Risk-Averse Trust Region Optimization for Reward-Volatility Reduction

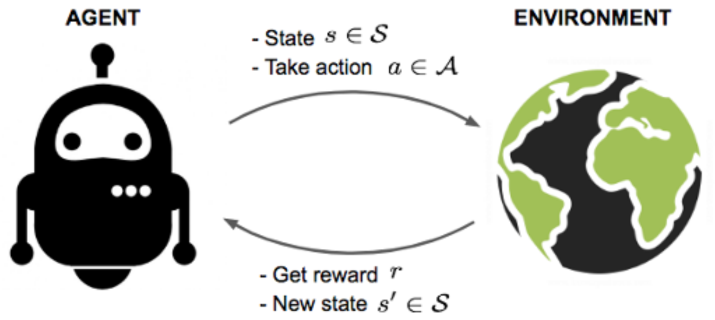
Lorenzo Bisi   Luca Sabbioni   Edoardo Vittori
Matteo Papini   Marcello Restelli

September 18, 2020

Outline

- Reinforcement Learning Intro
- Risk aversion in RL literature
- A new risk measure: reward volatility
- Policy Gradient Theorem for Volatility
- Safe guarantees and TRVO
- Experimental results

Reinforcement Learning intro



Reinforcement Learning intro

- Returns

$$G = \sum_{t=0}^{\infty} \gamma^t R_t$$

- Action-Value function

$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi[G|s_0 = s, a_0 = a] && (1) \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma V_\pi(S_{t+1})|s_t = s, a_t = a] && (2) \end{aligned}$$

- Value function

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[G|s_0 = s] && (3) \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma V_\pi(S_{t+1})|s_t = s] && (4) \end{aligned}$$

- Objective

$$J = \max_\pi \mathbb{E}_{s\sim\mu}[V_\pi(s)] \quad (5)$$

State of the art: Risk Averse RL - inherent risk

- Utility based
  - (Moldovan and Abbeel, 2012)
  - (Shen et al., 2014)
- Coherent Risk Measures
  - (Morimura et al., 2010)
  - (Tamar et al., 2017)
  - (Chow et al., 2017)
- Variance of the returns
  - (Sobel, 1982)
  - (Di Castro et al., 2012)
  - (Tamar and Mannor, 2013)
  - (Prashanth and Ghavamzadeh, 2014)
  - (Tamar et al., 2016)
- Risk averse RL in trading
  - (Moody and Saffell, 2001)

Reward Volatility (Bisi et al., 2020)

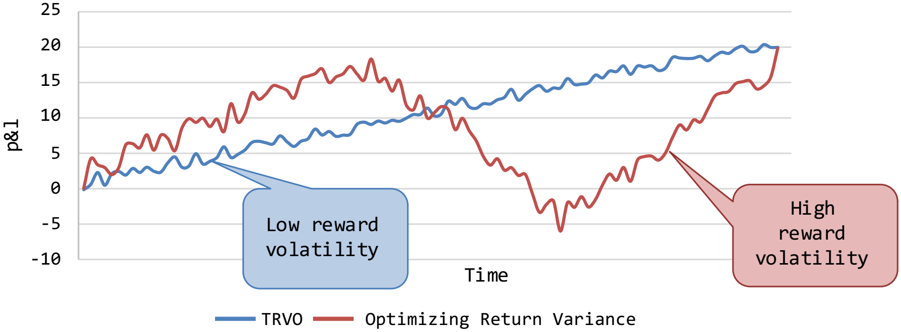
Reward Volatility

$$\nu_\pi^2 = (1-\gamma) \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \left(\mathcal{R}(s_t, a_t) - J_\pi \right)^2 \right]$$

Return variance

$$\sigma_\pi^2 := \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) - \frac{J_\pi}{1-\gamma} \right)^2 \right]$$

Variance/Volatility relation



$$\sigma_\pi^2 \leqslant \frac{\nu_\pi^2}{(1-\gamma)^2}$$

Mean-volatility objective

- Task defined through a *risk-aversion* coefficient λ:

$$\max_\pi \eta_\pi := J_\pi - \lambda \nu_\pi^2$$
$$= (1-\gamma) \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_t \gamma^t \underbrace{\left(R(s_t, a_t) - \lambda (R(s_t, a_t) - J_\pi)^2 \right)}_{R_\pi^\lambda(s_t, a_t)} \right]$$

Bellman Equation

- *Action-volatility* function and *State-volatility* function

$$X_\pi(s,a) := \mathop{\mathbb{E}}_{\substack{s_{t+1}\sim P(\cdot|s_t,a_t)\\ a_{t+1}\sim\pi(\cdot|s_{t+1})}}\left[\sum_{t=0}^{\infty}\gamma^t(\mathcal{R}(s_t,a_t)-J_\pi)^2|s,a\right]$$

$$W_\pi(s) := \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[X_\pi(s,a)]$$

- *Linear Bellman* Equation

$$X_\pi(s,a) = \big(R(s,a)-J_\pi\big)^2 + \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim\pi(\cdot|s')}}\big[X_\pi(s',a')\big].$$

Framework: parametric policies $\pi_{\boldsymbol{\theta}}$

Reward Volatility Policy Gradient Theorem

$$\nabla_{\boldsymbol{\theta}} \nu_\pi^2 = \mathop{\mathbb{E}}_{\substack{s \sim d_{\mu,\pi} \\ a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}} \left[\nabla \log \pi_{\boldsymbol{\theta}}(a|s) X_\pi(s,a) \right].$$

This enables the risk averse version of the classical REINFORCE algorithm.

$$\boldsymbol{\theta} \rightarrow \boldsymbol{\theta} + \nabla_{\boldsymbol{\theta}} \eta_\pi \tag{6}$$

State of the art: Safety - model risk

- Seminal paper
  - (Kakade and Langford, 2002)
- Stationary and stochastic policies
  - (Pirotta et al., 2013b)
  - (García and Fernández, 2015)
- Practical algorithms
  - (Schulman et al., 2015)
  - (Schulman et al., 2017)
- Gaussian, Lipschitz and Smoothing Policies
  - (Pirotta et al., 2013a)
  - (Papini et al., 2017)
  - (Pirotta et al., 2015)
  - (Papini et al., 2019)

Algorithm 1 Trust Region Volatility Optimization (TRVO)

Input: Initial parameters $\boldsymbol{\theta}_0$, batch size N, number of iterations K.

for $k = 0, \dots, K-1$ **do**

Collect N trajectories with $\boldsymbol{\theta}_k$ to obtain dataset $\mathcal{D}_N$

Compute estimates $\widehat{J}$

Estimate advantage values $A^{\lambda}_{\boldsymbol{\theta}_k}(s,a)$

Solve the constrained optimization problem

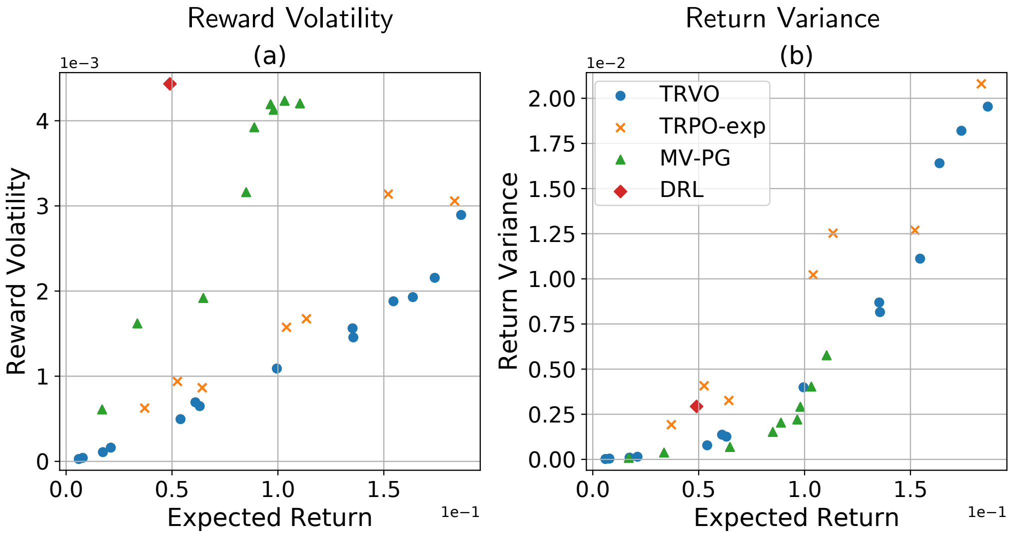
$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}\in\Theta} \left[L^{\lambda}_{k}(\boldsymbol{\theta}) - \frac{2\epsilon\gamma}{1-\gamma} D^{max}_{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}}) \right]$$

$$\text{where } \epsilon = \max_{s} \max_{a} \left| A^{\lambda}_{\boldsymbol{\theta}_k}(s,a) \right|$$

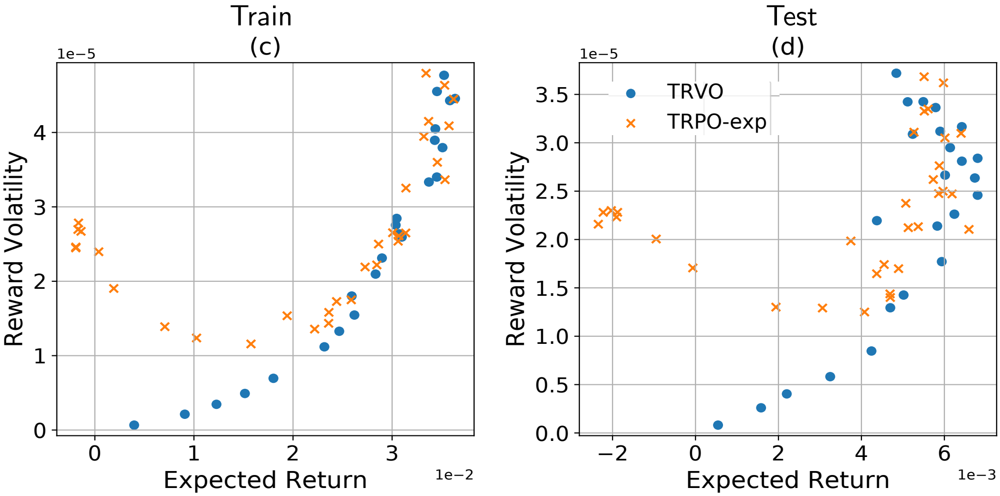
$$L^{\lambda}_{k}(\boldsymbol{\theta}) = \eta_{\boldsymbol{\theta}_k} + \mathop{\mathbb{E}}_{\substack{s\sim d_{\mu,\pi_k} \\ a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}} A^{\lambda}_{\boldsymbol{\theta}_k}(s,a)$$

end for

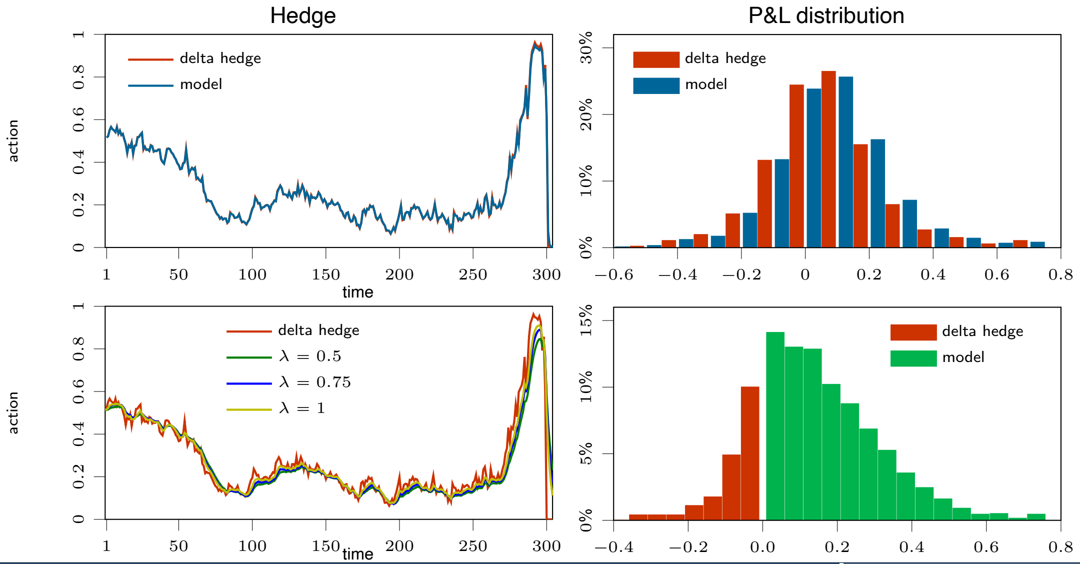
Experiments: S&P 500



Experiments: FX EURUSD & USPJPY



Option Hedging: Experimental Results



Conclusions

- risk averseness in reinforcement learning
- reward volatilty: a novel risk measure
- TRVO, a risk averse TRPO
- great experimental performance on trading and option hedging environments

Thank You for Your Attention!

References

Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. Risk-averse trust region optimization for reward-volatility reduction. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4583–4589. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *JMLR*, 18(1):6070–6120, 2017.

Dotan Di Castro, Aviv Tamar, and Shie Mannor. Policy gradients with variance related risk criteria. *ICML*, 1, 06 2012.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *JMLR*, 16: 1437–1480, 2015. URL `http://jmlr.org/papers/v16/garcia15a.html`.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

Petter N Kolm and Gordon Ritter. Modern perspectives on reinforcement learning in finance. *Modern Perspectives on Reinforcement Learning in Finance (September 6, 2019). The Journal of Machine Learning in Finance*, 1(1), 2019.

Teodor M. Moldovan and Pieter Abbeel. Risk aversion in Markov decision processes via near optimal Chernoff bounds. In *NeurIPS*, pages 3131–3139, 2012.

John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE transactions on neural Networks*, 12(4):875–889, 2001.

References (cont.)

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *ICML*, 2010.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Adaptive batch size for safe policy gradients. In *NeurIPS*, pages 3591–3600, 2017.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients, 2019.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *NeurIPS 26*, pages 1394–1402. Curran Associates, Inc., 2013a.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *ICML*, pages 307–315, 2013b.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2-3):255–283, 2015.

L. A. Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive reinforcement learning. *arXiv preprint arXiv:1403.6530*, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

References (cont.)

Yun Shen, Ruihong Huang, Chang Yan, and Klaus Obermayer. Risk-averse reinforcement learning for algorithmic trading. pages 391–398, March 2014. doi: 10.1109/CIFEr.2014.6924100.

Matthew J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4): 794–802, 1982.

Thomas Spooner and Rahul Savani. A natural actor-critic algorithm with downside risk constraints. *arXiv preprint arXiv:2007.04203*, 2020.

Aviv Tamar and Shie Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential Decision Making With Coherent Risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, July 2017. ISSN 0018-9286, 1558-2523. doi: 10.1109/TAC.2016.2644871. URL `http://ieeexplore.ieee.org/document/7797146/`.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Per-step reward: A new perspective for risk-averse reinforcement learning. *arXiv preprint arXiv:2004.10888*, 2020.

Backup Slides

Performance Difference

Performance Difference Lemma (Kakade and Langford, 2002)

$$J_{\widetilde{\pi}} - J_{\pi} = \int_{\mathcal{S}} d_{\mu,\widetilde{\pi}}(s) \int_{\mathcal{A}} \widetilde{\pi}(a|s) A_{\pi}(s,a) \mathrm{d}a \mathrm{d}s$$

Performance difference in mean-volatility optimization

$$\eta_{\widetilde{\pi}} - \eta_{\pi} = \int_{\mathcal{S}} d_{\mu,\widetilde{\pi}}(s) \int_{\mathcal{A}} \widetilde{\pi}(a|s) A_{\pi}^{\lambda}(s,a) \mathrm{d}a \mathrm{d}s + \lambda(1-\gamma)^2 (J_{\widetilde{\pi}} - J_{\pi})^2.$$

- While looking for the best $\widetilde{\pi}$ given π, $d_{\mu,\widetilde{\pi}}$ is unknown.
- Surrogate function: consider $d_{\mu,\pi}$ instead.

Mean-volatility Advantage

- Expected Return:
  $J_\pi = (1-\gamma)\int_{\mathcal{S}} \mu(s) V_\pi(s) ds$
- Volatility:
  $\nu_\pi^2 = (1-\gamma)\int_{\mathcal{S}} \mu(s) W_\pi(s) ds$
- Advantage:
  $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$
- Vola-Advantage:
  $B_\pi(s,a) = X_\pi(s,a) - W_\pi(s)$

Mean-volatility objective

$$\eta_\pi := J_\pi - \lambda \nu_\pi^2$$

$$A_\pi^\lambda(s,a) := A_\pi(s,a) - \lambda B_\pi(s,a) = \Big(\underbrace{Q_\pi(s,a) - \lambda X_\pi(s,a)}_{Q_\pi^\lambda(s,a)}\Big) - \Big(\underbrace{V_\pi(s) - \lambda W_\pi(s)}_{V_\pi^\lambda(s)}\Big)$$

Safe Improvement Bound

Adopting a surrogate function: $L_\pi^\lambda(\widetilde{\pi}) := \eta_\pi + \int_{\mathcal{S}} d_{\mu,\pi}(s) \int_{\mathcal{A}} \widetilde{\pi}(a|s) A_\pi^\lambda(s,a) \mathrm{d}a \mathrm{d}s$,

Let $\alpha = D_{KL}^{\max}(\pi, \widetilde{\pi}) = \max_s D_{KL}(\pi(\cdot|s), \widetilde{\pi}(\cdot|s))$

$$\epsilon_\lambda = \max_s \left| \mathop{\mathbb{E}}_{a\sim\widetilde{\pi}} \left[A_\pi^\lambda(s,a)\right] \right|, \quad \epsilon = \max_s \left| \mathop{\mathbb{E}}_{a\sim\widetilde{\pi}} \left[A_\pi(s,a)\right] \right|$$

Then:

$$\eta_{\widetilde{\pi}} \geqslant L_\pi^\lambda(\widetilde{\pi}) - \frac{2\gamma\epsilon_\lambda}{1-\gamma}\alpha + \lambda(1-\gamma)^2 M^2,$$

where

$$M := \max\left\{0, A_\pi^{\widetilde{\pi}} - \frac{2\epsilon\gamma}{1-\gamma}\alpha, -A_\pi^{\widetilde{\pi}} - \frac{\gamma}{1-\gamma}\alpha R_{\max}\right\},$$

$$A_\pi^{\widetilde{\pi}} := \int_{\mathcal{S}} d_{\mu,\pi}(s) \int_{\mathcal{A}} \widetilde{\pi}(a|s) A_\pi(s,a) \mathrm{d}a \mathrm{d}s.$$

Connections with Reward-Volatility

- Reward transformation in financial setting (Kolm and Ritter, 2019)
- Downside Risk Constraints (Spooner and Savani, 2020)
- MVPI: Fenchel duality and Block Coordinate Ascent (Zhang et al., 2020):
- ROSA *(under review)*: MDP transformation framework

Option Hedging: Experimental Results