Supplementary Material for *Lightning Does Not Strike Twice: Robust MDPs with Coupled Uncertainty*

A. Proof of Theorem 1

Proof. Define random variables X_s for $s \in \mathcal{S}$ as

$$X_s \triangleq \mathbf{1}_{(p_s, r_s) \neq (p_s^0, r_s^0)}.$$

That is, $X_s = 1$ if the parameters of state s deviate, and $X_s = 0$ otherwise. Thus, $\mathbb{E}X_s = \alpha_s$. Let $Y_s = X_s - \alpha_s$, a zero-mean random variable satisfying $|Y_s| \leq 1$. Bernstein's inequality (Bennett, 1962) gives that for any $c > 0$,

$$\Pr\left\{\sum_{s\in\mathcal{S}} Y_s \geq c\right\} \leq \exp\left\{-\frac{c^2/2}{\sum_{s\in\mathcal{S}} \mathbb{E}Y_s^2 + c/3}\right\}.$$

Note that $\mathbb{E}Y_s^2 = \mathrm{Var}(X_s) = \alpha_s(1-\alpha_s) \leq \alpha_s$, we thus have

$$\Pr\left\{\sum_{s\in\mathcal{S}} X_s \geq \sum_{s\in\mathcal{S}} \alpha_s + c\right\} \leq \exp\left\{-\frac{c^2/2}{\sum_{s\in\mathcal{S}} \alpha_s + c/3}\right\}.$$

Taking the r.h.s to be δ and using the quadratic formula to solved for the positive root of c we get:

$$c = \frac{1}{3}\log(1/\delta)\left(1 + \sqrt{1 + 18\frac{\sum_{s\in\mathcal{S}} \alpha_s}{\log(1/\delta)}}\right)$$

which completes the proof. □

B. Proof of Theorem 2

Proof. We show that $\mathcal{L}(\Pi^{HR})$ is NP hard by reduction from the vertex cover problem, which is known to be an NP-hard problem. Recall that the vertex cover problem is to determine the answer for the following question:

Decision Problem. *Given a non-directed graph $G(V, E)$ where V is the set of vertices and E is the set of edges, and $D \in [1 : |V|]$, does there exists $\mathcal{I} \subseteq V$ with $|\mathcal{I}| \leq D$, such that for each edge $e \in E$, at least one of its nodes belongs to $\mathcal{I}$?*

Given $G(V, E)$, we construct the following MDP $\mathcal{M}$. The state space $\mathcal{S} \triangleq V \bigcup s_{\text{end}}$. That is, each node of the graph corresponds to one state of $\mathcal{M}$, in addition, there exists a terminal state s_{end}. The action space is trivial: for each state, there is only one action. The reward is precisely known as $r(s) = 1$ for all $s \in V$, and $r(s_{\text{end}}) = 0$. That is, the decision maker get a reward 1 at any state other than the terminal one.

The nominal transition probabilities are the following:

$$\forall s \in V:\ p^0(s'|s) = \begin{cases} \frac{1}{\text{degree}(s)} & \text{if } (s, s') \in E; \\ 0 & \text{otherwise.} \end{cases}$$

$$p^0(s'|s_{\text{end}}) = \begin{cases} 1 & \text{if } s' = s_{\text{end}}; \\ 0 & \text{otherwise.} \end{cases}$$

Here degree(s) stands for the number of edges connected to s in $G(V,E)$. Thus, observe that for $s \in V$, the Markov chain is indeed a random walk on $G(V,E)$; while s_{end} is an absorption state.

To construct the uncertainty set, first denote

$$\overline{p}(s'|s) = \begin{cases} 1 & \text{if } s' = s_{\text{end}}; \\ 0 & \text{otherwise.} \end{cases}$$

For $s \in V$, the uncertainty set $\mathcal{U}_s$ is the line segment between p_s^0 and $\overline{p}_s$. There is no uncertainty for the terminal state s_{end}. Observe that for any s, the worst parameter realization in $\mathcal{U}_s$ is $\overline{p}_s$, which leads to the terminal state. The initial state is uniformly distributed over V, and the time horizon is 3.

We now show that the vertex cover problem is equivalent to $\mathcal{L}(\Pi^{HR})$ for the MDP $\mathcal{M}$. In particular, we show that there exists a vertex cover for $G(V,E)$ with at most D nodes, if and only if

$$\max_{\pi \in \Pi^{HR}} \min_{(p,r) \in \mathcal{U}_D} X(\pi, p, r) < \frac{2n - D}{n} + \frac{1}{n^2}.$$

Suppose there exists a D-node vertex cover $\mathcal{I}$. Consider the following parameter realization of $\mathcal{M}$: deviate the parameters of all $s \in \mathcal{I}$ to $\overline{p}_s$. Thus, from $s \in \mathcal{I}$, one will reach the terminal state in the next stage. Observe that such a parameter realization belongs to $\mathcal{U}_D$. Moreover, the total expected reward in this case is $(2n-D)/n$. This is because, any trajectory starting from $s \in \mathcal{I}$ will reach the terminal state at the second stage, thus having a total reward 1. Any trajectory starting from $s \notin \mathcal{I}$ will reach a state $s' \in \mathcal{I}$ due to the fact that $\mathcal{I}$ is a vertex cover, and then reach the terminal state at stage 3. Hence, its total reward is 2. Recall that the initial state is uniformly distributed, we have that the total expected reward under this parameter realization is $(2n - D)/n$. This implies that

$$\max_{\pi \in \Pi^{HR}} \min_{(p,r) \in \mathcal{U}_D} X(\pi, p, r) \leq \frac{2n - D}{n}.$$

On the other hand, suppose there is no D-node vertex cover. Consider an arbitrary subset of $\mathcal{S}$ with cardinality D, denoted by $\mathcal{I}'$, and allow the parameters of all $s \in \mathcal{I}'$ deviate. Recall the worst deviation for any state s is to take $\overline{p}_s$. Similar to previous arguments, a trajectory starting from $s \in \mathcal{I}'$ will have a reward 1. A trajectory starting from $s \notin \mathcal{I}'$ will have a reward at least 2. Moreover, since $\mathcal{I}'$ is not a vertex cover, there exists an edge $e \in E$, whose nodes, denoted s_1, s_2, do not belong to $\mathcal{I}'$. Consider the trajectories starting from s_1, and then reaches s_2. Since $s_2 \notin \mathcal{I}'$, the next state (after s_2) still belongs to V. Such trajectories have a total reward 3. Notice that the probability of such trajectories are at least $1/n^2$. Thus, when parameters of all $s \in \mathcal{I}'$ deviate, the total expected reward is at least $[(2n - D)/n] + (1/n^2)$. Notice that $\mathcal{I}'$ is arbitrary, which leads to

$$\max_{\pi \in \Pi^{HR}} \min_{(p,r) \in \mathcal{U}_D} X(\pi, p, r) \geq \frac{2n - D}{n} + \frac{1}{n^2}.$$

Thus, determining whether a vertex cover of size D exists can be reduced to determining $\mathcal{L}(\Pi^{HR})$. Notice that the action set is a singleton, hence $\Pi^{HR} = \Pi^{MR} = \Pi^{MD}$. The theorem thus follows. □

C. Proof of Theorem 3

Our main tool is the dual LP formulation of MDP, which we recall from (Puterman, 1994) (pp 224, Eq. 6.9.2): for any fixed $r \in \mathcal{U}_D$, solving the resulting MDP can be done by solving the following linear program on x:[1]

$$\begin{aligned} \max_{x \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}} \quad & \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) x(s,a) \\ \text{s.t} \quad & \sum_{a \in \mathcal{A}} x(s',a) - \sum_{\substack{s \in \mathcal{S} \\ a \in \mathcal{A}}} \gamma p^0(s'|s,a) x(s,a) = \alpha(s'), \quad \forall s' \in \mathcal{S}; \\ \text{and} \quad & x(s,a) \geq 0, \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}. \end{aligned}$$

[1] We investigate here the discounted-reward infinite horizon case. Note that a finite-horizon MDP can be trivially converted into a discounted-reward infinite horizon MDP by adding an absorbing terminal state, and appropriately inflating the rewards to offset the discount.

Observe that the feasible set does not depend on the reward parameter r. Hence, in the reward-uncertainty case, Problem 1 can be reformulated as the following robust LP:

$$\text{Maximize:}_{x\in\mathcal{X}} \quad \min_{r\in\mathcal{U}_D} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x(s,a),$$

where the feasible set by $\mathcal{X}$ is given by

$$\mathcal{X} \triangleq \Big\{ x\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \Big| \sum_{a\in\mathcal{A}} x(s',a) - \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \gamma p^0(s'|s,a)x(s,a) = \alpha(s)', \quad \forall s'\in\mathcal{S}; \\ x(s,a)\geq 0, \quad \forall s\in\mathcal{S}, \forall a\in\mathcal{A}. \Big\}$$

Here we give proof of Theorem 3.

Proof. Due to the LP formulation, it suffices to show that the following optimization problem is tractable

$$\begin{aligned} \text{Maximize:}_{x\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|},w\in\mathbb{R}} \quad & w \\ \text{Subject to:} \quad & w \leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x(s,a), \ \forall r\in\mathcal{U}_D; \\ & x\in\mathcal{X}. \end{aligned} \tag{C.1}$$

Observe that for any $r\in\mathcal{U}_D$, $\{(w,x)|w\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x(s,a)\}$ is convex. Hence their intersection, which is the feasible set to C.1, is a convex set. It is known that in the case of optimizing a linear objective in a convex set a sufficient condition for the optimization problem to be solvable in polynomial time is the existence of a polynomial-time *separation oracle* (Grötschel et al., 1988) of the feasible set.[2]

To establish the existence of a separation oracle, we first show that for a fixed (w^0,x^0), in polynomial time we can solve

$$\text{Minimize:}_{r\in\mathcal{U}_D} \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x^0(s,a).$$

Observe that

$$\begin{aligned} \min_{r\in\mathcal{U}_D}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x^0(s,a) &= \min_{I:I\subseteq\mathcal{S},|I|\leq D}\Big\{\sum_{s\in I}\big[\min_{r_s\in\mathcal{U}_s} r_s^\top x_s^0\big] + \sum_{s'\notin I} r_{s'}^{0\top}x_{s'}^0\Big\} \\ &= \min_{I:I\subseteq\mathcal{S},|I|\leq D}\Big\{\sum_{s\in I} r_s^{*\top}x_s^0 + \sum_{s'\notin I} r_{s'}^{0\top}x_{s'}^0\Big\}, \end{aligned}$$

where $r_s^* \triangleq \arg\min_{r_s\in\mathcal{U}_s} x_s^{0\top}r_s$. Since $\mathcal{U}_s$ is tractable, in polynomial time we can solve r_s^*. Furthermore, finding the optimal I can be done efficiently, by ordering the states according to $[r_s^* - r_s^0]^\top x_s^0$ and pick the D smallest states. Thus, in polynomial time we can solve

$$\text{Minimize:}_{r\in\mathcal{U}_D} \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x^0(s,a),$$

where the solution is denoted by $r^\triangle$.

We now construct a separation oracle by considering three possibilities:

- **Case 1** – $x^0\notin\mathcal{X}$: then any violated linear constraint is a separation oracle.
- **Case 2** – $x^0\in\mathcal{X}$ and $w^0\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r^\triangle(s,a)x^0(s,a)$: then one can conclude that (w^0,x^0) is feasible.

[2]Here, a separation oracle stands for a routine, such that given an arbitrary vector it can in polynomial time tell whether the vector belongs to the set, and further output a separation hyperplane if the vector does not belong to the set.

- **Case 3** – $x^0 \in \mathcal{X}$ and $w^0 > \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\triangle}(s,a)x^0(s,a)$: then the hyperplane

$$w - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\triangle}(s,a)x(s,a) = 0$$

  is a separation hyperplane.

Recall that $r^{\triangle}$ can be solved in polynomial time, the separation oracle can be constructed in polynomial time. This establishes the theorem. □

D. Theorem D.1, Statement and Proof

Theorem D.1. *Suppose that only the rewards are subject to uncertainty. Let*

$$\hat{\mathcal{U}}_D \triangleq \Bigg\{ r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \Big| \exists \alpha \in \mathbb{R}^{|\mathcal{S}|} : 0 \le \alpha \le 1 : \sum_{s\in S} \alpha_s \le D; \\ \exists \delta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} : \delta_s \in \Delta\mathcal{U}_s; r_s = r_s^0 + \alpha_s \delta_s \Bigg\},$$

$$\text{where} \quad \Delta\mathcal{U}_s \triangleq \{\delta_s \in \mathbb{R}^{|\mathcal{A}|} | r_s^0 + \delta_s \in \mathcal{U}_s\},$$

then Problem 1 is equivalent to

$$\text{Maximize:}_{x\in\mathcal{X}} \qquad \min_{r\in\hat{\mathcal{U}}_D} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)x(s,a). \tag{D.1}$$

Proof. It suffices to show that for any fixed x, we have

$$\min_{r\in\mathcal{U}_D} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)x(s,a) = \min_{r\in\hat{\mathcal{U}}_D} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)x(s,a). \tag{D.2}$$

We let x_s denote $\big(x(s,a_1), \cdots, x(s,a_{|\mathcal{A}|})\big)^\top$ for all s, and establish Equation (D.2) via the following algebraic manipulation,

$$\begin{aligned}
& \min_{r\in\hat{\mathcal{U}}_D} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)x(s,a) = \min_{r\in\hat{\mathcal{U}}_D} \sum_{s\in\mathcal{S}} r_s^\top x_s \\
= \quad & \min_{\alpha:0\le\alpha\le 1, \sum_s \alpha_s \le D} \Big[\min_{\{r_s | r_s = r_s^0 + \alpha_s\delta_s, \delta_s\in\Delta\mathcal{U}_s\}} \sum_{s\in\mathcal{S}} r_s^\top x_s \Big] \\
= \quad & \min_{\alpha:0\le\alpha\le 1, \sum_s \alpha_s \le D} \sum_{s\in\mathcal{S}} \Big[r_s^0 x_s + \alpha_s \min_{\delta_s\in\Delta\mathcal{U}_s} \delta_s^\top x_s \Big] \\
\overset{(a)}{=} \quad & \sum_{s\in\mathcal{S}} r_s^0 x_s + \min_{I: I\subseteq\mathcal{S}, |I|\le D} \Bigg\{ \sum_{s\in I} \min_{\delta_s\in\Delta\mathcal{U}_s} \delta_s^\top x_s \Bigg\} \\
= \quad & \min_{r\in\mathcal{U}_D} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)x(s,a),
\end{aligned}$$

where in (a) we use the fact that the optimal solution to an LP lies on a vertex of the feasible set. □

E. Reward-only Uncertainties with Polytopial Uncertainty Sets

By Theorem D.1 and duality of LP we get the following corollary:

Corollary E.1. *Suppose that only the reward parameters are subject to uncertainty. If for all $s \in \mathcal{S}$, $\mathcal{U}_s$ are polytopes defined as $\mathcal{U}_s = \{r_s | A_s r_s \ge c_s\}$, then Problem 1 can be formulated as a Linear Programming problem.*

Proof. In Theorem D.1 it was proved that when only reward is uncertain, Problem 1 can be re-written as

$$\max_{x\in\mathcal{X}} \min_{r\in\hat{\mathcal{U}}_D} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x(s,a),$$

which is further equivalent to

$$\max_{x\in\mathcal{X}} \min_{\alpha:0\le\alpha\le 1,\sum_{s\in\mathcal{S}}\alpha_s\le D} \sum_{s\in\mathcal{S}} \left[r_s^0 x_s + \alpha_s \min_{\delta_s\in\Delta\mathcal{U}_s} \delta_s^\top x_s\right].$$

By duality we have

$$\min_{\delta_s\in\Delta\mathcal{U}_s} \delta_s^\top x_s = \max_{z_s: A_s^\top z_s = x_s, z_s\ge 0} : \ [c_s - A_s r_s^0]^\top z_s.$$

Note that A_s, c_s and r_s^0 are fixed. Therefore the right-hand-side is a function of x_s, which we denote as $f_s(x_s)$. Thus we have

$$\min_{\alpha:0\le\alpha\le 1,\sum_{s\in\mathcal{S}}\alpha_s\le D} \sum_{s\in\mathcal{S}} \left[r_s^{0\top} x_s + \alpha_s \min_{\delta_s\in\Delta\mathcal{U}_s} \delta_s^\top x_s\right] = \sum_{s\in\mathcal{S}} r_s^{0\top} x_s + \min_{\alpha:0\le\alpha\le 1,\sum_s\alpha_s\le D} \sum_{s\in\mathcal{S}} \alpha_s f_s(x_s).$$

Again by duality, we have

$$\min_{\alpha:0\le\alpha\le 1,\sum_{s\in\mathcal{S}}\alpha_s\le D} \sum_{s\in\mathcal{S}} \alpha_s f_s(x_s) = \begin{cases} \max_{v,w}: & \sum_s v_s + Dw \\ \text{Subject to:} & v_s + w \le f_s(x_s),\ \forall s\in\mathcal{S}; \\ & v_s \le 0,\ \forall s\in\mathcal{S}; \\ & w\le 0. \end{cases}$$

By definition of $f_s(\cdot)$, this can be further simplified as

$$\begin{aligned} \max_{z,v,w}: \quad & \sum_s v_s + Dw \\ \text{Subject to:} \quad & v_s + w \le [c_s - A_s r_s^0]^\top z_s,\ \forall s\in\mathcal{S}; \\ & A_s^\top z_s = x_s,\ \forall s\in\mathcal{S}; \\ & z_s \ge 0,\ \forall s\in\mathcal{S}; \\ & v_s \le 0,\ \forall s\in\mathcal{S}; \\ & w\le 0. \end{aligned}$$

Hence, Problem 1 is equivalent to

$$\begin{aligned} \max_{x,v,w,z} \quad & \sum_s (r_s^{0\top} x_s + v_s) + Dw \\ \text{Subject to:} \quad & v_s + w \le [c_s - A_s r_s^0]^\top z_s,\ \forall s\in\mathcal{S}; \\ & A_s^\top z_s = x_s,\ \forall s\in\mathcal{S}; \\ & z_s \ge 0,\ \forall s\in\mathcal{S}; \\ & v_s \le 0,\ \forall s\in\mathcal{S}; \\ & w\le 0, \\ & \sum_{a\in\mathcal{A}} x(s',a) - \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \gamma p^0(s'|s,a)x(s,a) = \alpha(s)',\ \forall s'\in\mathcal{S}; \\ & x(s,a)\ge 0,\ \forall s\in\mathcal{S}, \forall a\in\mathcal{A}. \end{aligned}$$

□

F. Infinite-Horizon LDST MDPs

We first show that Problem 4 (Setup A) is solvable by augmenting the state-space with the number of deviations to go.

Theorem F.1. *Denote*

$$q(s,d,a,p,r,v) \triangleq \gamma \sum_{s'} p(s'|s,a)v(s',d) + r(s,a).$$

Then:

(a) The following set of equations,

$$v(s,d) = \begin{cases} \max\limits_{a\in\mathcal{A}} \min\left[q(s,d,a,p_s^0,r_s^0,v), \min\limits_{(p,r)\in\mathcal{U}_s} q(s,d-1,a,p,r,v)\right], & \text{if } d \geq 1; \\ \max\limits_{a\in\mathcal{A}} q(s,d,a,p_s^0,r_s^0,v), & \text{if } d = 0; \end{cases}$$

have a unique solution, denoted by v^*.

(b) Let

$$a^*(s,d) = \begin{cases} \arg\max\limits_{a\in\mathcal{A}} \min\left[q(s,d,a,p_s^0,r_s^0,v), \min\limits_{(p,r)\in\mathcal{U}_s} q(s,d-1,a,p,r,v)\right], & \text{if } d \geq 1; \\ \arg\max\limits_{a\in\mathcal{A}} q(s,d,a,p_s^0,r_s^0,v), & \text{if } d = 0; \end{cases}$$

Then the optimal policy to Problem 4 is to take $a^*(s,d)$ at state s, where stage-parameters are allowed to deviate at most d times from this stage on.

Proof. To prove the first statement, observe that for fixed s,d,a,p,r, the function $q(s,d,a,p,r,\cdot)$ is a γ contraction with respect to the ℓ_∞ norm. That is,

$$|q(s,d,a,p,r,v_1) - q(s,d,a,p,r,v_2)| \leq \gamma \|v_1 - v_2\|_\infty.$$

Notice that the property of being a γ-contraction is preserved under minimization and maximization, hence the first statement follows from Banach's fixed point theorem.

To show the second statement, we again consider a stochastic game with perfect information. The game is constructed in a same way as the one presented in the finite horizon case, except that the game will be played for infinite number of steps, and that future rewards will be discounted. It is easy to see that solving Problem 4 is equivalent to solving this game.

We now construct a Nash equilibrium. At the decision epoch $2t-1$ the decision maker makes his move, let his move be $a^*(s,d)$ at $(s,d) \in \overline{\mathcal{S}}_D$. At the decision epoch $2t$ Nature makes his move, let Nature take the following action at $(s,d,a) \in \overline{\mathcal{S}}_N$: do not deviate the parameters – i.e, pick parameter (p_s^0, r_s^0) – if either $d=0$, or $q(s,d,a,r^0,p^0,v^*) \leq \min_{(p,r)\in\mathcal{U}_s} q(s,d-1,a,p,r,v^*)$; otherwise, pick parameters $\arg\min_{(p,r)\in\mathcal{U}_s} q(s,d-1,a,p,r,v^*)$.

Observe that if Nature fixes this strategy, the stochastic game reduces to a standard infinite horizon, discounted-reward MDP, for the decision maker to maximize his total discounted reward. It is easy to check that the optimal strategy is take $a^*(s,d)$ at each $(s,d) \in \overline{\mathcal{S}}_D$ at the decision epoch $2t-1$.

On the other hand, if the decision maker fixes his strategy, then the optimal strategy of Nature solves an MDP with compact action space. Denote the state value for this Nature's MDP by $\tilde{v}$, then by Bellman's Equations we have:

$$\tilde{v}(s,d,a) = \begin{cases} \min\left[\gamma \sum\limits_{s'\in\mathcal{S}} p^0(s'|s,a)\tilde{v}(s',d) + r^0(s,a), \min\limits_{(p,r)\in\mathcal{U}_s} [\gamma \sum\limits_{s'\in\mathcal{S}} p(s'|s,a)\tilde{v}(s',d-1) + r(s,a)]\right], & \text{if } d \geq 1; \\ \gamma \sum\limits_{s'\in\mathcal{S}} p^0(s'|s,a)\tilde{v}(s',d) + r^0(s,a) & \text{if } d = 0; \end{cases}$$

where $\tilde{v}(s,d) \triangleq \tilde{v}(s,d,a^*(s,d))$.

Observe that $\tilde{v}(s,d) = v^*(s,d)$. Hence, the aforementioned strategy of Nature is optimal, if the decision maker fixed his strategy. Thus, the aforementioned pair of strategies is a Nash equilibrium, which implies the theorem holds. □

Problem 5 (Setup B) is more complicated – because the number of deviations is also discounted, the remaining budget of Nature is not necessarily an integer. Thus, state-augmentation leads to a stochastic game with a continuous state-space. Define the following quantity for $s \in \mathcal{S}$ and $d \in \mathbb{R}$,

$$v^*(s,d) \triangleq \max_{a_1 \in \mathcal{A}} \min_{(p_1,r_1)\in\mathcal{U}_{s_1}} \max_{a_2 \in \mathcal{A}} \min_{(p_2,r_2)\in\mathcal{U}_{s_2}} \cdots \max_{a_t \in \mathcal{A}} \min_{(p_t,r_t)\in\mathcal{U}_{s_t}} \cdots \mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t,a_t)\right\};$$

$$\text{Subject to:} \quad \sum_{t=1}^{\infty} \beta^{t-1} \mathbf{1}_{(p_t,r_t)\neq(p_t^0,r_t^0)} \leq d; \; s_1 = s.$$

Notice that v^* is the value function at state s with budget d. Following a similar contraction argument as the proof of Theorem F.1 (recall that v is uniformly bounded), we have the following theorem regarding the value function. Notice that due to discounting, the remaining budget that pass to the next stage will be increased by a factor of $1/\beta$.

Theorem F.2. *Denote*

$$q(s,d,a,p,r,v) \triangleq \gamma \sum_{s'} p(s'|s,a) v(s',d) + r(s,a).$$

Then $v^*(s,d)$ is the unique solution to the following set of equations on v,

$$v(s,d) = \begin{cases} \max_{a\in\mathcal{A}} \left[\min\left(q(s,d/\beta,a,p_s^0,r_s^0,v), \min_{(p,r)\in\mathcal{U}_s} q(s,(d-1)/\beta,a,p,r,v)\right)\right], & \text{if } d \geq 1; \\ \max_{a\in\mathcal{A}} q(s,d/\beta,a,p_s^0,r_s^0,v), & \text{if } d < 1. \end{cases}$$

However, as d can take infinitely many values, solving v^* exactly appears intractable. Instead, we approximate it using discretization. We first restrict d in $[0, 1/(1-\beta)]$, using the following lemma.

Lemma F.1. *If $d > 1/(1-\beta)$, then for any $s \in \mathcal{S}$,*

$$v^*(s,d) = v^*(s, 1/(1-\beta)).$$

Proof. This follows immediately from the definition of $v^*(s,d)$ and $\sum_{t=1}^{\infty} \beta^{t-1} = 1/(1-\beta)$. □

Next we discretize the interval $[0, 1/(1-\beta)]$. In particular, we fix $\kappa > 0$ and denote $\mathcal{K} \triangleq \{0, \kappa, \cdots, \lceil \frac{1}{(1-\beta)\kappa} \rceil \kappa\}$. Instead of allowing d to take any value in $[0, 1/(1-\beta)]$, we restrict d to $\mathcal{K}$. The next theorem provides guarantees for such approximation.

Theorem F.3. *Let $h : [0, 1/(1-\beta)] \mapsto \mathcal{K}$ be*

$$h(x) \triangleq \min\{y \in \mathcal{K} | y \geq x\}.$$

Let $\overline{v} : \mathcal{S} \times \mathcal{K} \mapsto \mathbb{R}$ be the unique solution to the following set of equations on v

$$\forall s \in \mathcal{S}, d \in \mathcal{K}: \qquad v(s,d) = \begin{cases} \max_{a\in\mathcal{A}} \left[\min\left(q(s,h(\frac{d}{\beta}),a,p_s^0,r_s^0,v), \min_{(p,r)\in\mathcal{U}_s} q(s,h(\frac{d-1}{\beta}),a,p,r,v)\right)\right], & \text{if } d \geq 1; \\ \max_{a\in\mathcal{A}} q(s,h(\frac{d}{\beta}),a,p_s^0,r_s^0,v), & \text{if } d < 1. \end{cases}$$

Then we have for all $s \in \mathcal{S}$ and $d \in \mathcal{K}$

$$v^*(s, d + \frac{\beta\kappa}{1-\beta}) \leq \overline{v}(s,d) \leq v^*(s,d).$$

To establish Theorem F.3, we need the following lemma.

Lemma F.2. *Suppose functions $v_1(s,d)$ and $v_2(s,d)$ are non-increasing over d, and for all $s \in \mathcal{S}, d \geq 0$, $v_1(s,d) \leq v_2(s,d)$. Let $d_1 \geq d_2$, then we have*

$$q(s,d_1,a,p,r,v_1) \leq q(s,d_2,a,p,r,v_2).$$

Proof. By definition

$$\begin{aligned} q(s, d_1, a, p, r, v_1) &= \gamma \sum_{s'} p(s'|s,a) v_1(s', d_1) + r(s,a) \\ &\le \gamma \sum_{s'} p(s'|s,a) v_2(s', d_1) + r(s,a) \\ &\le \gamma \sum_{s'} p(s'|s,a) v_2(s', d_2) + r(s,a) \\ &= q(s, d_2, a, p, r, v_2), \end{aligned}$$

where the first inequality follows from $v_1 \le v_2$, and the second inequality follows from the assumption that v_2 is non-increasing. □

Now we turn to prove Theorem F.3.

Proof. First notice that the existence and uniqueness of $\overline{v}(\cdot,\cdot)$ follows from a contraction argument, similar to that in the proof of Theorem F.1. Observe that for any $s \in \mathcal{S}$, $v^*(s,d)$ is non-increasing in d.

Define three operators Γ^*, Γ_* and $\overline{\Gamma}$ as the following:

$$\Gamma_*(v)(s,d) =$$
$$\begin{cases} \max_{a\in\mathcal{A}} \Big[\min(q(s, d/\beta + \kappa, a, p_s^0, r_s^0, v), \min_{(p,r)\in\mathcal{U}_s} q(s, (d-1)/\beta + \kappa, a, p, r, v)) \Big], & \text{if } d \ge 1; \\ \max_{a\in\mathcal{A}} q(s, d/\beta + \kappa, a, p_s^0, r_s^0, v), & \text{if } d < 1; \end{cases}$$
$$\Gamma^*(v)(s,d) =$$
$$\begin{cases} \max_{a\in\mathcal{A}} \Big[\min(q(s, d/\beta, a, p_s^0, r_s^0, v), \min_{(p,r)\in\mathcal{U}_s} q(s, (d-1)/\beta, a, p, r, v)) \Big], & \text{if } d \ge 1; \\ \max_{a\in\mathcal{A}} q(s, d/\beta, a, p_s^0, r_s^0, v), & \text{if } d < 1; \end{cases}$$
$$\overline{\Gamma}(v)(s,d) =$$
$$\begin{cases} \max_{a\in\mathcal{A}} \Big[\min(q(s, h(\frac{d}{\beta}), a, p_s^0, r_s^0, v), \min_{(p,r)\in\mathcal{U}_s} q(s, h(\frac{d-1}{\beta}), a, p, r, v)) \Big], & \text{if } d \ge 1; \\ \max_{a\in\mathcal{A}} q(s, h(\frac{d}{\beta}), a, p_s^0, r_s^0, v), & \text{if } d < 1. \end{cases}$$

Observe that $v^* = \lim_{n\to\infty} (\Gamma^*)^n(v)$ and $\overline{v} = \lim_{n\to\infty} \overline{\Gamma}^n(v)$ for any initial $v(\cdot,\cdot)$. Furthermore, following a contraction argument, $\lim_{n\to\infty} (\Gamma_*)^n(v)$ uniquely exists regardless of the choice of initial $v(\cdot,\cdot)$. We denote this limit as v_*. Notice that for any x, $x \le h(x) \le x + \kappa$. Thus, by Lemma F.2, we have

$$\Gamma_*(\overline{v})(s,d) \le \overline{\Gamma}(\overline{v})(s,d) \le \Gamma^*(\overline{v})(s,d).$$

Applying Lemma F.2 repeatedly, we have in the limit

$$v_*(s,d) \le \overline{v}(s,d) \le v^*(s,d). \tag{F.1}$$

Our last step is to bound $v_*(s,d)$. Notice that $v_*(s,d)$ is the optimal value to the following

$$\begin{aligned} &\max_{a_1\in\mathcal{A}} \min_{(p_1,r_1)\in\mathcal{U}_{s_1}} \max_{a_2\in\mathcal{A}} \min_{(p_2,r_2)\in\mathcal{U}_{s_2}} \cdots \mathbb{E}\left\{ \sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t, a_t) \right\}; \\ &\text{Subject to:} \quad \sum_{t=1}^{\infty} \beta^{t-1} \Big\{ \mathbf{1}_{(p_t,r_t)\neq(p_t^0,r_t^0)} - \beta\kappa \Big\} \le d. \end{aligned} \tag{F.2}$$

Simplifying the constraint, we can rewrite Equation (F.2) as

$$\begin{aligned} &\max_{a_1\in\mathcal{A}} \min_{(p_1,r_1)\in\mathcal{U}_{s_1}} \max_{a_2\in\mathcal{A}} \min_{(p_2,r_2)\in\mathcal{U}_{s_2}} \cdots \mathbb{E}\left\{ \sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t, a_t) \right\}; \\ &\text{Subject to:} \quad \sum_{t=1}^{\infty} \beta^{t-1} \mathbf{1}_{(p_t,r_t)\neq(p_t^0,r_t^0)} \le d + \frac{\beta\kappa}{1-\beta}, \end{aligned}$$

whose optimal value, by definition, is $v^*(s, d + \frac{\beta\kappa}{1-\beta})$. Hence we have

$$v^*(s, d + \frac{\beta\kappa}{1-\beta}) = v_*(s, d).$$

The theorem follows by inserting the inequality into Equation (F.1). □

We remark that to achieve the discretized value function $\overline{v}$, the decision maker can take the following strategy $\overline{a}(s, d)$,

$$\forall s \in \mathcal{S}, d \in \mathcal{K}: \qquad \overline{a}(s,d) = \begin{cases} \arg\max_{a\in\mathcal{A}} \left[\min\left(q(s, h(\frac{d}{\beta}), a, p_s^0, r_s^0, v), \min_{(p,r)\in\mathcal{U}_s} q(s, h(\frac{d-1}{\beta}), a, p, r, v)\right)\right], \text{ if } d \geq 1; \\ \arg\max_{a\in\mathcal{A}} q(s, h(\frac{d}{\beta}), a, p_s^0, r_s^0, v), \text{ if } d < 1. \end{cases}$$

Before concluding this section, we briefly discuss the computational complexity of the two setups. In Setup A, recall the algorithm repeatedly apply the finite-horizon algorithm (whose complexity is $O(TD|S||A|(|S| + M))$) and resort to the γ contraction. Thus to achieve an accuracy of ϵ, the complexity is $O(TD|S||A|(|S| + M)\log\epsilon/\log\gamma)$. For Setup B, since we discretize the possible remaining budget to a finite set with $1/((1-\beta)\kappa)$ elements, the total complexity is then $O(\frac{1}{(1-\beta)\kappa}D|S||A|(|S| + M)\log\epsilon/\log\gamma)$.

G. Adaptive Continuous Deviations

In this appendix we solve Problem 6. We let $v^*(s, d)$ denote the value function, i.e.,

$$\begin{aligned} v^*(s,d) \triangleq & \max_{a_1\in\mathcal{A}} \min_{(p_1,r_1)\in\mathcal{U}_{s_1}} \max_{a_2\in\mathcal{A}} \min_{(p_2,r_2)\in\mathcal{U}_{s_2}} \cdots \\ & \max_{a_t\in\mathcal{A}} \min_{(p_t,r_t)\in\mathcal{U}_{s_t}} \cdots \mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t, a_t)\right\}; \\ \text{Subject to:} \quad & \sum_{t=1}^{\infty} \beta^{t-1} b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) \leq d; \ s_1 = s. \end{aligned}$$

Then by contraction argument, we have that v^* is the unique solution to the following DP equation.

$$\begin{aligned} v(s,d) = & \max_{a\in\mathcal{A}} \min_{\substack{\alpha\in[0,1] \\ \alpha\leq d \\ (\delta_p,\delta_r)\in\Delta\mathcal{U}_s}} q\left(s, \frac{d-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right), \\ \text{where } q(s,d,a,p,r,v) \triangleq & \gamma \sum_{s'} p(s'|s,a) v(s', d) + r(s,a). \end{aligned}$$

The remaining budget d can take infinitely many values, which makes computing v^* intractable. Thus, we discretize it. In particular, let

$$d_{max} = \begin{cases} D & \text{if } \beta = 1; \\ \frac{1}{1-\beta} & \text{if } \beta < 1. \end{cases}$$

It is easy to see that $v^*(s, d) = v^*(s, d_{max})$ for all $d \geq d_{max}$. We fix $\kappa > 0$ and denote $\mathcal{K} \triangleq \{0, \kappa, \cdots, \lceil\frac{d_{max}}{\kappa}\rceil\kappa\}$. Instead of allowing d to take any value in $[0, d_{max}]$, we restrict d to $\mathcal{K}$. The following theorem provides guarantees to the quality of the approximation.

Theorem G.1. *Denote* $M = \sup_{s_1,a_1,r_1\in\mathcal{U}_{s_1}} r_1(s_1, a_1) - \inf_{s_2,a_2,r_2\in\mathcal{U}_{s_2}} r_2(s_2, a_2)$. *Let* $h : [0, d_{max}] \mapsto \mathcal{K}$ *be defined as*

$$h(x) = \min_{y\in\mathcal{K}, y\geq x} y.$$

Let $\overline{v}: \mathcal{S} \times \mathcal{K} \mapsto \mathbb{R}$ be the unique solution to the following set of equations

$$\forall s \in \mathcal{S}, d \in \mathcal{K}: \qquad v(s,d) = \max_{a \in \mathcal{A}} \inf_{\substack{\alpha \in [0,1] \\ \alpha \le d \\ (\delta_p, \delta_r) \in \Delta \mathcal{U}_s}} q\left(s, h\left(\frac{d-\alpha}{\beta}\right), a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right).$$

Then we have for all $s \in \mathcal{S}$ and $d \in \mathcal{K}$

$$v^*(s, d + \left\lceil \frac{\log(1/\kappa)}{\log(1/\gamma)} \right\rceil \beta\kappa) - \frac{\kappa M}{1-\gamma} \le \overline{v}(s,d) \le v^*(s,d).$$

If $\beta < 1$ we further have

$$v^*(s, d + \frac{\beta\kappa}{1-\beta}) \le \overline{v}(s,d) \le v^*(s,d).$$

Proof. Similar to the proof of Theorem F.3, we define three operators Γ^*, Γ_* and $\overline{\Gamma}$ as the following:

$$\Gamma_*(v)(s,d) = \left\{\max_{a \in \mathcal{A}} \left[\min_{\alpha \in [0,1], \alpha \le d, (\delta_p, \delta_r) \in \Delta \mathcal{U}_s} q\left(s, \frac{d-\alpha}{\beta} + \kappa, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \right] \right\},$$

$$\Gamma^*(v)(s,d) = \left\{\max_{a \in \mathcal{A}} \left[\min_{\alpha \in [0,1], \alpha \le d, (\delta_p, \delta_r) \in \Delta \mathcal{U}_s} q\left(s, \frac{d-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \right] \right\},$$

$$\overline{\Gamma}(v)(s,d) = \left\{\max_{a \in \mathcal{A}} \left[\inf_{\alpha \in [0,1], \alpha \le d, (\delta_p, \delta_r) \in \Delta \mathcal{U}_s} q\left(s, h\left(\frac{d-\alpha}{\beta}\right), a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \right] \right\}.$$

Observe that $v^* = \lim_{n\to\infty} (\Gamma^*)^n(v)$ and $\overline{v} = \lim_{n\to\infty} \overline{\Gamma}^n(v)$ for any initial v. Furthermore, $\lim_{n\to\infty}(\Gamma_*)^n(v)$ uniquely exists, which we denote as v_*. Notice that for any x, $x \le h(x) \le x + \kappa$. Thus, we have

$$v_*(s,d) \le \overline{v}(s,d) \le v^*(s,d). \tag{G.1}$$

Also note that $v_*(s,d)$ is the optimal value to the following

$$\max_{a_1 \in \mathcal{A}} \min_{(p_1,r_1) \in \mathcal{U}_{s_1}} \max_{a_2 \in \mathcal{A}} \min_{(p_2,r_2) \in \mathcal{U}_{s_2}} \cdots \max_{a_t \in \mathcal{A}} \min_{(p_t,r_t) \in \mathcal{U}_{s_t}} \cdots \mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t, a_t)\right\};$$

$$\text{Subject to:} \quad \sum_{t=1}^{\infty} \beta^{t-1}\Big\{b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) - \beta\kappa\Big\} \le d; \ s_1 = s. \tag{G.2}$$

When $\beta < 1$, we have

$$\sum_{t=1}^{\infty} \beta^{t-1}\Big\{b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) - \beta\kappa\Big\} = \sum_{t=1}^{\infty} \beta^{t-1} b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) - \frac{\beta\kappa}{1-\beta},$$

which leads to

$$v^*(s, d + \beta\kappa/(1-\beta)) \le v_*(s,d). \tag{G.3}$$

However, this bound becomes vacuous when $\beta = 1$. Hence we establish the following bound. Fix integer T, and consider the optimal value of the following problem, denoted as $\hat{v}(s,d)$,

$$\max_{a_1 \in \mathcal{A}} \min_{(p_1,r_1) \in \mathcal{U}_{s_1}} \cdots \mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t(s_t, a_t)\right\};$$

$$\text{Subject to:} \quad \sum_{t=1}^{\infty} \beta^{t-1}\Big\{b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) - \beta\kappa\Big\} \le d; \tag{G.4}$$

$$p_t = p_t^0, \ r_t = r_t^0, \ \forall t \ge T+1;$$

$$s_1 = s.$$

That is, Problem G.4 is essentially Problem G.2 with an additional constraint that parameter deviation is only allowed in the first T time steps. Recall that M is the maximal difference in the reward that the decision maker gets in one stage, hence we can bound $\hat{v}(s,d) - v_*(s,d)$ by the difference of rewards incurred in stage $T+1$ onwards, i.e.,

$$\hat{v}(s,d) - v_*(s,d) \leq \sum_{t=T+1}^{\infty} \gamma^{t-1} M \leq \gamma^T M/(1-\gamma).$$

Further note that for any allowed deviation of (G.4), we have

$$\sum_{t=1}^{\infty} \beta^{t-1} \Big\{ b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) \Big\} \leq d + T\beta\kappa,$$

which implies that

$$v^*(s, d + T\beta\kappa) \leq \hat{v}(s,d).$$

Therefore, we have that $v^*(s, d+T\beta\kappa) - \gamma^T M/(1-\gamma) \leq v_*(s,d)$. Take $T = \lceil \log(1/\kappa)/\log(1/\gamma) \rceil$, we have that

$$v^*(s, d + \lceil \log(1/\kappa)/\log(1/\gamma) \rceil \beta\kappa) - \kappa M/(1-\gamma) \leq v_*(s,d).$$

Inserting the above equation and Equation (G.3) into Equation (G.1) establishes the theorem. □

Furthermore, we can prove that $v^*(s,d)$ is Liptsitz continuous, which gives a simpler bound for $\overline{v}(s,d)$:

Theorem G.2. *Recall that $M = \sup_{s_1,a_1,r_1 \in \mathcal{U}_{s_1}} r_1(s_1,a_1) - \inf_{s_2,a_2,r_2 \in \mathcal{U}_{s_2}} r_2(s_2,a_2)$. We have that $v^*(s,d)$ is Liptsitz continuous with a coefficient $M/(1-\gamma)$.*

To prove the theorem, we first establish the following lemma.

Lemma G.1. *Denote $M = \sup_{s_1,a_1,r_1 \in \mathcal{U}_{s_1}} r_1(s_1,a_1) - \inf_{s_2,a_2,r_2 \in \mathcal{U}_{s_2}} r_2(s_2,a_2)$. Consider a function $v : \mathcal{S} \times \mathbb{R}^+ \mapsto \mathbb{R}$ and let $V = \sup_{s,d,s',d'} [v(s,d) - v(s',d')]$. If for all $s \in \mathcal{S}$, $v(s,\cdot)$ is non-increasing, and Liptsitz continuous with a coefficient $M + \gamma V$. Then v', defined as follows*

$$v'(s,d) = \max_{a \in \mathcal{A}} \left\{ \min_{\alpha \in [0,1], \alpha \leq d, (\delta_p, \delta_r) \in \Delta\mathcal{U}_s} q\left(s, \frac{d-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \right\},$$

is Liptsitz continuous (w.r.t. the second argurment) with a coefficient $M + \gamma V$. Furthermore

$$\sup_{s,d,s',d'} [v'(s,d) - v'(s',d')] \leq M + \gamma V.$$

Proof. The second claim follows easily from the definitions of V and M. We next prove the first claim. Let $L \triangleq M + \gamma V$. Note that for all $s \in \mathcal{S}$, $v'(s,\cdot)$ is also non-increasing, hence we consider

$$\begin{aligned} &v'(s,d) - v'(s,d+\epsilon) \leq \\ &\max_{a \in \mathcal{A}} \left\{ \min_{\alpha \in [0,1], \alpha \leq d, (\delta_p, \delta_r) \in \Delta\mathcal{U}_s} q\left(s, \frac{d-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \right. \\ &\qquad \left. - \min_{\alpha \in [0,1], \alpha \leq d+\epsilon, (\delta_p, \delta_r) \in \Delta\mathcal{U}_s} q\left(s, \frac{d+\epsilon-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \right\} \end{aligned}$$

Hence, fix a, we bound the following,

$$\begin{aligned} &\min_{\alpha \in [0,1], \alpha \leq d, (\delta_p, \delta_r) \in \Delta\mathcal{U}_s} q\left(s, \frac{d-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right) \\ &\qquad - \min_{\alpha \in [0,1], \alpha \leq d+\epsilon, (\delta_p, \delta_r) \in \Delta\mathcal{U}_s} q\left(s, \frac{d+\epsilon-\alpha}{\beta}, a, p_s^0 + \alpha\delta_p, r_s^0 + \alpha\delta_r, v\right). \end{aligned} \tag{G.5}$$

Let $(\alpha^*, \delta_p^*, \delta_r^*)$ minimizes the second term (hence $\alpha^* \le d+\epsilon$), and let $\alpha' = d\alpha^*/(d+\epsilon)$ which implies $\alpha' \le d$. Thus, $(\alpha', \delta_p^*, \delta_r^*)$ is a feasible choice for the first term. Hence, (G.5) is upper bounded by

$$
\begin{aligned}
& q\left(s, \frac{d-\alpha'}{\beta}, a, p_s^0 + \alpha'\delta_p^*, r_s^0 + \alpha'\delta_r^*, v\right) - q\left(s, \frac{d+\epsilon-\alpha^*}{\beta}, a, p_s^0 + \alpha^*\delta_p^*, r_s^0 + \alpha^*\delta_r^*, v\right) \\
=& \gamma \sum_{s'} [p^0(s'|s,a) + \alpha'\delta_p^*(s'|s,a)] v(s', \frac{d-\alpha'}{\beta}) + [r^0(s,a) + \alpha'\delta_r^*(s,a)] \\
& \quad - \gamma \sum_{s'} [p^0(s'|s,a) + \alpha^*\delta_p^*(s'|s,a)] v(s', \frac{d+\epsilon-\alpha^*}{\beta}) - [r^0(s,a) + \alpha^*\delta_r^*(s,a)] \\
=& \gamma \sum_{s'} [p^0(s'|s,a) + \alpha'\delta_p^*(s'|s,a)] \Big[v(s', \frac{d-\alpha'}{\beta}) - v(s', \frac{d+\epsilon-\alpha^*}{\beta}) \Big] \\
& \quad - (\alpha^* - \alpha') \left\{ \gamma \sum_{s'} \delta_p^*(s'|s,a) v(s', \frac{d+\epsilon-\alpha^*}{\beta}) + \delta_r^*(s,a) \right\}.
\end{aligned}
$$

We analyze the two terms separately.

$$
\begin{aligned}
& \gamma \sum_{s'} [p^0(s'|s,a) + \alpha'\delta_p^*(s'|s,a)] \Big[v(s', \frac{d-\alpha'}{\beta}) - v(s', \frac{d+\epsilon-\alpha^*}{\beta}) \Big] \\
\le& \gamma L \left| \frac{\epsilon + \alpha' - \alpha^*}{\beta} \right| \\
=& \frac{\gamma}{\beta} L\epsilon - \frac{\gamma}{\beta} L(\alpha^* - \alpha').
\end{aligned}
$$

Here the first inequality holds due to the fact that $p^0(\cdot|s,a) + \alpha'\delta_p^*(\cdot|s,a)$ defines a transition probability and hence is a vector on the probability simplex, i.e., $p^0(s'|s,a) + \alpha'\delta_p^*(s'|s,a) \ge 0$ and $\sum_{s'} p^0(s'|s,a) + \alpha'\delta_p^*(s'|s,a) = 1$, and the fact that $v(s,\cdot)$ is Liptsitz continuous with coefficient L. The last step holds because $\alpha^* - \alpha' = \epsilon\alpha'/(d+\epsilon) \le \epsilon$.

The second term is bounded as follows,

$$
\left| (\alpha^* - \alpha') \left\{ \gamma \sum_{s'} \delta_p^*(s'|s,a) v(s', \frac{d+\epsilon-\alpha^*}{\beta}) + \delta_r^*(s,a) \right\} \right| \le |\alpha^* - \alpha'| [\gamma V + M].
$$

Here, we use the fact that both $p^0(\cdot|s,a)$ and $p^0(\cdot|s,a) + \delta_p^*(\cdot|s,a)$ are transition probabilities, and hence

$$
\begin{aligned}
& \left| \sum_{s'} [\delta_p^*(s'|s,a) v(s', \frac{d+\epsilon-\alpha^*}{\beta})] \right| \\
= & \left| \sum_{s'} \{ [p^0(s'|s,a) + \delta_p^*(s'|s,a)] v(s', \frac{d+\epsilon-\alpha^*}{\beta}) \} - \sum_{s'} \{ p^0(s'|s,a) v(s', \frac{d+\epsilon-\alpha^*}{\beta}) \} \right| \\
\le & \sup_{s,d,s',d'} [v(s,d) - v(s',d')] = V.
\end{aligned}
$$

Combining the two terms we have

$$
\begin{aligned}
v'(s,d) - v'(s,d+\epsilon) &\le \frac{\gamma}{\beta} L\epsilon - \frac{\gamma}{\beta} L(\alpha^* - \alpha') + (\alpha^* - \alpha') \frac{M}{1-\gamma} \\
&= \frac{\gamma}{\beta} L\epsilon + (\alpha^* - \alpha')[M + \gamma V - \frac{\gamma}{\beta} L].
\end{aligned}
$$

Note that $\gamma \le \beta$ and $L = M + \gamma V$ implies that $[M + \gamma V - \frac{\gamma}{\beta} L] \ge 0$, and hence the right hand side is upper-bounded by

$$
\frac{\gamma}{\beta} L\epsilon + (\epsilon)[M + \gamma V - \frac{\gamma}{\beta} L] = (M + \gamma V)\epsilon,
$$

because $a^* - a' \le \epsilon$. This complete the proof. □

Now we turn to prove Theorem G.2.

Proof. Define the following operator Γ

$$\Gamma(v)(s,d) = \max_{a\in\mathcal{A}} \left\{ \min_{\alpha\in[0,1],\alpha\le d,(\delta_p,\delta_r)\in\Delta\mathcal{U}_s} q\left(s, \frac{d-\alpha}{\beta}, a, p_s^0+\alpha\delta_p, r_s^0+\alpha\delta_r, v\right)\right\}.$$

Note that $v^* = \lim_{n\to\infty}\Gamma^n(v)$ for any initial v. Further note that for any non-negative V_0, M_0, if $V_0 \le M_0/(1-\gamma)$, then $\gamma V_0 + M_0 \le M_0/(1-\gamma)$. Applying Lemma G.1 to the infinite sequence generated by applying Γ repetitively, starting from $v\equiv 0$, establishes the theorem. □

Combining Theorem G.1 and G.2, we have,

$$v^*(s,d) - \left[\left\lceil\frac{\log(1/\kappa)}{\log(1/\gamma)}\right\rceil\beta+1\right]\frac{M\kappa}{1-\gamma} \le \overline{v}(s,d) \le v^*(s,d).$$

H. Proof of Theorem 6

Proof. Let $X_t = b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t})$ and $Y_t = X_t - \alpha_t$, which is thus a zero-mean random variable satisfying $|Y_t| \le 1$. Bernstein's inequality gives that for any $c > 0$,

$$\Pr\left\{\sum_{t=1}^T Y_t \ge c\right\} \le \exp\left\{-\frac{c^2/2}{\sum_{t=1}^T \mathbb{E}Y_t^2 + c/3}\right\}.$$

Since $0 \le b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) \le 1$ we have that $\mathbb{E}X_t^2 \le \mathbb{E}X_t = \alpha_t$, thus $\mathbb{E}Y_t^2 \le \alpha_t(1-\alpha_t) \le \alpha_t$. From here the proof follows exactly the end of the proof to Theorem 1 in Appendix A, giving:

$$\Pr\left\{\sum_{t=1}^T b(p_t, r_t, p_t^0, r_t^0, \mathcal{U}_{s_t}) \ge \sum_{t=1}^T \alpha_t + \frac{1}{3}\log(1/\delta)\left(1+\sqrt{1+18\frac{\sum_{t=1}^T\alpha_t}{\log(1/\delta)}}\right)\right\} \le \delta$$

. □

References

Bennett, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Grötschel, M., Lovasz, L., and Schrijver, A. *The Ellipsoid Method and Combinatorial Optimization*. Springer, Heidelberg, 1988.

Puterman, Martin L. *Markov Decision Processes*. John Wiley & Sons, New York, 1994.