ICML 2026

RiskZero: Plan More to Risk Less

with a Learned Model

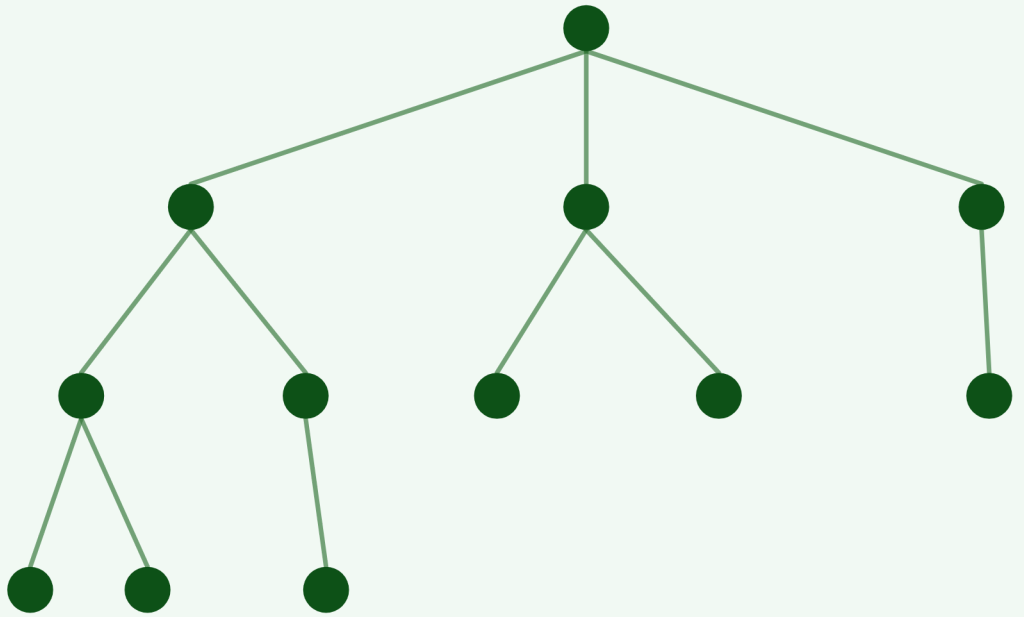
Planning-based Risk Aversion with MuZero-family algorithms.

Yousef Yassin · Junfeng Wen

🍁 Carleton University 🐦‍⬛

Faced with the unknown, humans plan.

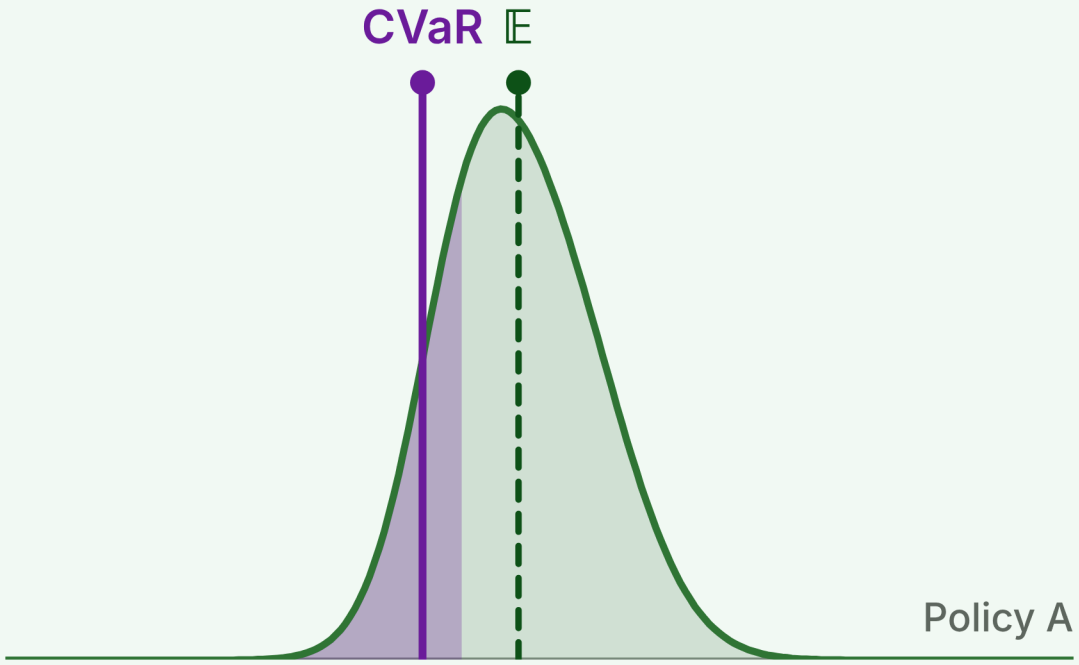
- Consider options + anticipate futures (right) → *behaviour*.
- Planning *agents* have also achieved remarkable success.
- **AlphaZero:** search + deep RL → superhuman performance.
- **MuZero:** generalized this idea with a *learned* world model.
- Restricted to maximizing ***average*** return.
- But *average (expectation* $\mathbb{E}$*)* can fail!

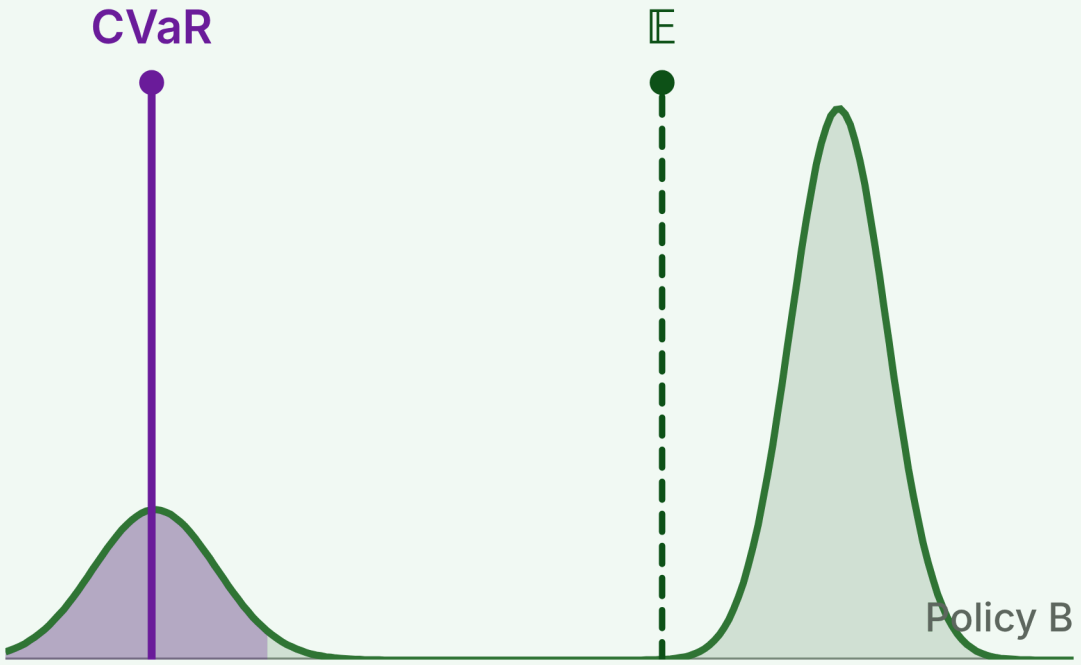


Monte Carlo Tree Search

Expectation can fail.

- Average = *risk-neutral*. Hides danger.
- Assumes recovery possible + multiple chances.
- Humans tend to prefer *reliability*.
- **Risk-Sensitive RL (RSRL):** optimize alternative criteria.
- **Conditional Value at Risk (CVaR)**: account for frequency + severity of rare and adverse events.

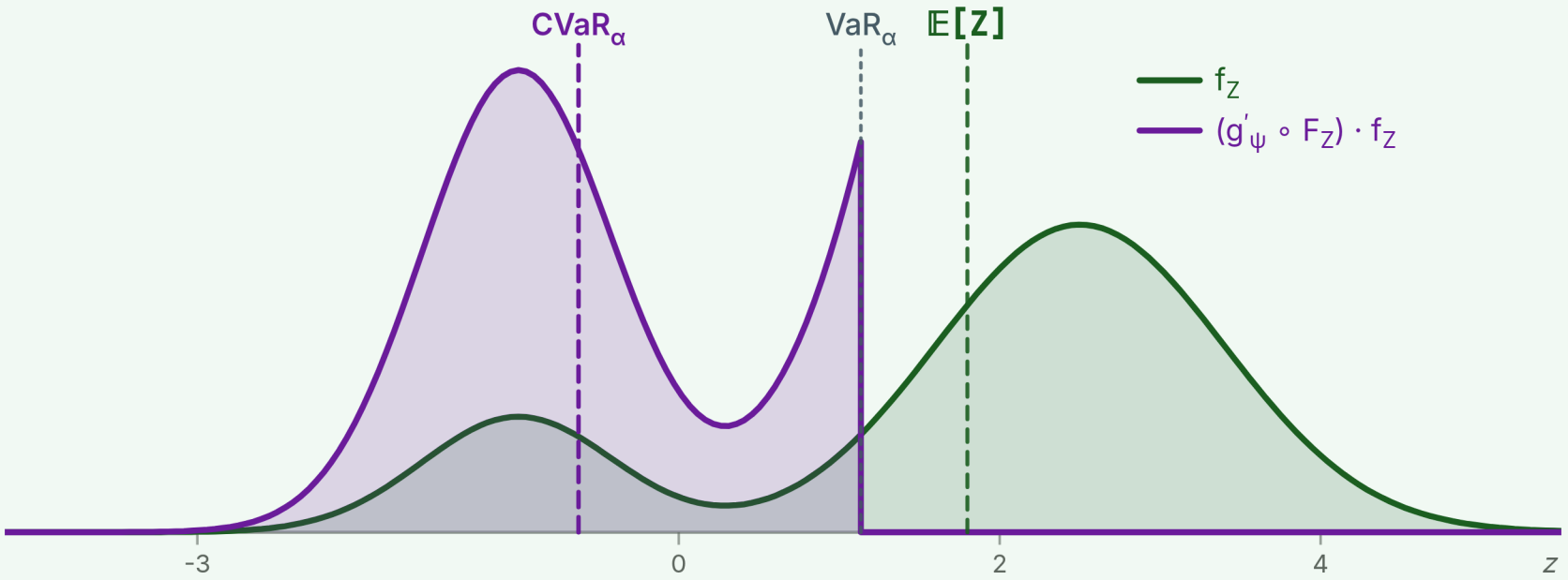




Risk-Sensitivity by Distortion.

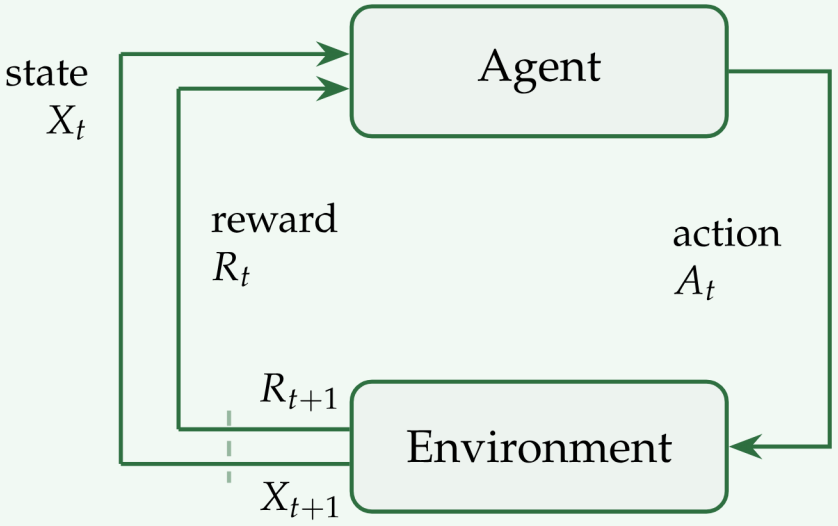
- Risk-sensitive criteria are *harder to optimize*.
- **Distributional RL** facilitates risk estimation via **distortion**.
- Simply optimizing for *future* risk → *biased*.
- Must account for what occurred in the *past* to act for the future.
- Optimization must be at the **trajectory level** → *sample-inefficient*.

$$\psi[Z] = \int_{-\infty}^{\infty} z\, g'_{\psi}(F_Z(z))\, f_Z(z)\, dz \qquad g_{\alpha\text{-CVaR}}(\omega) = \min\left\{\tfrac{\omega}{\alpha},\, 1\right\}, \qquad \alpha \in [0, 1]$$



Notation.

- MDP with finite horizon T
- state space $\mathcal{X}$, discrete action space $\mathcal{A}$
- random rewards $R(x, a)$, discount factor $\gamma \in [0, 1]$
- deterministic dynamics $x_{t+1} = M(x_t, a_t)$
- **Trajectory:** $\tau = x_0 a_0 x_1 \dots a_{T-1} x_T \in \mathcal{T}$
- **History:** $h_t = x_0 a_0 x_1 \dots a_{t-1} x_t \in \mathcal{H}$
- We write $h_t a \equiv h_t a x', \quad x' = M(x_t, a)$



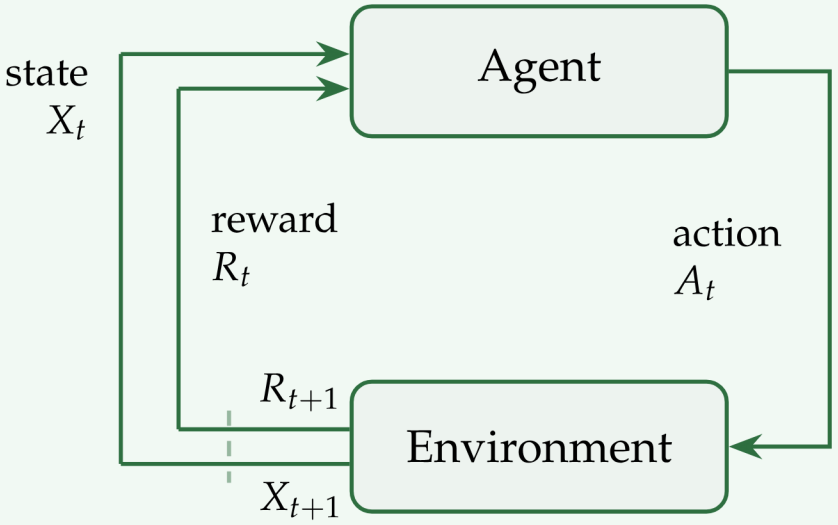
Notation and Objective.

- Agent acts w/ policy $\pi : \mathcal{H} \to \Delta_{\mathcal{A}}$
- **Return:** $Z^{\pi}(x_t) = \sum_{i \geq 0} \gamma^i R(x_{t+i}, a_{t+i})$
- **Standard RL:** $\max_{\pi} \mathbb{E}[Z^{\pi}(x_0)]$

(Our) RSRL Objective

$$\max_{\pi} \psi[Z^{\pi}(x_0)]$$

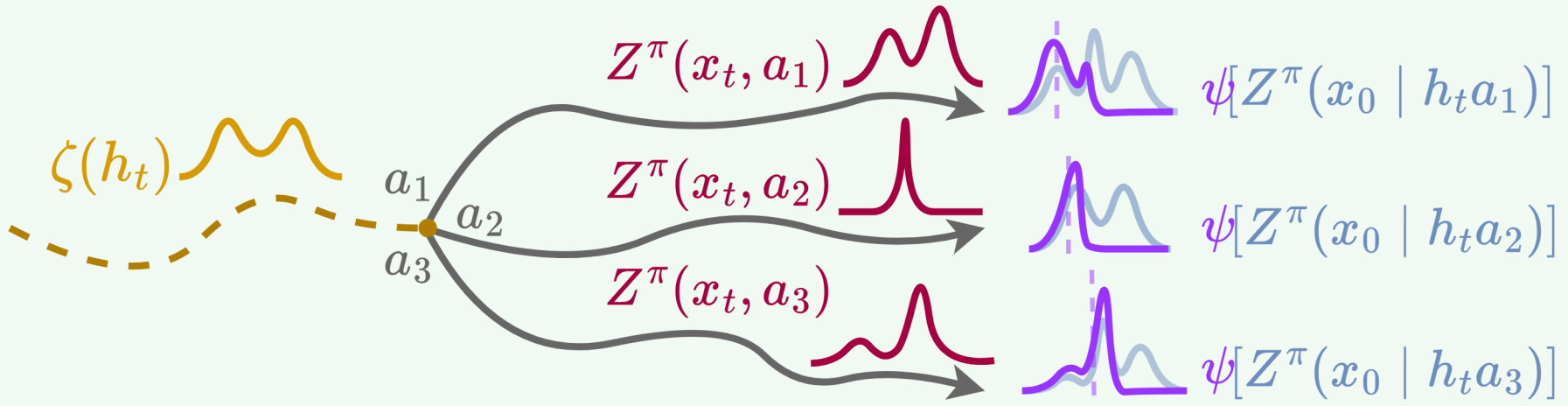
(restricted to *stationary* policies; refer to paper)



Risk-Sensitive Policy Improvement.

- Unbiased optimization requires the *trajectory*.
- **Trajectory Return:** $Z^\pi(x_0 \mid h_t) \overset{D}{:=} \sum_{i=0}^{t-1} \gamma^i R(x_i, a_i) + \gamma^t Z^\pi(x_t)$
- **Trajectory Risk:** $v_\psi^\pi(\tau) := \psi[Z^\pi(x_0 \mid \tau)]$
- Policy induces a distribution over τ with probabilities $d^\pi(\tau)$.
- **Expected Risk:** $v_\psi^\pi := \sum_\tau d^\pi(\tau) v_\psi^\pi(\tau)$
- **Insight:** optimizing true risk reduces to optimizing expected risk — just like regular RL!

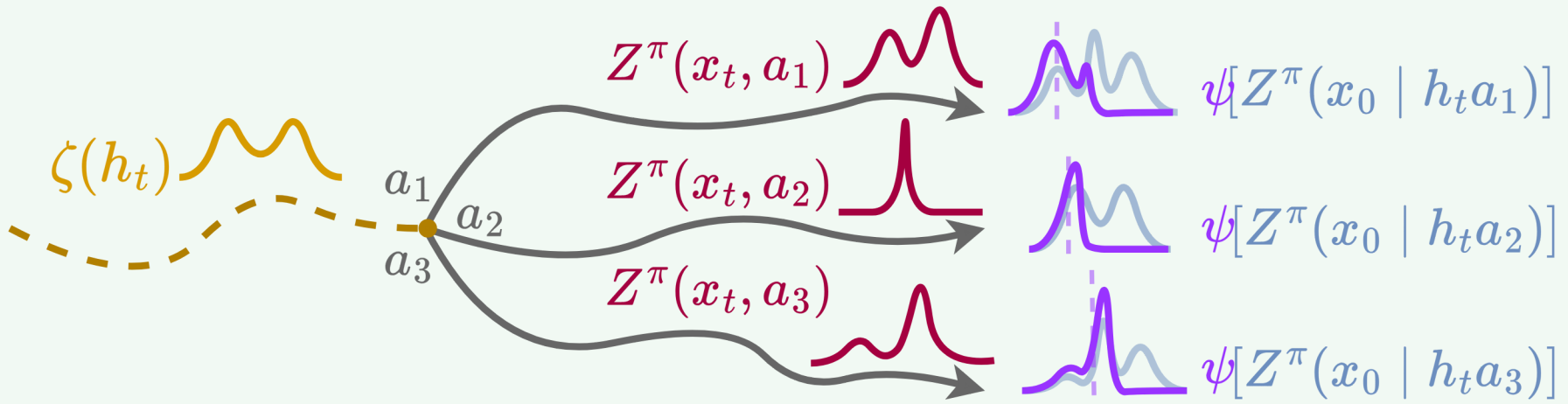
> **Proposition 3.3.** Let ψ be a coherent risk measure and Π_H the class of history-stationary policies. A deterministic policy $\pi^* \in \Pi_H$ is optimal in expected risk, $\pi^* \in \arg\max_{\pi \in \Pi_H} v_\psi^\pi(x_0)$, if and only if it is optimal in true risk, $\pi^* \in \arg\max_{\pi \in \Pi_H} \psi[Z^\pi(x_0)]$.



Optimizing Risk by Gumbel Planning.

- Use *Gumbel MuZero* as a surrogate optimizer with softmax policy π (logits θ).
- **Risk (action-)values:** $q_\psi^\pi(h,a) := v_\psi^\pi(ha), \quad v_\psi^\pi(h) := \sum_a \pi(a \mid h)\, q_\psi^\pi(h,a)$
- Sample $m < |\mathcal{A}|$ actions at root via gumbel $g \in \mathbb{R}^{|\mathcal{A}|}$ and Gumbel Top-K trick.
- MCTS with sequential halving to estimate q_ψ^π at root.
- **Non-root:** improved policy $\pi' = \mathrm{softmax}(\theta + \sigma(\widetilde{Q}_\psi))$, where $\widetilde{Q}_\psi(a) = \hat{q}_\psi^\pi(a)$ if visited, else $\hat{v}_\psi^\pi$.
- **Distill:** $\theta^{(t+1)} = \theta^{(t)} + \xi\, \sigma\left(\widetilde{Q}_\psi^{(t)}\right), \quad \xi > 0$
- **Root:** $a_t = \arg\max_{a \in \mathcal{A}_{\text{top-}m}} \left[g(a) + \theta_a + \sigma\left(\hat{q}_\psi^\pi(a)\right) \right]$
- Guarantees **policy improvement** $v_\psi^{\pi'}(x_0) \geq v_\psi^\pi(x_0)$.

> **Theorem 3.1.** Under a softmax policy with positive linear isomorphism σ and every history seen with positive probability, for all h: $v_\psi^{(t)}(h) \to v_\psi^*(h)$ as $t \to \infty$.

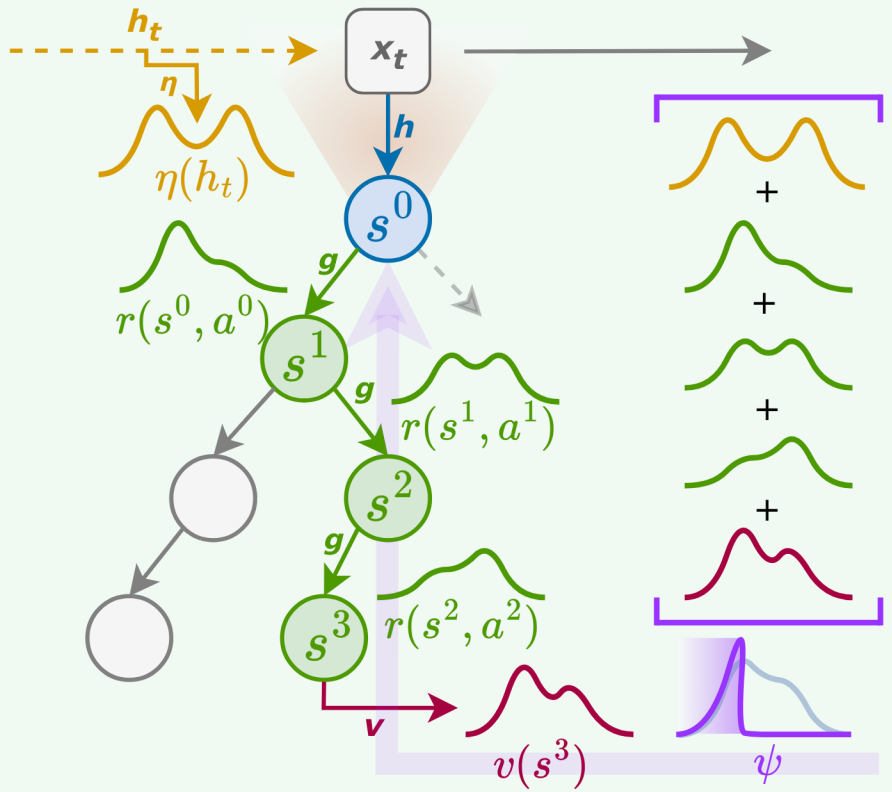


RiskZero.

Planning-based Risk-Aversion with a Learned Model

- Decompose trajectory return:
  accumulated return + ***imm. reward*** + ***future return***.
- Search in abstract MDP *(à la MuZero)*:
  - **representation**: $h_\phi : \mathcal{H} \to \mathcal{S}$ (maps h_t to abstract latent s_t^0)
  - **dynamics**: $g_\phi : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ (predicts abstract successor state)
  - **reward**: $r_\phi : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathbb{R}}$ (distributional)
  - **value**: $v_\phi : \mathcal{S} \to \Delta_{\mathbb{R}}$ (distributional)
  - **historical return**: $\eta_\phi : \mathcal{H} \to \Delta_{\mathbb{R}}$ (new — distributional)
  - **prior policy**: $\pi_\phi : \mathcal{S} \to \Delta_{\mathcal{A}}$ (maps abstract state to action probabilities)
- Sample trajectory return by inverse transform + summing:

$$Z(x_0 \mid h_{t+n}) \overset{D}{=} \zeta(h_t) + \sum_{i=0}^{n-1} \gamma^{t+i} R(x_i, a_i) + \gamma^{t+n} Z(x_{t+n})$$

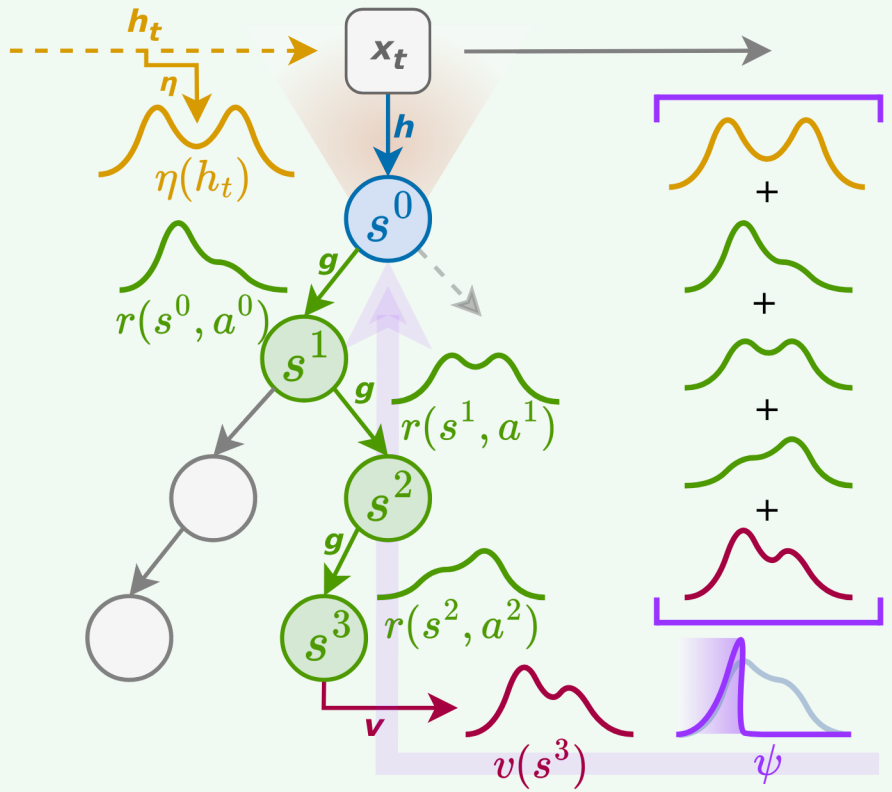


RiskZero.

Risk Sensitive Search

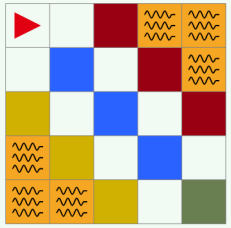
- **Initialize:** root latent $s_t^0 = h_\phi(x_{\leq t})$, hist. return $\eta_t = \eta_\phi(x_{\leq t}, a_{<t})$

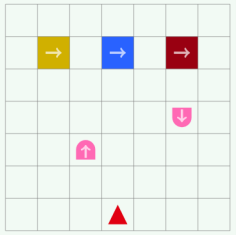
1. ***Selection.*** Draw D samples $\eta_{t,1:D} \sim \eta_t$
   - Descend tree, sample $r^k_{t,1:D} \sim r_\phi(s_t^k, a_t^k)$ until leaf $s_t^{\ell-1}$
2. ***Expansion.*** Sample $r^{\ell-1}_{t,1:D}$ and $v^\ell_{t,1:D} \sim v_\phi(s_t^\ell)$
   - Combine D trajectory samples:
     $z_j(x_0 \mid h_t^\ell) = \eta_{t,j} + \sum_{k=0}^{\ell-1} \gamma^{t+k} r^k_{t,j} + \gamma^{t+\ell} v^\ell_{t,j}$
     for $j \in [D]$, where $h_t^\ell := h_t a_t^1 s_t^1 \cdots a_t^{\ell-1} s_t^\ell$
3. ***Backup.*** Sort z_j (empirical quantiles), average by g_ψ^{-1}
   - Approximates $\psi[Z^\pi(x_0 \mid h_t^\ell)]$ by inverse transform sampling
   - Backup along path for estimates $\hat{q}^\pi_\psi(h, a)$

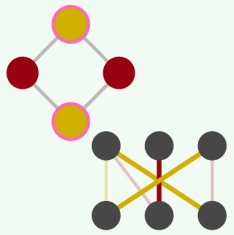


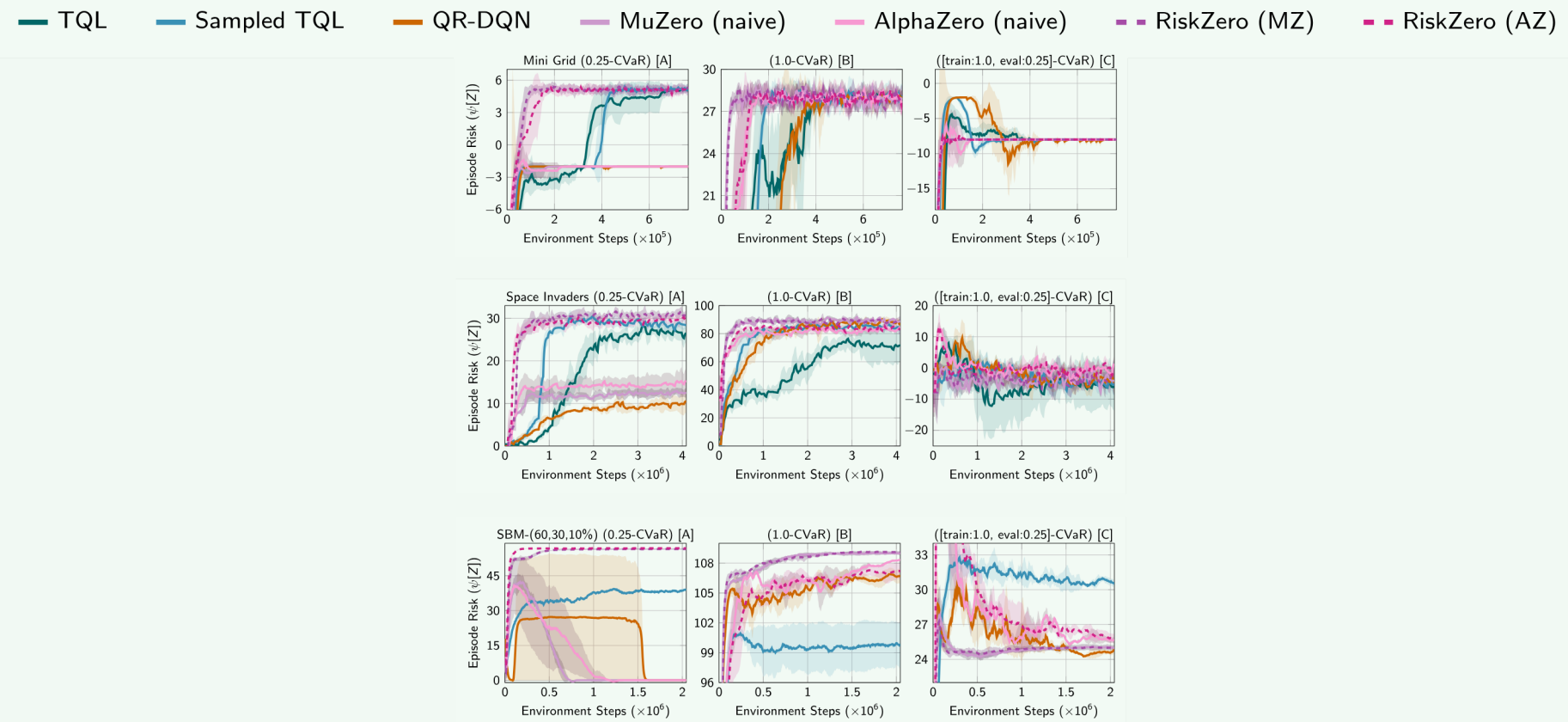
Experiments.

*RiskZero learns optimal risk-sensitive policies **efficiently**.*



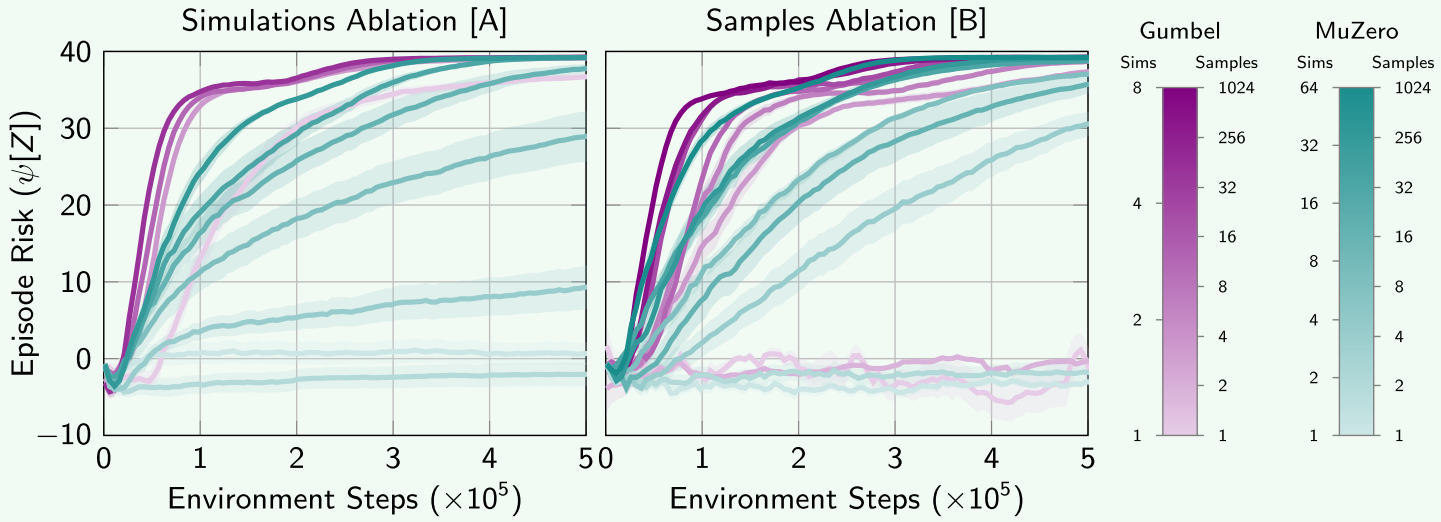


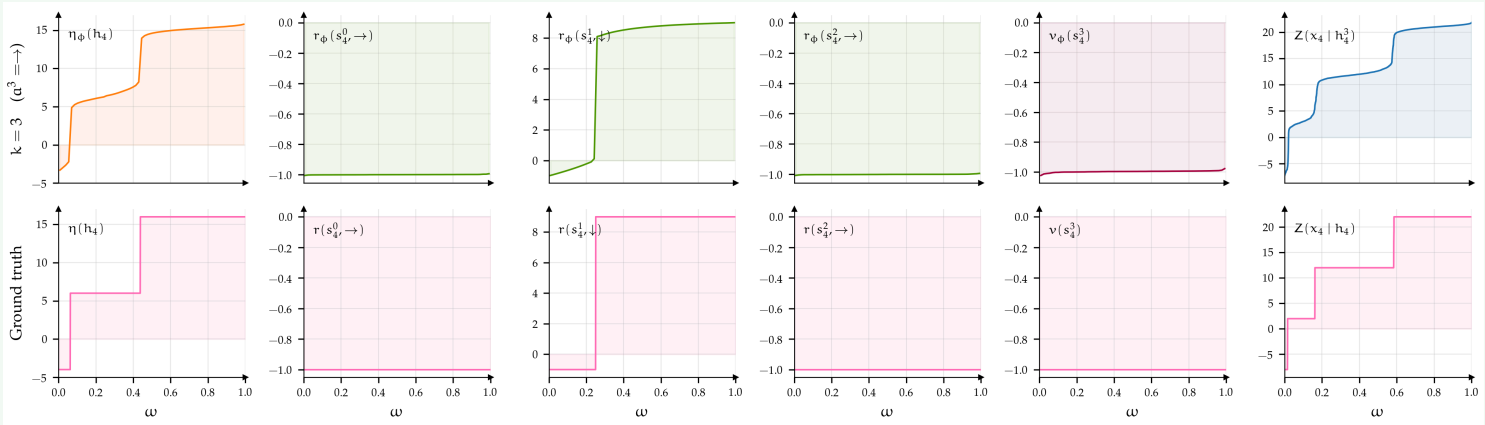




Ablations.

- ***Gumbel planning*** outperforms heuristic MCTS (↑ sample efficiency, ↓ quantile samples).
- *RiskZero* and inverse transform sampling recover the correct quantile distributions.



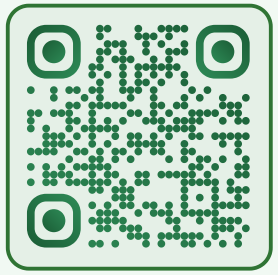


Conclusion.

- ***RiskZero***: first *MuZero*-family method for planning-based risk-aversion with a learned model.
- Achieved by learning and integrating trajectory-level distributions with Gumbel MCTS.
- Principled risk-sensitive planning with diverse coherent risk-measures.
- Provably converges to optimal risk-sensitive policies — theoretically and empirically.

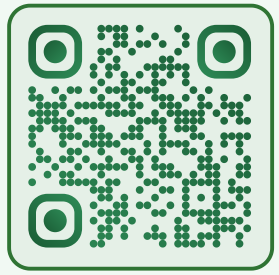
Limitations and Future Work

- Focus on ***stationary*** policies, in ***deterministic*** dynamics + ***discrete*** action spaces.
- Extension to stochastic dynamics + non-stationary policies!
- State-augmentation + ***Stochastic MuZero*** is a promising direction.
- Benefit from recent advances in **deep search** for:
  - robustness (***RobustZero***)
  - continuous action spaces (***Sampled MuZero***)
  - efficiency (***EfficientZero***)



🔗 paper

🔗 code



🔗 (better) slides

Thank You! ❤️