Measures of Variability for Risk-averse Policy Gradient

Yudong Luo[1,4], Yangchen Pan[2], Jiaqi Tan[3], Pascal Poupart[1,4]

yudong.luo@hec.ca

[1]**University of Waterloo,** [2]**University of Oxford,**
[3]**Simon Fraser University,** [4]**Vector Institute**

Table of Contents

Reinforcement Learning (RL)

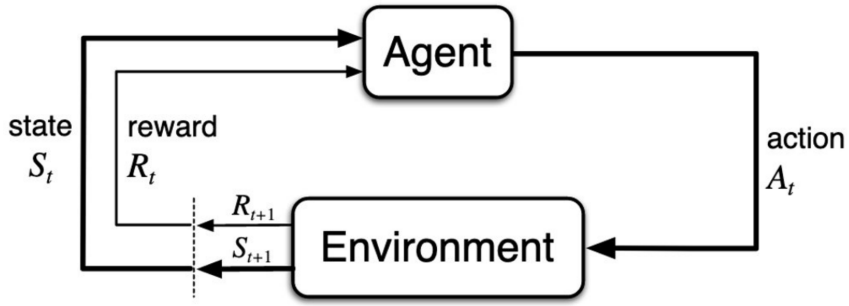


Figure 1: Markov Decision Process (MDP)

An agent interacts with environment using its policy $\pi(a|s)$.

- $\pi(a|s)$: mapping from state to action $\mathcal{S} \to \mathcal{A}$
- Stochastic policy: $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$

By interaction, a trajectory $\tau = (S_0, A_0, R_1, S_1, A_1, R_2, ...)$

- Total return random variable $G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + ...$

Reinforcement Learning (RL)

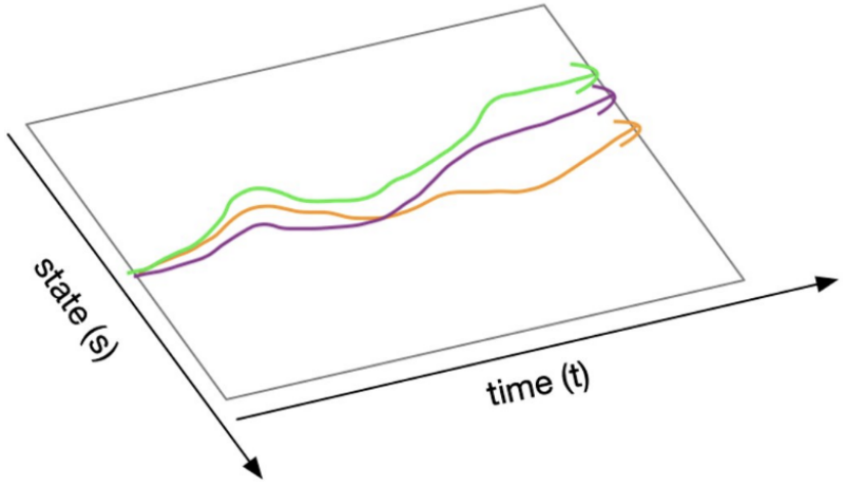


Figure 2: Random trajectories

- Traditional (Risk-neutral) RL: $\max_\pi \mathbb{E}[G_0]$
- Risk-averse RL (RARL): optimize $f(G_0)$, where f is a risk metric
  - Risk measures
  - Measures of variability

Risk Measures V.S. Measures of Variability

(Focus on) Risk Measures

- Value at Risk (VaR) (Chow et al., 2018)
- Conditional VaR (CVaR) (Bäuerle and Ott, 2011; Chow and Ghavamzadeh, 2014; Chow et al., 2015; Greenberg et al., 2022; Luo et al., 2024)
- Entropic risk measure (Fei et al., 2021; Hau et al., 2023)
- Expectile (Rowland et al., 2019; Marzban et al., 2023)

Measures of Variability (under explored)

- Variance (Tamar et al., 2012; La and Ghavamzadeh, 2013; Xie et al., 2018; Bisi et al., 2020; Zhang et al., 2021)
- Variance-related: STD (Yang et al., 2021), semi_STD (Tamar et al., 2015), semi_Variance (Ma et al., 2022)

Risk Measures V.S. Measures of Variability

Risk Measures: $\rho : \mathcal{X} \to (-\infty, +\infty]$

- (A) Law invariance: if $X \overset{d}{=} Y$, then $\rho(X) = \rho(Y)$ for all $X, Y \in \mathcal{X}$
- (A1) Positive homogeneity: $\rho(cX) = c\rho(X)$ for all $c > 0$ and $X \in \mathcal{X}$.
- (A2) Sub-additivity: $\rho(X + Y) \leq \rho(X) + \rho(Y)$ for all $X, Y \in \mathcal{X}$.
- (B1) Monotonicity: $\rho(X) \leq \rho(Y)$ if $X, Y \in \mathcal{X}$ and $X \leq Y$ $\mathbb{P}$-almost surely.
- (B2) Translation invariance: $\rho(X - c) = \rho(X) - c$ for all $c \in \mathbb{R}$ and $X \in \mathcal{X}$.

A risk measure is **coherent** if it satisfies (A1), (A2), (B1) and (B2) (Artzner et al., 1999).

Risk Measures V.S. Measures of Variability

Measures of Variability ν: $\mathcal{X} \to [0, \infty]$.

- (C1) Standardization: $\nu(m) = 0$ for all $m \in \mathbb{R}$.
- (C2) Location invariance: $\nu(X - m) = \nu(X)$ for all $m \in \mathbb{R}$ and $X \in \mathcal{X}$.

A measure of variability is **coherent** if it satisfies (C1), (C2), (A1) Positive homogeneity and (A2) Sub-additivity (Furman et al., 2017).

Risk Measures V.S. Measures of Variability

Risk Measures and Measures of Variability are also related.

- Rockafellar et al. (2006) demonstrated a one-to-one correspondence between deviation measures and "strictly expectation bounded risk measures".
- Mean-Semi_STD is a coherent risk measure (Tamar et al., 2015).

Table of Contents

Our Contribution

Policy gradient derivation and comparison of 9 measures of variability following seminal works of David (1998); Rockafellar et al. (2006). 4 metrics have not been previously studied in RARL.

- Variance
- Gini Deviation
- Mean Deviation
- Mean-Median Deviation
- Standard Deviation
- Inter-Quantile Range
- CVaR Deviation
- Semi_Variance
- Semi_STD

Problem:

$$\max_{\pi_\theta} \mathbb{E}[G_0] - \lambda \mathbb{D}[G_0] \tag{1}$$

pdf of X: $f_X(x;\theta)$, compute $\nabla_\theta \mathbb{D}[X]$

Metrics

Gini Deviation

$$\mathrm{GD}[X] = \frac{1}{2}\mathbb{E}[|X - X^*|] \;\; (X^* \text{ is i.i.d. copy of } X) \tag{2}$$

- Representation of variance could be misleading (based on the center of the underlying distribution) (Gini, 1912).
- Share some properties with Variance (Yitzhaki et al., 2003), e.g., consistent with convex order, represented by ordered statistics.

Variance

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{1}{2}\mathbb{E}[(X - X^*)^2] \;\; (X^* \text{ is i.i.d. copy of } X) \end{aligned} \tag{3}$$

Metrics

Mean Deviation

$$\mathrm{MD}[X] = \mathbb{E}[|X - \mathbb{E}[X]|] \tag{4}$$

- Portfolio management, since the problem can be reduced to a linear programming problem, instead of a quadratic programming in Variance.

Mean-Median Deviation

$$\mathrm{MMD}[X] = \mathbb{E}[|X - \mathrm{Median(X)}|] = \min_{x \in \mathbb{R}} \mathbb{E}[|X - x|] \tag{5}$$

- median is the minimum value of the L1 estimate.
- alternative to MD, since the median is more robust to outliers and skewed distributions.

Metrics

Inter Quantile Range

$$\mathrm{IQR}_\alpha[X] = F_X^{-1}(\alpha) - F_X^{-1}(1-\alpha), \alpha \in [\frac{1}{2}, 1) \tag{6}$$

- When $\alpha \to 1$, IQR recovers the full range of X.
- Plays a key role in the construction of a box plot (Spitzer et al., 2014)

CVaR Deviation

$$\mathrm{CD}[X] = \mathbb{E}[X] - \mathrm{CVaR}_\alpha^\vee(X) \tag{7}$$

- Deviation of the left tail value from the mean
- Mean-CVaR criteria is widely used in portfolio management (Yao et al., 2013) and also RL (Ying et al., 2022) to avoid potential losses.

Policy Gradient

Gini Deviation: $\mathrm{GD}[X] = \frac{1}{2}\mathbb{E}[|X - X^*|]$

Mean-Median Deviation: $\mathrm{MMD}[X] = \mathbb{E}[|X - \mathrm{Median(X)}|] = \min_{x\in\mathbb{R}} \mathbb{E}[|X - x|]$

Signed Choquet integral (Wang et al., 2020)

$$\Phi[X] = \int_0^1 F_X^{-1}(1-\alpha)d\,\boxed{h(\alpha)} \tag{8}$$

Gini Deviation: $h(\alpha) = -\alpha^2 + \alpha$

Mean-Median Deviation: $h(\alpha) = \min\{\alpha, 1-\alpha\}$

Policy Gradient (Gini Deviation)

Assume X is continuous, bounded $[-b, b]$, $\frac{\partial}{\partial \theta_i} q_\alpha(X;\theta)$, $\frac{\partial f_X(x;\theta)}{\partial \theta_i} / f_X(x;\theta)$

$$\nabla_\theta \mathrm{GD}[X] = \int_0^1 (2\alpha - 1) \nabla_\theta F_X^{-1}(\alpha) d\alpha = \int_0^1 (2\alpha - 1) \boxed{\nabla_\theta q_\alpha(X;\theta)} d\alpha. \tag{9}$$

Gradient of quantile

$$\boxed{\nabla_\theta q_\alpha(X;\theta)} = -\int_{-b}^{q_\alpha(X;\theta)} \nabla_\theta f_X(x;\theta) dx \cdot \left[f_X\left(q_\alpha(X;\theta);\theta\right) \right]^{-1}. \tag{10}$$

Then

$$\nabla_\theta \mathrm{GD}[X] = -\mathbb{E}_{x\sim X}\Big[\nabla_\theta \log f_X(x;\theta) \big(b + x - 2\mathbb{E}[\max\{X, x\}]\big) \Big]. \tag{11}$$

Unbiased estimator (sample set $\{x_i\}_{i=1}^n$)

$$\nabla_\theta \mathrm{GD}[X]_{[n]} = \frac{1}{n} \sum_{i=1}^n \nabla_\theta \log f_X(x_i;\theta) \big[\frac{2}{n-1} \sum_{j\neq i} \max\{x_j, x_i\} - (b + x_i) \big]. \tag{12}$$

Policy Gradient

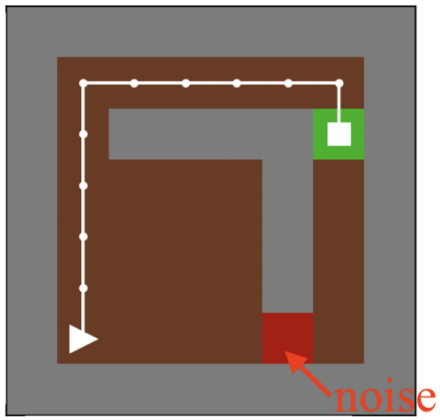
	PG unbiased	Require double sampling
CVaRDev	×	×
GiniDev	✓	×
IQR	×	×
MeanDev	✓	✓
MeanMedianDev	×	×
Variance	✓	✓
STD	✓	✓
Semi_Var	✓	✓
Semi_STD	✓	✓

Table 1: Summary of whether policy gradients (PG) are unbiased or require double sampling for different risk metrics.

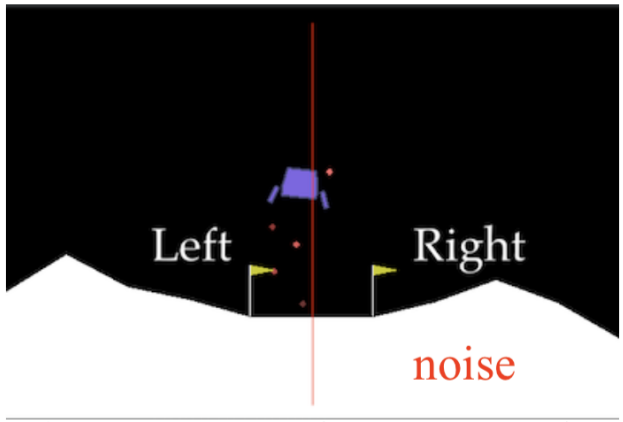
Table of Contents

Experiment Design

Risk-aversion can be clearly defined and verified.

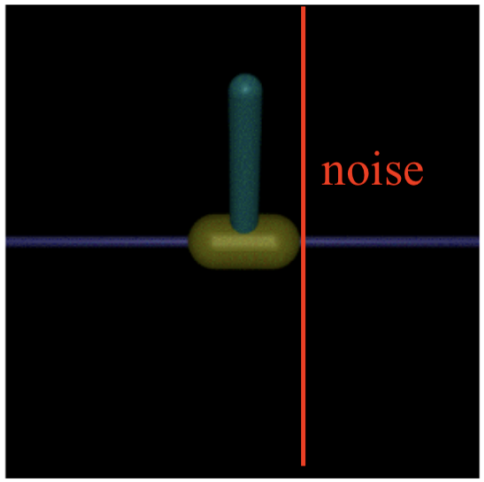


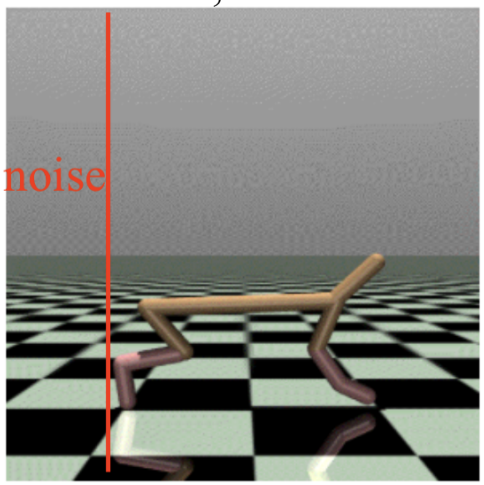
Tabular: Maze



Discrete Control: LunarLander

Continuous Control: InvertedPendulum, HalfCheetah





Results in LunarLander

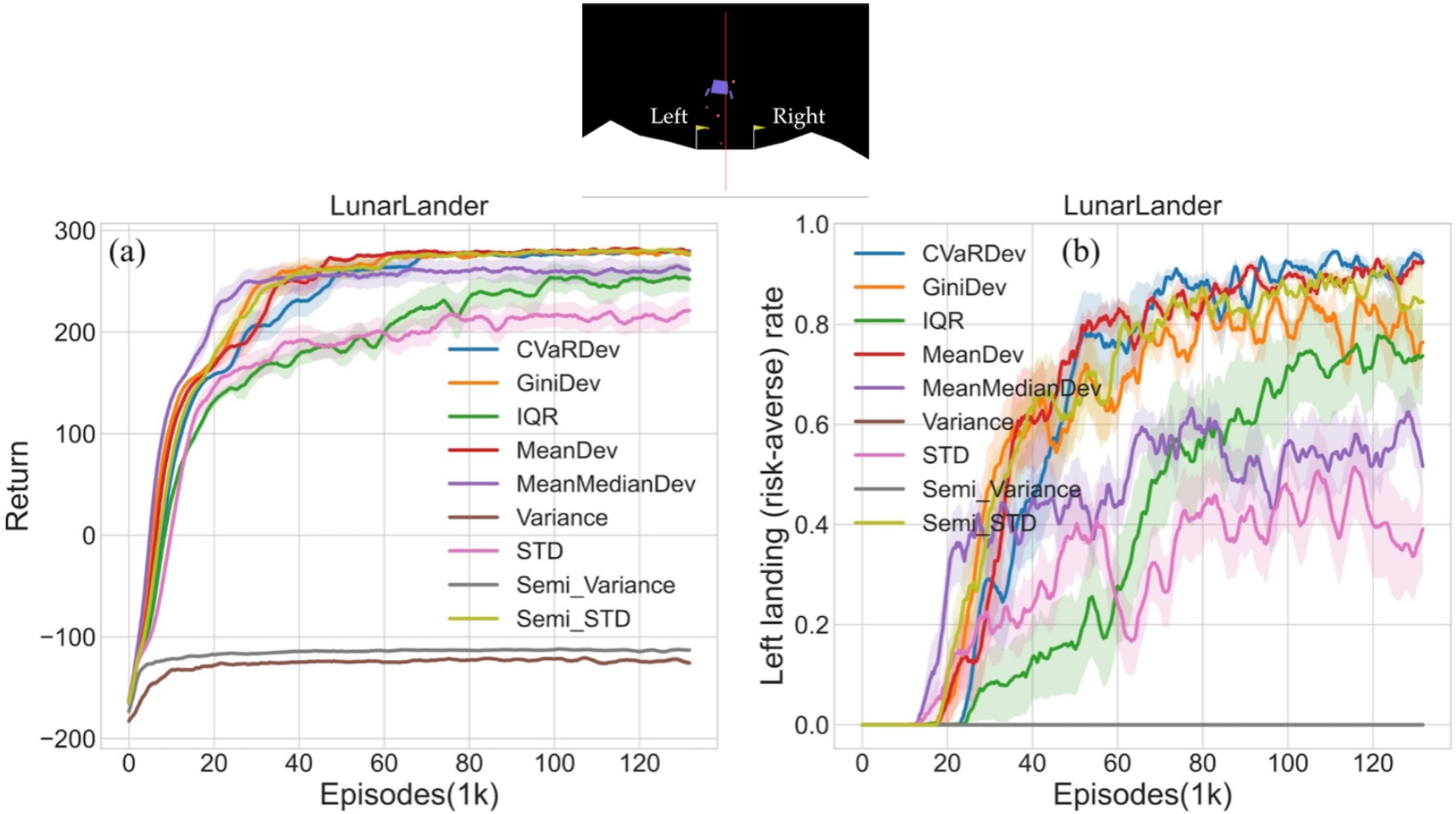


Figure 4: Return and risk-averse rate in LunarLander

Key Takeaway

- Variance-based (Variance, Semi_Variance) unstable due to quadratic term.
- STD-based (STD, Semi_STD) significantly better than variance-based by scaling the gradient. Semi_STD better than STD due to small gradient variance.
- Gini Deviation, CVaR Deviation consistent performance across different domains.
- Mean Deviation, Semi_STD also competitive, worse than GD and CD in some domain.

Thank you!

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. Risk-averse trust region optimization for reward-volatility reduction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4583–4589, 2020.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

HA David. Early sample measures of variability. *Statistical Science*, pages 368–377, 1998.

Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning (ICML)*, pages 3198–3207. PMLR, 2021.

Edward Furman, Ruodu Wang, and Ričardas Zitikis. Gini-type measures of risk and variability: Gini shortfall, capital allocations, and heavy-tailed risks. *Journal of Banking & Finance*, 83:70–84, 2017.

Corrado Gini. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini, 1912.

Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:32639–32652, 2022.

Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted mdps. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 47–76. PMLR, 2023.

Prashanth La and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013.

Yudong Luo, Yangchen Pan, Han Wang, Philip Torr, and Pascal Poupart. A simple mixture policy parameterization for improving sample efficiency of cvar optimization. *Reinforcement Learning Journal*, 1, 2024.

Xiaoteng Ma, Shuai Ma, Li Xia, and Qianchuan Zhao. Mean-semivariance policy optimization via risk-averse reinforcement learning. *Journal of Artificial Intelligence Research*, 75:569–595, 2022.

Saeed Marzban, Erick Delage, and Jonathan Yu-Meng Li. Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. *Quantitative finance*, 23(10):1411–1430, 2023.

R Tyrrell Rockafellar, Stan Uryasev, and Michael Zabarankin. Generalized deviations in risk analysis. *Finance and Stochastics*, 10:51–74, 2006.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

Michaela Spitzer, Jan Wildenhain, Juri Rappsilber, and Mike Tyers. Boxplotr: a web tool for generation of box plots. *Nature methods*, 11(2):121–122, 2014.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *International Conference on Machine Learning (ICML)*, pages 387–396, 2012.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 29, 2015.

Ruodu Wang, Yunran Wei, and Gordon E Willmot. Characterization, robustness, and aggregation of signed choquet integrals. *Mathematics of Operations Research*, 45(3):993–1015, 2020.

Tengyang Xie, Bo Liu, Yangyang Xu, Mohammad Ghavamzadeh, Yinlam Chow, Daoming Lyu, and Daesub Yoon. A block coordinate ascent algorithm for mean-variance optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 10639–10646, 2021.

Haixiang Yao, Zhongfei Li, and Yongzeng Lai. Mean–cvar portfolio selection: A nonparametric estimation framework. *Computers & Operations Research*, 40 (4):1014–1022, 2013.

ChengYang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk. In *International Joint Conference on Artificial Intelligence, IJCAI*, pages 3673–3680, 2022.

Shlomo Yitzhaki et al. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron*, 61(2):285–316, 2003.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 10905–10913, 2021.