Electronic Companion

EC.1. Proof of Theorem 1

Before proving the theorem, we need to derive some lemmas. Both approaches, stratifying and not stratifying, use a policy obtained from disturbance distributions estimated using data. Thus, the theoretical analysis requires understanding how well a policy that is optimal for one disturbance distribution performs for another disturbance distribution. We first introduce technical lemmas about two MDPs that differ only in disturbance distributions. These results are of interest by themselves in other contexts as well. The first lemma is from Cooper and Rangarajan (2012).

LEMMA EC.1 **(Lemma S-2 in Cooper and Rangarajan 2012)**. *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with constant ρ. Suppose also that F, G are distribution functions such that $||F - G|| \leq \epsilon$, $F(\alpha-) = G(\alpha-)$, and $F(\beta) = G(\beta)$, where $\alpha \leq \beta$. Then, $\left|\int_\alpha^\beta f(z)dF(z) - \int_\alpha^\beta f(z)dG(z)\right| \leq \epsilon\rho(\beta - \alpha)$.*

We use the above lemma to prove the (a) part of the theorem. For the (b) part, we use the below lemma.

LEMMA EC.2. *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is absolutely continuous and that its derivative f' is bounded almost everywhere on $[\alpha, \beta]$. Suppose also that F, G are distribution functions such that $||F - G|| \leq \epsilon$, $F(\alpha-) = G(\alpha-)$, and $F(\beta) = G(\beta)$. Then, $\left|\int_\alpha^\beta f(z)dF(z) - \int_\alpha^\beta f(z)dG(z)\right| \leq \epsilon \int_\alpha^\beta |f'(z)|dz$.*

The proof of this lemma is nearly identical to that of Lemma EC.1, so omitted. The proofs in this e-companion use the above lemmas to bound the impact of any difference in disturbance distribution on the value function in each stage. Here we prove the (a) part of the theorem and the (b) part can be proven similarly by using Lemma EC.2 instead of Lemma EC.1.

Consider two finite-horizon MDPs that differ only in disturbance distributions, and let $\{F_s\}$ and $\{G_s\}$ denote the two distribution functions. For $\Phi = F$ or G, we define $r^\Phi(s, a) \triangleq \int r(s, a, z)d\Phi_s(z)$. Let $V_t^\Phi(s) \triangleq \max_{a \in \mathcal{A}} W_t^\Phi(s, a)$, where $W_T^\Phi(s, a) \triangleq r^\Phi(s, a)$ and $W_t^\Phi(s, a) \triangleq$

$r^\Phi(s,a)+\int V_{t+1}^\Phi(\Psi(s,a,z))d\Phi_s(z)$ for $t<T$. Lastly, let $V_t^{F,G}(s)$ denote the expected total reward starting from state s at time t of a policy optimal for disturbance F evaluated in the model with disturbance G (in case of multiple optimal policies, we again apply a tie-breaking rule). That is, $V_T^{F,G}(s)\triangleq r^G(s,\pi_t^F(s))$ and $V_t^{F,G}(s)\triangleq r^G(s,\pi_t^F(s))+\int V_{t+1}^{F,G}(\Psi(s,\pi_t^F(s),z))dG_s(z)$ for $t<T$, where π^F is the optimal policy for F.

LEMMA EC.3. *Suppose that two sets of distribution functions $\{F_s\}$ and $\{G_s\}$ satisfy $F_s(\alpha_s-)=G_s(\alpha_s-)=0$ and $F_s(\beta_s)=G_s(\beta_s)=1$ for $s\in\mathcal{S}$, where α_s,β_s are constants satisfying $\alpha_s\le\beta_s$ for $s\in\mathcal{S}$. Suppose that $V_t^F(\Psi(s,a,\cdot))$ is Lipschitz continuous with constant $M_{t,s}$ and $r(s,a,\cdot)$ is Lipschitz continuous with constant l_s for $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Suppose also that $||F_s-G_s||\le\epsilon_s$ for $s\in\mathcal{S}$. Let $\lambda_{T,s}\triangleq(\beta_s-\alpha_s)l_s$ and $\lambda_{t,s}\triangleq(\beta_s-\alpha_s)(l_s+M_{t+1,s})$ for $t<T$. Then,*
(i) $||V_t^F-V_t^G||\le||W_t^F-W_t^G||\le\sum_{j=t}^T\max_{s\in\mathcal{S}}\epsilon_s\lambda_{j,s}$ for any t and
(ii) $||V_t^G-V_t^{F,G}||\le 2\sum_{j=t}^T\max_{s\in\mathcal{S}}\epsilon_s\lambda_{j,s}$ for any t.

The above lemma is an improved version of Proposition S-1 and S-2 in Cooper and Rangarajan (2012). In their paper, for example, the right hand side of (ii) is $2\sum_{k=t}^T\sum_{j=k}^T\max_{s\in\mathcal{S}}\epsilon_s\lambda_{j,s}$ in our notation.

Proof: We first show that $||V_t^F-V_t^G||\le||W_t^F-W_t^G||$. For $s\in\mathcal{S}$ and $a\in\mathcal{A}$, we have that

$$\begin{aligned}
&W_t^F(s,a)-W_t^G(s,a)\le|W_t^F(s,a)-W_t^G(s,a)|\\
\Leftrightarrow&\max_a W_t^F(s,a)\le\max_a|W_t^F(s,a)-W_t^G(s,a)|+\max_a W_t^G(s,a)\\
\Leftrightarrow&\max_a W_t^F(s,a)-\max_a W_t^G(s,a)\le\max_a|W_t^F(s,a)-W_t^G(s,a)|.
\end{aligned}$$

Since we can exchange F and G in the above, we have $|V_t^F(s)-V_t^G(s)|\le\max_a|W_t^F(s,a)-W_t^G(s,a)|\le||W_t^F-W_t^G||$ for any $s\in\mathcal{S}$, and therefore, $||V_t^F-V_t^G||\le||W_t^F-W_t^G||$.

We prove (i) by induction. For $t=T$, we have $W_T^\Phi(s,a)=r^\Phi(s,a)=\int r(s,a,z)d\Phi_s(z)$ for $\Phi=F,G$. By Lemma EC.1, $|W_T^F(s,a)-W_T^G(s,a)|\le\epsilon_s(\beta_s-\alpha_s)l_s$. Thus, $||W_T^F-W_T^G||\le\max_{s\in\mathcal{S}}\epsilon_s\lambda_{T,s}$. Therefore, we proved (i) for $t=T$.

For general t,

$$
\begin{aligned}
&|W_t^F(s,a) - W_t^G(s,a)| \\
&= \left| r^F(s,a) - r^G(s,a) + \int V_{t+1}^F(\Psi(s,a,z))dF_s(z) - \int V_{t+1}^G(\Psi(s,a,z))dG_s(z) \right| \\
&\le \left| \int r(s,a,z)dF_s(z) - \int r(s,a,z)dG_s(z) \right| \\
&\quad + \left| \int V_{t+1}^F(\Psi(s,a,z))dF_s(z) - \int V_{t+1}^F(\Psi(s,a,z))dG_s(z) \right| \\
&\quad + \left| \int V_{t+1}^F(\Psi(s,a,z))dG_s(z) - \int V_{t+1}^G(\Psi(s,a,z))dG_s(z) \right| \\
&\le \epsilon_s(\beta_s - \alpha_s)l_s + \epsilon_s(\beta_s - \alpha_s)M_{t+1,s} + \sum_{j=t+1}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s} \\
&\le \max_{s\in\mathcal{S}} \epsilon_s\lambda_{t,s} + \sum_{j=t+1}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s} = \sum_{j=t}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s},
\end{aligned}
$$

where the second inequality is from Lemma EC.1 and the induction hypothesis. Thus, (i) is proven.

For any policy π, let $V_t^{\pi,\Phi}$ be the value function of π under Φ. Then, we can show similarly as above that $||V_t^{\pi,F} - V_t^{\pi,G}|| \le \sum_{j=t}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s}$ for any t. Let π^F denote the optimal policy for F. For $s \in \mathcal{S}$ and any t,

$$
V_t^{F,G}(s) = V_t^{\pi^F,G}(s) \ge V_t^{\pi^F,F}(s) - \sum_{j=t}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s} = V_t^F(s) - \sum_{j=t}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s} \ge V_t^G(s) - 2\sum_{j=t}^{T} \max_{s\in\mathcal{S}} \epsilon_s\lambda_{j,s},
$$

and this proves (ii). □

Using the above lemma we can show the following.

LEMMA EC.4. *Suppose that for $s \in \mathcal{S}$, we have $||\hat{F}_s^X - F_s^X|| \le \zeta_s$ and $||\hat{F}_s^X - \hat{F}_s^Y|| \le \theta_s$, where ζ_s, θ_s are all positive constants. Then*

$$
||\tilde{V}_1^X - \tilde{V}_1^{\text{mix},X}|| \le 2\sum_{t=1}^{T} \max_{s\in\mathcal{S}} [\zeta_s + (1-\mu_s)\theta_s]\Lambda_{t,s}
$$

Proof: Because $||\hat{F}_s^X - F_s^X|| \le \zeta_s$ for $s \in \mathcal{S}$, by applying Lemma EC.3(ii) with $F = \hat{F}^X$ and $G = F^X$, we obtain $V_1^X(s) - \tilde{V}_1^X(s) \le 2\sum_{t=1}^{T} \max_{s\in\mathcal{S}} \zeta_s\Lambda_{t,s}$.

Now we will derive an upper bound of $V_1^X(s) - \tilde{V}_1^{\text{mix},X}(s)$. For $s \in \mathcal{S}$,

$$
||\hat{F}_s^{\text{mix}} - F_s^X|| \le ||\hat{F}_s^{\text{mix}} - \hat{F}_s^X|| + ||\hat{F}_s^X - F_s^X||
$$

$$= (1-\mu_s)||\hat{F}_s^Y - \hat{F}_s^X|| + ||\hat{F}_s^X - F_s^X||$$

$$\leq (1-\mu_s)\theta_s + \zeta_s.$$

By applying Lemma EC.3(ii) with $F = \hat{F}^{\text{mix}}$ and $G = F^X$, we obtain $V_1^X(s) - \tilde{V}_1^{\text{mix},X}(s) \leq 2\sum_{t=1}^{T} \max_{s\in\mathcal{S}}[\zeta_s + (1-\mu_s)\theta_s]\Lambda_{t,s}$.

Since $\tilde{V}_1^{\text{mix},X}(s)$ and $\tilde{V}_1^X(s)$ are value functions of some policy for the true disturbance distribution of Group X, they are no bigger than $V_1^X(s)$. Thus, we have

$$|\tilde{V}^X(s) - \tilde{V}^{\text{mix},X}(s)| \leq 2\sum_{t=1}^{T} \max_{s\in\mathcal{S}}[\zeta_s + (1-\mu_s)\theta_s]\Lambda_{t,s},$$

and the lemma is proven. □

The above lemma requires two inequalities regarding estimated distributions to be satisfied. One cannot check the first one ($||\hat{F}_s^X - F_s^X|| \leq \zeta_s$) in practice since we do not know the true distribution F^X, whereas we can check the second one which contains only estimated distributions. The next lemma fills the gap, leading to our main theorem.

LEMMA EC.5 **(Massart 1990)**. *Suppose that $Z^1, \ldots, Z^m$ are a random sample from a distribution function F. Let $\hat{F}$ denote the empirical distribution function estimated from the m observations. Then, for any $\epsilon > 0$, $P(||\hat{F} - F|| > \epsilon) \leq 2\exp(-2\epsilon^2 m)$.*

With this lemma, the rest of the proof uses a common technique called union bound. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the probability space on which the samples $\{X_s^i\}$ and $\{Y_s^i\}$ are defined. Let

$$A \triangleq \left\{\omega : \{X_s^i(\omega)\} \text{ satisfies } ||\hat{F}_s^X - F_s^X|| \leq \zeta_s \text{ for } s \in \mathcal{S}\right\}.$$

Then, the event A implies (7) by Lemma EC.4. The probability of A satisfies

$$P(A) = 1 - P(A^c) \geq 1 - \sum_{s\in\mathcal{S}} P(||\hat{F}_s^X - F_s^X|| > \zeta_s)$$

$$\geq 1 - 2\sum_{s\in\mathcal{S}} \exp(-2\zeta_s^2 m_s),$$

where we used Lemma EC.5 in the last inequality. By letting $\zeta_s = \sqrt{\frac{\log 2S/\delta}{2m_s}}$, the theorem is proven.

□

EC.2. Extension to Infinite-Horizon MDPs

In this section, we extend Bound(a) of Theorem 1 to infinite-horizon MDPs. Bound(b) can be extended similarly. In the infinite-horizon formulation, a policy depends only on state, but not on time. Also, we maximize the expected total discounted reward, i.e., $\max_\pi E\left[\sum_{t=1}^{T}\gamma^{t-1}r(S(t),\pi(S(t)),Z(t))\right]$, where $0<\gamma<1$.

Let $\hat{V}^X$ and $\hat{V}^{\text{mix}}$ denote the optimal value function of the MDP with disturbance distribution $\hat{F}^X$ and $\hat{F}^{\text{mix}}$, respectively. Let L_s be a Lipschitz constant of both $\hat{V}^X(\Psi(s,a,\cdot))$ and $\hat{V}^{\text{mix}}(\Psi(s,a,\cdot))$ for all $a\in\mathcal{A}$.

THEOREM EC.1. *For $s\in\mathcal{S}$, let $\theta_s=||\hat{F}_s^X-\hat{F}_s^Y||$. Then, the difference of the performance between stratifying two groups X and Y, and not stratifying for Group X satisfies*

$$||\tilde{V}^X-\tilde{V}^{\text{mix},X}||\le\frac{2}{1-\gamma}\max_{s\in\mathcal{S}}\left[\sqrt{\frac{\log 2S/\delta}{2m_s}}+(1-\mu_s)\theta_s\right]\Lambda_s \tag{EC.1}$$

with probability $1-\delta$, where $\Lambda_s\triangleq(\beta_s-\alpha_s)(l_s+\gamma L_s)$.

Proof: We first establish lemmas similar to Lemma EC.3 and Lemma EC.4 for the infinite-horizon formulation, and then the proof of the theorem follows immediately.

Consider two infinite-horizon MDPs that differ only in disturbance distributions. Let $\{F_s\}$ and $\{G_s\}$ denote the two distribution functions. For $\Phi=F$ or G, we define $r^\Phi(s,a)$ the same as in the previous section. Let V^Φ denote the optimal value function for the MDP with disturbance distribution Φ and define $W^\Phi(s,a)$ to be the state-action value function of Φ. Similarly as before, let $V^{F,G}(s)$ denote the expected total discounted reward starting from state s of a policy optimal for disturbance F evaluated in the model with disturbance G (in case of multiple optimal policies, we again apply a tie-breaking rule).

LEMMA EC.6. *Suppose that two sets of distribution functions $\{F_s\}$ and $\{G_s\}$ satisfy $F_s(\alpha_s-)=G_s(\alpha_s-)=0$ and $F_s(\beta_s)=G_s(\beta_s)=1$ for $s\in\mathcal{S}$, where α_s,β_s are constants satisfying $\alpha_s\le\beta_s$ for $s\in\mathcal{S}$. Suppose that $V^F(\Psi(s,a,\cdot))$ is Lipschitz continuous with constant M_s and $r(s,a,\cdot)$ is Lipschitz continuous with constant l_s for $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Suppose also that $||F_s-G_s||\le\epsilon_s$ for $s\in\mathcal{S}$. Then,*

(i) $||V^F - V^G|| \le ||W^F - W^G|| \le \frac{1}{1-\gamma}\max_{s\in\mathcal{S}}\epsilon_s\lambda_s$ *and*

(ii) $||V^G - V^{F,G}|| \le \frac{2}{1-\gamma}\max_{s\in\mathcal{S}}\epsilon_s\lambda_s$,

where $\lambda_s \triangleq (\beta_s - \alpha_s)(l_s + \gamma M_s)$.

Proof: First, we have that $||V^F - V^G|| \le ||W^F - W^G||$ as in the proof of Lemma EC.3.

Now we prove the second inequality of (i). For $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$
\begin{aligned}
&|W^F(s,a) - W^G(s,a)| \\
&= \left| r^F(s,a) - r^G(s,a) + \gamma \int V^F(\Psi(s,a,z))dF_s(z) - \gamma \int V^G(\Psi(s,a,z))dG_s(z) \right| \\
&\le \left| \int r(s,a,z)dF_s(z) - \int r(s,a,z)dG_s(z) \right| \\
&\quad + \gamma \left| \int V^F(\Psi(s,a,z))dF_s(z) - \int V^F(\Psi(s,a,z))dG_s(z) \right| \\
&\quad + \gamma \left| \int V^F(\Psi(s,a,z))dG_s(z) - \int V^G(\Psi(s,a,z))dG_s(z) \right| \\
&\le \epsilon_s(\beta_s - \alpha_s)l_s + \gamma\epsilon_s(\beta_s - \alpha_s)M_s + \gamma \int |V^F(\Psi(s,a,z)) - V^G(\Psi(s,a,z))|dG_s(z) \\
&\le \epsilon_s\lambda_s + \gamma||V^F - V^G||,
\end{aligned}
$$

where the second inequality is from Lemma EC.1. Thus, we have

$$
||V^F - V^G|| \le ||W^F - W^G|| = \max_{s,a}|W^F(s,a) - W^G(s,a)| \le \max_{s,a}\epsilon_s\lambda_s + \gamma||V^F - V^G||, \tag{EC.2}
$$

and therefore,

$$
||V^F - V^G|| \le \frac{\max_s \epsilon_s\lambda_s}{1-\gamma}.
$$

Also, by (EC.2),

$$
||W^F - W^G|| \le \max_s \epsilon_s\lambda_s + \gamma||V^F - V^G|| \le \max_s \epsilon_s\lambda_s + \frac{\gamma \max_s \epsilon_s\lambda_s}{1-\gamma} = \frac{\max_s \epsilon_s\lambda_s}{1-\gamma},
$$

and thus, (i) is proven.

For any policy π, let $V^{\pi,\Phi}$ be the value function of π under Φ. Then, we can show similarly that $||V^{\pi,F} - V^{\pi,G}|| \le \max_s \epsilon_s\lambda_s/(1-\gamma)$. Using these results, we obtain (ii) in the same way as we did at the end of the proof of Lemma EC.3. $\square$

LEMMA EC.7. *Suppose that for $s \in \mathcal{S}$, we have $||\hat{F}_s^X - F_s^X|| \leq \zeta_s$ and $||\hat{F}_s^X - \hat{F}_s^Y|| \leq \theta_s$, where ζ_s, θ_s are all positive constants. Then*

$$||\tilde{V}^X - \tilde{V}^{\text{mix},X}|| \leq \frac{2}{1-\gamma} \max_{s \in \mathcal{S}} [\zeta_s + (1-\mu_s)\theta_s]\Lambda_s,$$

where

$$\Lambda_s \triangleq (\beta_s - \alpha_s)(l_s + \gamma L_s). \tag{EC.3}$$

Given Lemma EC.6, the proof of the lemma is identical to the proof of Lemma EC.4, so omitted.

Given these two lemmas, the rest of the proof of Theorem EC.1 is identical to that of Theorem 1, so omitted.

EC.3. Computing Lipschitz Constants

The probabilistic bounds (7) and (EC.1) contain Lipschitz constants L_s and $L_{t,s}$. This section discusses how to compute those Lipschitz constants using Lipschitz constants of r and Ψ, and other parameters. Before we begin, it should be noted that the most straightforward way is computing $L_{t,s}$ and L_s directly after obtaining the corresponding value functions. Also, there can be other ways of deriving a Lipschitz constant than those presented here.

EC.3.1. Finite-Horizon Case

We first explain how to compute the Lipschitz constants for the finite-horizon case, because the infinite-horizon case builds upon it. In the bound for the finite-horizon case, $L_{t,s}$ is a Lipschitz constant of $\hat{V}_t^X(\Psi(s,a,\cdot))$ and $\hat{V}_t^{\text{mix}}(\Psi(s,a,\cdot))$ for any $a \in \mathcal{A}$. We only explain computing a Lipschitz constant of the former since that of the latter can be obtained similarly.

Let m and M be Lipschitz constants of $r(s,a,\cdot)$ and $\Psi(s,a,\cdot)$ for all s and a values, respectively. These constants exist because $\mathcal{S}$ and $\mathcal{A}$ are finite and the disturbances have a finite support. Let m' be a Lipschitz constant of $r(\cdot,a,z)$ for all a and z values. Let M' be a Lipschitz constant of $\Psi(\cdot,a,z)$ for all a and z values. Such constants exist again for the same reason. Let f be a constant satisfying $||\hat{F}_s^X - \hat{F}_{s'}^X|| \leq f|s-s'|$ for any $s, s' \in \mathcal{S}$. Such a constant can be found because $\mathcal{S}$ is finite

and $\hat{F}^X$ is available from data. In the rest of this subsection, we will omit the superscript X and the hat notation for brevity, but we emphasize that the following computation can be performed in practice since it uses $\hat{F}$.

For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\begin{aligned}
|r(s,a) - r(s',a)| &= \left| \int r(s,a,z) dF_s(z) - \int r(s',a,z) dF_{s'}(z) \right| \\
&\leq \left| \int r(s,a,z) dF_s(z) - \int r(s',a,z) dF_s(z) \right| + \left| \int r(s',a,z) dF_s(z) - \int r(s',a,z) dF_{s'}(z) \right| \\
&\leq \int |r(s,a,z) - r(s',a,z)| dF_s(z) + f|s-s'| m(\max\{\beta_s, \beta_{s'}\} - \min\{\alpha_s, \alpha_{s'}\}) \\
&\leq m'|s-s'| + f|s-s'| m(\max\{\beta_s, \beta_{s'}\} - \min\{\alpha_s, \alpha_{s'}\}) \leq [m' + fm(\beta - \alpha)]|s-s'|,
\end{aligned}$$

where the second inequality was obtained by Lemma EC.1, and we define $\beta \triangleq \max_s \beta_s$ and $\alpha \triangleq \min_s \alpha_s$. Thus, $r(\cdot, a)$ is Lipschitz continuous with constant $\bar{m} \triangleq m' + fm(\beta - \alpha)$.

We can easily show that

$$|V_t(s) - V_t(s')| \leq \max_a |W_t(s,a) - W_t(s',a)| \text{ for any } t \tag{EC.4}$$

using similar steps we used in the first part of the proof of Lemma EC.3. Since $W_T(s,a) = r(s,a)$, we can easily show that V_T is Lipschitz continuous with constant $\bar{m}$ using (EC.4).

Suppose that V_{t+1} is Lipschitz continuous with constant B_{t+1}. Then,

$$\begin{aligned}
|W_t(s,a) - W_t(s',a)| \leq{}& |r(s,a) - r(s',a)| + \left| \int V_{t+1}(\Psi(s,a,z)) dF_s(z) - \int V_{t+1}(\Psi(s',a,z)) dF_{s'}(z) \right| \\
\leq{}& \bar{m}|s-s'| + \left| \int V_{t+1}(\Psi(s,a,z)) dF_s(z) - \int V_{t+1}(\Psi(s',a,z)) dF_s(z) \right| \\
&+ \left| \int V_{t+1}(\Psi(s',a,z)) dF_s(z) - \int V_{t+1}(\Psi(s',a,z)) dF_{s'}(z) \right| \\
\leq{}& \bar{m}|s-s'| + \int |V_{t+1}(\Psi(s,a,z)) - V_{t+1}(\Psi(s',a,z))| dF_s(z) + f|s-s'| B_{t+1} M(\beta - \alpha) \\
\leq{}& \bar{m}|s-s'| + B_{t+1} M'|s-s'| + f|s-s'| B_{t+1} M(\beta - \alpha) \\
={}& \{\bar{m} + B_{t+1}[M' + fM(\beta - \alpha)]\}|s-s'|,
\end{aligned}$$

where the third inequality was obtained by Lemma EC.1. Thus, $W_t(\cdot, a)$ is Lipschitz continuous with constant $B_t \triangleq \bar{m} + B_{t+1}[M' + fM(\beta - \alpha)]$. By (EC.4), we have that V_t is also Lipschitz continuous with constant B_t. Therefore, $B_t M$ is a Lipschitz constant of $V_t(\Psi(s,a,\cdot))$.

EC.3.2. Infinite-Horizon Case

Recall that in the infinite-horizon case, L_s is a Lipschitz constant of both $\hat{V}^X(\Psi(s,a,\cdot))$ and $\hat{V}^{\text{mix}}(\Psi(s,a,\cdot))$ for any $a \in \mathcal{A}$. Again, we only explain how to compute a Lipschitz constant of the former since that of the latter can be obtained similarly. We define m, m', M, M', and f in the same way as in the previous subsection. We again omit the superscript X and the hat notation in the rest of this subsection for brevity. The following computation can be performed in practice since it uses $\hat{F}$, not the true distribution.

The first method to compute L_s is an extension of the finite-horizon case. Let $V^0(s) = 0$ for $s \in \mathcal{S}$ and V^n be the nth iterate of value iteration algorithm starting from V^0 (e.g., see Bertsekas 2012). Using the same arguments of the previous subsection, we can show that V^n is Lipschitz continuous with constant b_n, where b_n is defined by a recursive equation $b_n = \bar{m} + \gamma\kappa b_{n-1}$, $b_1 = \bar{m}$, and $\kappa \triangleq M' + fM \max_s(\beta - \alpha)$ (recall that γ is a discount factor). It is well known that V^n converges to the optimal value function V under suitable conditions which are satisfied by the problem formulation of this paper. If $\gamma\kappa < 1$, then b_n has a limit which equals $\bar{m}/(1-\gamma\kappa)$, and then, we can easily show that $\bar{m}/(1-\gamma\kappa)$ is a Lipschitz constant of V. To illustrate the condition $\gamma\kappa < 1$, consider the HIV treatment problem in Section 6.1. We name the four states of CD4 levels as 1, 2, 3, and 4. 0 denotes death and 5 is the terminal state after initiating the therapy. The state transition can be written as $\Psi(s, \text{'wait'}, z) = s - z$ and $\Psi(s, \text{'initiate'}, z) = 5$, where z represents the health state decrement. Then, $M = M' = 1$ and $\beta - \alpha = 4$. In addition, recall that f is a constant satisfying $||F_s - F_{s'}|| \leq f|s - s'|$ for any $s, s' \in \mathcal{S}$. Thus, we can find a value for f that is no greater than 1, because the left hand side of the inequality is no greater than 1 (possibly much smaller than 1) and we have $|s - s'| \geq 1$ for any s, s' such that $s \neq s'$. In sum, for the HIV example, if $\gamma\kappa = \gamma(1+4f) < 1$, then $\bar{m}/(1-\gamma\kappa)$ is a Lipschitz constant we need.

The other method is using the error bound of value iteration. There are well-known error bounds for value iteration (e.g., see Proposition 2.2.1 of Bertsekas 2012). After n iterations of value iteration, we obtain V^n which approximates V and its error bound at each state. Then, using V^n

and the error bounds, we can compute a constant L satisfying $|V(s) - V(s')| \leq L|s - s'|$ for all $s, s' \in \mathcal{S}$. In order to obtain a smaller value of L (which will lead to a tighter bound on the benefit of stratifying), one will have to run more iterations of value iteration, and the computational load will approach to that of solving the empirical MDP. However, we remind the reader that solving an empirical MDP does not achieve the goal of this paper, which is making the modeling decision of whether to stratify or not based on data, and making the decision might take more computation than solving one empirical MDP, as the bootstrapping method in Section 4 suggests.

EC.4. Proof of Theorem 2

Consider the following class of MDPs with five states, two actions, and three time periods. s_0 is an initial state. At s_0, taking a_1 and a_2 takes the controlled system to s_1 and s_2, respectively. s_0 is the only state where a choice of action matters. In other words, at all other states the two actions result in the same reward and transition. The transitions at s_1 and s_2 are written as $\Psi(s_i, a, 1) = s_3$ and $\Psi(s_i, a, 0) = s_4$ for $i = 1, 2$ and for any action a. So, the random disturbance is binary (0 or 1). At s_1, the random disturbance is 1 with probability p and at s_2 with probability q. In other words, at s_1, the system moves to s_3 with probability p and to s_4 with probability $1 - p$, and at s_2, it moves to s_3 with probability q and to s_4 with probability $1 - q$. At s_3 and s_4, we receive a reward of 1 and -1, respectively. Also, other than s_3 and s_4, there are no rewards. Thus, s_3 and s_4 are called good and bad, respectively. Note that the random disturbance matters only at s_1 and s_2. At s_0, we need to find out which of s_1 or s_2 is more likely to lead to the good state s_3. We denote such an MDP as $\mathcal{M}(p, q)$.

Let Group X's MDP be $\mathcal{M}(p_1, p_2)$ and Group Y's be $\mathcal{M}(p_3, p_4)$, where $p_1 < p_2$ and $p_3 > p_4$. More formally, the distributions of random disturbances are written as $P(X_{s_1} = 1) = p_1$, $P(X_{s_2} = 1) = p_2$, $P(Y_{s_1} = 1) = p_3$, $P(Y_{s_2} = 1) = p_4$. Thus, for Group X, a_2 is the optimal action. Let $\nu_{s_i} = \xi_{s_i} > 0$ for $i = 0, 1, 2$. Thus, $\mu_{s_i} = 1/2$ for $i = 0, 1, 2$. In other words, we get the same number of samples from the two groups at each state.

We first compute the bound and then the benefit. Since the reward does not depend on the random disturbance, we have $l_s = 0$ for all s. Note that s_0 is visited only in time period 1, s_1 and

s_2 in period 2, and s_3 and s_4 in period 3. Except at s_1 and s_2, state transition does not depend on the random disturbance. Thus, $L_{2,s_0}=0$. At s_1 and s_2, depending on the realized value of the random disturbance, we move either to the good state where we earn a reward of one or to the bad state where we get -1. Thus, we can easily obtain that $L_{3,s_1}=L_{3,s_2}=2$.

Let $\hat{p}_1^X$ denote the empirical probability of reaching the good state from s_1 estimated from Group X and define $\hat{p}_1^Y$ in the same way. Then,

$$\left[\sqrt{\frac{\log(2S/\delta)}{2m_{s_1}}}+(1-\mu_{s_1})\theta_{s_1}\right](\beta_{s_1}-\alpha_{s_1})(l_s+L_{3,s_1})=\sqrt{\frac{\log(2S/\delta)}{2m_{s_1}}}+|\hat{p}_1^X-\hat{p}_1^Y|,$$

where we used the fact that $\beta_{s_1}=1$, $\alpha_{s_1}=0$. As N goes to infinity,

$$\left[\sqrt{\frac{\log(2S/\delta)}{2m_{s_1}}}+(1-\mu_{s_1})\theta_{s_1}\right](\beta_{s_1}-\alpha_{s_1})(l_{s_1}+L_{3,s_1})\to|p_1-p_3|\text{ a.s.}$$

In the same way,

$$\left[\sqrt{\frac{\log(2S/\delta)}{2m_{s_2}}}+(1-\mu_{s_2})\theta_{s_2}\right](\beta_{s_2}-\alpha_{s_2})(l_{s_2}+L_{3,s_2})\to|p_2-p_4|\text{ a.s.}$$

Since $L_{2,s_0}=0$ and $l_s=0$ for all s, we have that $\Lambda_1=\Lambda_3=0$. Thus, the bound (Theorem 1, the version in which the maximization is limited to those states that can be visited in each time period) converges to $2\max\{|p_1-p_3|,|p_2-p_4|\}$ almost surely.

Now let us turn to the actual benefit. Under the event that $\hat{\pi}^X(s_0)=a_2$ and $\hat{\pi}^{\text{mix}}(s_0)=a_1$, the benefit of stratifying is

$$\tilde{V}_1^X(s_0)-\tilde{V}_1^{\text{mix},X}(s_0)=2p_2-1-2p_1+1=2(p_2-p_1).$$

Let $p_1=\epsilon'>0, p_2=1-\epsilon', p_3=1, p_4=0$ for $\epsilon'\in(0,1/2)$ whose value we will determine later. Since $p_1<p_2$, $P(\hat{\pi}^X(s_0)=a_2)=1$ when N goes to infinity. Also, because we get the same number of samples from the two groups at each state and $(p_1+p_3)/2>(p_2+p_4)/2$, we have $P(\hat{\pi}^{\text{mix}}(s_0)=a_1)=1$ when N goes to infinity. Therefore, the benefit of stratifying goes to $2(p_2-p_1)=2-4\epsilon'$ with probability one. We know that the bound converges to $2\max\{|p_1-p_3|,|p_2-p_4|\}=2-2\epsilon'$. Thus, the ratio of the benefit over the bound goes to $\frac{2-4\epsilon'}{2-2\epsilon'}$. By setting ϵ' so that $\epsilon=\epsilon'/(1-2\epsilon')$, the theorem is proven. $\square$

EC.5. The Challenge of Estimating the Benefit

In this section, we show why popular approaches in robust MDP and RL literature fall short of deriving upper and lower bounds of the benefit that are asymptotically tight for any instance. First, we consider an approach motivated by robust MDP literature and show that the approach leads us to an NP-hard problem. Second, we consider an approach that defines an estimator and derives an error bound of that estimator, which is popular in RL literature. Then, we show why a standard proof technique in the literature does not apply to our research question. The purpose of this section is to illustrate the difficulty of estimating the benefit.

The benefit of stratifying is the difference between the performance of stratifying and the performance of not stratifying. Recall that stratifying obtains a policy from the data of a group and evaluates it for the same group. The estimation of the performance of stratifying has been studied extensively in both operations research (OR) and RL literature, as it is essentially off-policy optimization (see Section 1.1). For example, one can obtain a confidence interval for stratifying using Corollary 1 and Theorem EC.5. Thus, we will focus on the case of not stratifying in this section.

EC.5.1. A robust MDP approach

Generally speaking, evaluating not stratifying is obtaining a policy from data generated from a disturbance distribution and then evaluating it using data from another disturbance distribution. We will use F^0 to denote the true distribution of the data we obtain a policy from and G^0 to denote the true distribution for which we want to evaluate the policy. Thus, F^0 corresponds to the true mixed distribution (the true distribution of the mix can be defined by determining the proportions of data from the two groups at each state, i.e., ν_s and ξ_s defined right before Theorem 2) and G^0 to the true distribution of Group X.

Consider the following approach motivated from robust MDP literature (Iyengar 2005, Mannor and Xu 2019, Nilim and El Ghaoui 2005, Wiesemann et al. 2013). Let $\mathcal{F}$ be an ambiguity set for the true distribution F^0 such that F^0 belongs to $\mathcal{F}$ with probability at least $1-\delta/2$. Let $\mathcal{G}$ be defined in the same way. Consider

$$\max \mathcal{V}^{\mathcal{F},\mathcal{G}}, \tag{EC.5}$$

where

$$\mathcal{V}^{\mathcal{F},\mathcal{G}} \triangleq \{V_1^{\pi,G}(s) \mid \pi \text{ is optimal for } F,\ F \in \mathcal{F}, G \in \mathcal{G}\},$$

being optimal for F means being optimal for the MDP with the disturbance distribution F, and we used the value function notation introduced in the e-companion EC.1. Note that this problem includes finding those policies optimal for $F \in \mathcal{F}$. If we can compute the maximum and its minimization counterpart, then this leads to a confidence interval of the value of not stratifying. Note that as the data size goes to infinity, $\mathcal{F}$ and $\mathcal{G}$ shrink to $\{F^0\}$ and $\{G^0\}$, respectively, and if F^0 has a unique optimal policy (or we apply a tie-breaking rule), then $\mathcal{V}^{\mathcal{F},\mathcal{G}}$ shrinks to a single value. Thus, if successful, this approach can lead to a bound of the benefit that is asymptotically tight for any instance.

Let us compare the above problem with existing robust MDP formulations. If there is no constraint on policy π or if $\mathcal{F} = \mathcal{G}$ in (EC.5), then the problem is equivalent to the optimistic version of robust MDP (optimizing the optimistic value) (Iyengar 2005), because

$$\max_{\pi} \max_{G \in \mathcal{G}} V^{\pi,G} = \max_{G \in \mathcal{G}} \max_{\pi} V^{\pi,G} = \max_{G \in \mathcal{G}} \max_{\pi \in \Pi^{\mathcal{G}}} V^{\pi,G},$$

where $\Pi^{\mathcal{G}}$ is the set of policies optimal for any $G \in \mathcal{G}$. On the other hand, if $\mathcal{F}$ is a singleton and the element has a unique optimal policy, then the problem (EC.5) reduces to the robust policy evaluation problem in Wiesemann et al. (2013).

To analyze the difficulty of (EC.5), we consider a simpler version in which $\mathcal{G}$ is a singleton. Thus, we fix G and consider the following:

$$\max V_1^{\pi^F,G}(s) \text{ s.t. } \pi^F \text{ is any optimal policy for } F, F \in \mathcal{F}. \tag{EC.6}$$

Note that this problem is finding a policy that optimizes for G among those optimal for any $F \in \mathcal{F}$. To the author's knowledge, both (EC.5) and (EC.6) have not been considered in both OR and RL literature. Again, a typical setting that has been studied extensively is where $\mathcal{F}$ is an uncertainty set for G in (EC.6).

We analyze (EC.6) under the standard assumptions of robust MDP literature. The past work assumes that when we visit the same state multiple times, uncertain parameters of that state can take different values. This is because the other case (uncertain parameters of a state stay the same over multiple visits) is intractable in general (Xu and Mannor 2012). Thus, we make the same assumption but note that this leads to a more conservative confidence interval. Also, we assume that the uncertainty set $\mathcal{F}$ is rectangular, meaning that the uncertain parameters at different states are uncoupled, as the past work in robust MDP literature does.

In addition, let $\Pi^{\mathcal{F}}$ denote the set of policies optimal for any $F \in \mathcal{F}$. We further simplify (EC.6) by assuming that $\Pi^{\mathcal{F}}$ is given:

$$\max_{\pi \in \Pi^{\mathcal{F}}} V_1^{\pi,G}(s). \tag{EC.7}$$

We need to introduce some definitions to analyze the difficulty of (EC.7). We say that a set of policies Π is t-rectangular if

$$\Pi = \Pi_1 \times \Pi_2 \times \ldots \times \Pi_T,$$

where the product operation is the Cartesian product, and Π_t for any t is the projection of policies in Π onto the action choices in time period t. We say that a set of policies Π is ts-rectangular if

$$\Pi = \bigotimes_{t=1}^{T} \bigotimes_{s \in \mathcal{S}} \Pi_{t,s},$$

where $\Pi_{t,s} \subset \mathcal{A}$ and the Cartesian product is defined similarly as above. If $\Pi^{\mathcal{F}}$ is ts-rectangular, then we can solve (EC.7) by backward induction. However, the following theorem shows that it is not the case in general.

THEOREM EC.2. *There exists a finite horizon MDP and a rectangular uncertainty set $\mathcal{F}$ for which $\Pi^{\mathcal{F}}$ is not t-rectangular.*

Proof: Consider a finite horizon MDP with $T = 2$, $\mathcal{S} = \{s_1, s_2, s_3\}$, and $\mathcal{A} = \{a_1, a_2\}$. The initial state is s_1. At s_1, the reward is always zero and taking action a_1 and a_2 deterministically leads to

s_2 and s_3, respectively. At s_2, there is a binary random disturbance Z_2, which is 1 with probability p_2 and 0 with $1-p_2$. The reward at s_2 is

$$r(s_2,a_1,Z_2)=\begin{cases}3 \text{ if } Z_2=1\\ \\ 1 \text{ otherwise}\end{cases}, \quad r(s_2,a_2,Z_2)=\begin{cases}0 \text{ if } Z_2=1\\ \\ 2 \text{ otherwise}\end{cases}$$

and the reward at s_3 is 2.5 deterministically for both actions. Note that at s_1 and s_3, the reward and transition are not affected by their random disturbances. Let $\mathcal{F}_i$ denote the uncertainty set for state s_i. Let $\mathcal{F}_1=\mathcal{F}_3=[0,1]$, $\mathcal{F}_2=[0.1,0.9]$, and $\mathcal{F}=\mathcal{F}_1\times\mathcal{F}_2\times\mathcal{F}_3$. Then, $\mathcal{F}$ is clearly s-rectangular. Calculations show that if $p_2\le 0.25$, then the optimal policy π is $\pi(s_1)=a_2$ and $\pi(s_2)=a_2$. If $0.25\le p_2\le 0.75$, then $\pi(s_1)=a_2$ and $\pi(s_2)=a_1$. Lastly, if $p_2\ge 0.75$, then $\pi(s_1)=a_1$ and $\pi(s_2)=a_1$. Therefore, the policy π such that $\pi(s_1)=a_1$ and $\pi(s_2)=a_2$ is never optimal, and thus, $\Pi^{\mathcal{F}}$ is not t-rectangular. □

This theorem implies that in the problem (EC.7), the set of actions one can take at a state in a time period may depend on those actions taken in the previous time periods, i.e., the set of available actions is history-dependent, even when the uncertainty set is rectangular. This means that in a dynamic programming approach, we need to include the history of actions taken so far in the state definition, which makes the state space grow exponentially in S and the horizon T.

Now we know that (EC.7) is not an MDP but belongs to the class of problems obtained from an MDP by relaxing the assumption that the set of available actions is Markov unless we change the state definition so that the state space grows exponentially. For simplicity of terminology, we call such problems MDPs without the Markov property of action space. The following result shows that an MDP without the Markov property of action space is NP-hard in general. Thus, the observation that dynamic programming approaches are not tractable for MDPs without the Markov property of action space is not because of the choice of approach but due to the intrinsic nature of the problem.

THEOREM EC.3. *Suppose that we are limited to deterministic policies. Then, determining if the optimal value of an MDP without the Markov property of action space is greater than or equal to a given constant is NP-hard.*

Proof: The proof is nearly identical to Section 3.1 of Littman 1994, thus omitted.

In sum, although we simplified the original problem (EC.5) of computing a confidence interval for the performance of not stratifying, we showed that the simplified problem lacks the structure that is essential for dynamic programming and as a result, it falls into a class of NP-hard problems. Though these results signal strongly about the difficulty of (EC.5), we also note that the results of this section do not imply directly that the problem (EC.5) is NP-hard. It is unknown if a polynomial reduction from an NP-complete problem to (EC.5) exists. The difficulty in deriving a polynomial reduction stems from the fact that the feasible region of (EC.5) is formed by policies that are optimal for any $F \in \mathcal{F}$, i.e., (EC.5) has a nested structure with two stages of optimization. Thus, to show that (EC.5) is NP-hard, it seems that one needs to find $\mathcal{F}$ such that the set of optimal policies for $F \in \mathcal{F}$ coincides with an NP-complete problem's feasible region. Note that if we find a F for each feasible solution of an NP-complete problem and form $\mathcal{F}$ using those Fs, then the construction of $\mathcal{F}$ takes an exponential amount of computation due to the combinatorial nature of NP-complete problems, and thus, it does not qualify as a polynomial reduction.

So far, we studied (EC.7) which assumes that $\Pi^{\mathcal{F}}$ is given but constructing $\Pi^{\mathcal{F}}$ is also not trivial. It is solving an optimization problem for a set of input parameters, which is called parametric programming in general. It is well known that parametric programming is intractable, even for linear programs (LPs) (Filippi and Romanin-Jacur 2002). Lee et al. (2019) studied parametric programming for LPs where the uncertainty is in the objective coefficients and the constant part of constraints ("b of $Ax = b$"). Since an MDP can be formulated as an LP, their method applies to MDPs. However, in LP formulations of MDPs, transition probabilities are coefficients of decision variables in constraints. To the author's knowledge, parametric programming for the parameters in the constraint coefficient matrix is yet to be studied.

EC.5.2. Deriving an error bound of an estimator

The second approach we consider is defining a natural estimator for the performance of not stratifying and deriving an error bound for the estimator. If the error bound goes to zero as the data

size goes to infinity, this can potentially lead to a bound of the benefit that is asymptotically tight for all instances. The exact value of the performance of not stratifying can be written as V^{π^{mix},F^X}, where π^{mix} is the optimal policy for the true distribution of the mix. A natural estimator for the performance of not stratifying is $\hat{V}^{\text{mix},X} = V^{\hat{\pi}^{\text{mix}},\hat{F}^X}$. We consider deriving an error bound of this estimator.

However, as emphasized above, $V^{\hat{\pi}^{\text{mix}},\hat{F}^X}$ is the performance of a policy that is optimal for an estimate of one distribution evaluated for an estimate of another distribution, and the convergence of such a quantity has not been studied in the literature. On the other hand, what has been studied extensively is the quantity $V^{\hat{\pi}^X,\hat{F}^X}$, in which case, there is only one true distribution. We first explain conceptually how these two are different and then illustrate why a common technique for the latter does not apply to the former.

It seems that deriving an error bound of $V^{\hat{\pi}^{\text{mix}},\hat{F}^X}$ requires a theoretical result characterizing how fast $\hat{\pi}^{\text{mix}}$ converges, i.e., the convergence rate of the empirical policy itself. We will roughly explain using an example. Let π^1 be an optimal policy for F^{mix} and π^2 be a near-optimal policy for F^{mix}. As the size of data increases, $\hat{F}^{\text{mix}}$ approaches F^{mix}. Suppose that $\hat{\pi}^{\text{mix}}$ equals π^1 sometimes and π^2 at other times as the data size increases. Although the values of π^1 and π^2 for F^{mix} are close, their performance for F^X may be vastly different. If both of the policies are optimal for F^{mix}, then $\hat{\pi}^{\text{mix}}$ may oscillate between the two policies, in which case $V^{\hat{\pi}^{\text{mix}},\hat{F}^X}$ does not converge, whereas $V^{\hat{\pi}^{\text{mix}},\hat{F}^{\text{mix}}}$ still converges. In Section 2.3, we assumed a tie-breaking rule for the case of multiple optimal policies, and our theoretical results hold for any tie-breaking rule. However, note that a tie-breaking rule does not address the issue being discussed because we do not know if the two policies are indeed tied for the true distribution F^{mix} and we may not observe ties when we evaluate the policies using an estimated distribution.

When analyzing the case of stratifying, i.e., analyzing $V^{\hat{\pi}^X,\hat{F}^X}$, most existing proofs use the following technique (e.g., Agarwal et al. (2020), Xiao et al. (2021)). We have

$$
\begin{aligned}
&|V^{\hat{\pi}^X,\hat{F}^X} - V^{\pi^X,F^X}| = |V^{\hat{\pi}^X,\hat{F}^X} - V^{\hat{\pi}^X,F^X} + V^{\hat{\pi}^X,F^X} - V^{\pi^X,F^X}| \\
&\leq |V^{\hat{\pi}^X,\hat{F}^X} - V^{\hat{\pi}^X,F^X}| + |V^{\hat{\pi}^X,F^X} - V^{\pi^X,F^X}| = |V^{\hat{\pi}^X,\hat{F}^X} - V^{\hat{\pi}^X,F^X}| + V^{\pi^X,F^X} - V^{\hat{\pi}^X,F^X},
\end{aligned}
$$

where π^X is the optimal policy for F^X. In the last expression, the absolute value is a policy evaluation error (an error caused by using the empirical distribution when evaluating a given policy), which can be bounded by Theorem EC.5. The last difference is bounded as follows:

$$\begin{aligned}0 \leq & V^{\pi^X,F^X} - V^{\hat{\pi}^X,F^X} = V^{\pi^X,F^X} - V^{\pi^X,\hat{F}^X} + V^{\pi^X,\hat{F}^X} - V^{\hat{\pi}^X,F^X} \\ & \leq V^{\pi^X,F^X} - V^{\pi^X,\hat{F}^X} + V^{\hat{\pi}^X,\hat{F}^X} - V^{\hat{\pi}^X,F^X},\end{aligned}$$

where the inequality is obtained because $\hat{\pi}^X$ is optimal for $\hat{F}^X$. Then, the last expression is the sum of two policy evaluation errors, which can again be bounded by Theorem EC.5 or similar results in off-policy evaluation literature (for references, see Section 1.1). This proof technique uses the fact that the policy $\hat{\pi}^X$ is optimal for the estimated distribution we are using to evaluate policies, but this does not hold when we apply the same technique to the case of not stratifying. Thus, this technique does not apply to the case of not stratifying.

EC.6. Proof of Theorem 3

We have

$$\begin{aligned}||\hat{F}_s^{\text{mix}} - F_s^X|| &\leq ||\mu_s \hat{F}_s^X + (1-\mu_s)\hat{F}_s^Y - F_s^X|| \\ &= ||\mu_s(\hat{F}_s^X - F_s^X) + (1-\mu_s)(\hat{F}_s^Y - F_s^Y) + (1-\mu_s)(F_s^Y - F_s^X)|| \\ &\leq \mu_s||\hat{F}_s^X - F_s^X|| + (1-\mu_s)||\hat{F}_s^Y - F_s^Y|| + (1-\mu_s)\kappa_s \\ &\leq \mu_s\sqrt{\frac{\log 4S/\delta}{2m_s}} + (1-\mu_s)\sqrt{\frac{\log 4S/\delta}{2n_s}} + (1-\mu_s)\kappa_s,\end{aligned}$$

where the last inequality holds with probability $1-\delta$ because of Lemma EC.5 and the union bound. Then, the rest of the proof is similar to that of Theorem 1. □

EC.7. Generalization to State-Action-Dependent Disturbances

In this appendix, we generalize the probabilistic upper bound on the benefit of stratifying for infinite-horizon MDPs (Theorem EC.1) to the case where the random disturbances may depend on both state and action. The probabilistic upper bounds for finite-horizon MDPs (Theorem 1) can

also be extended to the general case in a similar manner. Here we only present the infinite-horizon case.

Let us introduce some notation for the general case. Let $Z_{s,a}$ be the random disturbance where action a is taken at state s and let $F_{s,a}$ denote the distribution function. The single-period expected reward is $r(s,a) = \int r(s,a,z)dF_{s,a}(z)$. The definitions of V and W remain the same except that $F_{s,a}$ replaces F_s.

Two historical data sets $\{X^i_{s,a} : i = 1,\ldots,m_{s,a}, s \in \mathcal{S}, a \in \mathcal{A}\}$ and $\{Y^i_{s,a} : i = 1,\ldots,n_{s,a}, s \in \mathcal{S}, a \in \mathcal{A}\}$ are available for two groups. Let $\{\hat{F}^X_{s,a}\}$ and $\{\hat{F}^Y_{s,a}\}$ denote the sets of MLE distribution functions for the two groups. It should be noted that we have $|\mathcal{A}|$ times more distributions to estimate in this generalization, so we need possibly much more data to estimate them at similar accuracy. The mixed empirical distribution $\hat{F}^{\text{mix}}_{s,a}$ and $\mu_{s,a}$ are also defined similarly as in Section 2.3 by appending action a to every subscript s. We define $\tilde{V}^X(s)$ as before, that is, the expected total discounted reward of the policy optimal for the empirical distribution of Group X, evaluated for the true distribution of Group X. Also, the definition of $\tilde{V}^{\text{mix},X}(s)$ remains the same, which is the expected total discounted reward of not stratifying, evaluated for the true distribution of Group X. Then, the benefit of stratifying in this general case is defined the same as in (5).

Let us introduce additional notation needed to generalize Lemma EC.6. For any disturbance distribution $\Phi \triangleq \{\Phi_{s,a}\}$, we define $r^\Phi(s,a)$, $V^\Phi(s)$, and $W^\Phi(s,a)$ similarly as in the e-companion EC.2 by appending action to every subscript s. For two sets of distribution functions $\{F_{s,a}\}$ and $\{G_{s,a}\}$, $V^{F,G}(s)$ is also defined the same as in the e-companion EC.2.

LEMMA EC.8 **(Generalization of Lemma EC.6)**. *Suppose that two sets of distribution functions $\{F_{s,a}\}$ and $\{G_{s,a}\}$ satisfy $F_{s,a}(\alpha_{s,a}-) = G_{s,a}(\alpha_{s,a}-) = 0$ and $F_{s,a}(\beta_{s,a}) = G_{s,a}(\beta_{s,a}) = 1$, where $\alpha_{s,a}, \beta_{s,a}$ are constants satisfying $\alpha_{s,a} \le \beta_{s,a}$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Suppose that $V^F(\Psi(s,a,\cdot))$ is Lipschitz continuous with constant $M_{s,a}$ and $r(s,a,\cdot)$ is Lipschitz continuous with constant $l_{s,a}$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Suppose also that $||F_{s,a} - G_{s,a}|| \le \epsilon_{s,a}$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Let $\lambda_{s,a} \triangleq (\beta_{s,a} - \alpha_{s,a})(l_{s,a} + \gamma M_{s,a})$. Then,*

(i) $||V^F - V^G|| \le ||W^F - W^G|| \le \max_{s,a} \epsilon_{s,a}\lambda_{s,a}/(1-\gamma)$ *and*

(ii) $||V^G - V^{F,G}|| \le 2\max_{s,a} \epsilon_{s,a}\lambda_{s,a}/(1-\gamma)$.

The proof is nearly identical to that of Lemma EC.6, so omitted.

For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, let $l_{s,a}$ be the Lipschitz constant of $r(s,a,\cdot)$. $\hat{V}^X(\Psi(s,a,\cdot))$ is also Lipschitz continuous for $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Let $L_{s,a}$ be its Lipschitz constant.

LEMMA EC.9 **(Generalization of Lemma EC.7)**. *Suppose that for $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have $||\hat{F}^X_{s,a} - F^X_{s,a}|| \le \zeta_{s,a}$ and $||\hat{F}^X_{s,a} - \hat{F}^Y_{s,a}|| \le \theta_{s,a}$, where $\zeta_{s,a}, \theta_{s,a}$ are all positive. Then $||\tilde{V}^X - \tilde{V}^{\mathrm{mix},X}|| \le 2\max_{s,a}[\zeta_{s,a} + (1-\mu_{s,a})\theta_{s,a}]\Lambda_{s,a}/(1-\gamma)$, where $\Lambda_{s,a} \triangleq (\beta_{s,a} - \alpha_{s,a})(l_{s,a} + \gamma L_{s,a})$.*

Proof: The proof is nearly identical to that of Lemma EC.7. Because $||\hat{F}^X_{s,a} - F^X_{s,a}|| \le \zeta_{s,a}$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$, by applying Lemma EC.8(ii) with $F = \hat{F}^X$ and $G = F^X$, we have $V^X(s) - \tilde{V}^X(s) \le 2\max_{s,a} \zeta_{s,a}\lambda'_{s,a}/(1-\gamma)$, where $\lambda'_{s,a}$ is defined as $\lambda_{s,a}$ in Lemma EC.8, but with $L_{s,a}$ in place of $M_{s,a}$ and ζ_s in place of ϵ_s.

For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, $||\hat{F}^{\mathrm{mix}}_{s,a} - F^X_{s,a}|| \le ||\hat{F}^{\mathrm{mix}}_{s,a} - \hat{F}^X_{s,a}|| + ||\hat{F}^X_{s,a} - F^X_{s,a}|| = (1-\mu_{s,a})||\hat{F}^Y_{s,a} - \hat{F}^X_{s,a}|| + ||\hat{F}^X_{s,a} - F^X_{s,a}|| \le (1-\mu_{s,a})\theta_{s,a} + \zeta_{s,a}$. By Lemma EC.8(ii), we have $V^X(s) - \tilde{V}^{\mathrm{mix},X}(s) \le 2\max_{s,a}[\zeta_{s,a} + (1-\mu_{s,a})\theta_{s,a}]\Lambda_{s,a}/(1-\gamma)$. In the same way as in the proof of Lemma EC.7, we have $|\tilde{V}^X(s) - \tilde{V}^{\mathrm{mix},X}(s)| \le 2\max_{s,a}[\zeta_{s,a} + (1-\mu_{s,a})\theta_{s,a}]\Lambda_{s,a}/(1-\gamma)$. Thus, the lemma is proven. □

THEOREM EC.4 **(Generalization of Theorem EC.1)**. *For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, let $\theta_{s,a} = ||\hat{F}^X_{s,a} - \hat{F}^Y_{s,a}||$. Then, the difference of the performance between stratifying and not stratifying for Group X satisfies*

$$||\tilde{V}^X - \tilde{V}^{\mathrm{mix},X}|| \le \frac{2}{1-\gamma} \max_{s\in\mathcal{S},a\in\mathcal{A}} \left[\sqrt{\frac{\log 2SA/\delta}{2m_s}} + (1-\mu_{s,a})\theta_{s,a} \right] \Lambda_{s,a} \tag{EC.8}$$

with probability of at least $1-\delta$, where $\Lambda_{s,a}$ is defined in Lemma EC.9.

The proof is almost identical to that of Theorem EC.1, so omitted.

EC.8. Empirical Evaluation of an Obtained Policy

Evaluating a policy using batch transition data is an important part of the main question of this paper and also has been studied extensively in RL literature (see Section 1.1 for more explanations

and references). One of the popular methods is evaluating the given policy's performance using the empirical MDP (called the model-based estimator, e.g., see Pananjady and Wainwright 2020, Uehara et al. 2020). We provide a probabilistic bound on the error of the model-based estimator.

For the finite-horizon case, given a policy π, let $V_t^{\pi,X}(s)$ denote the expected total reward of π starting from state s in period t under the true disturbance distribution of Group X and $\hat{V}_t^{\pi,X}(s)$ denotes the same but evaluated for the empirical distribution of Group X. For the infinite-horizon case, $V^{\pi,X}(s)$ and $\hat{V}^{\pi,X}(s)$ are defined similarly. Let $L_{t,s}^{\pi}$ and L_s^{π} denote a Lipschitz constant of $\hat{V}_t^{\pi,X}(\Psi(s,a,\cdot))$ and $\hat{V}^{\pi,X}(\Psi(s,a,\cdot))$ for any $a \in \mathcal{A}$, respectively.

THEOREM EC.5. *(i)* (The finite-horizon case) *For any policy π, the difference between the two value functions of π for the empirical distribution and the true distribution is bounded as follows*

$$||\hat{V}_t^{\pi,X} - V_t^{\pi,X}|| \leq \sum_{j=t}^{T} \max_{s\in\mathcal{S}} \sqrt{\frac{\log 2S/\delta}{2m_s}} \Lambda_{j,s}^{\pi}$$

with probability of at least $1-\delta$, where

$$\Lambda_{T,s}^{\pi} \triangleq (\beta_s - \alpha_s) l_s \text{ and}$$

$$\Lambda_{t,s}^{\pi} \triangleq (\beta_s - \alpha_s)(l_s + L_{t+1,s}^{\pi}) \text{ for } t < T.$$

(ii) (The infinite-horizon case) *For any policy π, the same difference is bounded as follows*

$$||\hat{V}^{\pi,X} - V^{\pi,X}|| \leq \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \sqrt{\frac{\log 2S/\delta}{2m_s}} \Lambda_s^{\pi}$$

with probability of at least $1-\delta$, where

$$\Lambda_s^{\pi} \triangleq (\beta_s - \alpha_s)(l_s + \gamma L_s^{\pi}).$$

The proof is omitted as it is nearly identical to those of Theorem 1 and Theorem EC.1. This result is similar to Thomas et al. (2015) in that they both use concentration inequalities to give a probabilistic error bound, and thus, they give finite-sample guarantees. The above theorem and the results of Thomas et al. (2015) are not directly comparable, because they used transition data differently.

On the other hand, there are other approaches to estimate the performance of a policy using asymptotic theory. Dai et al. (2020) used an asymptotic result from empirical likelihood to derive a confidence interval of a policy's performance. Also, one can derive a confidence interval by using the fact that the model-based value estimate of a given policy follows a normal distribution asymptotically. The MLE transition probabilities follow a joint normal distribution asymptotically (Anderson and Goodman 1957), and because the estimated value of a policy can be considered a function of the estimated transition probabilities, its limiting distribution is also normal (Mannor et al. 2007, Serfling 1980). For infinite-horizon MDPs, Mannor et al. (2007) derive a closed-form formula approximating the variance, so one can compute a confidence interval. However, we note that these approaches use the asymptotic behavior of the estimated value of a policy, and in practice, we often have a fixed amount of data and cannot easily increase its size.

Given a policy and transition data, we can also empirically quantify the uncertainty of the policy evaluation, following steps similar to the bootstrapping method. Obtain a bootstrapped sample, compute the MLE disturbance distribution from the sample, evaluate the given policy using the estimated distribution, and repeat these steps B times. This can also be used for comparing two given policies empirically.

EC.9. Parameters of the HIV Therapy Problem and Additional Results

All parameters of this problem are from Shechter et al. (2008) except the three transition probability matrices for the Markov chain describing the natural evolution of the CD4 level. Recall that there are 5 states in the HIV therapy problem. 0 means death. States 1 to 4 correspond to the CD4 level 0-49, 50-199, 200-349, and ≥ 350, respectively.

The immediate rewards are in Table EC.1. These values are the quality-adjusted life months in different states of HIV before initiating HAART. They are from Schackman et al. (2002) and can be also found in Table 3 of Shechter et al. (2008).

CD4 level	0-49	50-199	200-349	≥ 350
Immediate Reward (months)	0.88	0.91	0.97	0.97

Table EC.1 Immediate Rewards of Waiting at Different CD4 Levels

The terminal rewards of initiating HAART are in Table EC.2. These values are obtained as explained in Shechter et al. (2008). First, the off-therapy monthly utilities in Table EC.1 are multiplied by a factor to represent on-therapy utilities per month. Shechter et al. (2008) consider three multiplicative factors, 0.9, 0.7, and 0.5. We used 0.7. Then, each of the on-therapy monthly utilities in different health states is multiplied by the expected lifetime after initiating HAART in that health state (the life expectancies are in Table 1 of Shechter et al. 2008).

CD4 level	0-49	50-199	200-349	≥ 350
Terminal Reward (months)	46.87	88.75	140.55	201.99

Table EC.2 Terminal Rewards for Initiating HAART from Different CD4 Levels

The three transition probability matrices for the Markov chain describing the natural evolution of health state are in Tables EC.3, EC.4, and EC.5.

State	Death	0-49	50-199	200-349	≥ 350
Death	1	0	0	0	0
0-49	0.01	0.99	0	0	0
50-199	0.004	0.006	0.99	0	0
200-349	0.002	0.002	0.006	0.99	0
≥ 350	0.001	0.002	0.003	0.004	0.99

Table EC.3 Transition Probability Matrix of the Natural Evolution of CD4 Level, Slow Pace

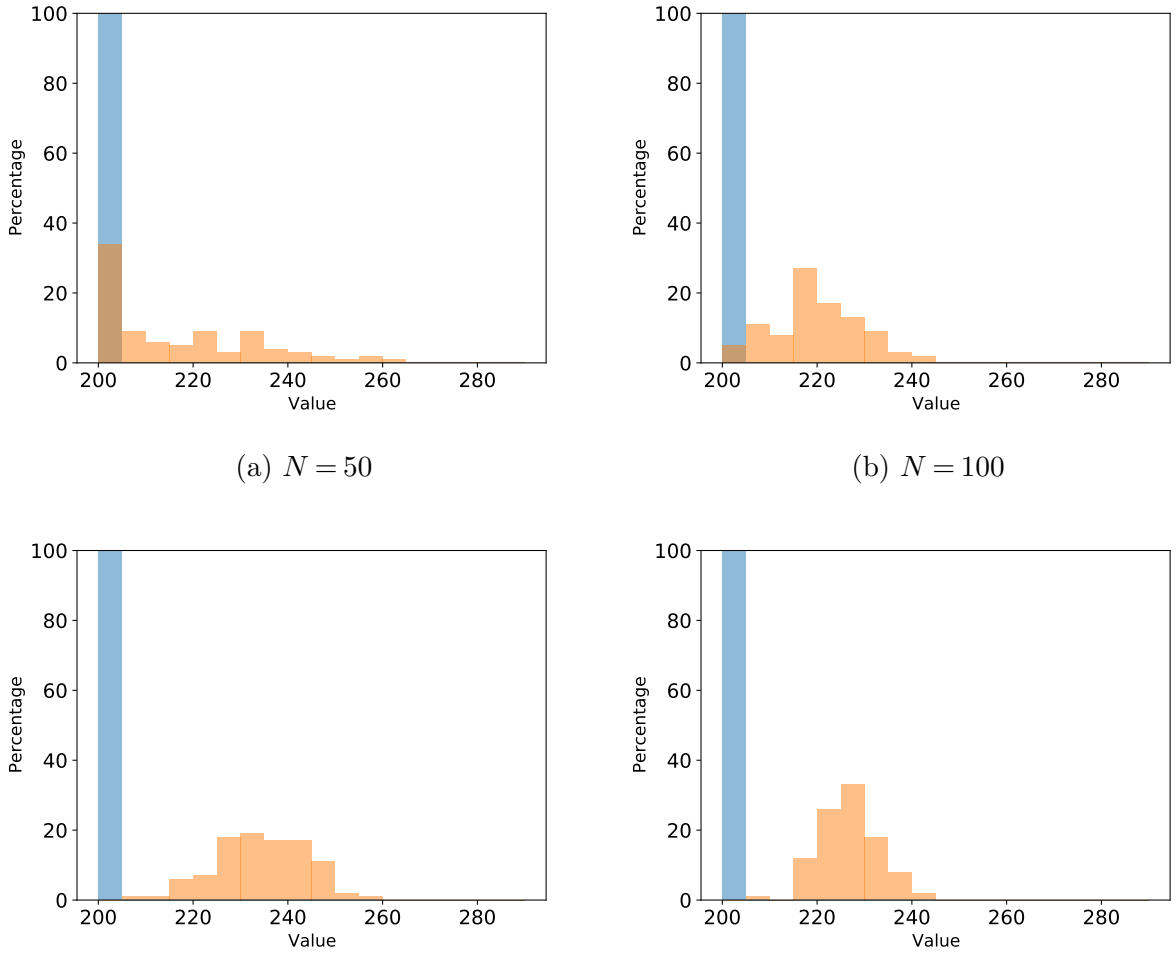
State	Death	0-49	50-199	200-349	≥ 350
Death	1	0	0	0	0
0-49	0.02	0.98	0	0	0
50-199	0.01	0.01	0.98	0	0
200-349	0.002	0.003	0.015	0.98	0
≥ 350	0.001	0.003	0.006	0.010	0.98

Table EC.4 Transition Probability Matrix of the Natural Evolution of CD4 Level, Medium Pace

State	Death	0-49	50-199	200-349	≥ 350
Death	1	0	0	0	0
0-49	0.03	0.97	0	0	0
50-199	0.016	0.014	0.97	0	0
200-349	0.002	0.004	0.024	0.97	0
≥ 350	0.001	0.004	0.009	0.016	0.97

Table EC.5 Transition Probability Matrix of the Natural Evolution of CD4 Level, Fast Pace

Figure EC.1 shows experimental results similar to those in Section 6.3 for the HIV therapy problem with sample sizes 50, 100, 200, and 500. We used P_2 and P_3 for this experiment as there is no benefit of separately modeling P_1 and P_2. The benefit was evaluated for an individual following P_3. In the plots, the value of not stratifying (blue bar) is concentrated at a single value. This is because estimated transition probability matrices of not stratifying are near the mid-point between P_2 and P_3, and for those transition patterns, the optimal policy is initiating the therapy in all states and all periods, which immediately terminates the process and receives a deterministic terminal reward. Thus, although an estimated transition probability matrix of not stratifying involves randomness, the optimal value of not stratifying is not affected by the randomness. On the other hand, stratifying (modeling a separate MDP for P_3, orange in the plot) results in waiting,



(a) $N = 50$ (b) $N = 100$

(c) $N = 200$ (d) $N = 500$

Figure EC.1 Distributions of the empirical value of stratifying and not stratifying for the HIV treatment problem with varying sample sizes, estimated by the bootstrapping method. Orange is stratifying and blue is not stratifying.

so its value is affected by the randomness, and thus, forms a distribution over multiple values. In Figure EC.1, we observe a similar impact of sample size. At $N = 50$ and 100, the two distributions overlap, but at $N = 200$ and 500, they are completely separated, clearly demonstrating the benefit of stratifying. The bootstrapping procedure took 5.69, 15.09, 15.33, and 35.47 seconds for sample sizes 50, 100, 200, and 500, respectively.

EC.10. Generating Random MDPs and the Stratification Selection Problems

In this section, we first explain how we generated each of the random MDP instances used in Section 6. Let Z be the size of the support of the random disturbance. For given values of (T, S, A, Z), we randomly determined the transition function Ψ by assigning a uniformly sampled state to each $\Psi(s, a, z)$. Then we randomly defined the rewards by assigning $r(s, a, z)$ a random number uniformly sampled from $[0, 1]$.

After setting Ψ and the reward, we generate MDPs that differ only in the random disturbance as follows. Determining the distributions of the random disturbance requires more care since this paper analyzes the impact of the difference of the distributions. We used different procedures when we generated two 'differently-behaving' MDPs and two 'similarly-behaving' MDPs. First, we explain how we generated two differently-behaving MDPs. We first generate one MDP as follows. Note that the support of the random disturbance is finite, so the random disturbance has a discrete distribution with Z probabilities. For each state, we do the following. We sample $Z/2$ numbers (if Z is odd, then we use rounding) from $N(\tau, \sigma)$ where $\tau \in (0, 1)$ and $\sigma > 0$. Then, we sample $Z/2$ numbers from $N(1-\tau, \sigma)$. Thus, we have Z random numbers in total and now we normalize the Z numbers so that they sum up to one and become a disturbance distribution for that state. To generate another MDP that is differently-behaving, at each state, we sample the first $Z/2$ numbers from $N(1-\tau, \sigma)$ and the rest from from $N(\tau, \sigma)$. Thus, the parameter τ adjusts how different the two MDPs are. Unless otherwise mentioned, we used $\tau = 0.9$ and $\sigma = 0.1$.

Given an MDP, we generated a similarly-behaving MDP as follows. The only difference is the random disturbance. At each state, we either add or subtract ϵ with probability 0.5 to or from each probability value in the disturbance distribution. Then, we normalize the resulting values so that they form a discrete distribution. We used $\epsilon = 0.1$ to generate a 'similarly-behaving' MDP and $\epsilon = 0.01$ for a 'very similarly-behaving' MDP.

Now let us explain how we generated the instances of the stratification choice problem used in Section 6.4. To generate an instance for a given τ value, we first generate two 'differently-behaving' MDPs using that τ value and call them Groups X and Z. Then, we generate an MDP

'behaving similarly' as X and call it Y. Also, we generate an MDP 'behaving similarly' as Z and call it W. To generate many instances with varying characteristics, we considered the following combinations of parameters. For T, we used 5, 10, and 50. For each value of T, we considered the following combinations for the (S, A, Z) triplet: $(5, 2, 3), (5, 2, 10), (10, 5, 5), (10, 5, 20), (20, 10, 10)$, and $(20, 10, 40)$. Thus, there are 18 combinations of (T, S, A, Z). For each combination, we generated 6 stratification choice problems randomly as described above. Thus, we generated $18 \cdot 6 = 108$ stratification choice problems for each τ value we used. For both σ and ϵ, we used 0.1.