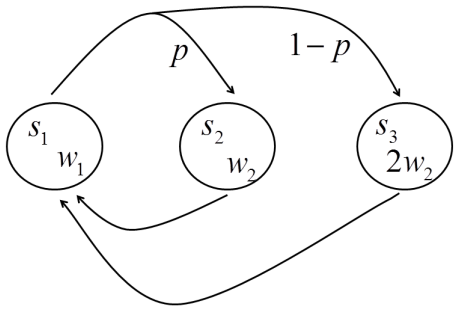
# Scaling Up Robust MDPs by Function Approximation - Supplementary Material

## 1 A Divergent Example

We show that even if Assumption 2 fails for a single state $\tilde{x}$, and for that state there is no approximation - i.e., there is a feature $\tilde{\phi}(x) = \mathbb{1}\{x = \tilde{x}\}$ that is orthogonal to all other features, iteratively applying $\Pi T^\pi$ may diverge.

Consider the following MDP with 3 states $\{s1, s2, s3\}$, zero rewards, and let the value function approximation be $(w_1, w_2, 2w_2)^T$.




The Bellman operator for some $v = (w_1, w_2, 2w_2)^T$ is

$$T^\pi v = \gamma \begin{pmatrix} pw_2 + (1-p)\,2w_2 \\ w_1 \\ w_1 \end{pmatrix}$$

Consider an exploration policy ($\hat{P}$) where $p = 0.5$, and therefore the steady state distribution of $s2$ and $s3$ are equal. Note that the only transition change (between the exploration policy and the true MDP) is in $s_1$, for which there is no approximation in the value function. The least squares regression of a vector $(x_1, x_2)$ onto $(w_2, 2w_2)$ gives $w_2 = \frac{1}{5}(x_1 + 2x_2)$, and therefore the projected Bellman operator is

$$\Pi T^\pi v = \gamma \begin{pmatrix} 2w_2 - pw_2 \\ \frac{3}{5}w_1 \\ \frac{6}{5}w_1 \end{pmatrix},$$

and in terms of $w$, we can write the result of applying $\Pi T^\pi$ as $w'$, and we have

$$w' = \gamma \begin{pmatrix} 0 & (2-p) \\ \frac{3}{5} & 0 \end{pmatrix} w.$$

The eigenvalues of the above matrix are $\pm\gamma\frac{\sqrt{15(2-p)}}{5}$, and we have that for $p < 2 - \frac{5}{3\gamma^2}$ (For example $p = 0.1, \gamma = 0.95$) the eigenvalues are outside the unit circle and the process of repeatedly applying $\Pi T^\pi$ diverges.

# 2 Complexity of Solving the Inner Problem using SAA

We state a result of Shapiro & Nemirovski (2005) that bounds the sample complexity of SAA. Recall that we approximate the solution of the problem

$$\inf_{\theta\in\Theta} \mathbb{E}_{\tilde{p}} \left[ \frac{p_\theta(x)}{\tilde{p}(x)} \phi(x)^\top w_k \right], \tag{1}$$

using the solution of the SAA

$$\inf_{\theta\in\Theta} \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{p_\theta(x_i)}{\tilde{p}(x_i)} \phi(x_i)^\top w_k. \tag{2}$$

Assume that $\mathbb{E}_{\tilde{p}} \left[ \frac{p_\theta(x)}{\tilde{p}(x)} \phi(x)^\top w_k \right]$ is convex in $\theta$, and that $\frac{p_\theta(x)}{\tilde{p}(x)} \phi(x)^\top w_k$ is Lipschitz continuous on $\Theta$ with constant $L$ independent of $x$. Let $D \doteq \sup_{\theta,\theta'\in\Theta} \|\theta' - \theta\|$ denote the diameter of $\Theta$. Then the following sample complexity result holds.

**Theorem 1.** *(Theorem 2 in Shapiro & Nemirovski 2005) For a sample size $N_s$ that satisfies*

$$N_s \geq \mathcal{O}(1) \left( \frac{DL}{\epsilon} \right)^2 \left[ n \log \left( \frac{DL}{\epsilon} \right) + \log \left( \frac{\mathcal{O}(1)}{\alpha} \right) \right]$$

*we are guaranteed that every $(\epsilon/2)$-optimal solution of the SAA problem (2) is an $\epsilon$-optimal solution of the true problem (1) with probability $1 - \alpha$.*

Theorem 1 bounds the number of samples required for constructing the SAA approximation (2). However, one still needs to solve the SAA problem. When $\mathbb{E}_{\tilde{p}} \left[ \frac{p_\theta(x)}{\tilde{p}(x)} \phi(x)^\top w_k \right]$ is convex and twice continuously differentiable, the SAA may be solved efficiently using, e.g., interior point methods (Boyd & Vandenberghe, 2004).

# 3 Proof of Proposition 6

**Proposition 2.** *The sequence $\{\pi_i\}$ generated by the general approximate robust policy iteration algorithm satisfies*

$$\limsup_{i\to\infty} \|V^{\pi_i} - V^*\|_\infty \leq \frac{\epsilon + 2\gamma\delta}{(1-\gamma)^2}.$$

*Proof.* The proof of Proposition 2.5.8 of Bertsekas (2012) holds provided that the operators $T^\pi$ and $T$ are both $\gamma$-contractions in the sup-norm and monotone. The contraction property was shown by Iyengar (2005). We now show the monotonicity.

Choose some policy $\pi$ and $\epsilon' > 0$. Let $v, v' \in \mathbb{R}^{|\mathcal{X}|}$ satisfy $v(x) \leq v'(x)$ for all $x$. Also let $\bar{p}_x \in \mathcal{P}(x, \pi(x))$ such that $\bar{p}_x^\top v' \leq \inf_{p \in \mathcal{P}(x,\pi(x))} p^\top v' + \epsilon'$. We have that for all $x$

$$\inf_{p \in \mathcal{P}(x,\pi(x))} p^\top v' + \epsilon' \geq \bar{p}_x^\top v' \geq \bar{p}_x^\top v \geq \inf_{p \in \mathcal{P}(x,\pi(x))} p^\top v,$$

where the second inequality holds since by definition $\bar{p}_x \geq 0$. Since $\epsilon'$ was arbitrary we conclude that $\inf_{p \in \mathcal{P}(x,\pi(x))} p^\top v \leq \inf_{p \in \mathcal{P}(x,\pi(x))} p^\top v'$, therefore $T^\pi v \leq T^\pi v'$, which proves the monotonicity of $T^\pi$. Since this holds for every $\pi$, it holds also for $T$. □

# 4 Optimal Stopping Problems

Consider the optimal stopping problem setting, and let $\hat{\pi}$ denote a policy that never chooses to terminate, i.e., $\hat{\pi}(x) = 0$, $\forall x$. We now show that if Assumption 2 is satisfied for $\pi = \hat{\pi}$, then it is immediately satisfied for all other policies.

**Proposition 3.** *Consider an optimal stopping problem, and let Assumption 2 hold for $\pi = \hat{\pi}$. Then, for every policy $\pi$ we have*

$$\gamma P(x'|x, \pi(x)) \leq \beta \hat{P}(x'|x, \hat{\pi}(x)), \quad \forall P \in \mathcal{P}, x \in \mathcal{X}, x' \in \mathcal{X}. \tag{3}$$

*Proof.* We prove by induction. Assume (3) holds for some $\pi$. Let $\tilde{\pi}$ be the same as $\pi$ for all states except $\tilde{x}$, for which $\pi(\tilde{x}) = 0$ and $\tilde{\pi}(\tilde{x}) = 1$. Then we have for all $x \neq \tilde{x}$ that $P(x'|x, \tilde{\pi}(x)) = P(x'|x, \pi(x)), \quad \forall P \in \mathcal{P}$. For $\tilde{x}$, a transition to a terminal state occurs without uncertainty, namely $P(x'|\tilde{x}, \tilde{\pi}(\tilde{x})) = 0 \quad \forall P \in \mathcal{P}, x' \in \mathcal{X}$, therefore (3) is satisfied with $\pi$ replaced by $\tilde{\pi}$.

Since (3) is assumed to hold for $\hat{\pi}$, by induction it holds for all $\pi$. □

# 5 Optimistic MDPs

Interestingly, as was recognized by Iyengar (2005), results on robust MDPs may be extended to optimistic MDPs. An optimistic MDP is similar to an RMDP, but the optimization goal is different. Here, instead of the worst case performance, we seek the most optimistic value $V^+(x) = \sup_\pi \left\{\sup_{P \in \mathcal{P}} V^{\pi,P}(x)\right\}$. In addition to obtaining risk-seeking policies, optimistic MDPs have been used for efficient exploration, driving algorithms such as UCRL2 Jaksch et al. (2010) by employing the principle of 'optimism in the face of uncertainty'. Our work may be important for large-scale implementations of such algorithms, by use of function approximation. We also conjecture that the performance gap due to uncertainty $V^+(x) - V(x)$ may be important for feature selection and model selection, tasks that are critical for truly large-scale applications.

For some $x$ and $u$ let us define the operator $\sigma^+_{\mathcal{P}(x,u)} : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}$ as (cf. the definition of $\sigma_{\mathcal{P}(x,u)}$)

$$\sigma^+_{\mathcal{P}(x,u)} v \doteq \sup \left\{ p^\top v : p \in \mathcal{P}(x,u)] \right\}.$$

All our results extend to optimistic MDPs, namely, by replacing the operator $\sigma_{\mathcal{P}(x,u)}$ with $\sigma^+_{\mathcal{P}(x,u)}$. We now show this for the proof of Proposition 3.

*Proof.* Fix $x \in \mathcal{X}$, and assume that $T^\pi y(x) \leq T^\pi z(x)$. Choose some $\epsilon > 0$, and $P_x \in \mathcal{P}$ such that

$$\mathbb{E}^{P_x} \left[ z(x') | \, x, \pi(x) \right] \geq \sup_{P \in \mathcal{P}} \mathbb{E}^P \left[ z(x') | \, x, \pi(x) \right] - \epsilon. \tag{4}$$

Also, note that by definition

$$\sup_{P\in\mathcal{P}} \mathbb{E}^{P}\left[y(x')\middle|\, x,\pi(x)\right] \geq \mathbb{E}^{P_x}\left[y(x')\middle|\, x,\pi(x)\right]. \tag{5}$$

Now, we have

$$\begin{aligned}
0 &\leq T^{\pi}z(x) - T^{\pi}y(x) \\
&\leq \left(\gamma\mathbb{E}^{P_x}\left[z(x')\middle|\, x,\pi(x)\right] + \gamma\epsilon\right) - \left(\gamma\mathbb{E}^{P_x}\left[y(x')\middle|\, x,\pi(x)\right]\right) \\
&= \gamma\mathbb{E}^{P_x}\left[z(x') - y(x')\middle|\, x,\pi(x)\right] + \gamma\epsilon \\
&\leq \beta\mathbb{E}^{\hat{P}}\left[\left|z(x') - y(x')\right|\middle|\, x,\pi(x)\right] + \gamma\epsilon,
\end{aligned}$$

where the second inequality is by (4) and (5), and the last inequality is by Assumption 2. Conversely, if $T^{\pi}y(x) \geq T^{\pi}z(x)$, following the same procedure we obtain $0 \leq T^{\pi}y(x) - T^{\pi}z(x) \leq \beta\mathbb{E}^{\hat{P}}\left[\left|z(x') - y(x')\right|\middle|\, x,\pi(x)\right] + \gamma\epsilon$, and we therefore conclude that $\left|T^{\pi}y(x) - T^{\pi}z(x)\right| \leq \beta\mathbb{E}^{\hat{P}}\left[\left|y(x') - z(x')\right|\middle|\, x,\pi(x)\right] + \gamma\epsilon$. Since $\epsilon$ was arbitrary, we have that $\left|T^{\pi}y(x) - T^{\pi}z(x)\right| \leq \beta\mathbb{E}^{\hat{P}}\left[\left|y(x') - z(x')\right|\middle|\, x,\pi(x)\right]$ for all $x$, and therefore

$$\left\|T^{\pi}y - T^{\pi}z\right\|_d \leq \beta\left\|\hat{P}\left|y - z\right|\right\|_d \leq \beta\left\|y - z\right\|_d,$$

where in last equality we used the well-known result that the state transition matrix $\hat{P}$ is contracting in the $d$-weighted Euclidean norm. □

We now show that Proposition 6 also holds.

*Proof.* We need to show the monotonicity property.

Choose some policy $\pi$ and $\epsilon' > 0$. Let $v, v' \in \mathbb{R}^{|\mathcal{X}|}$ satisfy $v(x) \geq v'(x)$ for all $x$. Also let $\bar{p}_x \in \mathcal{P}(x,\pi(x))$ such that $\bar{p}_x^{\top}v' \geq \sup_{p\in\mathcal{P}(x,\pi(x))} p^{\top}v' - \epsilon'$. We have that for all $x$

$$\sup_{p\in\mathcal{P}(x,\pi(x))} p^{\top}v' - \epsilon' \leq \bar{p}_x^{\top}v' \leq \bar{p}_x^{\top}v \leq \sup_{p\in\mathcal{P}(x,\pi(x))} p^{\top}v,$$

where the second inequality holds since by definition $\bar{p}_x \geq 0$. Since $\epsilon'$ was arbitrary we conclude that $\sup_{p\in\mathcal{P}(x,\pi(x))} p^{\top}v \geq \sup_{p\in\mathcal{P}(x,\pi(x))} p^{\top}v'$, therefore $T^{\pi}v \geq T^{\pi}v'$, which proves the monotonicity of $T^{\pi}$. Since this holds for every $\pi$, it holds also for $T$. □

# 6 Parameters for Option Trading Experiments

The parameters for the experiments in Section 6 were chosen to balance the different factors in the problem. Specifically, we chose

| Experiment | $Kput$ | $Kcall$ | $T$ | $\gamma$ | $f_u$ | $f_d$ | $p$ | $p_1$ | $p_2$ | $N_{data}$ | $N_{sim}$ | $\delta$ | $x_0$ | $N_{test}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Put option | 1 | 1.5 | 20 | 0.98 | 9/8 | 8/9 | 0.45 | | | 5 | 500 | 1 | 1 | 50,000 |
| Put and call | 1 | 1.5 | 20 | 0.98 | 9/8 | 8/9 | 0.45 | | | 3 | 500 | 1 | 1.25 | 50,000 |
| Model misspecification | 1 | 1.5 | 20 | 0.98 | 9/8 | 8/9 | | 0.3 | 0.6 | 3 | 500 | 1 | 1.25 | 50,000 |

We used 2-Dimensional Gaussian RBF features with a uniform spacing $\Delta x = 0.4$, $\Delta t = 6$, and widths $\sigma_x = 0.235$ and $\sigma_t = 3.535$. The outputs of the RBFs were normalized.

# 7 Experiments with a Geometric Brownian Motion Model

In this section we consider the option trading domain of Section 6, where the price model follows a Geometric Brownian Motion (GBM), a popular model for stock price fluctuations. In continuous time, a GBM obeys the stochastic differential equation $dx_t = \mu x_t dt + \sigma x_t dW_t$, where $\mu$ is the risk free interest rate, $\sigma$ is the stock volatility, and $W$ is a standard Brownian motion. In discrete time, a GBM trajectory may be simulated by $x_{t+\Delta t} = x_t \exp\left\{(\mu - \sigma^2/2)\Delta t + \sigma\sqrt{\Delta t}\omega\right\}$, where $\omega \sim \mathcal{N}(0,1)$. Thus, $x_{t+1}/x_t$ has a lognormal distribution

$$\frac{x_{t+\Delta t}}{x_t} \sim \ln\mathcal{N}\left(\Delta t(\mu - \sigma^2/2), \sigma^2\Delta t\right).$$

In practice, the volatility is not known, but estimated from data. Thus, we construct the uncertainty set as the 95% confidence intervals for the estimated volatility. Our empirical evaluation proceeds as follows. In each experiment, we generate $N_{data}$ trajectories of length $T$ from the true model $M$ with parameters $\mu$ and $\sigma$ where $\mu$ is the risk-free interest rate. From these trajectories we estimate the volatility $\hat{\sigma}$, and the 95% confidence intervals $\hat{\sigma}_-$ and $\hat{\sigma}_+$ using the Matlab function `lognfit`, which constructs our uncertain model $M_{robust}$. We also build a model without uncertainty $M_{nominal}$ by setting $\hat{\sigma}_- = \hat{\sigma}_+ = \hat{\sigma}$. Using $\hat{\sigma}$, we then simulate $N_{sim}$ trajectories of length $T$ (this corresponds to a policy that never executes the option), where $x_0 = K + \epsilon$, and $\epsilon$ is uniformly distributed in $[-\delta, \delta]$. These trajectories are used as input data for the ARPI algorithm of Section 4. For solving the inner problem, we use the SAA method of Section 3.4 with $N_s$ samples, where we set $\tilde{p}$ to the lognormal distribution corresponding to $\hat{\sigma}_+$. The deterministic sampled problem was solved using Matlab's `fminbnd` method.

In Figure 1 we plot the tail distribution of the total reward $R$ (from 20 independent experiments) obtained by $\pi_{robust}$ and $\pi_{nominal}$ for the put option scenario (cf. Section 6.2.2 of the main text). The results are similar to the case of the Bernoulli price fluctuation model. These results confirm that our method scales robust MDPs to truly large scale domains.

The parameters for this experiment were chosen to balance the different factors in the problem. Specifically, we chose

| $Kput$ | $T$ | $\mu$ | $\sigma$ | $\Delta t$ | $\gamma$ | $N_s$ | $N_{data}$ | $N_{sim}$ | $\delta$ | $x_0$ | $N_{test}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 0.0025 | 3 | 0.01 | 0.9975 | 50 | 5 | 500 | 1 | 1 | 50,000 |

The RBFs were the same as in the experiments in the main text.

# References


Bertsekas, D. P. *Dynamic Programming and Optimal Control, Vol II.* Athena Scientific, fourth edition, 2012.

Boyd, S. P. and Vandenberghe, L. *Convex optimization.* Cambridge Univ Pr, 2004.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Shapiro, A. and Nemirovski, A. On complexity of stochastic programming problems. In *Continuous optimization*, pp. 111–146. Springer, 2005.

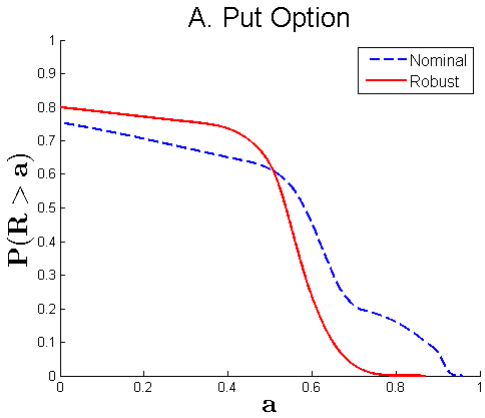



Figure 1: Performance of robust vs. nominal policies. The tail distribution (complementary cumulative distribution function) of the total reward $R$ for the put option scenario, obtained from 20 independent experiments.