On Dynamic Programming Decompositions of Static Risk Measures in Markov Decision Processes

Jia Lin Hau, Erick Delage, Mohammad Ghavamzadeh, Marek Petrik

Summary

Primal methods

Augment state space to keep track of target value or time.[Wu, 1999], [Lin, 2003], [Bauerle, 2011], [Hau, 2023]

Robust methods

Define **MDP** for risk measure with augmented state space to keep track of risk levels. These popular dynamic program (DP) methods for solving risk averse MDP [Chow, 2015], [Jin, 2019], [Li, 2022], [Ni, 2022] are thought to be optimal with sufficiently discretized risk levels for policy optimization.

Contribution

Policy Evaluation

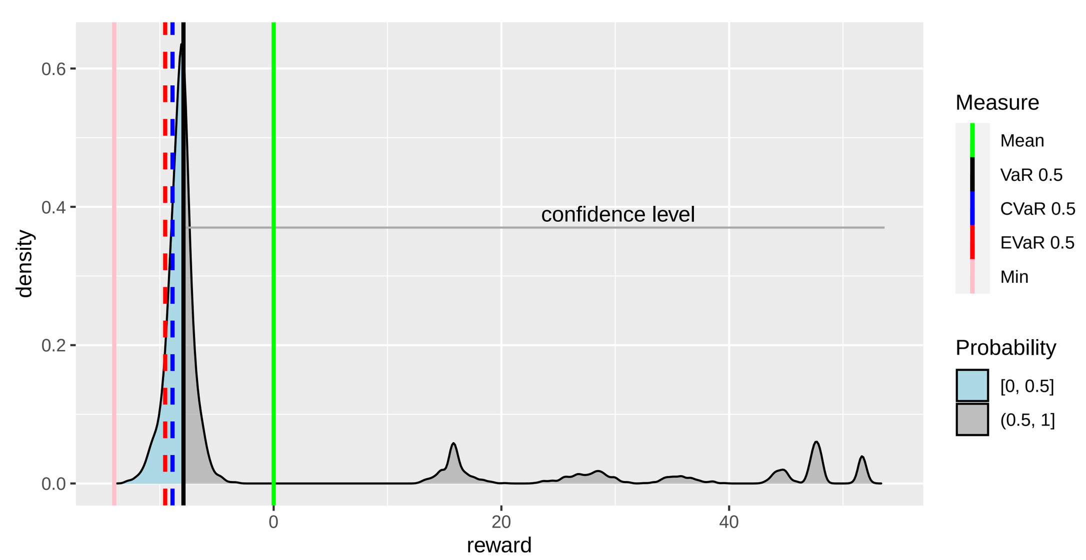
- Prove a new correct VaR and EVaR decomposition.

Policy Optimization

- Prove robust methods are suboptimal for CVaR [Chow 2015] and EVaR [Ni 2022] MDP.
- Propose optimal DP for robust methods Value-at-Risk (VaR) MDP.

Risk Averse MDPs

Risk Measures



$$\mathrm{VaR}_\alpha[\tilde{x}] = \sup\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} < z] \le \alpha\} = \inf\{z \in \mathbb{R} \mid \mathbb{P}[\tilde{x} \le z] > \alpha\}$$

$$\mathrm{CVaR}_\alpha[\tilde{x}] = \sup_{z\in\mathbb{R}} \left(z - \alpha^{-1}\mathbb{E}[z-\tilde{x}]_+\right) = \inf_{\xi\in\Delta_m} \left\{\boldsymbol{\xi}^\top \boldsymbol{x} \mid \alpha\cdot\boldsymbol{\xi} \le \boldsymbol{q}\right\}$$

$$\mathrm{EVaR}_\alpha[\tilde{x}] = \sup_{\beta>0} -\frac{1}{\beta}\log\left(\alpha^{-1}\mathbb{E}\left[\exp(-\beta\cdot\tilde{x})\right]\right) = \inf_{\xi\in\Delta_m} \left\{\boldsymbol{\xi}^T \boldsymbol{x} \mid \mathrm{KL}(\boldsymbol{\xi}\|\boldsymbol{q}) \le -\log\alpha\right\}$$

Risk Averse Objectives

- Maximizes the risk measure $\psi[\cdot]$ of the total reward in a Markov decision process (MDP)

$$\max_{\pi\in\Pi} \psi\left[\sum_{t=0}^{T} r^\pi(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})\right]$$

  Assume: Rewards $r(s,a,s') \in \mathbb{R}$, transition probabilities $P(s,a) \in \triangle^S$ and finite horizon.

- If $\psi[r(\tilde{s},\tilde{a},\tilde{s}')]$ and $\max_{\pi\in\Pi}\psi[r^\pi(\tilde{s},\tilde{a},\tilde{s}')]$ can be written in term of $\psi[r(s,\tilde{a},\tilde{s}'|\tilde{s}=s)]$ and $\max_{\pi\in\Pi}\psi[r^\pi(s,\tilde{a},\tilde{s}'|\tilde{s}=s)]$ then exist DP for evaluation and policy optimization respectively.

Challenges: $\psi \in$ VaR, CVaR and EVaR, do not satisfy *tower property*: $\psi[X] = \psi[\,\psi[X \mid Y]\,]$.

$$\max_{\pi\in\Pi} \psi\left[\sum_{t=\tau}^{T} r^\pi(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})\right] \ne \max_{a\in A} \psi_{s'}\left[r(s,a,s') + \max_{\pi\in\Pi}\psi\Big[\sum_{t=\tau+1}^{T} r^\pi(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1}) | \tilde{s}_{\tau+1} = s'\Big]\right]$$

Extended Conditional Coherent Risk Measure

Proposition 3.1 : Lemma 22 in [Pflug, 2016] Suppose that $\pi \in \Pi$ and $\tilde{s} \sim \hat{p}$, $\tilde{a} \sim \pi(\tilde{s})$, $\tilde{s}' \sim p_{s,a}$. Then,

$$\mathrm{CVaR}_\alpha[r(\tilde{s},\tilde{a},\tilde{s}')] = \min_{\zeta\in\mathcal{Z}_C} \sum_{s\in\mathcal{S}} \zeta_s \,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s} = s],$$

where the state s on the right-hand side is not random and $\mathcal{Z}_C = \{\zeta \in \Delta_S \mid \alpha\cdot\zeta \le \hat{p}\}$.

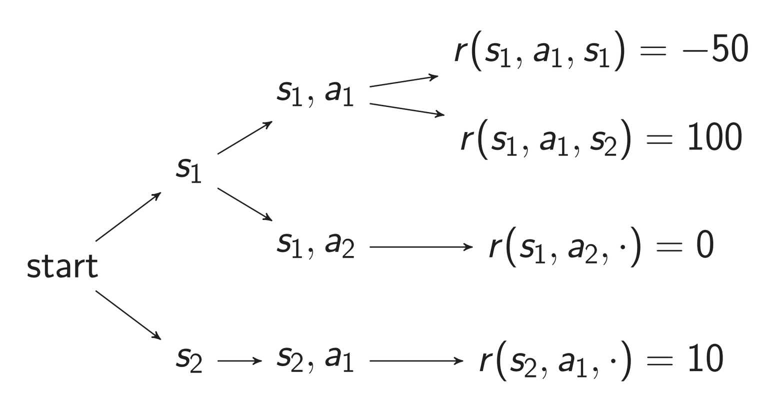
CVaR Decomposition Fails in Optimization

Theorem 3.2 : There exists an MDP and a risk level $\alpha \in [0,1]$ such that

$$\begin{aligned}\max_{\pi\in\Pi} \mathrm{CVaR}_\alpha[r^\pi(\tilde{s},\tilde{a},\tilde{s}')] &= \max_{\pi\in\Pi}\min_{\zeta\in\mathcal{Z}_C} \sum_{s\in\mathcal{S}} \zeta_s\,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[r^\pi(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \\ &< \min_{\zeta\in\mathcal{Z}_C}\max_{\pi\in\Pi} \sum_{s\in\mathcal{S}} \zeta_s\,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[r^\pi(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s] \\ &= \min_{\zeta\in\mathcal{Z}_C} \sum_{s\in\mathcal{S}} \zeta_s \max_{\pi\in\Pi}\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[r^\pi(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s]\end{aligned}$$

Let $\theta_\pi(\zeta) = \sum_{s\in\mathcal{S}} \zeta_s\,\mathrm{CVaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[r^\pi(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s]$ and $\alpha = 0.5$.

Figure right below plot the function $\min_{\zeta\in\mathcal{Z}_C}\max_{\pi\in\Pi}\theta_\pi(\zeta)$ and $\max_{\pi\in\Pi}\min_{\zeta\in\mathcal{Z}_C}\theta_\pi(\zeta)$ for the 2 states simple example in left below to illustrate the inaccurate approximation.



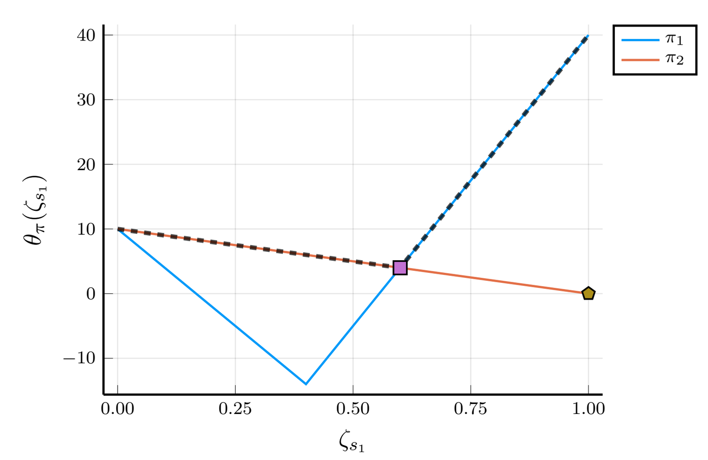
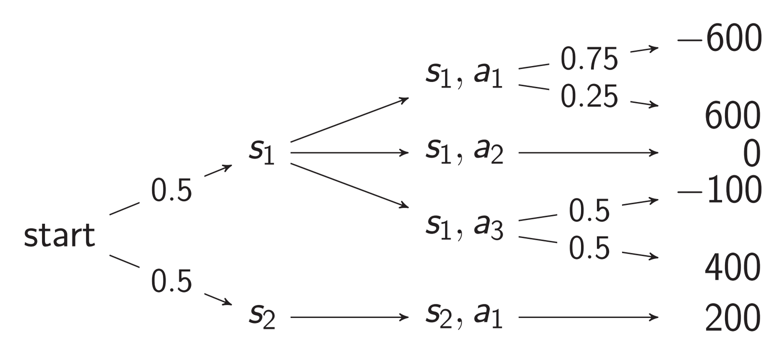
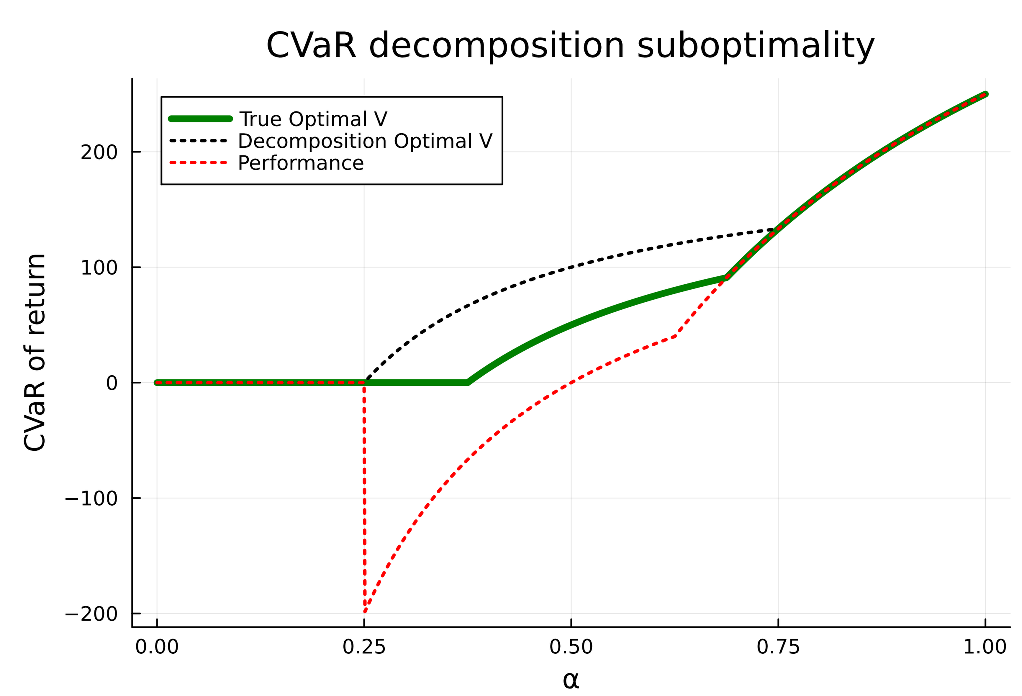


Figure below demonstrate an example and illustrate the sub-optimality of a policy for CVaR decomposition.





EVaR Decomposition Fails in Optimization

Theorem 4.2: EVaR evaluation decomposition Given $\alpha \in (0,1]$, we have that

$$\mathrm{EVaR}_\alpha[r(\tilde{s},\tilde{a},\tilde{s}')] = \inf_{\zeta\in(0,1]^S,\,\xi\in\mathcal{Z}'_E(\zeta)} \sum_{s\in\mathcal{S}} \xi_s\,\mathrm{EVaR}_{\zeta_s}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s],$$

where $\mathcal{Z}'_E(\zeta) = \left\{\boldsymbol{\xi} \in \Delta_S \mid \boldsymbol{\xi} \ll \hat{\boldsymbol{p}},\ \sum_{s\in\mathcal{S}}\xi_s\big(\log(\xi_s/\hat{p}_s) - \log(\zeta_s)\big) \le -\log\alpha\right\}$.

Note : EVaR Decomposition also fails in optimization with the same reason in Theorem 3.2.

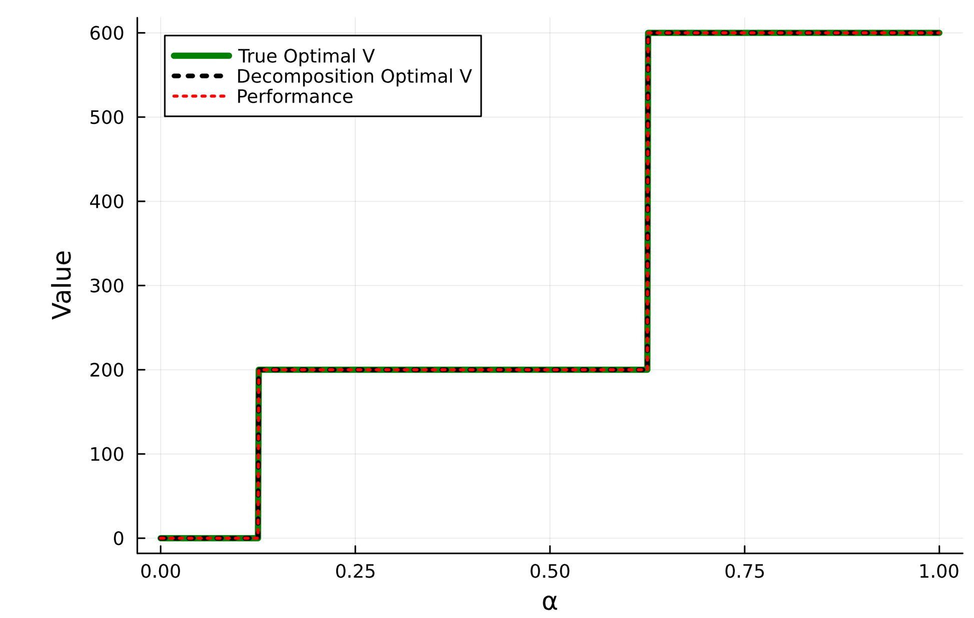
Value at Risk (VaR) MDP

Theorem 5.1: VaR policy evaluation Given any finite MDP for risk level $\alpha \in [0,1]$ we have

$$\mathrm{VaR}_\alpha[r(\tilde{s},\tilde{a},\tilde{s}')] = \sup_{\zeta\in\Delta_N}\left\{\min_{s} \mathrm{VaR}_{\alpha\zeta_s\hat{p}_s^{-1}}[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s],\ \alpha\cdot\boldsymbol{\zeta} \le \hat{\boldsymbol{p}}\right\},$$

Theorem 5.2: VaR policy optimization Given any finite MDP with $\alpha \in [0,1]$, we have

$$\begin{aligned}\max_{\pi\in\Pi}\mathrm{VaR}_\alpha^{\tilde{a}\sim\pi(\tilde{s})}[r(\tilde{s},\tilde{a},\tilde{s}')] &= \max_{\pi\in\Pi}\sup_{\zeta\in\Delta_S}\left\{\min_{s\in\mathcal{S}}\mathrm{VaR}^{\tilde{a}\sim\pi}_{\alpha\zeta_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right] \mid \alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}\right\} \\ &= \sup_{\zeta\in\Delta_S}\left\{\max_{\pi\in\Pi}\min_{s\in\mathcal{S}}\mathrm{VaR}^{\tilde{a}\sim\pi}_{\alpha\zeta_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right] \mid \alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}\right\} \\ &= \sup_{\zeta\in\Delta_S}\left\{\min_{s\in\mathcal{S}}\max_{\pi\in\Pi}\mathrm{VaR}^{\tilde{a}\sim\pi}_{\alpha\zeta_s\hat{p}_s^{-1}}\left[r(s,\tilde{a},\tilde{s}') \mid \tilde{s}=s\right] \mid \alpha\cdot\boldsymbol{\zeta}\le\hat{\boldsymbol{p}}\right\}\end{aligned}$$



Value at Risk (VaR) Value Iteration

$$q^\star_\tau(s,\alpha,a) = \max_{\pi\in\Pi}\mathrm{VaR}^\pi_\alpha\left[\sum_{t=\tau}^{T} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right] = \sup_{\zeta\in\Delta_S}\left\{\min_{s'\in\mathcal{S}}\left[r(s,a,s') + \max_{a'\in A} q^\star_{\tau+1}\Big(s', \frac{\alpha\zeta_{s'}}{p_{sas'}}, a'\Big)\right]\right\}$$

References

1. Congbin Wu and Yuanlie Lin. Minimizing risk models in Markov decision processes with policies depending on target values. Journal of mathematical analysis and applications, 1999.
2. Yuanlie Lin, Congbin Wu, and Boda Kang. Optimal models with maximizing probability of first achieving target value in the preceding stages. Science in China Series A: Mathematics, 2003.
3. Nicole Bauerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. Mathematical Methods of Operations Research, 2011.
4. Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: A CVaR optimization approach. In Neural Information Processing Systems (NIPS), 2015.
5. Georg Ch Pflug and Alois Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. Mathematics of Operations Research, 2016.
6. I Ge Jin, Bastian Schurmann, Richard M Murray, and Matthias Althoff. Risk-aware motion planning for automated vehicle among human-driven cars. In American Control Conference (ACC), 2019.
7. Xiaocheng Li, Huaiyang Zhong, and Margaret L. Brandeau. Quantile Markov decision processes. Operations Research, 2022.
8. Xinyi Ni and Lifeng Lai. Risk-sensitive reinforcement learning via Entropic-VaR optimization. In Asilomar Conference on Signals, Systems, and Computers, IEEE, 2022.
9. Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted MDPs. In Artificial Intelligence and Statistics (AISTATS), 2023.