Catastrophic–risk–aware reinforcement learning with extreme–value–theory–based policy gradients

José Garrido

Concordia University, Montreal, Canada
jose.garrido@concordia.ca

Insurance Data Science Conference
Bayes Business School, University of London, June 19–20, 2025

https://arxiv.org/abs/2406.15612v1

Research funded by the Natural Sciences and Engineering Research Council of Canada (NSERC)

Collaborators

Parisa Davar
Concordia University
and Deloitte, Montreal

Frédéric Godin
Concordia U., Montreal

Motivation

- Importance of proactive risk management: Highlighted by the Covid–19 pandemic, the 2008 financial crisis, and shocks in the economy.
- Address rare risk focus: Identify, measure, and mitigate to avoid financial ruin.
- Heavy–tailed patterns: Highly rare events occur when data exhibits heavy–tail distribution.

Motivation

- Our Approach:
  - Integrating risk–averse policy gradient RL and EVT for tail risk optimisation to mitigate catastrophic risks.
  - EVT: Focuses on modelling rare events.
  - First to integrate these two methods.
- Evaluation:
  - Simulated data from heavy tailed distributions,
  - Address a hedging problem when options are very expensive.

Table of Contents

Table of Contents

Related work – I

- Recent interest in RL: risk–sensitive RL, integrating risk considerations into reinforcement learning (RL).
- Survey by Prashanth et al. (2022) categorizes risk–sensitive RL techniques into two settings:
  1. Maximising returns while considering risk as a constraint.
  2. Directly incorporating risk as an objective in the optimisation process.

In the second setting, the agent aims to minimize risks due to the stochastic environment, leading to a risk–averse RL method.

Related work – II

Various risk measurement methods in risk–sensitive RL:

- Mean–variance: La and Ghavamzadeh (2013) and Tamar et al. (2012).
- Cumulative prospect theory: Prashanth et al. (2016) and Jie et al. (2018).
- Percentile performance: Chow et al. (2018).
- CVaR: Policy gradient is the most popular approach for CVaR optimisation in RL (Greenberg et al., 2022).

Related work – III

- In previous papers CVaR is usually estimated by the sample average.
- Troop et al. (2022): Estimate CVaR by EVT, integrating with risk–averse multi–armed bandit problem.
- Bader et al. (2018): EVT with automated threshold selection method.

Table of Contents

Risk measures: VaR and CVaR

Value at Risk (VaR)

Let X denote a random loss. VaR at confidence level α is calculated as:

$$VaR_\alpha(X) = \inf\{x \in \mathbb{R} | F_X(x) \geq \alpha\}, \tag{1}$$

where F_X is the cumulative distribution function (CDF) of X.

Conditional Value at Risk (CVaR)

Assume that X is absolutely continuous. The CVaR of X at confidence level α is given by

$$CVaR_\alpha(X) = \mathbb{E}[X | X \geq VaR_\alpha(X)] = \frac{1}{1-\alpha} \int_\alpha^1 VaR_\gamma(X) d\gamma. \tag{2}$$

Sample average – I

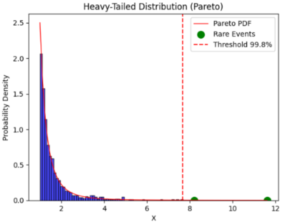
Estimating methods for CVaR: Sample average (SA) and Extreme value theory (EVT)

- Sample average (SA): Empirical average of exceedance above a threshold.

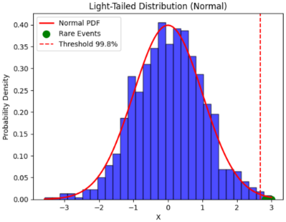
$$\widehat{CVaR}_{\alpha,n}(x) = \frac{\sum_{i=1}^{n} X_i 1_{\{X_i \geq \hat{q}_{\alpha,n}\}}}{\sum_{j=1}^{n} 1_{\{X_j \geq \hat{q}_{\alpha,n}\}}}, \tag{3}$$

  where $\hat{q}_{\alpha,n}(X)$ represents the empirical distribution quantiles.

- Cons: Imprecise estimates when α is close to 1. This is particularly apparent in heavy–tailed distributions.



(a) Pareto distribution



(b) Normal distribution

EVT: Fisher–Tippett's and Pickands–Balkema–de Haan's theorems provide a practical method to approximate CVaR, see McNeil et al. (2015):

$$\hat{c}_{u,\alpha} = \begin{cases} u + \frac{\hat{\sigma}_u}{1-\hat{\xi}_u}\left(1 + \frac{1}{\hat{\xi}_u}\left[\left(\frac{1-\hat{F}(u)}{1-\alpha}\right)^{\hat{\xi}_u} - 1\right]\right), & \text{if } \xi \neq 0, \\ u + \hat{\sigma}_u\left[\log\left(\frac{1-\hat{F}(u)}{1-\alpha}\right) + 1\right], & \text{if } \xi = 0, \end{cases} \tag{4}$$

where $(\hat{\xi}, \hat{\sigma})$ represents the MLE parameter estimates, and α denotes the confidence level such that $\alpha > \hat{F}(u)$.

Extreme value theory – II

- Selecting a suitable threshold u is a challenging problem in EVT.
- Bader et al. (2018) automated threshold selection:
  - Choose a fixed set of candidate thresholds $u_1 < \ldots < u_k$.
  - There are k_i excess samples over each threshold.
  - Anderson–Darling (AD) statistic:

$H_0^{(i)}$: The distribution of the n_i exceedances above u_i follows a GPD.

Table of Contents

RL and risk averse policy gradient

Reinforcement learning has achieved substantial attention in finance:

- Option pricing and hedging.
- Portfolio optimisation.
- Robo–advising.

For a comprehensive overview, see Hambly et al. (2023).

Markov decision process (MDP) – I

For a definition of MDP refer to Puterman (2014):

Markov decision process (MDP)

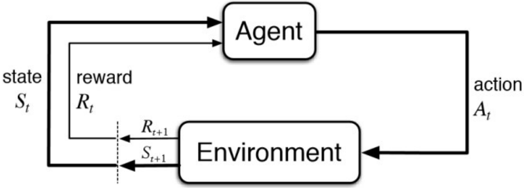
MDP involves a tuple (S, A, R, P, γ) where

1. S is a state space,
2. A is an action space,
3. R is the set of rewards,
4. P is the matrix of transition probabilities between states characterizing the evolution of states and rewards:

$$P : S \times R \times S \times A \to [0, 1],$$

5. γ is a discount factor.

How does a MDP work?



The agent–environment interaction (Sutton and Barto, 2018).

- he agent follows policy π to choose an actions.
- This leads to the following sequence: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots$.
- Main goal of RL: Find the optimal policy to optimise the objective function (reward/risk).

Risk averse policy gradient method

- Risk averse policy gradient method: Directly finds the optimal policy.
- The optimal policy is approximated using a parameterised policy with parameters $\theta \in \mathbb{R}^d$.
- Objective: Minimise $J(\theta) : \theta \to \mathbb{R}$.

$$\theta^* = \arg\min_{\theta \in \Theta} J(\theta). \tag{5}$$

This is addressed using a stochastic gradient descent method.

Table of Contents

Simulated data

- Simplified problem:
  - One–dimensional policy and single action.
  - Given distribution for the cost: GPD or Burr distribution.
  - Parametric relationship between cost and action (policy).
  - The agent following a policy, selects an action that incurs 2000 independent cost.

Problem description

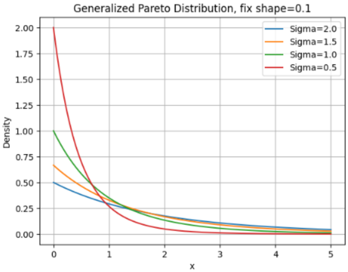
- Risk–averse policy gradient method:
  - To find the optimal policy.
  - Objective function: CVaR.
  - Employs EVT with automated threshold selection for CVaR estimation.
  - Finite differences for CVaR gradient estimation:

$$\widehat{\nabla J(\theta)} \approx \frac{\widehat{J}(\theta+\epsilon)-\widehat{J}(\theta)}{\epsilon}, \text{ where } \epsilon > 0.$$

    - Estimating gradient of the estimated CVaR.
  - $\alpha = 0.998$.

Generalized Pareto distribution (GPD)

$$g_{\xi,\sigma,\mu}(x) = \begin{cases} \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{\frac{-1}{\xi}-1}, & \text{if } \xi \neq 0, \\ \frac{1}{\sigma} e^{\frac{-(x-\mu)}{\sigma}}, & \text{if } \xi = 0. \end{cases}$$



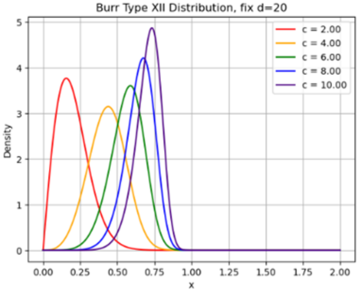
With parameters: shape (ξ), scale (σ), and location (μ).

As the scale σ decreases, the density becomes lighter–tailed, so CVaR decreases.

In our case, $\mu = 0$, $\xi > 0$ are fixed, and σ is considered a function of the action (policy).

Burr

$$f_{c,d}(x) = cd\frac{x^{c-1}}{(1+x^c)^{d+1}}$$

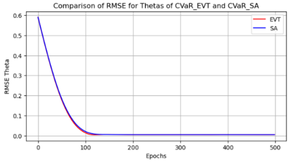


Characterised by two shape parameters $c > 0$ and $d > 0$.

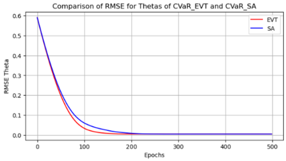
As c decreases, the density becomes lighter–tailed, so CVaR decreases.

In our case, d is predefined and c is considered a function of the action (policy).

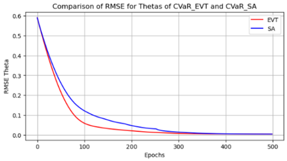
Experimental results for the GPD distribution – I



(c) $\xi = 0.4$

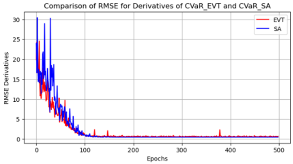


(d) $\xi = 0.6$

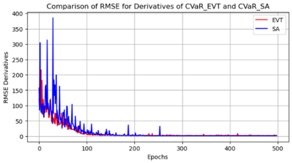


(e) $\xi = 0.8$

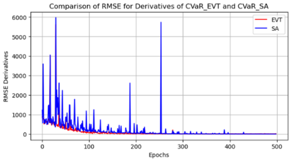
Policy convergence for the GPD distribution.



(f) $\xi = 0.4$

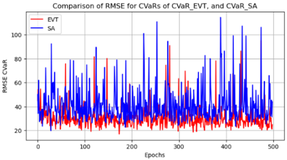


(g) $\xi = 0.6$

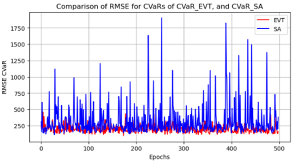


(h) $\xi = 0.8$

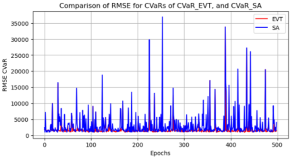
CVaR gradient convergence for the GPD distribution.



(i) $\xi = 0.4$



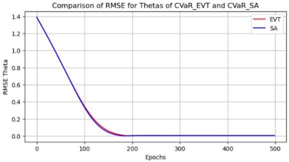
(j) $\xi = 0.6$



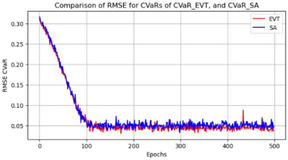
(k) $\xi = 0.8$

CVaR convergence for the GPD distribution.

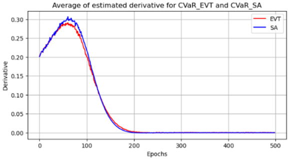
Experimental results for the Burr distribution



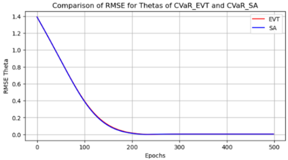
(l)



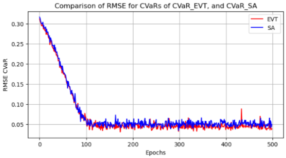
(m)



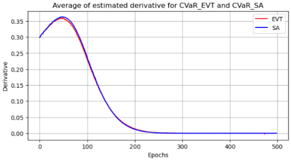
(n)



(o)



(p)



(q)

Left: Policy, Middle: CVaR, Right: CVaR gradient convergence for the Burr distribution when $d = 20, 40$.

Table of Contents

Delta–gamma hedging – I

- Hedging: Offset potential losses by taking an opposite position in a related assets.
- Delta: Option price sensitivity to the underlying asset's price, S.
- Gamma: Second–order sensitivity of option price to S.
- Hedging error:

$$\text{Hedging error} = \max(S_T - K, 0) - V_T,$$

where V_T is the portfolio value at time T.

Delta–gamma hedging – II

- Rolling options strategy:
  - Close and open positions at the beginning and end of each period.
  - Gamma hedge option C on stock S using stock S and option D.
  - The replication portfolio includes θ^s shares of S, θ^D of option D on S, and cash:

$$\begin{cases} Cash_i : & V_i - (\theta_i^s S_i + \theta_i^D D_i^b), \\ V_{i+1} : & \theta_i^s S_{i+1} + \theta_i^D D_{i+1}^e + \text{Cash}_i \; e^{rdt}, \end{cases} \tag{6}$$

where D^e is the option price at the end, and D^b is the option price at the beginning.

Market set–up

- Gamma hedging an at-the-money European call option (short position):
  - with $k = S_0 = 1000$, $T = 0.5$, $\mu = 0.1$, $\sigma = 0.25$, and $r = 0.02$.
- Exponential normal inverse Gaussian (NIG)–Lévy model:

$$S_t = S_0 e^{\sum_{k=1}^{t} Z_k}, \tag{7}$$

$$B_t = e^{rt}, \tag{8}$$

where Z_k is a NIG–Lévy process.

  - NIG distribution: A class of Lévy processes with semi–heavy tails.

Problem description – I

Challenge:

- Options are more expensive here than usual, so costly to fully gamma hedge.
- Fully gamma hedge, it is not optimal to minimise CVaR of hedging error.

Solution:

- Hedge a portion ($K\%$) of the gamma.

Method:

- Find the optimal K (policy):

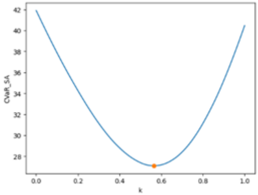
$$\min_{k} \mathrm{CVaR}_{\alpha}(C_T - V_T^k), \tag{9}$$

- Estimate CVaR: EVT with automated threshold selection and SA.

Problem description – II

- Increase options cost in Exponential NIG–Lévy Model, simulate paths with parameters:
  - $\alpha = 15$, $\beta = -10.8$, $\delta = 1$, and $\mu = 6.7 \times 10^{-3}$.
- Simulate 1000 and 2000 weekly paths of NIG Lévy process for underlying stock S.
- Rolling–over strategy on ATM European call option with $T = 0.1$ on S.

Evaluation

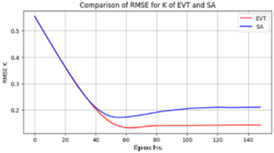


$\text{CVaR}_\alpha(C_T - V_T^k)$ with respect to 500 values of $k \in (0, 1)$ for 1,000,000 weekly paths

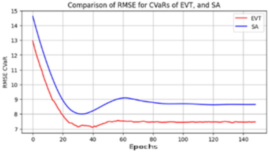
k	CVaR
0.565130	27.088763

Optimal values of policy k and minimum CVaR

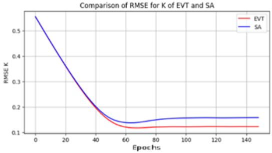
Risk–averse policy gradient experiment



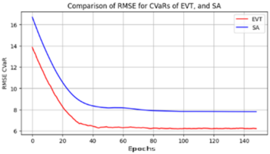
(r) Policy k, $n = 1000$



(s) Min CVaR, $n = 1000$



(t) Policy k, $n = 2000$



(u) Min CVaR, $n = 2000$

RMSE of convergence of policy k and corresponding minimum CVaR for two different values of n.

Conclusions

- Integrated policy gradient RL and EVT for tail risk optimisation to mitigate catastrophic risks.
- Experimental results show risk assessment for very extreme events are unstable, we still have some estimated risk error.
- We able to identify the optimal policy parameter. Also, the approximations of the gradient of the estimated CVaR, with respect to policy, converge.
- Less sample data: EVT outperforms SA in heavy–tail distributions for large α.

Table of Contents

References I

Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, 12(1):310–329.

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

Greenberg, I., Chow, Y., Ghavamzadeh, M., and Mannor, S. (2022). Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:32639–32652.

Hambly, B., Xu, R., and Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503.

References II

Jie, C., Prashanth, L., Fu, M., Marcus, S., and Szepesvári, C. (2018). Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882.

La, P. and Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. *Advances in neural information processing systems*, 26.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

Prashanth, L., Fu, M. C., et al. (2022). Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends® in Machine Learning*, 15(5):537–693.

References III

Prashanth, L., Jie, C., Fu, M., Marcus, S., and Szepesvári, C. (2016). Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pages 1406–1415. PMLR.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tamar, A., Di Castro, D., and Mannor, S. (2012). Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396.

Troop, D., Godin, F., and Yu, J. Y. (2022). Best-arm identification using extreme value theory estimates of the CVaR. *Journal of Risk and Financial Management*, 15(4):172.