Hedging an Options Book with Reinforcement Learning

Petter Kolm
Courant Institute, NYU

Kolm and Ritter (2019a), "Dynamic Replication and Hedging: A Reinforcement Learning Approach," *Journal of Financial Data Science*, Winter 2019, 1 (1), pp. 159-171.

Kolm and Ritter (2019b), "Modern Perspectives on Reinforcement Learning in Finance," *SSRN working paper*.

Swissquote Conference 2019 on AI in Finance, EPFL
November 8, 2019

Baron Schwartz 

@xaprb

Follow

When you're fundraising, it's AI
When you're hiring, it's ML
When you're implementing, it's linear regression
When you're debugging, it's printf()

9:52 PM - 14 Nov 2017

4,426 Retweets **10,093** Likes

69 4.4K 10K

Outline

Background & motivation

Replication & hedging

- Replicating and hedging an option position is fundamental in finance
- The core idea of the seminal work by Black-Scholes-Merton (BSM):
  - In a complete and frictionless market there is a continuously rebalanced dynamic trading strategy in the stock and riskless security that perfectly replicates the option (Black and Scholes (1973), Merton (1973))
- In practice continuous trading of arbitrarily small amounts of stock is infinitely costly and the replicating portfolio is adjusted at discrete times
  - Perfect replication is impossible and an optimal hedging strategy will depend on the desired trade-off between replication error and trading costs

Related work I

While a number of articles have considered hedging in discrete time or transaction costs alone,

- Leland (1985) was first to address discrete hedging under transaction costs
  - His work was subsequently followed by others[1]
  - The majority of these studies treat proportionate transaction costs
- More recently, several studies have considered option pricing and hedging subject to both permanent and temporary market impact in the spirit of Almgren and Chriss (1999), including Rogers and Singh (2010), Almgren and Li (2016), Bank, Soner, and Voß (2017), and Saito and Takahashi (2017)
- Halperin (2017) applies reinforcement learning to options but the approach is specific to the BSM model and does not consider transaction costs

Related work II

- Buehler et al. (2018) evaluate NN-based hedging under coherent risk measures subject to proportional transaction costs

[1] See, for example, Figlewski (1989), Boyle and Vorst (1992), Henrotte (1993), Grannan and Swindle (1996), Toft (1996), Whalley and Wilmott (1997), and Martellini (2000).

What we do

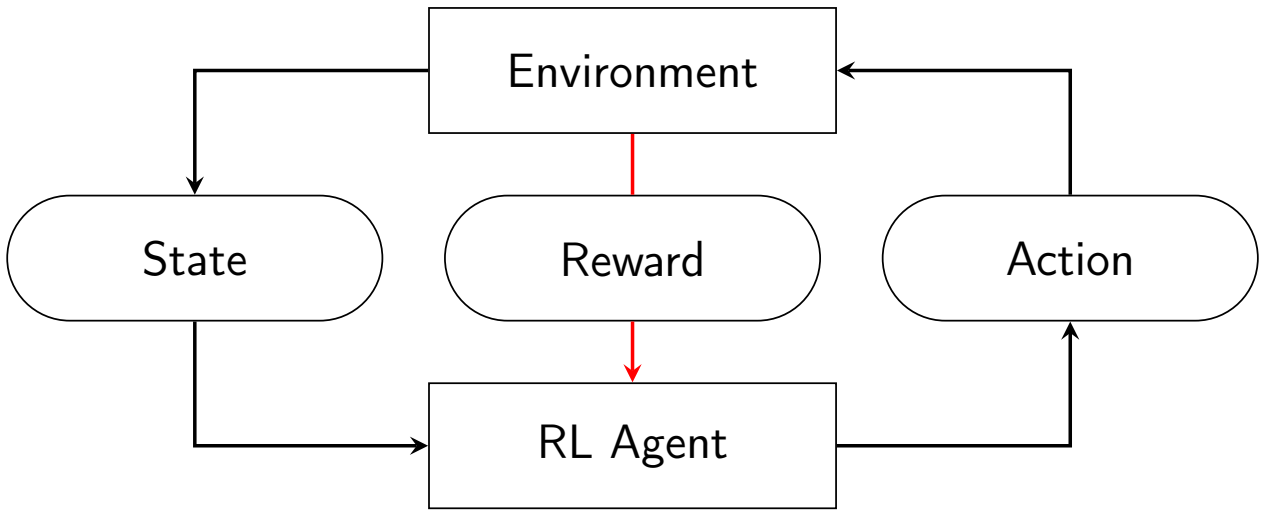
In these articles we:

- Show how to build a reinforcement learning (RL) system which can learn how to optimally hedge an option (or other derivative securities) in a fully realistic setting
  - Discrete time
  - Nonlinear transaction costs
  - Round-lotting
- Method allows the user to "plug-in" any option pricing and simulation library, and then train the system with no further modifications
- The system learns how to optimally trade-off trading costs versus hedging variance for that security
  - Uses a continuous state space
  - Relies on nonlinear regression techniques to the "sarsa targets" derived from the Bellman equation
- Method extends in a straightforward way to arbitrary portfolios of derivative securities

Reinforcement learning

Brief introduction to RL I

- RL agent interacts with its environment. The "environment" is the part of the system outside of the agent's direct control
- At each time step t, the agent observes the current state of the environment s_t and chooses an action a_t from the action set
- This choice influences both the transition to the next state, as well as the reward R_t the agent receives



Brief introduction to RL II

- A policy π is a way of choosing an action a_t, conditional on the current state s_t
- RL is the search for policies which maximize the expectation of the cumulative reward G_t

$$\mathbb{E}[G_t] = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots]$$

  where γ is discount factor (such that the infinite sum converges)
- Mathematically speaking, RL is a way to solve multi-period optimal control problems
- Standard texts on RL includes Sutton and Barto (2018) and Szepesvari (2010)

Brief introduction to RL III

- The action-value function expresses the value of starting in state s, taking an arbitrary action a, and then following policy π thereafter

$$q_\pi(s, a) := \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \tag{1}$$

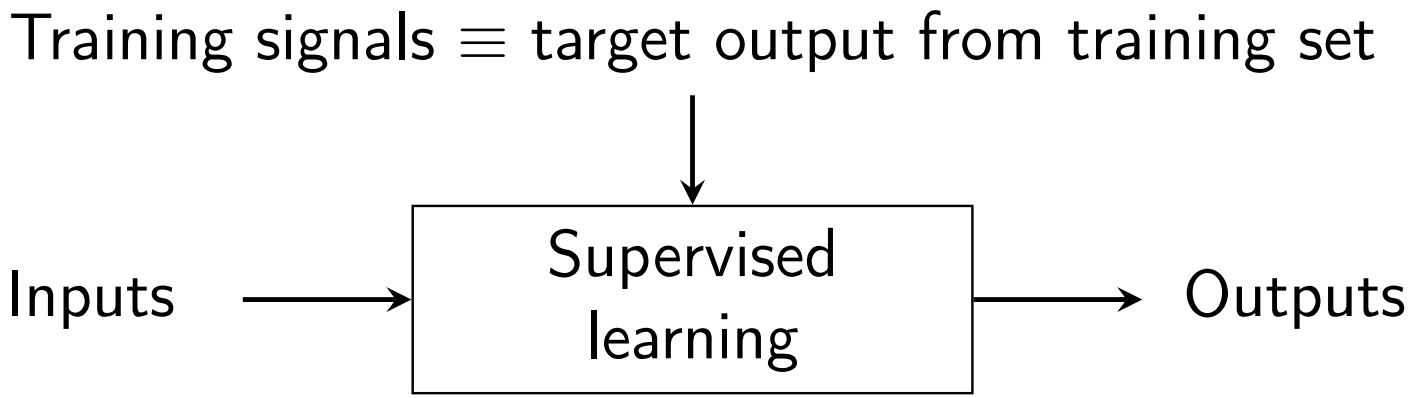
  where $\mathbb{E}_\pi$ denotes the expectation under the assumption that policy π is followed
- If we knew the q-function corresponding to the optimal policy, q_*, we would know the optimal policy itself, namely
  - We choose a in the action set that maximizes $q_*(s_t, a)$

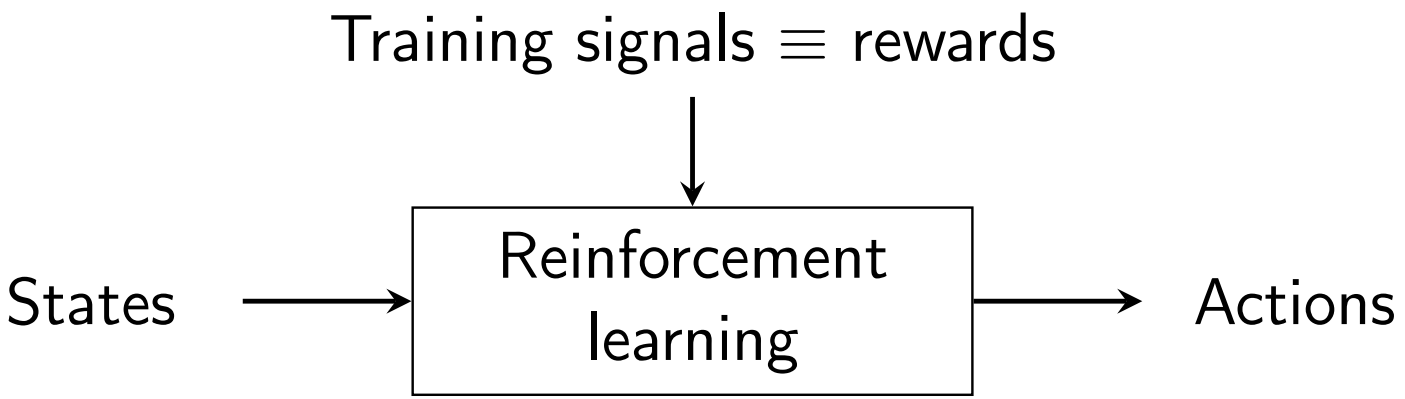
  This is called the *greedy policy*
- Hence the problem is reduced to finding q_*, or producing a sequence of iterates that converges to q_*
- Methods for producing those iterates are based on the Bellman equations

A visual example

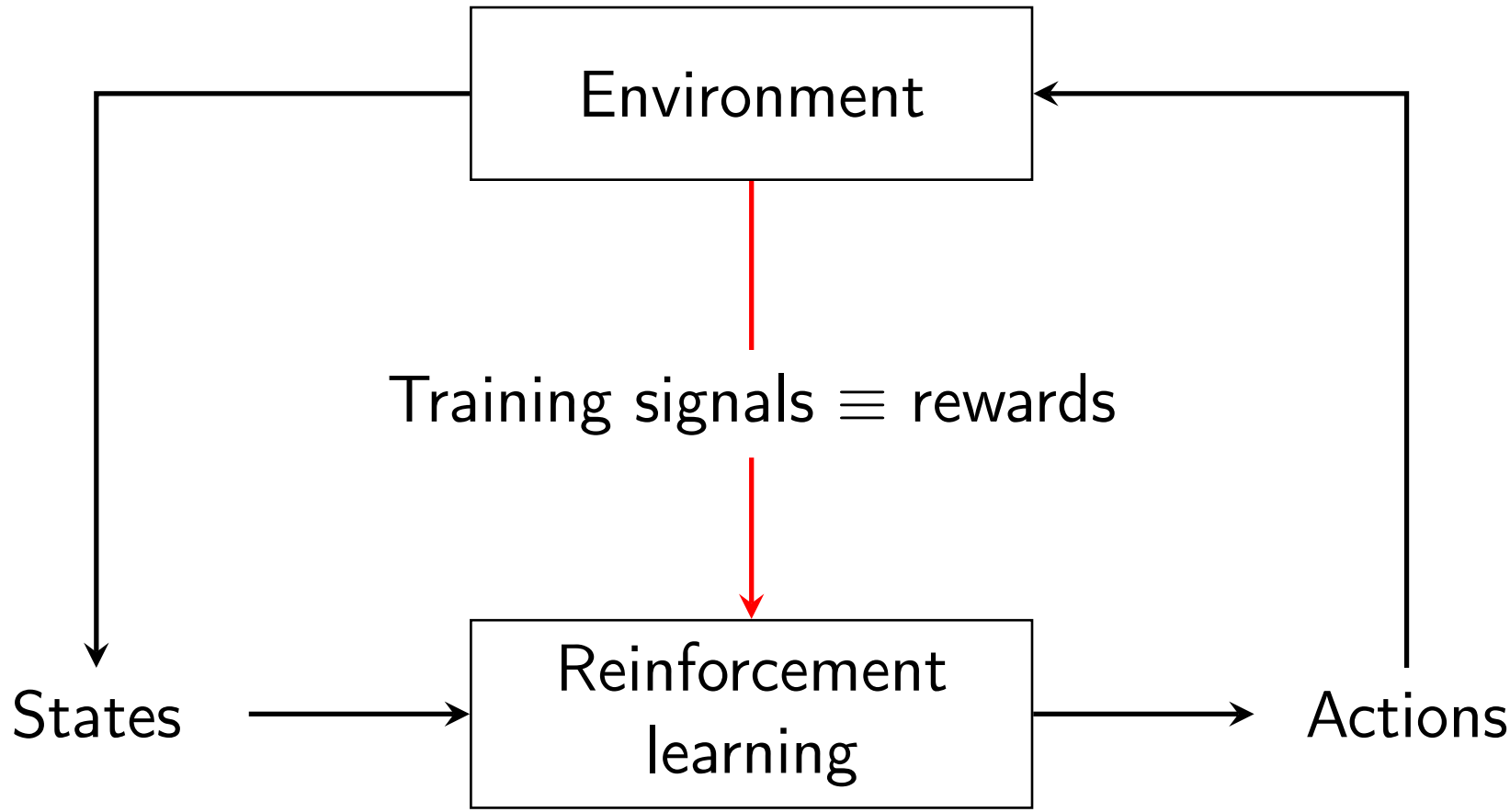
YouTube example

Supervised learning vs. RL?





RL has a feedback loop



Is RL worth the trouble?

RL is often harder than supervised learning

- Joint learning and planning/optimizing from correlated samples
- The distributions of the data changes with the choice of actions
- Need access to the environment for training

So when may it be a good idea to use RL?

- Data comes in the form of trajectories (i.e. non-IID sequences)
- We need to make a sequence of decisions (i.e. non-IID decisions)
- We are able to observe feedback to state or choice of actions. This information can be partial and/or noisy
- There is a gain to be made by optimizing action choice over a portion of the trajectory (i.e. time consistency needed for the Bellman equation to hold)

Other ML techniques cannot easily deal with these situations

Common challenges when solving RL problems

- Specifying the model
  - Representing the state
  - Choosing the set of actions
  - Designing the reward
- Acquiring data for training
  - Exploration / exploitation
  - High cost actions
  - Time-delayed reward
- Function approximation (random forests, CNNs)
- Validation / confidence measures

Reinforcement learning for hedging

Automatic hedging in theory I

- We define automatic hedging to be the practice of using trained RL agents to handle hedging
- With no trading frictions and where continuous trading is possible, there may be a dynamic replicating portfolio which hedges the option position perfectly, meaning that the overall portfolio (option minus replication) has zero variance
- With frictions and where only discrete trading is possible the goal becomes to minimize variance and cost
  - We will use this to define the reward

Automatic hedging in theory II

- This suggest we can seek the agent's optimal portfolio as the solution to a mean-variance optimization problem with risk-aversion κ

$$\max \left(\mathbb{E}[w_T] - \frac{\kappa}{2} \mathbb{V}[w_T] \right) \tag{2}$$

where the final wealth w_T is the sum of individual wealth increments δw_t,

$$w_T = w_0 + \sum_{t=1}^{T} \delta w_t$$

We will let wealth increments include trading costs

Automatic hedging in theory III

- In the random walk case, this leads to solving

$$\min_{\text{permissible strategies}} \sum_{t=0}^{T} \left(\mathbb{E}[-\delta w_t] + \frac{\kappa}{2} \mathbb{V}[\delta w_t] \right) \tag{3}$$

  where

$$-\delta w_t = c_t$$

  and c_t is the total trading cost paid in period t (including commissions, bid-offer spread cost, market impact cost, and other sources of slippage)

- With an appropriate choice of reward function the problem of maximizing this mean-variance problem can be recast as a RL problem

Automatic hedging in theory IV

- We choose the reward in each period to be[2]

$$R_t := \delta w_t - \frac{\kappa}{2}(\delta w_t)^2 \tag{4}$$

- Thus, training reinforcement learners with this kind of reward function amounts to training automatic hedgers who tradeoff costs versus hedging variance

[2] See Ritter (2017) for a general discussion of reward functions in trading.

Automatic hedging in practice I

- Simplest possible example: A European call option with strike price K and expiry T on a non-dividend-paying stock
- We take the strike and maturity as fixed, exogenously-given constants. For simplicity, we assume the risk-free rate is zero
- The agent we train will learn to hedge this specific option with this strike and maturity. It is not being trained to hedge any option with any possible strike/maturity

- For European options, the state must minimally contain (1) the current price S_t of the underlying, (2) the time $\tau := T - t > 0$ remaining to expiry, and (3) our current position of n shares
- The state is thus naturally an element of[3]

$$\mathcal{S} := \mathbb{R}_+^2 \times \mathbb{Z} = \{(S, \tau, n) \mid S > 0, \tau > 0, n \in \mathbb{Z}\}.$$

Automatic hedging in practice II

- The state *does not* need to contain the option Greeks, because they are (nonlinear) functions of the variables the agent has access to via the state
  - We expect the agent to learn such nonlinear functions on their own
- A key point: This has the advantage of not requiring any special, model-specific calculations that may not extend beyond BSM models

[3] If the option is American, then it may be optimal to exercise early just before an ex-dividend date. In this situation, the state must be augmented with one additional variable: The size of the anticipated dividend in period $t + 1$.

Let's put RL at a disadvantage

- The RL agent is at a disadvantage: It does not know any of the following information:
  - the strike price K
  - that the stock price process is a geometric Brownian motion (GBM)
  - the volatility of the price process
  - the BSM formula
  - the payoff function $(S - K)_+$ at maturity
  - any of the Greeks

  Thus, it must infer the relevant information, insofar as it affects the value function, by interacting with a simulated environment

Simulation assumptions I

- We simulate a discrete BSM world where the stock price process is a geometric Brownian motion (GBM) with initial price S_0 and daily lognormal volatility of σ/day
- We consider an initially at-the-money European call option (struck at $K = S_0$) with T days to maturity
- We discretize time with D periods per day, hence each "episode" has $T \cdot D$ total periods
- We require trades (hence also holdings) to be integer numbers of shares
- We assume that our agent's job is to hedge one contract of this option
- In the specific examples below, the parameters are $\sigma = 0.01, S_0 = 100, T = 10$, and $D = 5$. We set the risk-aversion, $\kappa = 0.1$

Simulation assumptions II

- T-costs: For a trade size of n shares we define

$$\text{cost}(n) = \text{multiplier} \times \text{TickSize} \times (|n| + 0.01n^2)$$

  where we take TickSize $= 0.1$. With multiplier $= 1$, the term TickSize $\times |n|$ represents a cost, relative to the midpoint, of crossing a bid-offer spread that is two ticks wide. The quadratic term is a simplistic model for market impact
- All of the examples were trained on a single CPU, and the longest training time allowed was one hour
- Baseline agent = RL agent trained in a friction-less world

Example: Baseline agent (discrete & no t-costs)

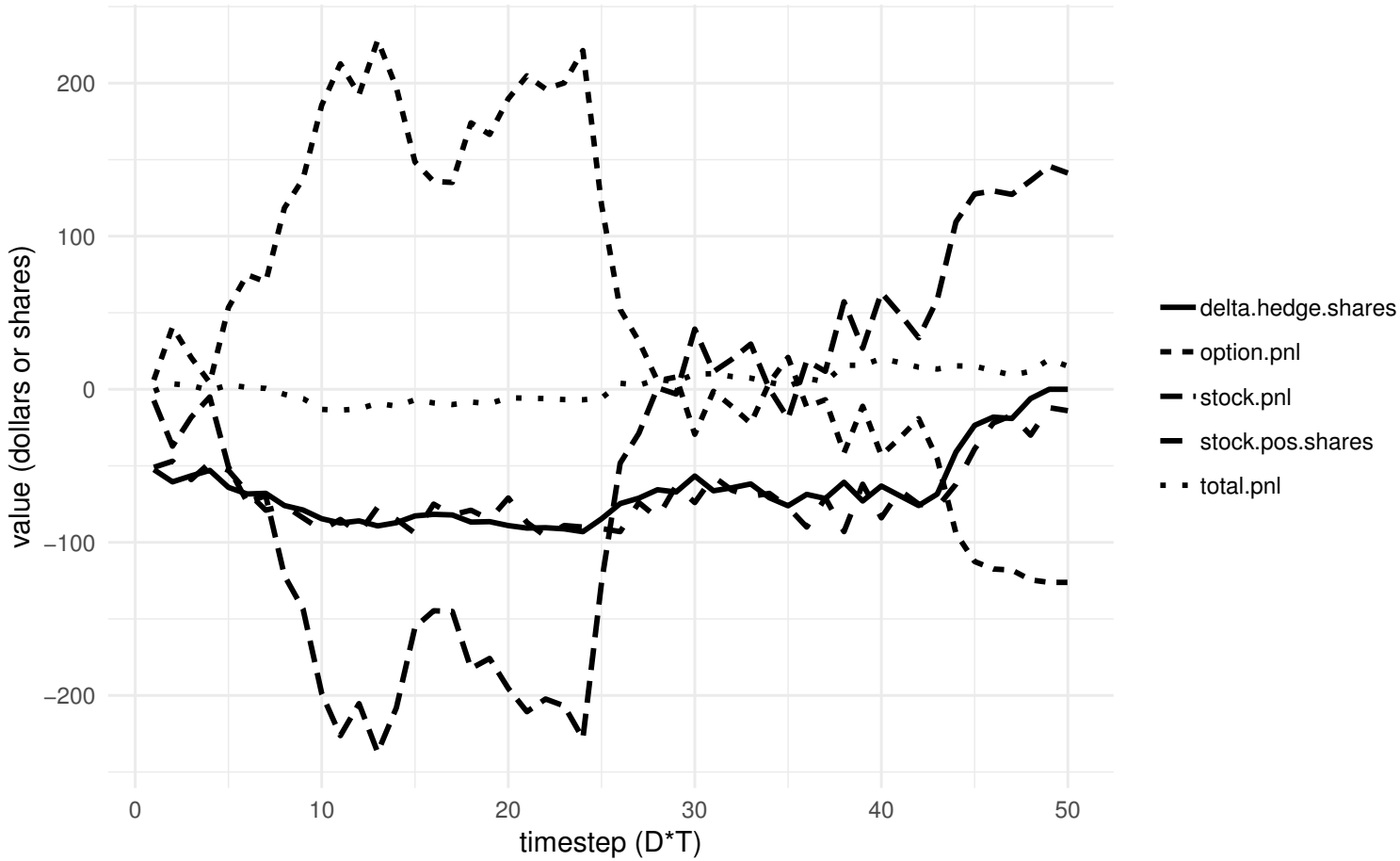


Figure 1: Stock & options P&L roughly cancel to give the (relatively low variance) total P&L. The agent's position tracks the delta

Example: Baseline agent (discrete & t-costs)

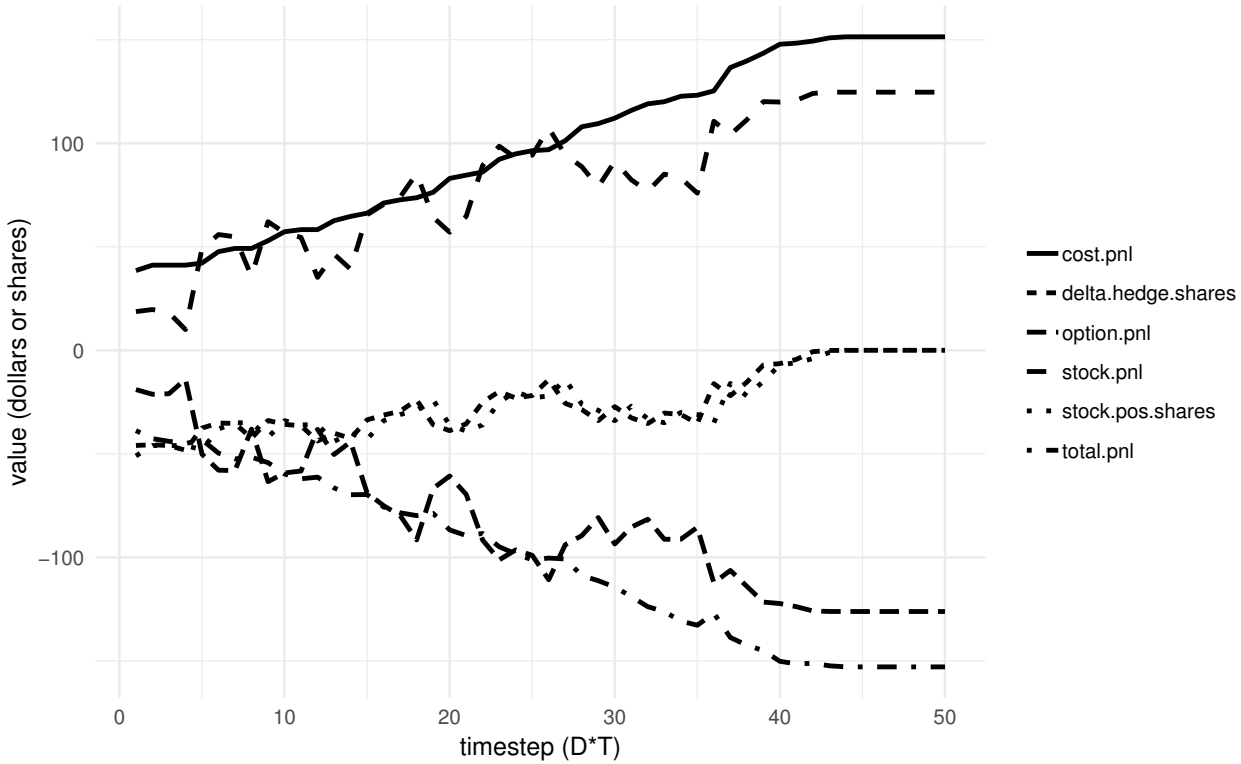
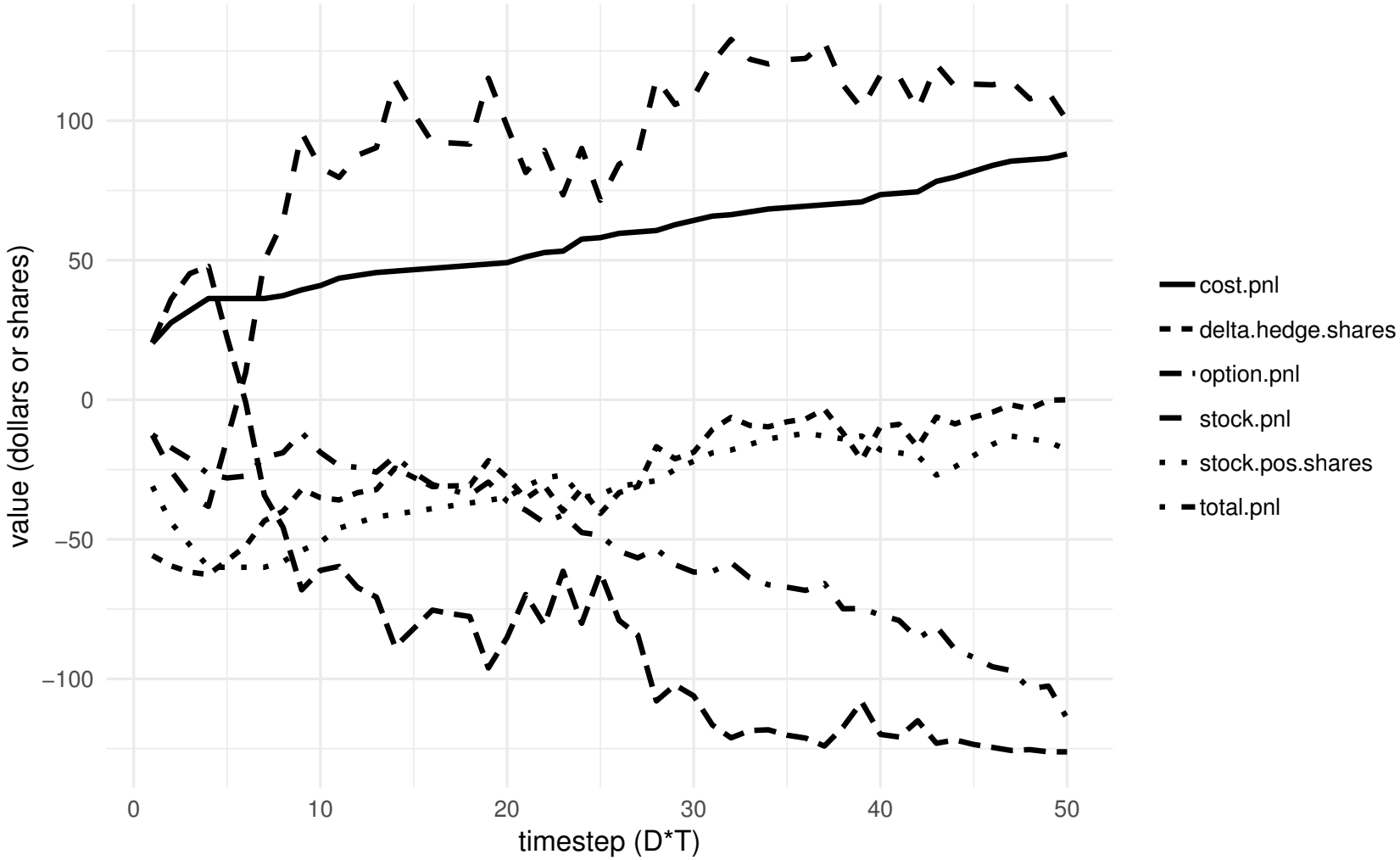


Figure 2: Stock & options P&L roughly cancel to give the (relatively low variance) total P&L. The agent trades so that the position in the next period will be the quantity $-100 \cdot \Delta$ rounded to shares

Example: T-cost aware agent (discrete & t-costs)



Kernel density estimates of total cost & volatility

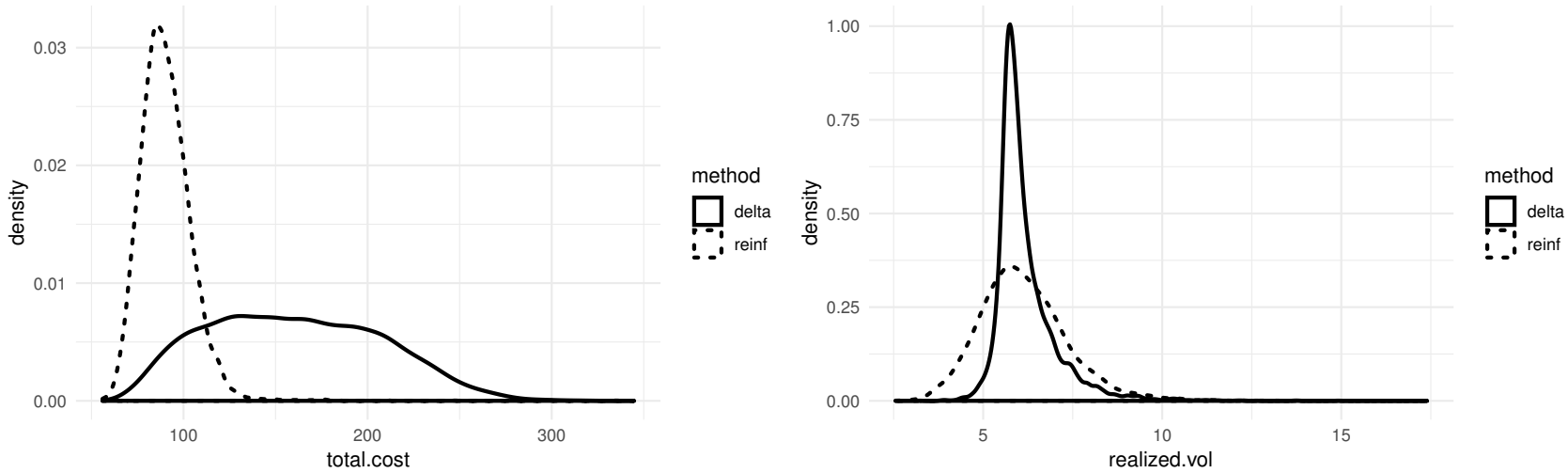


Figure 3: Kernel density estimates for total cost (left panel) and volatility of total P&L (right panel) from $N = 10,000$ out-of-sample simulations. The "reinf" policy achieves much lower cost (t-statistic $= -143.22$) with no significant difference in volatility of total P&L.

Kernel density estimates of total P&L

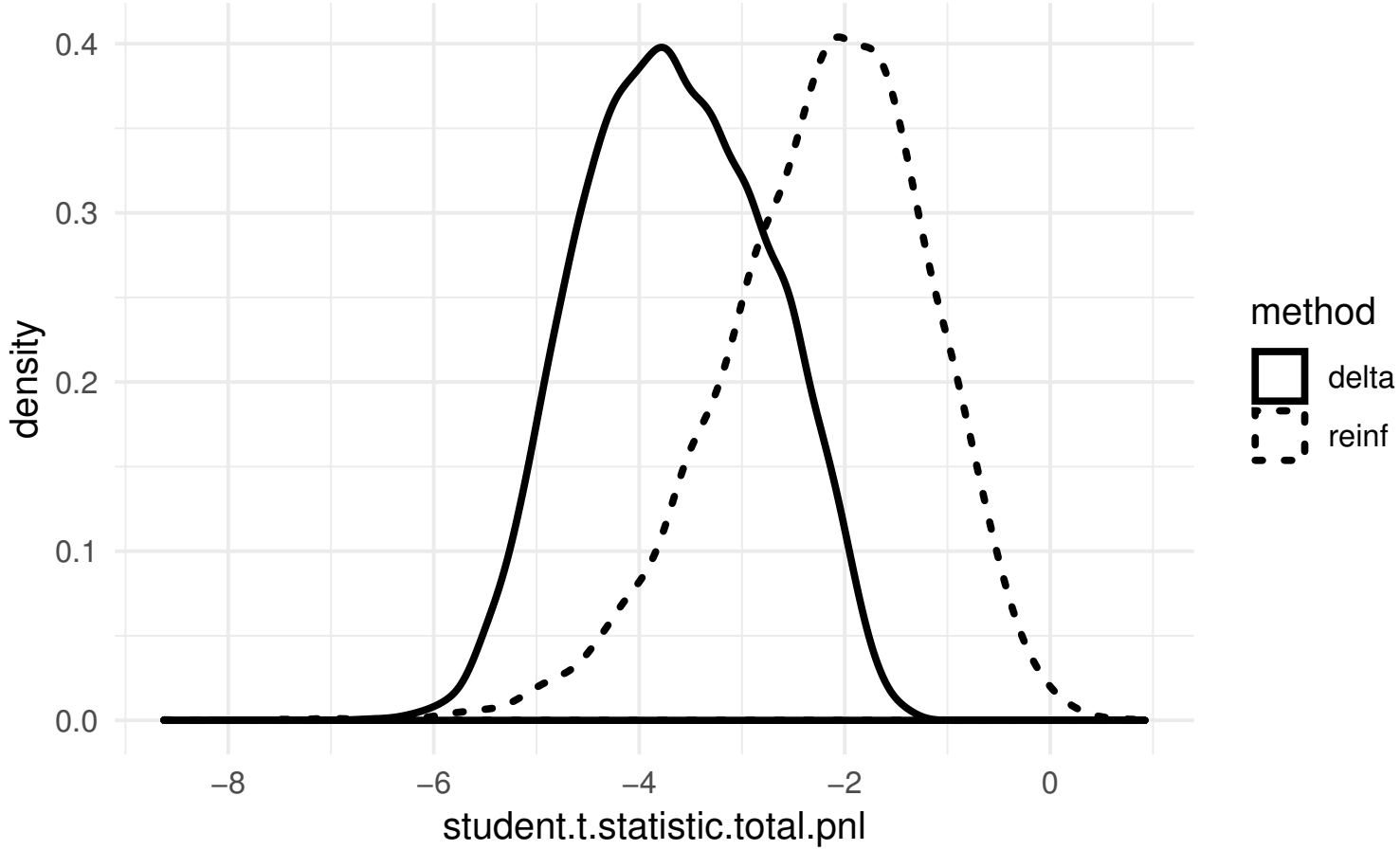


Figure 4: Kernel density estimates of the t-statistic of total P&L for each of our out-of-sample simulation runs, and for both policies represented above ("delta" and "reinf"). The "reinf" method is seen to outperform in the sense that the t-statistic is much more often close to zero and insignificant.

Conclusions I

We have introduced a RL system that hedges an option under realistic conditions of discrete trading and nonlinear t-costs

- The approach does not depend on the existence of perfect dynamic replication. The system learns to optimally trade off variance and cost, as best as possible using whatever securities it is given as potential candidates for inclusion in the replicating portfolio
- It accomplishes this without the user providing any of the following information:
  - the strike price K
  - the fact that the stock price process is a geometric Browning motion
  - the volatility of the price process
  - the Black-Scholes-Merton formula
  - the payoff function $(S - K)_+$ at maturity
  - any of the Greeks

Conclusions II

- A key strength of the RL approach: It does not make any assumptions about the form of t-costs. RL learns the minimum variance hedge subject to whatever t-cost function one provides. All it needs is a good simulator, in which t-costs and options prices are simulated accurately

References I

Almgren, Robert and Neil Chriss (1999). "Value under liquidation". In: *Risk* 12.12, pp. 61–63.

Almgren, Robert and Tianhui Michael Li (2016). "Option hedging with smooth market impact". In: *Market Microstructure and Liquidity* 2.1, p. 1650002.

Bank, Peter, H Mete Soner, and Moritz Voß (2017). "Hedging with temporary price impact". In: *Mathematics and Financial Economics* 11.2, pp. 215–239.

Black, Fischer and Myron Scholes (1973). "The pricing of options and corporate liabilities". In: *Journal of Political Economy* 81.3, pp. 637–654.

Boyle, Phelim P and Ton Vorst (1992). "Option replication in discrete time with transaction costs". In: *The Journal of Finance* 47.1, pp. 271–293.

Buehler, Hans et al. (2018). "Deep hedging". In: *arXiv:1802.03042*.

Figlewski, Stephen (1989). "Options arbitrage in imperfect markets". In: *The Journal of Finance* 44.5, pp. 1289–1311.

Grannan, Erik R and Glen H Swindle (1996). "Minimizing transaction costs of option hedging strategies". In: *Mathematical Finance* 6.4, pp. 341–364.

Halperin, Igor (2017). "QLBS: Q-Learner in the Black-Scholes (-Merton) Worlds". In: *arXiv:1712.04609*.

Henrotte, Philippe (1993). "Transaction costs and duplication strategies". In: *Graduate School of Business, Stanford University*.

References II

Kolm, Petter and Gordon Ritter (2019a). "Dynamic Replication and Hedging: A Reinforcement Learning Approach". In: *The Journal of Financial Data Science* 1.1, pp. 159–171.

— (2019b). "Modern Perspectives on Reinforcement Learning in Finance". In: *SSRN*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3449401.

Leland, Hayne E (1985). "Option pricing and replication with transactions costs". In: *The Journal of Finance* 40.5, pp. 1283–1301.

Martellini, Lionel (2000). "Efficient option replication in the presence of transactions costs". In: *Review of Derivatives Research* 4.2, pp. 107–131.

Merton, Robert C (1973). "Theory of rational option pricing". In: *The Bell Journal of Economics and Management Science*, pp. 141–183.

Ritter, Gordon (2017). "Machine Learning for Trading". In: *Risk* 30.10, pp. 84–89.

Rogers, Leonard CG and Surbjeet Singh (2010). "The cost of illiquidity and its effects on hedging". In: *Mathematical Finance* 20.4, pp. 597–615.

Saito, Taiga and Akihiko Takahashi (2017). "Derivatives pricing with market impact and limit order book". In: *Automatica* 86, pp. 154–165.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. Second edition, in progress. MIT press Cambridge.

Szepesvari, Csaba (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers.

References III

Toft, Klaus Bjerre (1996). "On the mean-variance tradeoff in option replication with transactions costs". In: *Journal of Financial and Quantitative Analysis* 31.2, pp. 233–263.

Whalley, A Elizabeth and Paul Wilmott (1997). "An asymptotic analysis of an optimal hedging model for option pricing with transaction costs". In: *Mathematical Finance* 7.3, pp. 307–324.