 0 \ , \tag{42}$$

If (42) holds for any non-parent j, then G-SCOPE will only add parents from Φ_i. For (42) to hold, it is sufficient to have

$$\widehat{\alpha}^*(k) - \widehat{\alpha}(j, v, v_j, a) > 0 \qquad \forall j \in [D] \setminus \Phi_i, v, v_j, a \ , \tag{43}$$

By applying the triangle inequality, we obtain

$$\begin{aligned}\alpha(k) = \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \leq \\ \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v, v_k), a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v, v_k), a)\|_1 + \\ \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 + \\ \widehat{\alpha}(k) \ ,\end{aligned} \tag{44}$$

and

$$\begin{aligned}\widehat{\alpha}(j) = \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a)\|_1 \leq \\ \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a)\|_1 + \\ \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a)\|_1 + \\ \alpha(j) \ .\end{aligned} \tag{45}$$

By applying equations 44 and 45, Lemma 3 (with our choice of $N(\epsilon, \delta_1)$) and Assumption 1,

$$\begin{aligned}\widehat{\alpha}^*(k) - \widehat{\alpha}(j, v, v_j, a) \geq \\ &\alpha^*(k) \\ &- \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v^*, v_k^*), a^*) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v^*, v_k^*), a^*)\|_1 \\ &- \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v^*, a^*) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v^*, a^*)\|_1 \\ &\quad - \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a)\|_1 \\ &\quad - \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \\ &\quad - \alpha(j, v, v_j, a) \\ \geq \quad &\alpha^*(k) - \alpha(j, v, v_j, a) - 4\epsilon \\ \geq \quad &C_1 - 4\epsilon > 0\end{aligned}$$

with probability at least $1 - 4\delta_1$ (union bound) for a particular v, v_j, a if $C_1 > 4\epsilon$. This holds for all j, v, v_j, a with probability at least $1 - (2 + 2(D - m)A\Gamma^{|\hat{\Phi}_i|+1})\delta_1$ (union bound again).

This also means all variables in Φ_i^s will all be detected by `G-SCOPE` , as at each iteration at least one strong parent k has a score over the threshold. Using the triangle inequality, the same bounds on

$$\| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v^*, v_k^*), a^*) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v^*, v_k^*), a^*)\|_1 \tag{46}$$

$$\|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v^*, a^*) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v^*, a^*)\|_1 \tag{47}$$

as above, the following inequality derived from Assumption 1:

$$\alpha^*(k) \geq \max_{j \in [D] \setminus \Phi_i} \alpha^*(j) + C_1 \geq C_1 \ , \tag{48}$$

and the fact that $C_1 > 4\epsilon + C_2$, we have

$$\begin{aligned}\widehat{\alpha}^*(k) \geq \\ &\alpha^*(k) \\ &\quad - \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v^*, v_k^*), a^*) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{k\}) = (v^*, v_k^*), a^*)\|_1 \\ &\quad - \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v^*, a^*) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v^*, a^*)\|_1 \\ \geq \quad &C_1 - 2\epsilon > C_2 + 2\epsilon \ ,\end{aligned}$$

with probabilities $1 - 2\delta_1$.

Notice that we made Assumption 1 much stronger than needed as we demanded the strong parent to stand out for all its possible realizations. For the proof, we only need to ensure that at least one realization verifying Assumption 1 is seen enough times to make sure a strong parent is preferred. Alternatively, we could modify assumption 1 to bound the probability of not acquiring enough samples for a particular realization that has a sufficiently large score. This would have negligible impact on the bounds of this lemma, the assumption would be weaker, but its presentation in the body of the paper would be more complex.

Case 2 (`G-SCOPE` has added all strong parent variables):
Now, we bound the probability that `G-SCOPE` adds a non parent variable j if all strong parents variable Φ_i^s have already been added, that is, $\Phi_i^s \subseteq \hat{\Phi}_i \subseteq \Phi_i$:

$$\begin{aligned}\widehat{\alpha}(j) \leq & \\ & \|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i \cup \{j\}) = (v, v_j), a)\|_1 + \\ & \| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 + \\ & \underbrace{\alpha(j)}_{\leq C_2 \text{ because of Assumption 2}} \leq C_2 + 2\epsilon,\end{aligned}$$

with probability at least $1-2\delta_1$ (according to Lemma 3) for a particular v, v_j, a and with probability at least $1-2(D-m)A\Gamma^{|\hat{\Phi}_i|+1}\delta_1$ for all v, v_j, a.

Combining Case 1 & 2:
These two points must hold for all stages of the algorithm.

- They must also hold for each iteration building $\hat{\Phi}_i$. Iterations in the first step correspond to all strong parents, and $\Phi_i^{w,1} \in \Phi_i^w$, the weak parents added in step 1 (before all strong parents are included). The number of iterations in the second point is at most all remaining weak parents $\Phi_i^{w,2} \subseteq \Phi_i^w \backslash \Phi_i^{w,1}$ added in the second step, plus one (when the algorithm stops). Note that the probability the first point holds is only $1-(2+2(D-m)A\Gamma^{|\hat{\Phi}_i|+1})\delta_1$ and not $1-(4+2(D-m)A\Gamma^{|\hat{\Phi}_i|+1})\delta_1$ because we are using the same two bounds involving k twice.
- They must hold for all D target variable i.

Let $\kappa = 2(D-m)A\Gamma^{m+1} \geq 2(D-m)A\Gamma^{|\hat{\Phi}_i|+1}$. Using the union bound, these points hold for all stages of the algorithm with at least probability

$$\begin{aligned}1 - \max_i \left((|\Phi_i^s| + |\Phi_i^{w,1}|)(2+\kappa) + (|\Phi_i^{w,2}|+1)\kappa\right)\delta_1 D &\geq 1 - (\max_i |\Phi_i|(2+\kappa)+\kappa)D\delta_1 && (49)\\ &\geq 1 - (2m + (m+1)\kappa)D\delta_1 && (50)\\ &\geq 1 - 2D\left[m + (m+1)(D-m)A\Gamma^{m+1}\right]\delta_1 && (51)\end{aligned}$$

Transitioning from Probabilities over $\hat{\Phi}_i$ to Probabilities over Φ_i:
We define Φ_i^k to be the union of $\hat{\Phi}_i$ with the first k variables in $\Phi_i \setminus \hat{\Phi}_i$ to be added greedily (according to the true probabilities) for the specific (w, a) pair. Also, denote $w = (v, \bar{v}_1^{|\Phi_i \setminus \hat{\Phi}_i|})$.

$$\begin{aligned}&\| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i) = (v, \bar{v}), a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \\ &\leq \sum_{k=1}^{|\Phi_i \setminus \hat{\Phi}_i|} \| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i^k) = (v, \bar{v}_1^k), a) - \Pr(\underline{Y}(i)|\underline{X}(\Phi_i^{k-1}) = (v, \bar{v}_1^{k-1}), a)\|_1 \\ &\quad + \| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \ .\end{aligned} \tag{52}$$

The inequality is due to the triangle inequality - we observe the quality of adding each additional parent, and are left with the estimation error on v. Since the parents were added greedily, by Assumption 3 we can form a bound for the

sum. Since we have enough samples of v (it's in Θ_i) the second term is small with high probability (by Lemma 3):

$$\begin{aligned}
&\leq m\| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i^1) = (v,\bar{v}_1), a) - \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 + m^2 C_3 + \epsilon \ , \\
&\leq m\| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i^1) = (v,\bar{v}_1), a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\Phi_i^1) = (v,\bar{v}_1), a)\|_1 \\
&\quad + m\| \Pr(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \\
&\quad + m\|\widehat{\Pr}(\underline{Y}(i)|\underline{X}(\Phi_i^1) = (v,\bar{v}_1), a) - \widehat{\Pr}(\underline{Y}(i)|X(\hat{\Phi}_i) = v, a)\|_1 + m^2 C_3 + \epsilon \ .
\end{aligned} \tag{53}$$

Where the inequality holds from the triangle inequality. Similar to before, the first two summands can be bounded by ϵ with probability $1-\delta_1$. The third summands is bounded by the algorithm - since $\bar{v}_1$ was not added to $\hat{\Phi}_i$, and there were enough samples from it ($N(w,a) \geq N(\epsilon,\delta_1)$), it is necessarily smaller than the threshold $2\epsilon + C_2$, with probability $1-2\delta_1$ for a specific i and $1-2D\delta_1$ for all of them. Therefore, the difference is bounded by: $(4m+1)\epsilon + mC_2 + m^2C_3$ for these probabilities.

Everything together:

$$\| \Pr(\underline{Y}(i)|\underline{X}(\Phi_i) = (v,\bar{v}), a) - \widehat{\Pr}(\underline{Y}(i)|\underline{X}(\hat{\Phi}_i) = v, a)\|_1 \leq (4m+1)\epsilon + mC_2 + m^2C_3 \tag{54}$$

with at least probability (union bound)

$$1 - 2D\left[m + (m+1)(D-m)A\Gamma^{m+1}\right]\delta_1 - 2D\delta_1 = 1 - 2D(m+1)\left[1 + (D-m)A\Gamma^{m+1}\right]\delta_1 \tag{55}$$

which is lower bounded by $1 - 2D(m+1)(D+1-m)A\Gamma^{m+1}\delta_1$ or $1 - 2D(D+1)^2A\Gamma^{m+1}\delta_1$

□

B.5. Proof of Theorem 1

Theorem 1. *Suppose Assumptions 1, 2 and 3 hold. Let* $\frac{C_1}{4} > \epsilon + \frac{C_2}{4}, \epsilon > 0, \delta_1 > 0$, *and* $m = \max_{i\in[D]} |\Phi_i|$, *then there exists*

$$H(\epsilon,\delta_1) = O\left(\frac{\Gamma^2}{\delta_1\epsilon^2}\ln\left(\frac{\Gamma}{\delta_1}\right)\right)$$

such that if G-SCOPE *is given* H *trajectories, with probably at least* $1-2AD(m+2)(D+1-m)\Gamma^{m+1}\delta_1$, G-SCOPE *returns an evaluation of* π *satisfying:*

$$|\nu - \tilde{\nu}| \leq \delta^* T + \epsilon^* D T^2 \tag{56}$$

where

$$\begin{aligned}
\epsilon^* &= (4m+1)\epsilon + mC_2 + m^2C_3 \\
\delta^* &= T\sum_{i=1}^{D}\psi_i A\Gamma^m\delta_1 \\
\psi_i &= \max_{(v,a)\in F_i}\frac{\sum_{t=1}^{T}\Pr(\underline{X}_t(\Phi_i) = v, a_t = a|\pi)}{\sum_{t=1}^{T}\Pr(\underline{X}_t(\Phi_i) = v, a_t = a|\pi_b)} \ .
\end{aligned} \tag{57}$$

Proof. 1. By Lemma 4, given

$$H(\epsilon,\delta') \geq \frac{4AD\Gamma^m}{\delta'}\left(\frac{2\Gamma^2}{\epsilon^2}\ln\left(\frac{4AD\Gamma^{m+1}}{\delta'}\right) + \ln\left(\frac{2AD\Gamma^m}{\delta'}\right)\right)$$

trajectories there is a partition of Γ into more (set K) and less likely (v,a) pairs with probability at least $1-\delta'$. Pairs in set K are seen at least $N(\epsilon,\delta_1)$ times.

2. Since these pairs in K are seen at least $N(\epsilon,\delta_1)$ times, Lemma 5 provides a bound on the estimation error on the conditional transition probabilities in the FMDP constructed by G-SCOPE that holds with probability at least $1 - 2D(m+1)(D+1-m)A\Gamma^{m+1}\delta_1$.

3. This FMDP is therefore an $D\epsilon^*$-induce MDP with respect to the original MDP and K (Definition 4, Proposition 3 and Lemma 4).

4. Therefore, $A4(D\epsilon^*, \delta^*, \pi)$ is verified with probability at least (union bound on steps 1 and 2)

$$1 - (1 + (m+1)(D+1-m)\Gamma)\delta' \geq 1 - (m+2)(D+1-m)\Gamma\delta'$$

for

- $\epsilon^* = (4m+1)\epsilon + mC_2 + m^2C_3$,
- $\delta^* = T\sum_{i=1}^{D} \psi_i \delta'/2D$.

5. These values are then substituted into the simulation Lemma, and we replace $\delta' = 2AD\Gamma^m\delta_1$ (equation 32) to obtain the specified result.

□

C. Flat Model

We can find the bounds for a flat model, by assuming we have only one factor with $\Gamma' = \Gamma^D$ possible values. We can make the following changes and then substitute into the theorem:

$$\begin{aligned} D \leftarrow 1, m \leftarrow 1 \\ C_2 \leftarrow 0, C_3 \leftarrow 0 \\ \Gamma \leftarrow \Gamma^D \end{aligned} \tag{58}$$

Theorem 1. *Let $\epsilon > 0, \delta_1 > 0$, then there exists*

$$H(\epsilon, \delta_1) = O\left(\frac{D\Gamma^{2D}}{\delta_1\epsilon^2} \ln\left(\frac{\Gamma}{\delta_1}\right)\right)$$

such that if `G-SCOPE` *is given H trajectories, with probably at least $1 - 6A\Gamma^{2D}\delta_1$,* `G-SCOPE` *returns an evaluation of π satisfying:*

$$|\nu - \tilde{\nu}| \leq \delta^* T + \epsilon^* T^2 \tag{59}$$

where

$$\begin{aligned} \epsilon^* &= 5\epsilon \\ \delta^* &= T\psi A\Gamma^D\delta_1 \\ \psi &= \max_{(v,a)} \frac{\sum_{t=1}^{T} \Pr(\underline{X}_t = v, a_t = a|\pi)}{\sum_{t=1}^{T} \Pr(\underline{X}_t = v, a_t = a|\pi_b)} \quad . \end{aligned} \tag{60}$$

While the sample complexity grew exponentially, the probability of error and approximation quality did not change substantially.

D. Tree Implementation

The actual implementation we used is not greedy parent set increase, but rather a decision tree based variation. A different tree is generated for each action and output variable. In each node, we find the variable and realization which produces the largest difference described by the original algorithm, and split according to that variable - a new subtree for each possible realization ($|\Gamma|$ subtrees). Similarly to the set variation, we do not develop nodes without enough samples. When the tree is constructed, we can simply use it to simulate the model, and there is no need to build the parent set. We summarize the main differences between the tree and the greedy set variations:

1. **Implementation** - The tree variation is easy to embed in any standard decision tree code - we only change the decision rule, and provide the minimum number of samples per node.

2. **Sample complexity** - If not all realizations require all parents, the tree variation requires substantially less samples. Consider for example 3 binary variables, X_1, X_2, X_3. If the output variable is a noisy $X_1X_2 \oplus \overline{X}_1X_3$, then in the tree implementation we consider either $(X_1 = 1, X_2 = b)$ or $(X_1 = 0, X_3 = b)$ realizations. However, with the greedy set variation considers all possible realizations of these variables, which overall increases the number of samples.

3. **Statistical power** - The main disadvantage of the tree variation, is that in the worst case scenario where all parents affect all realizations, we will perform multiple tests on the same parent deciding whether to add it or not. If the total number of parents is $|\Phi_i|$, then so will be the tree depth. Since a test is performed in every node, in the worst case the tree is full, leading to $|\Gamma|^{|\Phi_i|}$ many nodes where in each a test is performed. Thus, applying the union bound as was done in the proof of Lemma 5 will result in an additional exponential dependency. That is the reason this implementation was not considered in the paper.

We used three parameters in the experiments:

1. Number of samples needed to perform a split - set to 2.

2. Minimum amount of L1 error needed to be introduced to split on a variable - set to 0.001.

3. Maximum tree depth - set to 15 to avoid running out of memory.

Finally, decision tree regression was used to learn the rewards in the Atari domain. This may also account for some of the error.

References

Abbeel, P. and Ng, A. Y. Exploration and apprenticeship learning in reinforcement learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 1997.

Li, L. *A Unifying Framework for Computational Reinforcement Learning Theory*. PhD thesis, Rutgers University, 2009.

Osband, I. and Van Roy, B. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pp. 604–612, 2014.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs, 2003.