A Technical Results and Proofs

A.1 Proofs of Results in Section 3

Proof of Theorem 3.1. The result can be derived as:

$$\begin{aligned}\mathbb{P}_{\tilde{P}\sim f}\left[\hat{\rho}\le\rho(\hat{\pi},\tilde{P})\right] &\overset{\text{(a)}}{=} \mathbb{P}_{\tilde{P}\sim f}\left[\rho(\hat{\pi},\tilde{P})\ge\max_{\pi\in\Pi}\min_{P\in\hat{\mathcal{P}}}\rho(\pi,P)\right]\\ &\overset{\text{(b)}}{=} \mathbb{P}_{\tilde{P}\sim f}\left[\rho(\hat{\pi},\tilde{P})\ge\min_{P\in\hat{\mathcal{P}}}\rho(\hat{\pi},P)\right]\\ &\overset{\text{(c)}}{\ge} \mathbb{P}_{\tilde{P}\sim f}\left[\tilde{P}\in\hat{\mathcal{P}}\right]\overset{\text{(d)}}{\ge} 1-\delta\ .\end{aligned}$$

The equality (a) follows from the definition of $\hat{\rho}$, the inequality (b) follows from $\hat{\pi}\in\Pi$ and is optimal, (c) follows because $\rho(\hat{\pi},\tilde{P})\ge\min_{P\in\hat{\mathcal{P}}}\rho(\hat{\pi},P)$ whenever $\tilde{P}\in\hat{\mathcal{P}}$, and (d) follows from the theorem's hypothesis. □

Proof of Theorem 3.2. Let $\hat{\mathcal{P}}=\mathcal{P}(\boldsymbol{w},\psi)$ and let $\hat{\rho}$ and $\hat{\pi}$ be the optimal return and policy for $\hat{\mathcal{P}}$ respectively. We start by establishing the following bound:

$$\hat{\rho}\ge\max_{\pi\in\Pi}\rho(\pi,\tilde{P})-\frac{\beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)}{1-\gamma}\ ,$$

where

$$\beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)=\max_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}\beta_{\hat{\boldsymbol{z}}}^{s,a}(\boldsymbol{w},\psi)\ .$$

Let $\hat{\boldsymbol{v}}\in\mathbb{R}^S$ be the optimal robust value function that satisfied $\hat{\boldsymbol{v}}=\mathfrak{L}\hat{\boldsymbol{v}}$ for the ambiguity set $\hat{\mathcal{P}}=\mathcal{P}(\boldsymbol{w},\psi)$. We use $\hat{\mathcal{P}}$ as a shorthand for $\mathcal{P}(\boldsymbol{w},\psi)$ throughout the proof. Recall that $\hat{\rho}=\boldsymbol{p}_0^\top\hat{\boldsymbol{v}}$. We also use $\mathfrak{T}_\pi^P$ to represent the Bellman evaluation operator for a policy $\pi\in\Pi$ and a transition function P defined for each $s\in\mathcal{S}$ as:

$$(\mathfrak{T}_\pi^P\, v)_s=P(s,\pi(s))^\top(\boldsymbol{r}_{s,a}+\gamma\cdot\boldsymbol{v})\ .$$

It is well known that $\mathfrak{T}_\pi^P\, v$ is a contraction, is monotone, and has a unique fixed point. Let $\tilde{v}$ be the unique fixed point of $\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}$:

$$\tilde{\boldsymbol{v}}=\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\tilde{\boldsymbol{v}}\ ,$$

where $\tilde{\pi}\in\arg\max_{\pi\in\Pi}\rho(\pi,\tilde{P})$. Note that it is well known that:

$$\boldsymbol{p}_0^\top\tilde{\boldsymbol{v}}=\rho(\tilde{\pi},\tilde{P})\ .$$

Now suppose that $\tilde{P}\in\hat{\mathcal{P}}$, which holds with probability $1-\delta$ according to Assumption 1. Then it is easy to see that:

$$\boldsymbol{p}_0^\top\hat{\boldsymbol{v}}=\min_{P\in\hat{\mathcal{P}}}\rho(\pi,P)\le\rho(\pi,\tilde{P})\le\boldsymbol{p}_0^\top\tilde{\boldsymbol{v}}\ .$$

Therefore:

$$0\le\boldsymbol{p}_0^\top\tilde{\boldsymbol{v}}-\boldsymbol{p}_0^\top\hat{\boldsymbol{v}}\le\left\|\tilde{\boldsymbol{v}}-\hat{\boldsymbol{v}}\right\|_\infty\ .$$

We are now ready to establish the probabilistic bound which is based on bounding the Bellman residual as follows:

$$\begin{aligned}(\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}}-\hat{\boldsymbol{v}})_s &\overset{\text{(a)}}{=} (\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}}-\mathfrak{L}\hat{\boldsymbol{v}})_s \overset{\text{(def)}}{=} \tilde{P}(s,\tilde{\pi}(a))^\top\hat{\boldsymbol{z}}_{s,\tilde{\pi}(s)}-\min_{P\in\hat{\mathcal{P}}}P(s,\hat{\pi}(a))^\top\hat{\boldsymbol{z}}_{s,\hat{\pi}(a)}\\ &\overset{\text{(b)}}{\le} \tilde{P}(s,\tilde{\pi}(a))^\top\hat{\boldsymbol{z}}_{s,\tilde{\pi}(s)}-\min_{P\in\hat{\mathcal{P}}}P(s,\tilde{\pi}(a))^\top\hat{\boldsymbol{z}}_{s,\tilde{\pi}(a)}\\ &\le\max_{a\in\mathcal{A}}\left(\tilde{P}(s,a)^\top\hat{\boldsymbol{z}}_{s,a}-\min_{P\in\hat{\mathcal{P}}}P(s,a)^\top\hat{\boldsymbol{z}}_{s,a}\right)\\ &\overset{\text{(c)}}{\le}\max_{a\in\mathcal{A}}\left(\max_{P\in\hat{\mathcal{P}}}P(s,a)^\top\hat{\boldsymbol{z}}_{s,a}-\min_{P\in\hat{\mathcal{P}}}P(s,a)^\top\hat{\boldsymbol{z}}_{s,a}\right)\\ &\le\max_{a\in\mathcal{A}}\beta_{\hat{\boldsymbol{z}}}^{s,a}(\boldsymbol{w},\psi)\ .\end{aligned}$$

(a) follows from $\hat{\boldsymbol{v}}$ being the fixed point of $\mathfrak{L}$, (b) follows from the optimality of $\hat{\pi}$: $\hat{\pi}(s) \in \arg\max_{a\in\mathcal{A}} \min_{\boldsymbol{p}\in\hat{\mathcal{P}}_{s,a}} \boldsymbol{p}^{\mathsf{T}}\boldsymbol{z}_{s,a}$, and (c) follows from $\tilde{P} \in \hat{\mathcal{P}}$. The rest follows by algebraic manipulation. Applying the inequality above to all states, we get:

$$\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}} - \hat{\boldsymbol{v}} \le \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)\cdot\mathbf{1}\ . \tag{12}$$

We can now use the standard dynamic programming bounding technique to bound $\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty$ as follows:

$$\mathbf{0} \overset{(a)}{\le} \tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}} \overset{(b)}{=} \tilde{\boldsymbol{v}} - \mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}} + \mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}} - \hat{\boldsymbol{v}} \overset{(12)}{\le} \tilde{\boldsymbol{v}} - \mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}} + \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)\cdot\mathbf{1} \overset{(c)}{\le} \mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\tilde{\boldsymbol{v}} - \mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}} + \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)\cdot\mathbf{1}\ .$$

We have (a) because $\hat{\boldsymbol{v}} \le \tilde{\boldsymbol{v}}$ because $\mathfrak{L}\tilde{\boldsymbol{v}} \le \tilde{\boldsymbol{v}}$ and thus $\tilde{\boldsymbol{v}} \ge \mathfrak{L}\mathfrak{L}\tilde{\boldsymbol{v}} \ge \ldots \ge \mathfrak{L}\ldots\mathfrak{L}\tilde{\boldsymbol{v}} \ge \hat{\boldsymbol{v}}$ because $\hat{\boldsymbol{v}}$ is the fixed point of $\mathfrak{L}$ and $\mathfrak{L}$ is monotone. (b) we add $\mathbf{0}$, (c) $\tilde{\boldsymbol{v}}$ is the fixed point of $\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}$.

Next, apply L_∞ norm to all sides, which is possible because the values are non-negative:

$$\begin{aligned}
\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty &\le \left\|\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\tilde{\boldsymbol{v}} - \mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}\hat{\boldsymbol{v}} + \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)\cdot\mathbf{1}\right\|_\infty \\
\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty &\le \gamma\cdot\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty + \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi) \\
\|\tilde{\boldsymbol{v}} - \hat{\boldsymbol{v}}\|_\infty &\le \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)/(1-\gamma)\ .
\end{aligned}$$

The first step follows by triangle inequality, and the second step follows from $\mathfrak{T}_{\tilde{\pi}}^{\tilde{P}}$ being a γ contraction in the L_∞ norm.

To prove the bound on $y^\star$ and $\hat{v}$, we show that $y^\star \le \zeta$ where $\zeta = \hat{\rho} + \beta_{\hat{\boldsymbol{z}}}(\boldsymbol{w},\psi)/(1-\gamma)$. Suppose to the contrary that $y^\star > \zeta$. Realize that $y^\star$ optimal in (1) must satisfy:

$$\mathbb{P}_{\tilde{P}\sim f}\left[\max_{\pi\in\Pi}\rho(\pi,\tilde{P}) \ge y^\star\right] \ge 1-\delta\ , \tag{13}$$

because $\max_{\pi\in\Pi}\rho(\pi,\tilde{P}) \ge \rho(\pi^\star,\tilde{P})$ for $\pi^\star$ optimal in (1). Recall also that from the first part of the theorem:

$$\mathbb{P}_{\tilde{P}\sim f}\left[\max_{\pi\in\Pi}\rho(\pi,\tilde{P}) \ge \zeta\right] \le \delta\ . \tag{14}$$

We now derive a contradiction as follows:

$$\delta \overset{(14)}{\ge} \mathbb{P}_{\tilde{P}\sim f}\left[\max_{\pi\in\Pi}\rho(\pi,\tilde{P}) \ge \zeta\right] \overset{(a)}{\ge} \mathbb{P}_{\tilde{P}\sim f}\left[\max_{\pi\in\Pi}\rho(\pi,\tilde{P}) \ge y^\star\right] \overset{(13)}{\ge} 1-\delta\ .$$

Here (a) follows from the assumption $y^\star > \zeta$. Then $\delta \ge 1-\delta$ is a contradiction with $\delta < 0.5$. Finally, $0 \le y^\star - \hat{\rho}$ follows directly from the optimality of $y^\star$ and Theorem 3.1, which proves the theorem. □

A.2 Proof of Results in Section 4

Proof of Lemma 4.1. We omit the s,a subscripts to simplify the notation. By relaxing the non-negativity constraints on $\boldsymbol{p}$ and using substitution $\boldsymbol{q}_1 = \boldsymbol{p}_1 - \bar{\boldsymbol{p}}$ and $\boldsymbol{q}_2 = \boldsymbol{p}_2 - \bar{\boldsymbol{p}}$, we get the following upper bound:

$$\begin{aligned}
\beta_{\boldsymbol{z}}^{s,a}(\boldsymbol{w},\psi) &= \max_{\boldsymbol{p}_1,\boldsymbol{p}_2}\left\{(\boldsymbol{p}_1-\boldsymbol{p}_2)^{\mathsf{T}}\boldsymbol{z} \mid \boldsymbol{p}_1,\boldsymbol{p}_2 \in \mathcal{P}_{s,a}(\boldsymbol{w},\psi)\right\} \\
&= \max_{\boldsymbol{p}_1,\boldsymbol{p}_2}\left\{(\boldsymbol{p}_1-\boldsymbol{p}_2)^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{p}_1-\bar{\boldsymbol{p}}\|_{\boldsymbol{w}} \le \psi,\ \|\boldsymbol{p}_2-\bar{\boldsymbol{p}}\|_{\boldsymbol{w}} \le \psi,\ \boldsymbol{p}_1 \in \Delta^S,\ \boldsymbol{p}_2 \in \Delta^S\right\} \\
&\le \max_{\boldsymbol{p}_1,\boldsymbol{p}_2\in\mathbb{R}^S}\left\{(\boldsymbol{p}_1-\boldsymbol{p}_2)^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{p}_1-\bar{\boldsymbol{p}}\|_{\boldsymbol{w}} \le \psi,\ \|\boldsymbol{p}_2-\bar{\boldsymbol{p}}\|_{\boldsymbol{w}} \le \psi,\ \mathbf{1}^{\mathsf{T}}\boldsymbol{p}_1 = 1,\ \mathbf{1}^{\mathsf{T}}\boldsymbol{p}_2 = 1\right\} \\
&= \max_{\boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^S}\left\{(\boldsymbol{q}_1-\boldsymbol{q}_2)^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{q}_1\|_{\boldsymbol{w}} \le \psi,\ \|\boldsymbol{q}_2\|_{\boldsymbol{w}} \le \psi,\ \mathbf{1}^{\mathsf{T}}\boldsymbol{q}_1 = 0,\ \mathbf{1}^{\mathsf{T}}\boldsymbol{q}_2 = 0\right\} \\
&= \max_{\boldsymbol{q}_1\in\mathbb{R}^S}\left\{\boldsymbol{q}_1^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{q}_1\|_{\boldsymbol{w}} \le \psi,\ \mathbf{1}^{\mathsf{T}}\boldsymbol{q}_1 = 0\right\} + \max_{\boldsymbol{q}_2\in\mathbb{R}^S}\left\{\boldsymbol{q}_2^{\mathsf{T}}(-\boldsymbol{z}) \mid \|\boldsymbol{q}_2\|_{\boldsymbol{w}} \le \psi,\ \mathbf{1}^{\mathsf{T}}\boldsymbol{q}_2 = 0\right\}\ .
\end{aligned}$$

The last equality follows because the the optimization problems over $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are independent. From the absolute homogeneity of the $\|\cdot\|_{\boldsymbol{w}}$ we have that:

$$\max_{\boldsymbol{q}_2 \in \mathbb{R}^S} \left\{ \boldsymbol{q}_2^{\mathsf{T}}(-\boldsymbol{z}) \mid \|\boldsymbol{q}_2\|_{\boldsymbol{w}} \le \psi,\, \mathbf{1}^{\mathsf{T}}\boldsymbol{q}_2 = 0 \right\} = \max_{\boldsymbol{q}_2 \in \mathbb{R}^S} \left\{ \boldsymbol{q}_2^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{q}_2\|_{\boldsymbol{w}} \le \psi,\, \mathbf{1}^{\mathsf{T}}\boldsymbol{q}_2 = 0 \right\},$$

and therefore:

$$\beta_{\boldsymbol{z}}^{s,a}(\boldsymbol{w},\psi) \le 2 \cdot \max_{\boldsymbol{q} \in \mathbb{R}^S} \left\{ \boldsymbol{q}^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{q}\|_{\boldsymbol{w}} \le \psi,\, \mathbf{1}^{\mathsf{T}}\boldsymbol{q} = 0 \right\}.$$

Substituting $\boldsymbol{q} = \boldsymbol{p} - \bar{\boldsymbol{p}}$ we get:

$$\beta_{\boldsymbol{z}}^{s,a}(\boldsymbol{w},\psi) \le 2 \cdot \max_{\boldsymbol{p} \in \mathbb{R}^S} \left\{ \boldsymbol{p}^{\mathsf{T}}\boldsymbol{z} \mid \|\boldsymbol{p} - \bar{\boldsymbol{p}}\|_{\boldsymbol{w}} \le \psi,\, \mathbf{1}^{\mathsf{T}}\boldsymbol{p} = 1 \right\} - 2 \cdot \boldsymbol{z}^{\mathsf{T}}\bar{\boldsymbol{p}}. \tag{15}$$

We can reformulate the optimization problem on the right-hand side of (15), again using variable substitution $\boldsymbol{q} = \boldsymbol{p} - \bar{\boldsymbol{p}}$:

$$\begin{aligned} \max_{\boldsymbol{q} \in \mathbb{R}^S} \quad & 2 \cdot (\boldsymbol{q} + \bar{\boldsymbol{p}})^{\mathsf{T}}\boldsymbol{z} - 2 \cdot \boldsymbol{z}^{\mathsf{T}}\bar{\boldsymbol{p}} \\ \text{s.t.} \quad & \|\boldsymbol{q}\|_{\boldsymbol{w}} \le \psi \\ & \mathbf{1}^{\mathsf{T}}(\boldsymbol{q} + \bar{\boldsymbol{p}}) = 1 \implies \mathbf{1}^{\mathsf{T}}\boldsymbol{q} = 0. \end{aligned}$$

Canceling out $\bar{\boldsymbol{p}}^{\mathsf{T}}\boldsymbol{z}$, we continue with:

$$\begin{aligned} 2 \cdot \max_{\boldsymbol{q} \in \mathbb{R}^S} \quad & \boldsymbol{q}^{\mathsf{T}}\boldsymbol{z} \\ \text{s.t.} \quad & \|\boldsymbol{q}\|_{\boldsymbol{w}} \le \psi \\ & \mathbf{1}^{\mathsf{T}}\boldsymbol{q} = 0. \end{aligned}$$

By applying the method of Lagrange multipliers, we obtain:

$$\begin{aligned} \min_{\lambda \in \mathbb{R}} \max_{\boldsymbol{q} \in \mathbb{R}^S} \quad & \boldsymbol{q}^{\mathsf{T}}\boldsymbol{z} - \lambda \cdot (\boldsymbol{q}^{\mathsf{T}}\mathbf{1}) = \boldsymbol{q}^{\mathsf{T}}(\boldsymbol{z} - \lambda \cdot \mathbf{1}) \\ \text{s.t.} \quad & \|\boldsymbol{q}\|_{\boldsymbol{w}} \le \psi. \end{aligned}$$

Letting $\boldsymbol{x} = \frac{\boldsymbol{q}}{\psi}$, we get:

$$\begin{aligned} \min_{\lambda \in \mathbb{R}} \max_{\boldsymbol{x} \in \mathbb{R}^S} \quad & \psi \cdot \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{z} - \lambda \cdot \mathbf{1}) \\ \text{s.t.} \quad & \|\boldsymbol{x}\|_{\boldsymbol{w}} \le 1. \end{aligned}$$

Given the definition of the *dual norm*, $\|\boldsymbol{z}\|_\star = \sup\{\boldsymbol{z}^{\mathsf{T}}\boldsymbol{x} \mid \|\boldsymbol{x}\| \le 1\}$, we have:

$$\begin{aligned} \beta_{\boldsymbol{z}}^{s,a}(\boldsymbol{w},\psi) &\le 2 \cdot \min_{\lambda \in \mathbb{R}} \psi \cdot \|\boldsymbol{z} - \lambda \cdot \mathbf{1}\|_\star \\ &\le 2 \cdot \psi \cdot \|\boldsymbol{z} - \lambda \cdot \mathbf{1}\|_\star. \end{aligned}$$

□

Proof of Lemma 4.2. Assume we are given a set of positive weights $\boldsymbol{w} \in \mathbb{R}^n_{++}$ for the following weighted L_1 optimization problem:

$$\begin{aligned} \max_{\boldsymbol{x} \in \mathbb{R}^S} \quad & \boldsymbol{z}^{\mathsf{T}}\boldsymbol{x} \\ \text{s.t.} \quad & \|\boldsymbol{x}\|_{1,\boldsymbol{w}} \le 1. \end{aligned} \tag{16}$$

We have:

$$\begin{aligned} \boldsymbol{x}^{\mathsf{T}}\boldsymbol{z} &= \sum_{i=1}^n x_i \cdot z_i \le \sum_{i=1}^n |x_i \cdot z_i| \\ &\overset{(a)}{\le} \sum_{i=1}^n |x_i| \cdot |z_i| = \sum_{i=1}^n w_i \cdot |x_i| \cdot \frac{1}{w_i} \cdot |z_i| \\ &\le \max_{i=1,\dots,n} \left\{ \frac{1}{w_i} \cdot |z_i| \right\} \cdot \sum_{i=1}^n w_i |x_i| = \max_{i=1,\dots,n} \left\{ \frac{1}{w_i} \cdot |z_i| \right\} \cdot \|\boldsymbol{x}\|_{1,\boldsymbol{w}} \\ &\overset{(b)}{\le} \max_{i=1,\dots,n} \left\{ \frac{1}{w_i} |z_i| \right\} = \|\boldsymbol{z}\|_{\infty,\frac{1}{\boldsymbol{w}}}. \end{aligned}$$

Here, (a) follows from the Cauchy-Schwarz inequality, and (b) follows from the constraint $\|\boldsymbol{x}\|_{1,\boldsymbol{w}} \leq 1$ of (16). □

Proof of Proposition 4.3. We use the notation $1/\boldsymbol{w}$ to denote an elementwise inverse of $\boldsymbol{w}$ such that $(1/\boldsymbol{w})_i = 1/w_i, i \in \mathcal{S}$. Note that for weighted L_1-constrained sets $q = \infty$, and for the L_∞-constrained sets $q = 1$. The value $\bar{\lambda}$ in (7) is fixed ahead of time and does not change with $\boldsymbol{w}$. Recall that the constraint $\sum_{i=1}^{S} w_i^2 = 1$ serves to normalize $\boldsymbol{w}$ in order to preserve the desired robustness guarantees with *the same* ψ. This is because scaling both $\boldsymbol{w}$ and ψ simultaneously by an identical factor leaves the ambiguity set unchanged. We adopt the constraint from an approximation of the guarantee by linearization of the upper bound using Jensen's inequality. Next, omitting terms that are constant with respect to $\boldsymbol{w}$ simplifies the optimization to:

$$\boldsymbol{w}^\star \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^S_{++}} \left\{ \left\|\boldsymbol{z} - \bar{\lambda}\mathbf{1}\right\|_{q,\frac{1}{\boldsymbol{w}}} \;:\; \sum_{i=1}^{S} w_i^2 = 1 \right\} . \tag{17}$$

For $q = \infty$, the nonlinear optimization problem in (17) is convex and can be solved *analytically.* Let $b_i = \left|z_i - \bar{\lambda}\right|$ for $i = 1, \ldots, S$, then (17) turns to:

$$\min_{t,\boldsymbol{w}\in\mathbb{R}^S_{++}} \left\{ t \;:\; t \geq b_i/w_i,\; \sum_{i=1}^{S} w_i^2 = 1 \right\} . \tag{18}$$

The constraints $\boldsymbol{w} > \mathbf{0}$ cannot be active since otherwise $1/w_i$ results in undefined division by zero and can be safely ignored. Then, the convex optimization problem in Equation (18) has a linear objective, $S + 1$ variables ($\boldsymbol{w}$'s and t), and $S + 1$ constraints. All constraints are active, therefore, in the optimal solution $\boldsymbol{w}^\star$ (Bertsekas, 2003) which must satisfy:

$$w_i^\star = b_i \big/ \sqrt{\textstyle\sum_{j=1}^{S} b_j^2} . \tag{19}$$

Since $\sum_i w_i^2 = 1$ implies $\sum_i b_i^2/t^2 = 1$, we conclude that $t = \sqrt{\sum_i b_i^2}$. For $q = 1$, the equivalent optimization of (18) becomes:

$$\min_{\boldsymbol{w}>\mathbf{0}} \left\{ \sum_{i=1}^{S} b_i/w_i \;:\; \sum_{i=1}^{S} w_i^2 = 1 \right\} . \tag{20}$$

Again, the inequality constraints on weights $\boldsymbol{w} > 0$ can be relaxed. Using the necessary optimality conditions (and a Lagrange multiplier), one solution for the optimal weights $\boldsymbol{w}$ are:

$$w_i^\star = b_i^{1/3} \big/ \sqrt{\textstyle\sum_{j=1}^{S} b_j^{2/3}} . \tag{21}$$

□

A.3 Proof of Results in Section 5

Proof of Proposition 5.2. The algorithm is an instance of the Sample Average Approximation (SAA) scheme. The result, therefore, is a direct consequence of Theorem 4.2 in (Petrik and Russel, 2019) and Theorem 5.3 in (Shapiro et al., 2014). □

A.4 Proof of Results in Section 6

We need several auxiliary results before proving the results.

Theorem A.1 (Weighted L_∞ error bound (Hoeffding))**.** *Suppose that $\bar{\boldsymbol{p}}_{s,a}$ is the empirical estimate of the transition probability obtained from $n_{s,a}$ samples for some $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then:*

$$\mathbb{P}_{\bar{\boldsymbol{p}}_{s,a}} \left[\left\|\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}^\star_{s,a}\right\|_{\infty,\boldsymbol{w}} \geq \psi_{s,a} \right] \leq 2 \sum_{i=1}^{S} \exp\left(-2 \frac{\psi_{s,a}^2 n_{s,a}}{w_i^2} \right) . \tag{22}$$

Proof. First, we will express the weighted L_∞ distance between two distributions $\bar{\boldsymbol{p}}$ and $\boldsymbol{p}^\star$ in terms of an optimization problem. Let $\mathbf{1}_i \in \mathbb{R}^{\mathcal{S}}$ be the indicator vector for an index $i \in \mathcal{S}$:

$$\begin{aligned} \left\|\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}^\star_{s,a}\right\|_{\infty,\boldsymbol{w}} &= \max_{\boldsymbol{z}} \left\{ \boldsymbol{z}^\mathsf{T} W (\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}^\star_{s,a}) : \|\boldsymbol{z}\|_1 \leq 1 \right\} \\ &= \max_{i\in\mathcal{S}} \Big\{ \mathbf{1}_i W (\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}^\star_{s,a}), -\mathbf{1}_i W (\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}^\star_{s,a}) \Big\} . \end{aligned}$$

Here, weights are on the diagonal entries of W. Using the expression above, we can bound the probability in the lemma as follows:

$$\begin{aligned}\mathbb{P}\left[\|\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a}\|_{\infty,\boldsymbol{w}}\ge\psi\right]&=\mathbb{P}\left[\max_{i\in\mathcal{S}}\left\{\mathbf{1}_iW(\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a}),-\mathbf{1}_iW(\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a})\right\}\ge\psi_{s,a}\right]\\&\overset{(a)}{\le}S\max_{i\in\mathcal{S}}\mathbb{P}\left[\mathbf{1}_iW(\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a})\ge\psi_{s,a}\right]+S\max_{i\in\mathcal{S}}\mathbb{P}\left[-\mathbf{1}_iW(\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a})\ge\psi_{s,a}\right]\\&\overset{(b)}{\le}2\sum_{i=1}^{S}\exp\left(-2\frac{\psi_{s,a}^2n}{w_i^2}\right)\ .\end{aligned}$$

Here, (a) follows from union bound, and (b) follows from Hoeffding's inequality since $\mathbf{1}_i^\mathsf{T}\bar{\boldsymbol{p}}\in[0,1]$ for any $i\in\mathcal{S}$ and its mean is $\mathbf{1}_i^\mathsf{T}\boldsymbol{p}^\star$. □

Now we describe a proof of error bound in (23) on the weighted L_1 distance between the estimated transition probabilities $\bar{\boldsymbol{p}}$ and the true one $\boldsymbol{p}^\star$ over each state $s\in\mathcal{S}=\{1,\ldots,S\}$ and action $a\in\mathcal{A}=\{1,\ldots,A\}$. The proof is an extension to Lemma C.1 (L_1 error bound) in (Petrik and Russel, 2019).

Theorem A.2 (Weighted L_1 error bound (Hoeffding))**.** *Suppose that $\bar{\boldsymbol{p}}_{s,a}$ is the empirical estimate of the transition probability obtained from $n_{s,a}$ samples for some $s\in\mathcal{S}$ and $a\in\mathcal{A}$. If the weights $\boldsymbol{w}\in\mathbb{R}^S_{++}$ are sorted in a non-increasing order $w_i\ge w_{i+1}$, then:*

$$\mathbb{P}_{\bar{\boldsymbol{p}}_{s,a}}\left[\left\|\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a}\right\|_{1,\boldsymbol{w}}\ge\psi_{s,a}\right]\le2\sum_{i=1}^{S-1}2^{S-i}\exp\left(-\frac{\psi_{s,a}^2n_{s,a}}{2w_i^2}\right)\ . \tag{23}$$

Proof. Let $\boldsymbol{q}_{s,a}=\bar{\boldsymbol{p}}_{s,a}-\boldsymbol{p}^\star_{s,a}$. To shorten notation in the proof, we omit the s,a indexes when there is no ambiguity. We assume that all weights are non-negative. First, we will express the $L_{1,\boldsymbol{w}}$ norm of $\boldsymbol{q}$ in terms of an optimization problem. It is worth noting that $\mathbf{1}^\mathsf{T}\boldsymbol{q}=0$. Let $\mathbf{1}_{\mathcal{Q}_1},\mathbf{1}_{\mathcal{Q}_2}\in\mathbb{R}^{\mathcal{S}}$ be the indicator vectors for some subsets $\mathcal{Q}_1,\mathcal{Q}_2\subset\mathcal{S}$ where $\mathcal{Q}_2=\mathcal{S}\setminus\mathcal{Q}_1$. According to Lemma 4.2 we have:

$$\begin{aligned}\|\boldsymbol{q}\|_{1,w}&=\max_{\boldsymbol{z}}\left\{\boldsymbol{z}^\mathsf{T}\boldsymbol{q}:\|\boldsymbol{z}\|_{\infty,\frac{1}{w}}\le1\right\}\\&=\max_{\mathcal{Q}_1,\mathcal{Q}_2\in2^{\mathcal{S}}}\left\{\mathbf{1}_{\mathcal{Q}_1}^\mathsf{T}W\boldsymbol{q}+\mathbf{1}_{\mathcal{Q}_2}^\mathsf{T}W(-\boldsymbol{q}):\mathcal{Q}_2=\mathcal{S}\setminus\mathcal{Q}_1\right\}\ .\end{aligned}$$

Here weights are on the diagonal entries of W. Using the expression above, we can bound the probability as follows:

$$\begin{aligned}\mathbb{P}\left[\max_{\mathcal{Q}_1,\mathcal{Q}_2\in2^{\mathcal{S}}}\left\{\mathbf{1}_{\mathcal{Q}_1}^\mathsf{T}W\boldsymbol{q}+\mathbf{1}_{\mathcal{Q}_2}^\mathsf{T}W(-\boldsymbol{q})\right\}\ge\psi\right]&\overset{(a)}{\le}\mathbb{P}\left[\max_{\mathcal{Q}_1\in2^{\mathcal{S}}}\left\{\mathbf{1}_{\mathcal{Q}_1}^\mathsf{T}W\boldsymbol{q}\right\}\ge\frac{\psi}{2}\right]+\mathbb{P}\left[\max_{\mathcal{Q}_2\in2^{\mathcal{S}}}\left\{\mathbf{1}_{\mathcal{Q}_2}^\mathsf{T}W(-\boldsymbol{q})\right\}\ge\frac{\psi}{2}\right]\\&\le\sum_{\mathcal{Q}_1\in2^{\mathcal{S}}}\mathbb{P}\left[\mathbf{1}_{\mathcal{Q}_1}^\mathsf{T}W\boldsymbol{q}\ge\frac{\psi}{2}\right]+\sum_{\mathcal{Q}_2\in2^{\mathcal{S}}}\mathbb{P}\left[\mathbf{1}_{\mathcal{Q}_2}^\mathsf{T}W(-\boldsymbol{q})\ge\frac{\psi}{2}\right]\\&=\sum_{\mathcal{Q}_1\in2^{\mathcal{S}}}\mathbb{P}\left[\mathbf{1}_{\mathcal{Q}_1}^\mathsf{T}W(\bar{\boldsymbol{p}}-\boldsymbol{p}^\star)\ge\frac{\psi}{2}\right]+\sum_{\mathcal{Q}_2\in2^{\mathcal{S}}}\mathbb{P}\left[\mathbf{1}_{\mathcal{Q}_2}^\mathsf{T}W(-\bar{\boldsymbol{p}}+\boldsymbol{p}^\star)\ge\frac{\psi}{2}\right]\\&\overset{(b)}{\le}\sum_{\mathcal{Q}_1\in2^{\mathcal{S}}}\exp\left(-\frac{\psi^2n}{2\|\mathbf{1}_{\mathcal{Q}_1}^\mathsf{T}W\|_\infty^2}\right)+\sum_{\mathcal{Q}_2\in2^{\mathcal{S}}}\exp\left(-\frac{\psi^2n}{2\|\mathbf{1}_{\mathcal{Q}_2}^\mathsf{T}W\|_\infty^2}\right)\\&\overset{(c)}{=}2\sum_{i=1}^{S-1}2^{S-i}\exp\left(-\frac{\psi^2n}{2w_i^2}\right)\ .\end{aligned}$$

(a) follows from union bound, and (b) follows from Hoeffding's inequality. (c) follows by $\mathcal{Q}_1^c=\mathcal{Q}_2$ and sorting weights $\boldsymbol{w}=\{w_1,\ldots,w_n\}$ in non-increasing order. □

Proof of Theorem 6.1. The result follows from Lemma A.1 in (Petrik and Russel, 2019) and Theorem A.1 by algebraic manipulation. □

Proof of Theorem 6.2. The result follows from Lemma A.1 in (Petrik and Russel, 2019) and Theorem A.2 by algebraic manipulation. □

A.5 Bernstein Concentration Inequalities

Theorem A.3 (Weighted L_1 error bound (Bernstein))**.** *Suppose that $\bar{\boldsymbol{p}}_{s,a}$ is the empirical estimate of the transition probability obtained from $n_{s,a}$ samples for some $s \in \mathcal{S}$ and $a \in \mathcal{A}$. If the weights $\boldsymbol{w} \in \mathbb{R}^S_{++}$ are sorted in non-increasing order $w_i \geq w_{i+1}$, then the following holds when using Bernstein's inequality:*

$$\mathbb{P}\left[\left\|\bar{\boldsymbol{p}}_{s,a} - \boldsymbol{p}^\star_{s,a}\right\|_{1,\boldsymbol{w}} \geq \psi_{s,a}\right] \leq 2 \sum_{i=1}^{S-1} 2^{S-i} \exp\left(-\frac{3\psi^2 n}{6w_i^2 + 4\psi w_i}\right)$$

where $\boldsymbol{w} \in \mathbb{R}^S_{++}$ is the vector of weights. The weights are sorted in non-increasing order.

Proof. The proof is similar to the proof of Theorem A.2 until section b. The proof continues from section (b) as follows:

$$\begin{aligned}
&\overset{(b)}{\leq} \sum_{\mathcal{Q}_1 \in 2^{\mathcal{S}}} \exp\left(-\frac{3\psi^2 n}{24\sigma^2 + 4c\psi}\right) + \sum_{\mathcal{Q}_2 \in 2^{\mathcal{S}}} \exp\left(-\frac{3\psi^2 n}{24\sigma^2 + 4c\psi}\right) \\
&\overset{(c)}{\leq} \sum_{\mathcal{Q}_1 \in 2^{\mathcal{S}}} \exp\left(-\frac{3\psi^2 n}{6\left\|\mathbf{1}^\mathsf{T}_{\mathcal{Q}_1} W\right\|^2_\infty + 4\psi\left\|\mathbf{1}^\mathsf{T}_{\mathcal{Q}_1} W\right\|_\infty}\right) + \sum_{\mathcal{Q}_2 \in 2^{\mathcal{S}}} \exp\left(-\frac{3\psi^2 n}{6\left\|\mathbf{1}^\mathsf{T}_{\mathcal{Q}_2} W\right\|^2_\infty + 4\psi\left\|\mathbf{1}^\mathsf{T}_{\mathcal{Q}_2} W\right\|_\infty}\right) \\
&\overset{(d)}{=} 2 \sum_{i=1}^{S-1} 2^{S-i} \exp\left(-\frac{3\psi^2 n}{6w_i^2 + 4\psi w_i}\right) .
\end{aligned}$$

Here (b) follows from Bernstein's inequality where σ^2 is the mean of variance of random variables, and c is their upper bound (Devroye et al., 2013). In the weighted case, with conservative estimate of variance $\sigma^2 = \left\|\mathbf{1}^\mathsf{T}_{\mathcal{Q}_1} W\right\|^2_\infty/4$, and $c = \left\|\mathbf{1}^\mathsf{T}_{\mathcal{Q}_1} W\right\|_\infty$, because the random variables are drawn from *Bernoulli* distribution with the maximum possible variance of 1/4. (d) follows by sorting weights $\boldsymbol{w}$ in non-increasing order. □

B Detailed Experimental Results

B.1 Experimental Setup

We assess L_1- and L_∞-bounded ambiguity sets, both with weights and without weights. We compare Bayesian credible regions with frequentist Hoeffding- and Bernstein-style sets. We start by assuming a true underlying model that produces simulated datasets containing 20 samples for each state and action. The frequentist methods construct ambiguity sets directly from the datasets. Bayesian methods combine the data with a prior to compute a posterior distribution and then draw 20 samples from the posterior distribution to construct a Bayesian ambiguity set.

B.2 RiverSwim MDP Graph

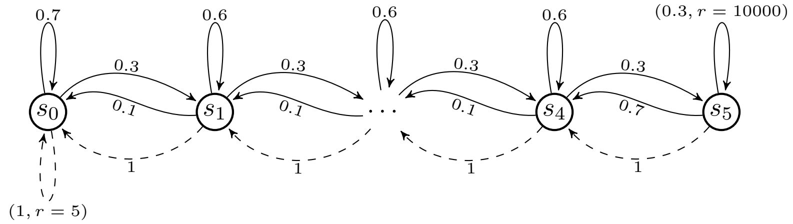

Figure 2: RiverSwim problem with six states and two actions (left-dashed arrow, right-solid arrow). The agent starts in either states s_1 or s_2.

B.3 Full Empirical Results

Tables 3 to 6 report the high-confidence lower bound on the return for the domains that we investigate. The column denotes the confidence $1 - \delta$ and the algorithm used to compute the weights $\boldsymbol{w}$ for the ambiguity set: "Unif.w" corresponds to $\boldsymbol{w} = \mathbf{1}$, "Analyt.w" corresponds to weights computed by Algorithm 2, and "SOCP.w" corresponds to weights computed by solving (8). The rows indicate which norm was used to define the ambiguity set (L_1 or L_∞) and whether Bayesian (B) or frequentist (H) guarantees were used. Note that the SOCP formulation is limited to the L_1 ambiguity sets.

	$\delta = 0.5$			$\delta = 0.05$		
Method	Unif.w	Analyt.w	SOCP.w	Unif.w	Analyt.w	SOCP.w
$L_1 B$	33887	**51470**	48620	25252	**47284**	43504
$L_\infty B$	33887	**48258**	-	25252	**43247**	-
L_1 H	16354	**33116**	30268	12555	**29472**	26398
L_∞ H	20055	**40166**	-	15184	**35955**	-

Table 3: The return with performance guarantees for the RiverSwim experiment. The return of the nominal MDP is 63080.

	$\delta = 0.5$			$\delta = 0.05$		
Method	Unif.w	Analyt.w	SOCP.w	Unif.w	Analyt.w	SOCP.w
$L_1 B$	-38.1	**-22.7**	-26.8	-42.0	**-23.7**	-28.4
$L_\infty B$	-38.1	**-22.6**	-	-42.0	**-23.5**	-
L_1 H	-86.8	**-33.2**	-47.9	-115.0	**-34.5**	-53.1
L_∞ H	-62.9	**-29.5**	-	-74.8	**-32.6**	-

Table 4: The return with performance guarantees for the Machine Replacement experiment. The return of the nominal MDP is -16.79.

	$\delta = 0.5$			$\delta = 0.05$		
Method	Unif.w	Analyt.w	SOCP.w	Unif.w	Analyt.w	SOCP.w
$L_1 B$	-25706	**-12151**	-12668	-25741	**-12200**	-12704
$L_\infty B$	-26782	**-15468**	-	-26795	**-15623**	-
L_1 H	-27499	**-27034**	-27409	-27501	**-27047**	-27421
L_∞ H	-27465	**-27143**	-	-27473	**-27184**	-

Table 5: The return with performance guarantees for the Population experiment. The return of the nominal MDP is -4127.

	$\delta = 0.5$			$\delta = 0.05$		
Method	Unif.w	Analyt.w	SOCP.w	Unif.w	Analyt.w	SOCP.w
$L_1 B$	3.75	**15.7**	10.9	3.64	**15.0**	10.6
$L_\infty B$	3.04	**20.2**	-	2.87	**19.8**	-
L_1 H	-8.91	**1.58**	-6.18	-8.94	**0.89**	-7.74
L_∞ H	-8.37	**5.83**	-	-8.63	**4.90**	-

Table 6: The return with performance guarantees for the Inventory Management experiment. The return of the nominal MDP is 163.1.

	$\delta = 0.5$			$\delta = 0.05$		
Method	Unif.w	Analyt.w	SOCP.w	Unif.w	Analyt.w	SOCP.w
$L_1 B$	3.83	**8.28**	4.21	3.82	**8.25**	4.20
$L_\infty B$	3.81	**7.78**	-	3.78	**7.71**	-
L_1 H	2.81	**3.44**	2.87	2.80	**3.42**	2.85
L_∞ H	3.18	**3.94**	-	3.15	**3.92**	-

Table 7: The return with performance guarantees for the Cart-Pole experiment. The return of the nominal MDP is 11.11.