# Dynamically Augmented CVaR for MDPs and Uncertainty Quantifications for Robust MDPs Characterizing Risk

**Eugene A. Feinberg**
Department of Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY 11794-3600
eugene.feinberg@stonybrook.edu

**Rui Ding**
Department of Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY 11794-3600
rrd051488@gmail.com

## Abstract

CVaR optimization is an important topic, and there are additional complications for defining and optimizing CVaR for sequential decision-making. This paper investigates the relation between CVaR optimization for an MDP with total discounted costs and a specially constructed Robust MDPs (RMDPs). This RMDP was introduced to the literature 10 years ago. It was proposed for efficient calculations of optimal CVaR values and is broadly used for this purpose. About two years ago it was understood that these calculations lead to lower bounds of the minimal static CVaR. This paper provides additional links between static CVaR and the RMDP. It shows that the optimal value of static CVaR is another characteristic for the RMDP rather than its value. Based on this understanding, this paper introduces the Dynamically augmented CVaR (DCVaR) risk measure, which is more natural than static CVaR. Unlike static CVaR, DCVaR does not sufferer from time inconsistency. In addition, DCVaR's value function is equal to the value function of the RMDP, and it can be efficiently computed by value iterations. Optimal policies minimizing DCVaR exist, and they can be computed efficiently by the algorithm proposed three years ago. DCVaR has certain similarities with nested CVaR, and DCVaR can be viewed as a flexible version of nested CVaR in which tail risk levels can be adjusted depending on achieved gains or losses.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

# 1 Introduction

Conditional Value-at-Risk (CVaR), also sometimes called Average Value-at-Risk (AVaR), is one of the most important and popular risk measures [13–15, 22]. However, for sequential decision-making and, in particular for Markov Decision Processes (MDPs), the situation is more complicated because of several reasons. One of them is that, while a risk measure is defined for a probability space, MDPs deal with families of probabilities depending on strategies. Static CVaR optimization [1] deals with optimization of CVaRs for probabilities defined by strategies, but this approach leads to computational and time inconsistency issues. Another popular approach, nested CVaR, is based on nested optimization of CVaR applied to backward equations [16, 21].

Chow et al. [2] introduced a Robust Markov Decision Process (RMDP) for computing optimal CVaR values by value iterations. The states of this RMDP are the original states of the MDP augmented with risk levels, which change over the history of the process. This is an elegant approach motivated by the CVaR decomposition theorem [12], and it was later used in numerous papers. How et al. [7] showed that these value iterations do not compute values of static CVaR. Here we introduce a natural risk measure, called Dynamically augmented CVaR (DCVaR), whose optimal values were computed in [2].

This paper studies the RMDP introduced by Chow et al. [2] and characterizes the minimal value of static CVaR in terms of this RMDP. Though its value function is not equal to the minimal values of static CVaR, the minimal value of static CVaR is the value function for the class of nonrandomized risk-independent policies for the Decision Maker (DM) in this RMDP, when the DM does not know risk levels assigned by Nature (the second player in the RMDP), and Nature plays a policy, which is the worst for the DM if Nature knows the policy played by the DM. This characterization clarifies the gap between the minimal value of static CVaR and supports claims on time inconsistency of static CVaR for MDPs [10, 17–21] since, in order to achieve CVaR, Nature should know future decisions of the DM.

For DM's policy, DCVaR is the worst expected total discounted costs for the DM in the RMDP defined in [2] if Nature plays its optimal policy. Thus different uncertainty quantifications for this RMDP lead to different meaningful risk measures. DCVaR provides a lower bound for static CVaR, and the optimal value of DCVaR is equal to the optimal value for the RMDP computed in [2]. An algorithm for computing optimal DCVaR policies is presented in [3–5]. There are some similarities between optimality equations for DCVaR and nested CVaR, and nested CVaR looks like a projection of DCVaR, when the risk levels should remain unchanged during the history of the process.

# 2 Static CVaR, Robust MDPs, and Dynamically Augmented CVaR

For an MDP $(\mathbb{X}, \mathbb{A}, A(\cdot), c, p)$ with finite state and action sets $\mathbb{X}$ and $\mathbb{A}$, the optimal value of static CVaR is

$$\mathrm{CVaR}_\alpha(Z_N; x) := \inf_{\pi\in\Pi} \mathrm{CVaR}_\alpha(Z_N; P_x^\pi), \qquad x \in \mathbb{X},$$

where $\mathrm{CVaR}_\alpha(Z_N; P_x^\pi)$ is CVaR of the total $N$-horizon discounted costs $Z_N$ with the tail risk level $\alpha \in [0, 1]$ for an initial state $x \in \mathbb{X}$ and a policy $\pi$ from the set of all policies $\Pi$; $N = 1, 2, \ldots, \infty$.

The following theorem states the existence of nonrandomized optimal policies and convergence of finite-horizon optimal values of static CVaRs to the optimal infinite-horizon value.

**Theorem 2.1.** *For every $N = 1, 2, \ldots$ or for $N = \infty$, there exist a nonrandomized optimal policy $\phi \in \Pi$ for the CVaR optimization problem for which $\mathrm{CVaR}_\alpha(Z_N; P_x^\phi) = \mathrm{CVaR}_\alpha(Z_N; x)$ for all $x \in \mathbb{X}$. In addition,*

$$\mathrm{CVaR}_\alpha(Z_\infty; x) = \lim_{N\to\infty} \mathrm{CVaR}_\alpha(Z_N; x), \qquad x \in \mathbb{X}. \tag{2.1}$$

Let us consider the RMDP introduced in Chow et al. [2]. This RMDP is defined by a tuple $(\mathbf{X}, \mathbb{A}, \mathbb{B}, A(\cdot), B(\cdot,\cdot,\cdot), c, q)$, where the state space is $\mathbf{X} := \mathbb{X} \times [0,1]$, action space is $\mathbb{A}$, uncertainty space is $B := \mathbb{R}^M$ with $M$ being the number of states in $\mathbb{X}$, action sets for the DM at states $(x,y) \in \mathbf{X}$ are $A(x,y) := A(x)$, uncertainty sets for Nature are

$$B(x,y,a) := \mathcal{U}(x,y,a) \cap \{b \in \mathbb{R}^M :\, b_{x'} = 0 \text{ if } p(x'|x,a) = 0\},$$

where, for $x \in \mathbb{X}$, $y \in [0,1]$, $a \in A(x)$,

$$\mathcal{U}(x,y,a) := \{b \in \mathbb{R}^M : b_{x'} \geq 0, y b_{x'} \leq 1, x' = 1, 2, \dots, M, \sum_{x'=1}^{M} b_{x'} p(x'|x,a) = 1\}, \qquad (2.2)$$

one-step costs $c((x,y), a, b, (x',y')) := c(x,a,x')$ for $(x,y), (x',y') \in \mathbf{X}$, $a \in A(x)$, $b \in B(x,y,a,)$, and transition probabilities $q(x', D|x,y,a,b) := b_{x'} p(x'|x,a) \delta_{y b_{x'}}(D)$, $x, x' \in \mathbb{X}$, $y \in [0,1]$, $D \in \mathcal{B}([0,1])$, $a \in A(x)$, $b \in \mathcal{U}(x,y,a)$, where $\delta_z(\cdot)$ is the Dirac measure on the interval $[0,1]$ concentrated at the point $z \in [0,1]$. States for this RMDPs are denoted by $(x,y) \in \mathbf{X}$, and we shall use the notation $v_N(x,y,\pi^A,\pi^{\mathbb{B}}) := v_N((x,y),\pi^A,\pi^{\mathbb{B}})$ for the expected total discounted payoffs over the horizon $N$. Value functions for this RMDP are denoted by $v_N(x,y)$. Let $\Pi^{\mathbb{A}}$ be the set of policies for the DM, and $\Pi^{\mathbb{B}}$ be the set of policies for Nature. Then $\Pi \subset \Pi^{\mathbb{A}}$, where $\Pi$ is the set of policies for the MDP. General results on RMDPs can be found in [5, 8, 9, 11].

The following theorem establishes the relations between the RMDP introduced in [2] and CVaR of the total discounted cost. It illustrates time inconsistency of static CVaR for an MDP: Nature may need to use information about future decisions of the DM in order to achieve static CVaR.

**Theorem 2.2.** *For each DM's policy $\pi \in \Pi$ in the MDP and for each $N = 1, 2, \dots$ or $N = \infty$, there exists a nonrandomized policy for Nature $\phi^{\mathbb{B}} = \phi^{\mathbb{B}}[N] \in \Pi^{\mathbb{B}}$ such that*

$$v_N(x,\alpha,\pi,\phi^{\mathbb{B}}) = \max_{\pi^{\mathbb{B}} \in \Pi^{\mathbb{B}}} v_N(x,\alpha,\pi,\pi^{\mathbb{B}}) \quad \text{for all } x \in \mathbb{X},\ \alpha \in [0,1].$$

The following theorem establishes the relations between the RMDP introduced in [2] and CVaR of the total discounted cost.

**Theorem 2.3.** *For each nonrandomized policy $\phi \in \Pi$ for the DM, for each initial risk level $\alpha \in [0,1]$, and for each $N = 1, 2, \dots$ or $N = \infty$,*

$$\mathrm{CVaR}_\alpha(Z_N; P_x^\phi) = \sup_{\pi^{\mathbb{B}} \in \Pi^{\mathbb{B}}} v_N(x,\alpha,\phi,\pi^{\mathbb{B}}), \qquad x \in \mathbb{X}.$$

The following statement characterizes the optimal value of static CVaR in terms of the RMDP.

**Corollary 2.4.** *For every $N = 1, 2, \dots$ or $N = \infty$, every $x \in \mathbb{X}$, and every $\alpha \in [0,1]$,*

$$\mathrm{CVaR}_\alpha(Z_N; x) = \min_{\phi \in \Pi_{NR}} \max_{\pi^{\mathbb{B}} \in \Pi^{\mathbb{B}}} v_N(x,\alpha,\phi,\pi^{\mathbb{B}}),$$

*where $\Pi_{NR}$ is the set of nonrandomized policies for the DM in the MDP, $\Pi_{NR} \subset \Pi \subset \Pi^{\mathbb{A}}$.*

For $N = 0, 1, \dots$ or $N = \infty$, for $x \in \mathbb{X}$, and for $\alpha \in [0,1]$, the gap between the optimal value of CVaR and the value of the RMDP is

$$\begin{aligned} \Delta_N(x,\alpha) : \ &= \ \mathrm{CVaR}_\alpha(Z_N, x) - v_N(x,\alpha) \\ = \min_{\phi \in \Pi_{NR}} \max_{\pi^{\mathbb{B}} \in \Pi^{\mathbb{B}}} v_N(x,\alpha,\phi,\pi^{\mathbb{B}}) \ &- \ \min_{\phi^{\mathbb{A}} \in \Pi^{\mathbb{A}}} \sup_{\pi^{\mathbb{B}} \in \Pi^{\mathbb{B}}} v_N(x,\alpha,\phi^{\mathbb{A}},\pi^{\mathbb{B}}) \geq 0. \end{aligned}$$

It is shown in [7] that this gap can be positive for $N > 1$; see also [6] for additional comments.

Since optimal policies for Nature do not depend on policies of the DM, the following definition removes the time inconsistency explained above Theorem 2.2.

**Definition 2.5.** *For a policy $\pi^{\mathbb{A}} \in \Pi^{\mathbb{A}}$, initial state $x \in \mathbb{X}$, risk level $\alpha \in [0,1]$, and time horizon $N = 1,2,\dots$ or $N = \infty$, the Dynamically augmented CVaR (DCVaR) is*

$$\mathrm{DCVaR}_\alpha(Z_N; x, \pi^{\mathbb{A}}) := \sup_{\pi^{\mathbb{B}} \in \Pi^{\mathbb{B}}_*} v_N(x, \alpha, \pi^{\mathbb{A}}, \pi^{\mathbb{B}}),$$

*where $\Pi^{\mathbb{B}}_*$ is the set of optimal policies for Nature.*

Since $\Pi \subset \Pi^{\mathbb{A}}$, this definition also applies to policies $\pi^{\mathbb{A}} \in \Pi$. If the DM plays a nonrandomized history-dependent policy $\phi \subset \Pi$, then Theorem 2.3 and Definition 2.5 imply that

$$\mathrm{DCVaR}_\alpha(Z_N; x, \phi) \le \mathrm{CVaR}_\alpha(Z_N; P^\phi_x).$$

Thus, in addition to being a more natural objective function than CVaR, DCVaR can be used for establishing performance guarantees for CVaR. In addition,

$$v_N(x, \alpha) = \min_{\pi^{\mathbb{A}} \in \Pi^{\mathbb{A}}} \mathrm{DCVaR}_\alpha(Z_N; x, \pi). \tag{2.3}$$

This is true because every optimal policy in the RMDP for the DM minimizes DCVaR, and for each $\pi^{\mathbb{B}}_* \in \Pi^{\mathbb{B}}_*$

$$v_N(x, \alpha) = \min_{\pi^{\mathbb{A}} \in \Pi^{\mathbb{A}}} v_N(x, \alpha, \pi^{\mathbb{A}}, \pi^{\mathbb{B}}_*) \le \min_{\pi^{\mathbb{A}} \in \Pi^{\mathbb{A}}} \mathrm{DCVaR}_\alpha(Z_N; x, \pi) \le v_N(x, \alpha),$$

where the equality follows from the existence and definition of an optimal policy for the DM in the RMDP, and the inequalities follow from the definition of the DCVaR and from the properties of the RMDP. Thus all inequalities in the last formula hold in the form of equalities.

**Theorem 2.6.** *For $N = 1, 2, \dots$ or $N = \infty$, $x \in \mathbb{X}$, and $\alpha \in [0,1]$, there is a nonrandomized risk-independent policy $\phi \in \Pi$ such that*

$$\mathrm{DCVaR}_\alpha(Z_N; P^\phi_x) = v_N(x, \alpha) = \min_{\pi^{\mathbb{A}} \in \Pi^{\mathbb{A}}} \mathrm{DCVaR}_\alpha(Z_N; P^{\pi^{\mathbb{A}}}_x) = \min_{\pi \in \Pi} \mathrm{DCVaR}_\alpha(Z_N; P^\pi_x).$$

The fundamental difference between a policy $\pi \in \Pi \subset \Pi^{\mathbb{A}}$ and a policy $\pi^{\mathbb{A}} \in \Pi^{\mathbb{A}}$ is that $\pi$ is DM's policy from the MDP, and it does not know risk levels assigned by Nature, while $\pi^{\mathbb{A}}$ is DM's policy from the RMDP, and it knows risk levels assigned by Nature. An algorithm for computing an optimal nonrandomized policy $\phi \in \Pi$ whose existence is stated in Theorem 2.6 is presented in [3–5]. This algorithm is based on learning estimations of risk levels from the history of the process, and these estimations are sufficient for making optimal decisions. These estimations are based on properties of specific mass transfer problems.

For $N = 1, 2, \dots$ or $N = \infty$, $x \in \mathbb{X}$, and $y \in [0,1]$, the optimality equation for the DCVaR is

$$v_N(x, y) = \min_{a \in A(x)} \max_{b \in B(x,y,a)} \sum_{x' \in \mathbb{X}} [c(x, a, x') + \beta v_{N-1}(x', y b_{x'})] b_{x'} p(x'|x, a), \tag{2.4}$$

where $\beta \in [0,1)$ is the discount factor. Values of $v_N(x,y)$ can be computed by value iterations starting with the terminal value $v_0(x,y)$ being continuous in $y \in [0,1]$. In addition, $v_N(x,y) \to v_\infty(x,y)$ uniformly in $y$ as $N \to \infty$, and $v_\infty$ is the unique bounded solution of (2.4) with $N = \infty$. The functions $v_N(x,y)$ are continuous in $y \in [0,1]$.

There is a relation between DCVaR and nested CVaR [16–22]. For nested CVaR, values $v_N(\cdot)$ and $v_{N-1}(\cdot)$ are functions of one variable $x \in \mathbb{X}$, and the tail risk level $y$ is fixed, that is $y = \alpha$. If we set $y \equiv \alpha$, then (2.4) is the optimality equation for nested CVaR. Similarly to nested CVaR, DVaR optimality equations also minimize CVaR of optimality operators, but tail risk levels are not constant. According to the algorithm presented in [3-5], the DVaR risk level depends on the value function and previous gains and losses. Contrary to this, for nested CVaR, the risk level $\alpha$ is constant. In this sense, DCVaR is a more flexible risk measure than nested CVaR.

**Acknowledgement** This research was partially supported by the U.S. Office of Naval Research (ONR) under Grant N000142412608.

## References

[1] Bäuerle, N., Ott, J., (2011) Markov decision processes with average-value-at-risk criteria. *Math. Meth. Oper. Res.* 74: 361-379.

[2] Chow, Y., Tamar, A., Mannor, S., Pavone, M., (2015) Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*, MIT Press, Cambridge, MA, USA, 1522–1530.

[3] Ding, R., (2023) Risk, Model Uncertainty, and Statistical Divergence: Estimation, Optimization, and Decision-Making with Financial Applications. Ph.D. Thesis, Department of Applied Mathematics and Statistics, Stony Brook University.

[4] Ding, R., Feinberg, E.A., (2022) CVaR optimization for MDPs: Existence and computation of optimal policies. *SIGMETRICS Perform. Eval. Rev.* 50(2): 39-41.

[5] Feinberg, E.A., Ding, R., (2025) Dynamically Augmented CVaR for MDPs , arXiv:2211.07288.

[6] Godbout, M., Durand, A., (2025) On the fundamental limitations of dual static CVaR decompositions in Markov decision processes, *Arxiv 2507.14005v1*.

[7] Hau, J.L., Delage, E., Ghavamzadeh, M., Petrik, M., (2023) On dynamic program decompositions of static risk measures in Markov decision processes. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

[8] Iyengar, G., (2005) Robust dynamic programming. *Math. Oper. Res.* 30(2): 257-280.

[9] Jaśkiewicz, A., Nowak, A.S., (2018) Zero-sum stochastic games. In: *Basar T., Zaccour G. (eds) Handbook of Dynamic Game Theory* (Springer, Cham), 215-279.

[10] Kang, B., Filar, J., (2006) Time consistent dynamic risk measures. *Math. Meth. Oper. Res.* 63: 169-186.

[11] Nilim, A., El Ghaoui, L., (2005) Robust control of Markov decision processes with uncertain transition matrices. *Ope. Res.* 53(5): 780-798.

[12] Pflug, G., Pichler, A., (2016) Time-consistent decisions and temporal decomposition of coherent risk functionals. *Math. Oper. Res.* 41: 682-699.

[13] Rockafellar, R.T., Uryasev, S., (2000) Optimization of conditional value-at-risk. *J. Risk*, 2: 21-42.

[14] Rockafellar, R.T., Uryasev, S., (2002) Conditional value-at-risk for general loss distributions. *J. Bank. Financ.* 26: 1443-1471.

[15] Rockafellar, R.T., Uryasev, S., (2013) The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surv. Oper. Res. Manag. Sci.* 18: 33-53.

[16] Ruszczyński, A., (2010) Risk-averse dynamic programming for Markov decision processes. *Math. Prog.* 125(2, Ser. B): 235-261.

[17] Ruszczyński, A., Shapiro, A., (2006) Conditional risk mappings. *Math. Oper. Res.* 31(3): 544-561.

[18] Ruszczyński, A., Shapiro, A., (2006) Optimization of convex risk functions. *Math. Oper. Res.* 31(3): 433-452.

[19] Shapiro, A., (2009) On a time consistency concept in risk averse multistage stochastic programming. *Oper. Res. Lett.* 37: 143-147.

[20] Shapiro, A., (2012) Time consistency of dynamic risk measures. *Oper. Res. Lett.* 40: 436-439.

[21] Shapiro. A., (2021) Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. *Eur. J. Oper. Res.* 288: 1-13.

[22] Shapiro, A., Dentcheva, D., Ruszczyński, A., (2021) *Lectures on Stochastic Programming: Modeling and Theory, 3rd ed.,* SIAM, Philadelphia, PA.