Mean-Variance Efficient Reinforcement Learning by Expected Quadratic Utility Maximization

Masahiro Kato, Cyberagent, Inc.
Kei Nakgawa, Nomura Asset Management
Kenshi Abe, Cyberagent, Inc.
Tetsuro Morimura, Cyberagent, Inc.

Abstract

■ Standard Reinforcement Learning (RL):

• Maximize the expected cumulative reward:

➢ The higher the expected cumulative reward, the larger the variance.

• Consider a risk-averse agent, who wants to control the variance.

■ Reinforcement Learning (RL) considering the Mean-Variance Tradeoff:

• Including the variance of the cumulative reward into the objective.

1. Mean-Variance Controlled RL

■ Consider a constraint optimization problem:

• Maximize the expected cumulative reward under a variance constraint, or

• Minimize the variance under an expected cumulative reward constraint.

Ex: Let us denote the parametric policy by π_θ.

Then, for cumulative reward G and a constant $\eta > 0$,

$$\max_\theta \mathbb{E}_{\pi_\theta}[G] \qquad \text{s.t. } \mathbb{V}_{\pi_\theta}(G) = \eta$$

• See Tamar et al. (2012) and Prashanth et al. (2013) for more details.

■ Constraint optimization problem has a difficulty in computation.

■ Double sampling issue:

• We need double sampling to compute the gradient of the variance term.

↑ Existence of the variance is a source of the difficulty.

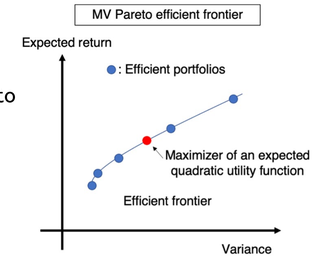
2. Mean-Variance Efficient RL

■ Consider obtaining a mean-variance Pareto efficient policy; that is,

• we cannot increase the expected cumulative reward $\mathbb{E}_{\pi_\theta}[G]$ without increasing the variance $\mathbb{V}_{\pi_\theta}(G)$, and

• we cannot decrease the variance $\mathbb{V}_{\pi_\theta}(G)$ without decreasing the expected cumulative reward $\mathbb{E}_{\pi_\theta}[G]$.

■ The solution of MV-controlled RL corresponds to a policy on the Pareto efficient frontier.



3. Expected Quadratic Utility Maximization

■ A policy maximizing the expected quadratic utility function is known as an efficient policy in economics and finance.

→ We extend the classical result to the RL setting.

■ We call our proposed method RL for expected quadratic utility function maximization (EQUMRL).

■ We maximize the following objective (expected utility function):

$$\mathbb{E}_{\pi_\theta}[u(G;\alpha,\beta)] = \alpha\mathbb{E}_{\pi_\theta}[G] - \frac{1}{2}\beta\mathbb{E}_{\pi_\theta}[G^2],$$

where $u(G;\alpha,\beta)$ denotes the quadratic utility function with parameters $\alpha > 0$ and $\beta \geq 0$.

■ Interpretations:

• A trained policy is mean-variance efficient.

• Maximizing the expected quadratic utility function is equal to the minimizing the difference from the targeted reward $\zeta > 0$:

$$\arg\max_\theta \mathbb{E}_{\pi_\theta}[u(G;\alpha,\beta)] = \arg\min_\theta \mathbb{E}_{\pi_\theta}[(\zeta - G)^2].$$

■ The objective function does not include the variance.

→ We do not suffer from the double sampling issue.

4. REINFORCE-based EQUMRL

■ EQUMRL can be used in combination with a variety of RL methods.

■ EQUMRL with the REINFORCE algorithm.

• Update the parameters with the following gradient:

$$\widehat{\nabla}\mathbb{E}_{\pi_\theta}[u(G;\,\alpha,\beta)] = \left(\alpha\hat{G} - \frac{1}{2}\beta(\hat{G}^2)\right)\sum_{t=1}^{n}\nabla_\theta \log\pi_\theta(S_t,A_t)$$

where $\hat{G} = \sum_{t=1}^{n}\gamma^{t-1}r(S_t,A_t)$ and $\hat{G}^2 = \left(\sum_{t=1}^{n}\gamma^{t-1}r(S_t,A_t)\right)^2$

Algorithm 1 REINFORCE-based EQUMRL

Initialize the policy parameter θ_0;
for $k = 1, 2, \ldots$ **do**
  Generate $\{(S_t^k, A_t^k, r(S_t^k, A_t^k))\}_{t=1}^n$ on π_θ;
  Update policy parameters $\theta_{k+1} \leftarrow \theta_k + \eta_i \widehat{\nabla}\mathbb{E}_{\pi_{\theta_k}}\left[u^{\text{trajectory}}(G;\alpha,\beta)\right]$.
end for

5. Synthetic Portfolio Management

■ Synthetic datasets of portfolio management.

• Compare our proposed method with the methods proposed by Tamar et al. (2012) and Xie et al. (2018).

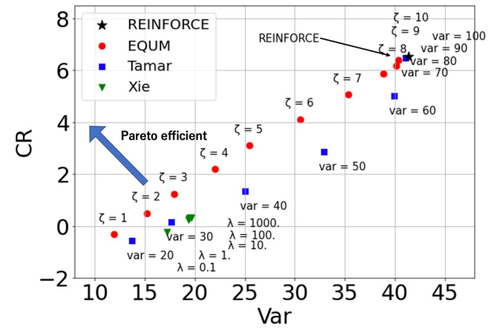


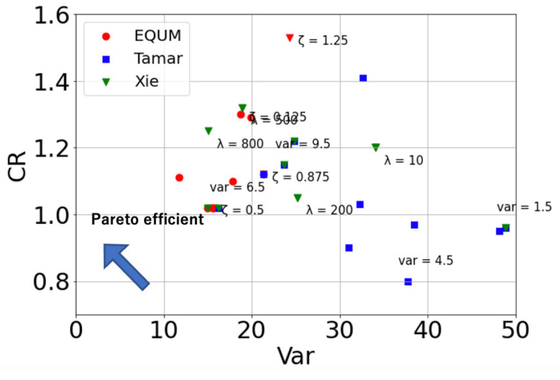
Figure 2: MV efficiency of the portfolio management experiment. Higher CRs and lower Vars methods are MV Pareto efficient.

6. Real-World Portfolio Management

■ Fama-French datasets.

• Real-world financial datasets.

• Compare our proposed method with the methods proposed by Tamar et al. (2012) and Xie et al. (2018).



References

Markowitz, H. Portfolio selection: efficient diversification of investments. Yale university press, 1959.
Tamar, A., Di Castro, D., and Mannor, S. Policy gradients with variance related risk criteria. In Proceedings of the 29th International Conference on International Conference on Machine Learning, pp. 1651–1658, 2012.
Xie, T., Liu, B., Xu, Y., Ghavamzadeh, M., Chow, Y., Lyu, D., and Yoon, D. A block coordinate ascent algorithm for mean–variance optimization. Advances in Neural Information Processing Systems, 31:1065–1075, 2018.