Beyond CVaR: Leveraging Static Spectral Risk Measures for Enhanced Decision-Making in Distributional RL

Mehrdad Moghimi, Hyejin Ku

YORK U

Paper

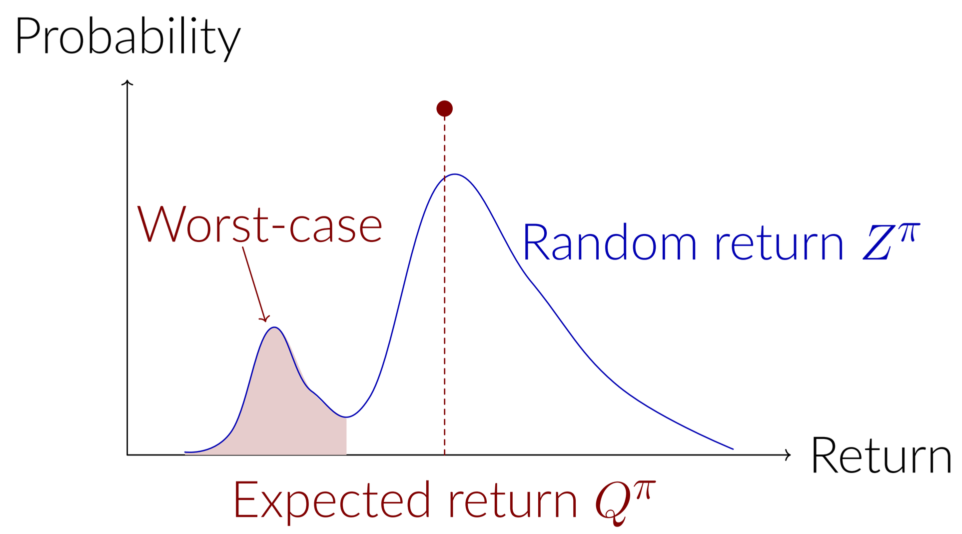
Code

ICML International Conference On Machine Learning

Motivation: Beyond Expected Return

In safety-critical domains (finance, robotics, healthcare, etc), maximizing the **expected return** is insufficient as it ignores rare but catastrophic outcomes.

- **Risk-Neutral RL:** Maximize expectation $\max_\pi \mathbb{E}[Z^\pi]$.
- **Risk-Sensitive RL:** Maximize a risk measure of the return $\max_\pi \rho(Z^\pi)$



Desirable Properties of a Risk Measure

The choice of risk measure ρ is critical. An ideal objective should possess:

- **Generality:** Expressiveness to capture diverse risk preferences beyond a single type (e.g., more than just CVaR).
- **Time-Consistency:** Ensures that an optimal plan remains optimal at all future decision points, avoiding self-contradictory actions. A policy $\pi^* = (a_0^*, \dots, a_T^*)$ is time-consistent if, for any $t = 1, \dots, T$, the shifted policy $\overrightarrow{\pi}^* = (a_t^*, \dots, a_T^*)$ is optimal for
$$\max_{\pi \in \boldsymbol{\pi}} \rho_{t,T}\left(Z_{t,T}^\pi\right)$$
- **Interpretability:** The agent's objective should be clear, and its evolving risk preferences at intermediate steps must be identifiable.

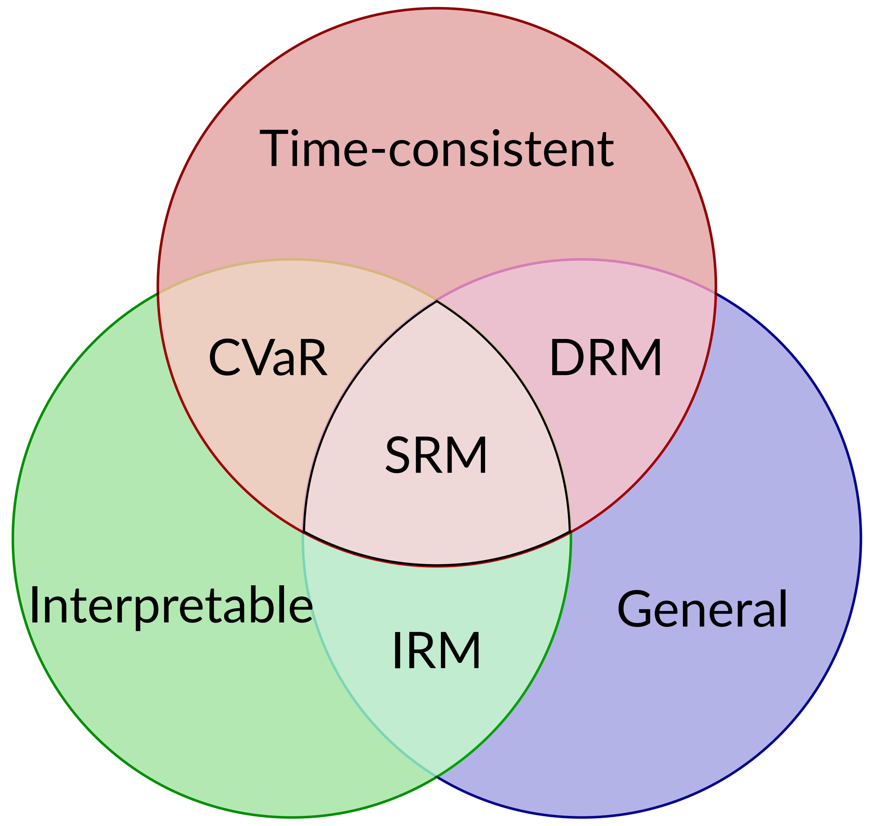
Existing methods often fail to satisfy all three properties simultaneously.

Our Approach: Static Spectral Risk Measures (SRMs)

We propose optimizing for **Spectral Risk Measures**, a general class of coherent risk measures that can be defined as weighted averages of CVaRs. Our method, **QR-SRM**, achieves all three desirable properties. An SRM is defined by a spectrum function ϕ or a probability measure μ:

$$\mathrm{SRM}_\phi(Z) = \int_0^1 F_Z^{-1}(u)\phi(u)\,\mathrm{d}u = \int_0^1 \mathrm{CVaR}_\alpha(Z)\,\mu(\mathrm{d}\alpha)$$

- Dynamic Risk Measures (DRMs) are time-consistent but hard to interpret.
- CVaR is interpretable but lacks generality.
- Combining general risk measures naively with distributional RL (referred to as Iterative Risk Measures or IRM) results in general and interpretable, but time-inconsistent solutions.



Policy Optimization

We leverage the dual representation of SRMs, which separates the optimization into two alternating steps:

$$\max_\pi \mathrm{SRM}_\phi(G^\pi) = \max_{h \in \mathcal{H}} \left(\max_\pi \mathbb{E}\left[h(G^\pi)\right] + \int_0^1 \hat{h}(\phi(u))\,\mathrm{d}u \right)$$

1. **Inner Optimization:** For a fixed function h, solve a distributional RL problem on an MDP with an extended state space $(\mathcal{X} \times \mathcal{S} \times \mathcal{C})$ to find the optimal policy. The distributional Bellman operator and the greedy action are:
$$G_{k+1,l}(x,s,c,a) \overset{\mathcal{D}}{=} R(x,a) + \gamma\, G_{k,l}(x', s', c', \mathrm{a}_{k,l}(x', s', c'))$$
$$\mathrm{a}_{G,h}(x,s,c) = \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[h\left(s + c\,G(x,s,c,a)\right)\right]$$
2. **Outer Optimization:** Use the learned return distribution G to update the function h in closed-form, refining the objective.
$$h_{\mu,Z}(z) = \int_0^1 F_Z^{-1}(\alpha) + \frac{1}{\alpha}\left(z - F_Z^{-1}(\alpha)\right)^- \mu(\mathrm{d}\alpha)$$

Convergence: Suppose $J(\pi, h) = \mathbb{E}\left[h(G^\pi)\right] + \int_0^1 \hat{h}(\phi(u))\mathrm{d}u$. If $\pi_{k,l}$ denotes the greedy policy extracted from $G_{k,l}$ and h_l, then for all $x \in \mathcal{X}, s \in \mathcal{S}, c \in \mathcal{C}$, and $a \in \mathcal{A}$,

$$J(\pi_{k,l}, h_l) \geq \max_{\pi \in \boldsymbol{\pi}} J(\pi, h_l) - \phi(0)c\gamma^{k+1} G_{\mathrm{MAX}}$$

Additionally, $J(\pi_l^*, h_l)$ is bounded and monotonically increases as l increases and provides a lower bound for our objective.

Time-Consistent Interpretation

Decomposition Theorem (Pflug & Pichler 2016): A law-invariant and coherent risk measure ρ has the following decomposition

$$\rho(Z) = \sup_{\tilde{\xi}} \mathbb{E}\left[\tilde{\xi} \cdot \rho_{\tilde{\xi}}(Z \mid \mathcal{F}_t)\right]$$

where the supremum is among all feasible non-negative $\mathcal{F}_t$-measurable random variables satisfying $\mathbb{E}\left[\tilde{\xi}\right] = 1$. Moreover, if ξ^α is the optimal dual variable to compute the CVaR at level α, i.e. $\mathbb{E}\left[\xi^\alpha Z\right] = \mathrm{CVaR}_\alpha(Z)$ and $0 \leq \xi^\alpha \leq 1/\alpha$, $\xi_t^\alpha = \mathbb{E}\left[\xi^\alpha \mid \mathcal{F}_t\right]$, and $\xi = \int_0^1 \xi_t^\alpha \mu(\mathrm{d}\alpha)$, the conditional risk measure is given by

$$\rho_\xi(Z \mid \mathcal{F}_t) = \int_0^1 \mathrm{CVaR}_{\alpha\xi_t^\alpha}(Z \mid \mathcal{F}_t) \frac{\xi_t^\alpha \mu(\mathrm{d}\alpha)}{\int_0^1 \xi_t^\alpha \mu(\mathrm{d}\alpha)}.$$

Theorem: For any SRM defined with probability measure μ, if ξ^α is the optimal dual variable to compute the CVaR at level α, i.e. $\mathbb{E}\left[\xi^\alpha G\right] = \mathrm{CVaR}_\alpha(G)$, $\lambda_\alpha = F_G^{-1}(\alpha)$ and F_{G_t} is the CDF of G_t, we can calculate $\xi_t^\alpha = \mathbb{E}\left[\xi^\alpha \mid \mathcal{F}_t\right]$ with:

$$\xi_t^\alpha = F_{G_t}\left(\frac{\lambda_\alpha - s_t}{c_t}\right)/\alpha$$

and derive the risk level and the weight of CVaRs, at a later time step with $\alpha\xi_t^\alpha$ and $\xi_t^\alpha \mu(\mathrm{d}\alpha) / \int_0^1 \xi_t^\alpha \mu(\mathrm{d}\alpha)$.

Intermediate Risk Preferences: An Example

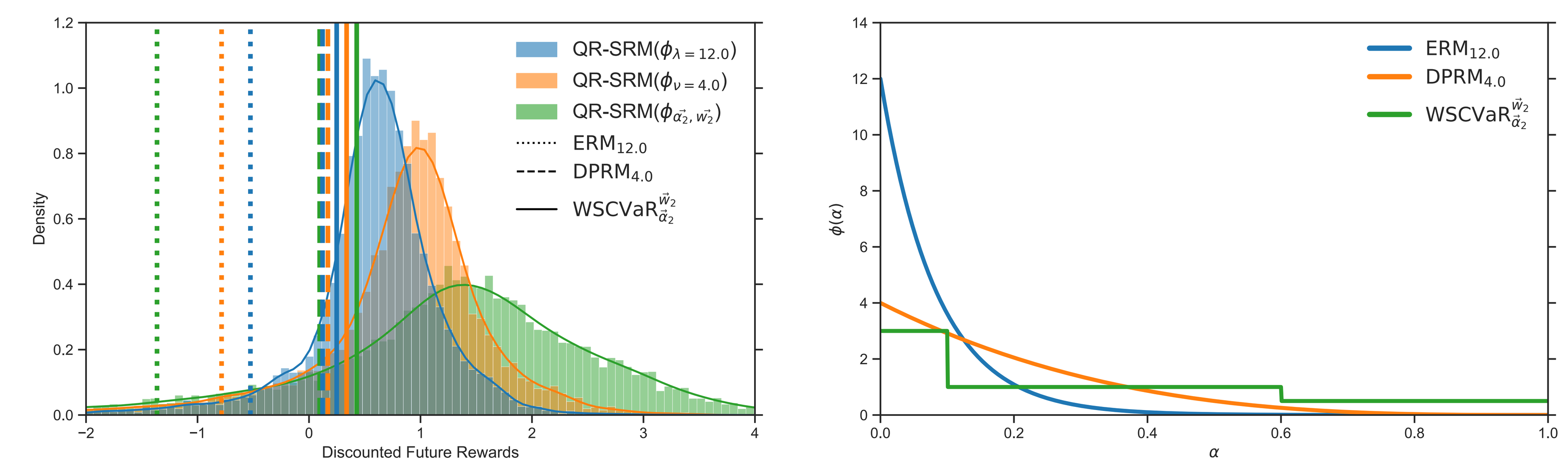
- Assume the optimal policy is for:
$$\rho(G) = 0.6 \cdot \mathrm{CVaR}_{0.25}(G) + 0.4 \cdot \mathrm{CVaR}_{0.8}(G)$$
- At time t, let $s_t = 5$ and $c_t = 0.8$
- $\lambda_{0.25} = 12 \quad \Rightarrow \quad 0.25 \cdot \xi_t^{0.25} = F_{G_t}((12-5)/0.8) = \mathbf{0.3} \Rightarrow \xi_t^{0.25} = 0.3/0.25 = \mathbf{1.2}$
- $\lambda_{0.8} = 39 \quad \Rightarrow \quad 0.8 \cdot \xi_t^{0.8} = F_{G_t}((39-5)/0.8) = \mathbf{1.0} \Rightarrow \xi_t^{0.8} = 1.0/0.8 = \mathbf{1.25}$
- Normalizing factor: $\xi = 0.6 \cdot 1.2 + 0.4 \cdot 1.25 = 1.22$
- Therefore, at time t, the policy is effectively optimized for:
$$\rho_\xi(G_t) = \frac{0.6 \cdot 1.2}{1.22} \cdot \mathrm{CVaR}_{0.3}(G_t) + \frac{0.4 \cdot 1.25}{1.22} \cdot \mathrm{CVaR}_{1.0}(G_t)$$

$\hat{\tau}$	$q_{\hat{\tau}}$	$\xi^{0.25}$	$\xi^{0.8}$	$q_{\hat{\tau},t}$
5%	7	4	1.25	5
15%	9	4	1.25	6
25%	12	2	1.25	8
35%	20	0	1.25	14
45%	21	0	1.25	15
55%	27	0	1.25	17
65%	30	0	1.25	21
75%	32	0	1.25	25
85%	39	0	0	28
95%	46	0	0	35

Experiment: Mean-Reversion Trading

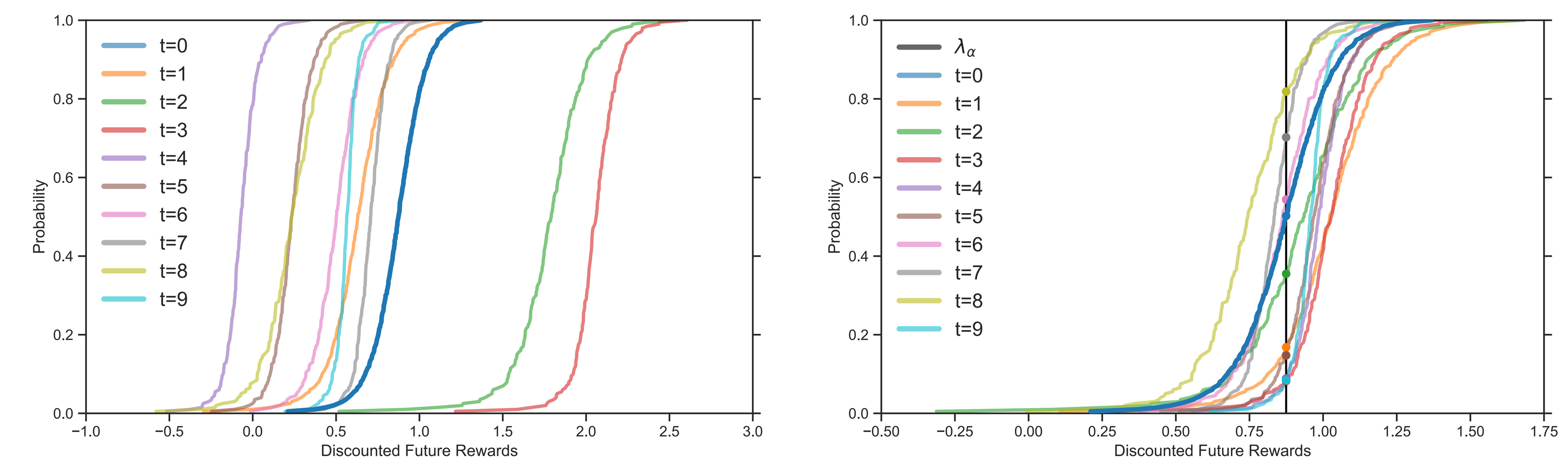
We evaluate QR-SRM in an algorithmic trading task where the asset price follows an Ornstein-Uhlenbeck process. The agent learns a policy to buy or Tsell assets to maximize a risk-adjusted return. We test a variety of complex risk measures beyond CVaR, including:

WSCVaR: $\phi_{\vec{\alpha},\vec{w}}(u) = \sum_i w_i \frac{1}{\alpha_i} 1_{[0,\alpha_i]}(u)$, ERM : $\phi_\lambda(u) = \frac{\lambda e^{-\lambda u}}{1 - e^{-\lambda}}$, DPRM: $\phi_\nu(u) = \nu(1-u)^{\nu-1}$



Visualizing Time-Consistent Interpretation

For a policy trained on $\mathrm{CVaR}_{0.5}$, we see how the initial risk level $\lambda_{0.5}$ (vertical line) maps to a different α−quantile of the aligned future return distribution at each step. This explicitly visualizes the evolving, time-consistent risk objective in action.



moghimi@yorku.ca

hku@yorku.ca