Planning and Learning in Average Risk-aware MDPs

Weikai Wang   Erick Delage

GERAD & HEC Montréal

Mila - Québec AI Institute

NeurIPS 2025

Motivations

- **Average (cost/reward) MDPs**
- **Risk-awareness** and **Dynamic risk measures**
- **Relative value iteration** (RVI) and **Q-learning** for **average risk-aware MDPs** (ARMDP)

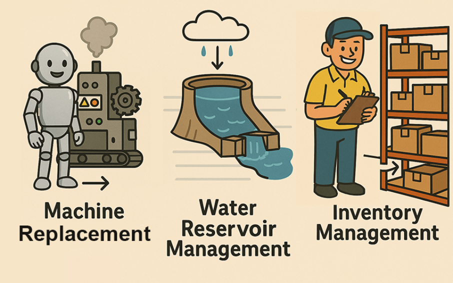

Figure: *Real-world continuing tasks benefiting from risk-aware strategies*

Our Contributions

- **Planning:** RVI algorithm for ARMDPs with general dynamic risk measures; proven convergence and optimality.
- **Learning:** Model-free Q-learning with multi-level Monte Carlo (MLMC); proven convergence and optimality.
- **UBSR Q-learning:** Off-policy Q-learning for utility-based shortfall risk (UBSR).
- **Experiments:** Validate analysis and demonstrate preference-aware policies in benchmark environments.

Average Risk-aware MDPs

Average cost MDP problem:

$$\bar{J}^* := \inf_{\boldsymbol{\pi}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} c^{\pi}(X_t)\right]. \tag{ACMDP}$$

T**-stage risk-aware total cost problem** with dynamic risk map $\mathcal{R}$:

$$J_T(\boldsymbol{\pi}) := c^{\pi_0}(X_0) + \mathcal{R}^{\pi_0}_{X_0}(c^{\pi_1}(X_1) + \cdots + \mathcal{R}^{\pi_{T-1}}_{X_{T-1}}(c^{\pi_T}(X_T))\cdots).$$

The infinite-horizon **average risk-aware MDP problem**:

$$J^* := \inf_{\boldsymbol{\pi}} J_\infty(\boldsymbol{\pi}) := \inf_{\boldsymbol{\pi}} \limsup_{T\to\infty} \frac{1}{T} J_T(\boldsymbol{\pi}). \tag{ARMDP}$$

Average Risk Optimality Equation

Theorem 2: (Theorem 5.9[1]) Under certain assumptions, there exists a unique $g^* \in \mathbb{R}$ and an $h^* \in \mathcal{L}(\mathcal{X})$ satisfying the **average risk optimality equation** (AROE):

$$g + h(x) = \min_{a \in \mathcal{A}} \{c(x,a) + \mathcal{R}_{x,a}(h)\}, \quad \forall x \in \mathcal{X}. \qquad \text{(AROE)}$$

Moreover, g^* solves the ARMDP, i.e., $g^* = J^* = J_\infty(\boldsymbol{\pi}^*)$ for a deterministic Markov policy π^*.

[1] Y. Shen, W. Stannat, and K. Obermayer, Risk-sensitive Markov control processes, *SIAM J. Control and Optim.*, 51(5): 3652–3672, 2013.

Average Risk-aware Relative Value Iteration

Risk-neutral RVI algorithm:

$$V_{n+1}(x) := \min_{a\in\mathcal{A}} \mathbb{E}\left[c(x,a) + V_n\right] - f(V_n), \quad \forall x \in \mathcal{X},$$

where f (resp. $\tilde{f}$) is some functional of value functions (resp. Q-factors) satisfying proper conditions (e.g. $f(V_n) = V_n(x_0)$).
Our **risk-aware RVI algorithm** replaces the expectation to a risk map: $\forall x \in \mathcal{X}$,

$$\begin{aligned} V_{n+1}(x) &= \min_{a\in\mathcal{A}} \mathcal{R}_{x,a}(c(x,a) + V_n) - f(V_n) \\ &=: \mathcal{G}(V_n)(x) - f(V_n). \end{aligned} \tag{1}$$

Average Risk-aware Relative Value Iteration

Risk-aware relative Q-factor iteration: $\forall (x,a) \in \mathcal{K}$,

$$\begin{aligned} Q_{n+1}(x,a) &= \mathcal{R}_{x,a}(c(x,a) + \min_{a' \in \mathcal{A}} Q_n(x,a')) - \tilde{f}(Q_n) \\ &=: \mathcal{H}(Q_n)(x,a) - \tilde{f}(Q_n), \end{aligned} \quad (2)$$

where $\mathcal{H}$ is called the **risk-aware Bellman optimality operator for Q-factors**.

Our Theorem 3.2 and 3.4 show that under certain conditions, the risk-aware RVI (1) and RQI (2) algorithms converge to a solution to the AROE, hence solves the ARMDP.

Average Risk-aware Q-learning

Average risk-aware Q-learning algorithm: if we can have an unbiased estimator for $\mathcal{H}$,

$$\begin{aligned} Q_{n+1}(x,a) =& Q_n(x,a) \\ &+ \gamma(n)\left(\hat{\mathcal{H}}(Q_n)(x,a) - \tilde{f}(Q_n) - Q_n(x,a)\right), \end{aligned} \tag{3}$$

where $\gamma(n)$ is some step size.
Our Theorem 4.5 shows that if $\hat{\mathcal{H}}$ is an unbiased estimator for $\mathcal{H}$, under certain assumptions, then almost surely, algorithm (3) converges to a solution to the AROE and the greedy policy converges to an optimal stationary policy to the ARMDP.

Constructing an Unbiased Estimator Using MLMC

One way of constructing an unbiased estimator $\hat{\mathcal{H}}$ is using the **Multilevel Monte Carlo** (MLMC) method.
Our Theorem 4.10 shows that, under certain conditions, the risk-aware MLMC Q-learning algorithm converges almost surely to a solution of the AROE for three classes of (possibly non-coherent) dynamic risk measures. This generalizes the result of Q-learning algorithm for average distributionally robust MDPs[2].

[2] Y. Wang, A. Velasquez, G. K. Atia, A. Prater-Bennette, and S. Zou, Model-free robust average-reward reinforcement learning, in *ICML*, 2023.

An Off-policy Q-learning Algorithm for UBSR

For UBSR, the AROE can be equivalently rewritten as a root finding problem: $\forall (x,a) \in \mathcal{K}$, for the loss function ℓ of UBSR,

$$\mathbb{E}\left[\ell\Big(c(x,a) + \min_{a' \in \mathcal{A}} q(\cdot, a') - f(q) - q(x,a)\Big)\right] = 0.$$

This motivates the following **UBSR Q-learning algorithm**:

$$\begin{aligned} Q_{n+1}(x,a) =& Q_n(x,a) + \gamma(n)\ell\Big(c(x,a) + \min_{a' \in \mathcal{A}} Q_n(x', a') \\ & - \tilde{f}(Q_n) - Q_n(x,a)\Big). \end{aligned} \tag{4}$$

No need of resampling in MLMC. However, the proof of convergence remains an open question.

Experiments

We evaluate our algorithms (1), (3), and (4) on a randomly generated MDP under the expectile risk measure (the only coherent case of UBSR) using the same amount of data.

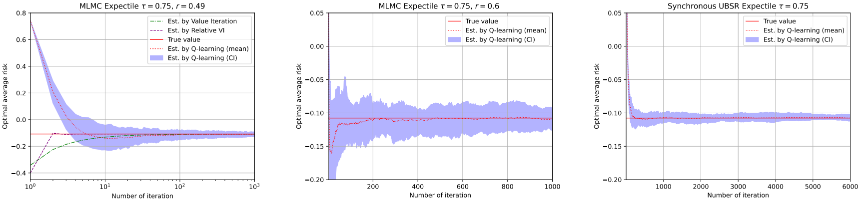

Figure: MLMC Q-learning with parameter $r = 0.49$ (log scale), $r = 0.6$, UBSR Q-learning.

Takeaways

- **Planning:** Risk-aware variant of RVI corresponds to average MDPs with dynamic risk measures.
- **Estimation:** MLMC yields an unbiased estimator for average risk-aware Bellman operators.
- **Learning:** UBSR Q-learning achieves higher efficiency than MLMC Q-learning.