A Conditional Value-at-Risk Approach for Uncertain Markov Decision Processes

Yao-Liang Yu YAOLIANG@CS.UALBERTA.CA
Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
Yuxi Li YUXI@CS.UALBERTA.CA
Dale Schuurmans DALE@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8 Canada

Abstract

We study the application of stochastic programming to handle parameter uncertainty in Markov Decision Processes. In particular, we adopt *Conditional Value-at-Risk* as an appropriate objective for handling parameter uncertainty, and show how stochastic programming can naturally bridge between minimax and averaging approaches in this case.

1. Introduction

Markov Decision Processes (MDPs) are a powerful modeling methodology used widely in sequential decision-making tasks. When model parameters are perfectly known, dynamic programming (DP) can be used to find an optimal policy. Unfortunately, since model parameters are usually estimated from data, uncertainty is inevitable and it has been shown that this uncertainty can lead to a strong bias in the underlying optimal policy (Mannor et al., 2007).

Perhaps the most common approach to handling parameter uncertainty in MDP planning is minimax; *i.e.*, account for the worst-case (Nilim & Ghaoui, 2005; Iyengar, 2005). A natural alternative is to average the uncertainty (Mannor et al., 2007). However, minimax is often overly conservative while both methods lack a straightforward mechanism for tuning one's optimistic/pessimistic attitude towards uncertainty.

We propose a conditional value-at-risk (CVaR) approach for handling uncertainty in MDPs. Value-at-risk (VaR) and CVaR are two popular risk measures in finance (Rockafellar & Uryasev, 2000). These objectives are both controlled by a percentile value that reflects one's tolerance of uncertainty, and thus both provide a principled objective for planning in the face of MDP uncertainty. Recently, VaR has been applied to uncertain MDPs to achieve policies that avoid this form of risk (Delage & Mannor, 2009). Though widely used, VaR is very difficult to optimize and in fact creates an NP-Hard planning problem for MDPs with uncertainty (Delage & Mannor, 2009).

Instead, by employing CVaR to handle uncertainty in MDPs we uncover many advantages: (1) in many cases CVaR yields a convex problem and thus is much easier to handle; (2) VaR is upper bounded by CVaR, thus optimizing CVaR always provides a feasible solution for the VaR counterpart; and (3) the CVaR approach connects previous minimax and averaging approaches and hence provides a unified framework for dealing with parameter uncertainty in MDPs.

2. Uncertain MDPs

An MDP is usually described by the quintuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. The standard planning problem is to find an optimal (stationary) policy Π such that the expected total discounted return is maximized:

$$\max_{\Pi} \quad \ell(\Pi; \mathcal{P}, r) \quad := \quad \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r_{s_t} | s_0 \sim q, \Pi\right) \tag{1}$$

where q is an initial distribution on states. Though problem (1) is a nonlinear function of the policy, it can be efficiently solved by dynamic programming, provided that the model parameters $\mathcal{P}, r$ are perfectly known. Unfortunately, this is rarely true in practice.

2.1. Previous Approaches

When parameter uncertainty needs to be taken into account, one can either be pessimistic by choosing the minimax approach (Nilim & Ghaoui, 2005; Iyengar, 2005):

$$\max_{\Pi} \quad \min_{\mathcal{P}, r} \quad \ell(\Pi; \mathcal{P}, r) \tag{2}$$

or be neutral by averaging (Mannor et al., 2007):

$$\max_{\Pi} \quad \mathbb{E}_{\mathcal{P}, r}\, \ell(\Pi; \mathcal{P}, r) \tag{3}$$

However, both approaches lack a way of tuning our attitude towards the uncertainty. By adopting VaR as the uncertainty measure, Delage & Mannor (2009) considered the percentile problem:

$$\max_{\Pi, y} \quad y \tag{4}$$

$$\text{s.t.} \quad \mathbb{P}_{\mathcal{P},r}\left[\ell(\Pi;\mathcal{P},r) \geq y\right] \;\geq\; 1-\epsilon \tag{5}$$

Here ϵ is a percentile parameter reflecting our uncertainty tolerance. The VaR approach guarantees that with probability $1-\epsilon$ the resulting performance will be at least y^{VaR}, where y^{VaR} is the maximum value obtained in (4). It is easy to see that by setting $\epsilon = 0$, problem (4) becomes problem (2). Unfortunately, it is well-known in the finance literature that VaR is hard to optimize since the probability constraint (5) is usually non-convex (Rockafellar & Uryasev, 2000). In fact, it has been proved in (Delage & Mannor, 2009) that (4)-(5) is NP-Hard in general.

2.2. CVaR Approach

Unlike VaR, CVaR is a coherent risk measure that is convex in many cases (as long as the loss function is convex) (Rockafellar & Uryasev, 2000). We propose to employ CVaR as uncertainty measure, which yields the following stochastic programming problem:

$$\max_{\Pi} \quad \frac{1}{\epsilon}\,\mathbb{E}_{\mathcal{P},r}\left[\,\ell(\Pi;\mathcal{P},r) \;\middle|\; \ell(\Pi;\mathcal{P},r) \leq y^{VaR}\,\right] \tag{6}$$

The normalization factor $\frac{1}{\epsilon}$ ensures that the optimal value of (6) will always be **no larger** than y^{VaR}. Additionally, the percentile parameter ϵ now plays a clearer role: As ϵ approaches 0, like the VaR approach, problem (6) will also become the minimax approach (2). On the other hand, if we set $\epsilon = 1$, problem (4)-(5) will be meaningless and $y^{VaR} = \infty$. However, the proposed problem (6) is still well defined and in fact, it becomes problem (3)! Thus, we see that the CVaR approach (6) naturally *bridges* the pessimistic attitude (2) and the neutral approach (3).

However, it might seem that CVaR is "harder" to optimize than VaR since y^{VaR}, the optimal value of (4), is involved in problem (6). Fortunately, an elegant alternative has been provided in (Rockafellar & Uryasev, 2000) to solve CVaR problems. Following their work, problem (6) can be equivalently reformulated as:

$$\max_{\Pi, y} \quad y - \frac{1}{\epsilon}\,\mathbb{E}_{\mathcal{P},r}[y - \ell(\Pi;\mathcal{P},r)]_+ \tag{7}$$

where $[x]_+$ is used to denote $\max\{x, 0\}$. It is easy to see that as long as $\ell(\Pi;\mathcal{P},r)$ is concave, problem (7) will be a convex optimization problem and in general it belongs to stochastic programming due to the expectation operator in its objective function. When the expectation is hard to solve analytically, Monte Carlo sampling can be an efficient way to approximate it.

2.3. Gaussian Reward Uncertainty

We give a brief example of how to apply the CVaR approach (6) to uncertain MDPs. For simplicity, assume the transition probability matrix $\mathcal{P}$ is known and deterministic, while the reward r follows a Gaussian prior distribution, $\mathcal{N}(\mu, \Sigma)$.

Proposition 1 *For any $\epsilon \in (0, 1]$, the CVaR approach (6) with Gaussian reward uncertainty and deterministic transition probabilities can be solved by the following second order cone programming (SOCP):*

$$\max_{\rho} \quad \rho^T\mathbf{1}\mu - \kappa_\epsilon\|\rho^T\mathbf{1}\Sigma^{\frac{1}{2}}\|_2 \tag{8}$$

$$\text{s.t.} \quad \rho \geq 0, \quad \rho^T\mathbf{1} = q^T + \gamma\rho^T\mathcal{P} \tag{9}$$

where $\kappa_\epsilon = \frac{1}{\sqrt{2\pi}\epsilon}\exp(-\frac{(\Phi^{-1}(\epsilon))^2}{2})$ and $\Phi(\cdot)$ is the c.d.f. of $\mathcal{N}(0, 1)$.

Note that the VaR approach (4)-(5) reduces to a similar SOCP in this setting. However, our CVaR approach differs from VaR in the constant κ_ϵ and the allowable range of the percentile parameter ϵ, specifically, (0, 1] versus (0, 0.5]. In particular, we see that when $\epsilon = 1$, κ_ϵ will be 0 and CVaR approach becomes the averaging approach (3).

3. Conclusion

We propose to employ the CVaR as the risk measure to take into account the parameter uncertainty in MDPs. Our approach can be regarded as a bridge connecting pessimistic and neutral attitudes towards uncertainty.

References

Delage, E., & Mannor, S. (2009). Percentile optimization for markov decision processes with parameter uncertainty. *Operations Research*, To Appear.

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

Mannor, S., Simester, D., Sun, P., & Tsitsiklis, J. N. (2007). Bias and variance approximation in value function estimates. *Management Science*, *53*, 308–322.

Nilim, A., & Ghaoui, L. E. (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, *53*, 780–798.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, *2*, 493–517.