Model-Free Option Pricing with Reinforcement Learning

Igor Halperin

NYU Tandon School of Engineering

Columbia U.- Bloomberg Workshop on Machine Learning in Finance 2018[1]

[1] I would like to thank Ali Hirsa and Gary Kazantsev for their kind invitation, and Peter Carr and the workshop participants for their interest and very helpful comments and discussions. All errors are mine. Some of the pictures are not mine.

Economics ended up with the theory of **rational expectations**, which maintains that there is a single optimum view of the future, that which corresponds to it, and eventually all the market participants will converge around that view. This postulate is **absurd**, but it is needed in order to allow economic theory to model itself on **Newtonian Physics**.

George Soros

Writing is easy. All you have to do is cross out the wrong words.

Mark Twain

One should avoid solving more difficult intermediate problems when solving a target problem.

Vladimir Vapnik, *Statistical Learning Theory*, 1998

Ptolemy's Epicycles on Wall Street

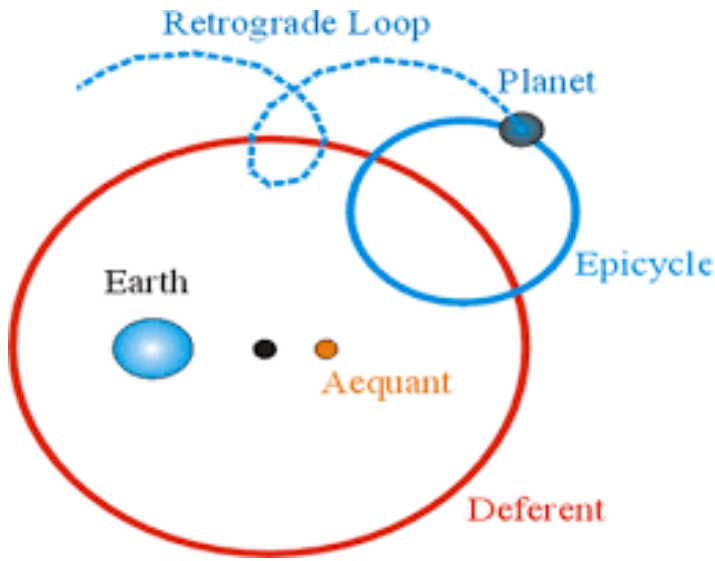


Figure: Ptolemy's geo-centric planetary model: the observed planetary motion is a motion on an epicycle whose center moves on a larger orbit. For more, watch a TEDx talk by Ben Vigoda, CEO of Gamalon, https://youtu.be/PCs3vsoMZfY.

In this presentation:

- Where do we stand in both industry and academia after the Black-Scholes-Merton groundbreaking work of 1973?
- Are there any *viable* (meaningful *and* tractable) alternatives?
- Insights from Physics and Reinforcement Learning
- Model-free option pricing and hedging by Reinforcement Learning (Q-learning and Fitted Q Iteration)
- Summary

PHYSICISTS IN FINANCE

Though the challenges of "quantitative finance" are diverse and often exhilarating, success for the erstwhile physicist is not at all assured. What factors are involved in making the transition to finance?

Joseph M. Pimbley

To a physicist facing significant difficulty in the job market, the allure of a career in finance is obvious: The industry has numerous opportunities that demand the physicist's quantitative skills, and pay handsomely for them. Those contemplating such a move, however, need to look beyond these immediate considerations, for the culture of finance differs markedly from that of physics, having different goals and philosophies, work styles, even dress codes. To be successful on Wall Street, the physicist must willingly adapt to Wall Street's ways.

To add precision to the phrase "physicists in finance," I am using "physicists" to denote PhD recipients and "finance" to refer to the disciplines that require the great-

directly into finance following graduate school or from a postdoctoral position. Less common are the émigrés from full-time "legitimate" physicist positions. Certainly this observation implies that one cause of the physics-to-finance transition is the shortage of jobs in physics, especially for those just starting their careers. But it is regrettable that younger physicists, who have not had the opportunity to explore their chosen disciplines and their abilities on their own, are more likely to shift career goals. Older colleagues, by contrast, have orchestrated successful research projects with lasting contributions, and are therefore much better equipped to contemplate leaving the physics profession. They know what they are forsaking

Figure: "To be successful on Wall Street, the physicist must willingly adapt to Wall Street's ways."
http://www.maxwell-consulting.com/Physicists_Finance_low_mem.pdf

- "In finance, you would be mostly solving a diffusion equation with various boundary conditions."
- Econophysics! (J.-P. Bouchaud, E.Stanley, and others)

In this presentation: Option pricing without PDEs or a model?

Sounds like a scientific blasphemy or maybe as Luddites?

Figure: Luddites (1811-1816) protested the use of machinery in a "fraudulent and deceitful manner" to get around standard labour practices.

Simplicity and beauty in scientific theories

"Everything should be made as simple as possible, but not simpler." (Albert Einstein)

"A physical law must possess mathematical beauty." (Paul Dirac).

"Let's trivialize the problem..." (Lev Landau).

Partial Differential Equations == simplicity and beauty?

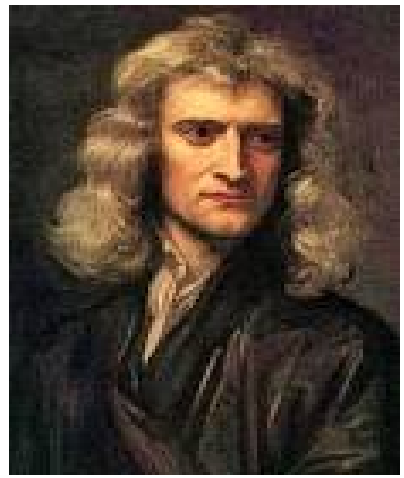


Figure: Isaac Newton

- Newtonian mechanics...
- The diffusion equation...
- The Black-Scholes equation...
- The Schrödinger equation...

Competitive market equilibrium = physics of Newton and Boltzmann

- D. Duffie, "Black, Scholes and Merton - Their Central Contributions to Economics" (1997)
- P. Bernstein, "Capital Ideas: the improbable origins of modern Wall Street", Wiley 2005
- Duffie: "While there are important alternatives, a current basic paradigm for valuation, in both academia and in practice, is that of **competitive market equilibrium**". The price is the price that equates total demand to total supply.

Competitive market equilibrium in Finance

- Three Nobel Prizes in Economics for work based on the paradigm of market equilibrium:
  - Modigliniani-Miller (1958): irrelevance of capital structure for the market value of a corporation.
  - The Capital Asset Pricing Model (CAPM) of William Sharpe (1964).
  - The Black-Scholes option pricing theory (1973) (no-arbitrage as a weaker form of market equilibrium)
- Market equilibrium theories "model themselves on Newtonian physics" (G. Soros).
- More precisely, they describe a thermodynamics equilibrium of statistical mechanics of Ludwig Boltzmann (1844-1906).

Does the Black-Scholes model pass Einstein's test?

Two key elements :

- Option pricing by replication (dynamic hedging)
- Taken to the continuous-time limit $\Delta t \to 0$

Together, these two steps produce the celebrated Black-Scholes equation

$$\frac{\partial C_t}{\partial t} + rS_t \frac{\partial C_t}{\partial S_t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 C_t}{\partial S_t^2} - rC_t = 0 \qquad (1)$$

Isn't it simple and beautiful?

"I applied the Capital Asset Pricing Model to *every moment* in a warrant's like, for every possible stock orice and warrant value... I stared at the differential equation for many many months... (Fisher Black).

Black-Scholes model: the main take-aways:

- Data requirements: two numbers: the current stock price S_t and stock volatility σ (plus parameters for an option)
- The option price is *unique* and given by a solution of the BS equation.
- The optimal option hedge (the amount of stock in a replicating portfolio) is obtained *after* the option price is computed.
- Options are **redundant** (= perfectly replicable in terms of stocks and cash - why bother?) and have **instantaneously zero risk**!!
- (What does it even mean? Time in Finance is fundamentally discrete...)
- "When people are seeking profits, equilibrium will prevail" (Fisher Black).

Black-Scholes model as a model of fake markets?

- Arbitrage pricing gives you an equilibrium price, so that you should not trade below it, and you should not trade above it.
- It only forgot to explain why you should trade at the equilibrium price itself!
- What is the rationale of having entirely redundant financial instruments?
- Options are **not** redundant because they carry **risk**!

"I certainly hope you are wrong, Herr Professor!"

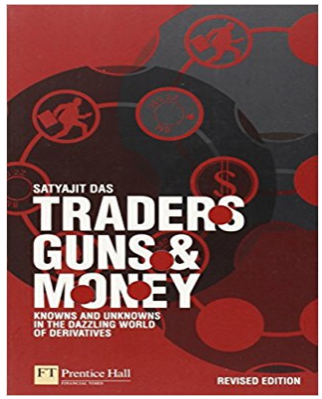


Figure: "A German bank hired a professor from a leading university to help quantify its risk. After some months of extensive analysis, the professor has concluded that the bank had "absolutely no risk". The bank's head of trading responded: "I certainly hope you are wrong, Herr Professor. If you are correct then we can't be making any money!"

What is the **main** problem with the Black-Scholes model?

- Does not match **option** price data?
- Does not match **stock** price data (stock prices are not lognormal)?
- Transaction costs are neglected?
- Discrete hedging?
- Real markets are incomplete?
- **Risk** has disappeared?
- Question: what is a **minimal** change to the BS model so that a new model
  - Will be more useful/meaningful
  - Will have the same or similar level of tractability as the BS model

"Match the market" mantras

The main problem of 'risk-neutral' Quantitative Finance is that it mixes together two problems with the Black-Scholes model:

- It does not incorporate **risk** in option
- The real-world stock price dynamics are not log-normal
- Risk-neutral models ignore the first problem and pursue the second one (in the "risk-neutral" measure!)
- The end result are "match the market" mantras:
  - *Parametric mantras*: Stochastic volatility models, jump-diffusion, Levy models, etc.
  - *Non-parametric mantras:* Local volatility models, MaxEnt, non-parametric Bayes

Ptolemy's epicycles of "risk-neutral" Mathematical Finance

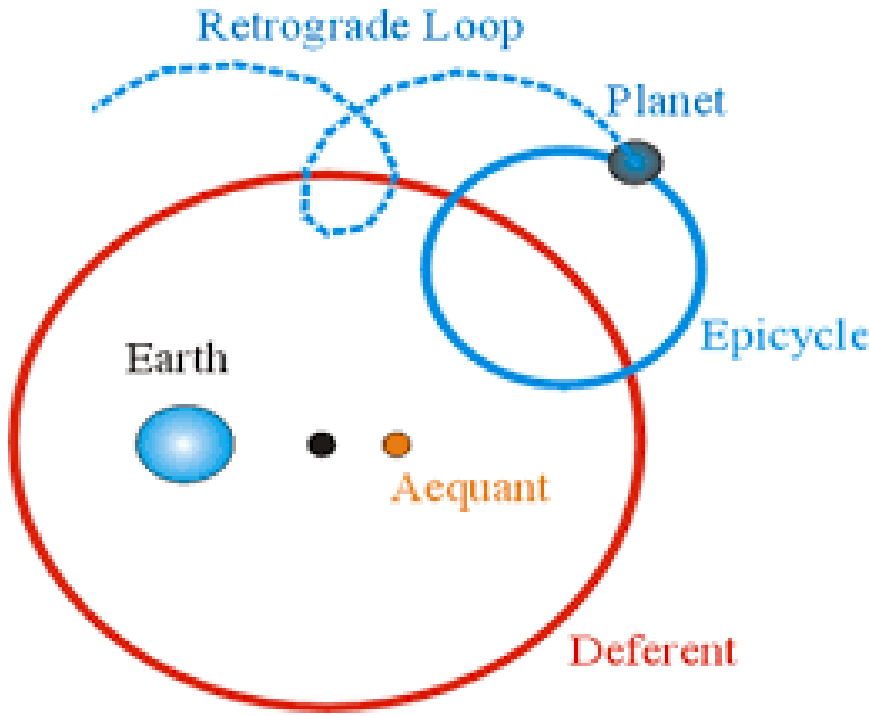


Figure: Ptolemy's model explained 'imperfections' of motion of planets by postulating that the apparently irregular movements were a combination of several regular circular motions seen in perspective from a stationary Earth. Ptolemy had separately fitted model parameters for more than 40 heavenly bodies.

Cargo cult of "risk-neutral" Mathematical Finance

This theory is not even wrong...

Figure: Cult members worshiped certain unspecified Americans having the name "John Frum" who they claimed had brought cargo to their island during World War II and who they identified as being the spiritual entity who would provide cargo to them in the future. (https://en.wikipedia.org/wiki/Cargo_cult)

Control question: model building by subtraction

Mark Twain's approach to Quantitative Finance:
What are the wrong words that should be crossed out when trying to improve on the Black-Scholes model? Select all correct answers:

1. No arbitrage pricing
2. Risk-less hedges
3. "Risk-neutral" option valuation
4. The continuous-time limit
5. PDE's
6. All of the above

Correct answers: ?

Risk is a science of fluctuations

- In the Markowitz portfolio theory: risk is $\mathcal{R}_t = \lambda \text{Var}\left[\Pi_t\right]$
- In statistical mechanics, there are models for both equilibrium and non-equilibrium fluctuations
- The BS model **neglects fluctuations**. This is equivalent to a **thermodynamic limit** in equilibrium statistical mechanics, where all fluctuations die off.
- Market equilibrium models are models where entropy is maximized and does not fluctuate: they are models of a 'heat death' of the Universe as an equilibrium system in a thermodynamic limit.
- A $50Bn question: Is it a right limit to use as a reference point (or a 'first approximation') to describe a risky business?

"Premature continuous time limit" in the BS model?

- The continuous-time limit is taken from the start
- As pricing by replication becomes exact in this limit, all risk is instantaneously eliminated
- To re-install *option risk* as a first-class citizen of a model, we need to revert back to a discrete-time setting!
- This view will show that the BS equation is just a PDE for a **mean** of the option value in the **mathematical** limit $\Delta t \to 0$
- This limit makes a perfect sense *mathematically* but not *financially*, as it looses risk: the option magically becomes risk-less!
- But shouldn't risk in the option be the *original* purpose, a part of option valuation?

Pricing and hedging as sequential risk minimization

- Falls within the class of incomplete market models
- Keep the time discrete!
- Hedging amounts to a **sequential risk minimization**
- Hedged Monte Carlo (HMC) of Potters and Bouchaud (2001) (https://arxiv.org/abs/cond-mat/0008147).
- Similar to American Monte Carlo of Longstaff and Schwartz (2001), but done for hedging of a European option under a real (physical) measure
- Previous work by Follmer and Schweitzer (1989) ("Hedging by Sequential Regression: an Introduction to the Mathematics of Option Trading", ASTIN *Bulletin* 18 , 147-160, 1989.
- Practical implementation and extensions: V. Kapoor *et. al*, "Optimal Dynamic Hedging of Equity Options: Residual-Risks, Transaction-Costs, & Conditioning" (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1530046), A. Grau, Ph.D. thesis (2007).

Discrete-time MDP model for option pricing and hedging

> This work is in parts original and in parts deep. Unfortunately, the original parts are not deep, and the deep parts are not original.
>
> Oscar Wilde

To define risk in an option, we follow a local risk minimization approach.

Take the view of a seller of a European option (e.g. a put option) with maturity T and the terminal payoff of $H_T(S_T)$.

To hedge the option, the seller sets up a replicating (hedge) portfolio Π_t made of the stock S_t and a risk-free bank deposit B_t. The value of hedge portfolio at any time $t \leq T$ is

$$\Pi_t = u_t S_t + B_t \tag{2}$$

where u_t is a stock position at time t, taken to hedge risk in the option.

Pricing and hedging as value maximization

Let's use one-step variances of the hedge portfolio to specify a *risk-averse* price for the option seller (here λ is a Markowitz-like risk aversion parameter):

$$C_0^{(ask)}(S,u) = \mathbb{E}_0\left[\Pi_0 + \lambda \sum_{t=0}^{T} e^{-rt}\mathrm{Var}_t\left[\Pi_t\right] \;\; S_0 = S, u_0 = u\right]$$

The option seller should *minimize* this option price to be competitive. Equivalently, we can flip the sign and formulate the problem as maximization of the value function $V(S,u) \equiv -C_0^{(ask)}(S,u)$.

This MDP model can be solved using Dynamic Programming (DP), similar to the HMC method. Can use either simulated or real-world data!

Vapnik's principle

"One should avoid solving more difficult intermediate problems when solving a target problem"

V. Vapnik, *Statistical Learning Theory*, (1998)

Reinforcement Learning for option pricing

- RL solves the same problem as DP, i.e. it finds an optimal policy.
- Unlike DP, RL does *not* assume that transition probabilities and reward function are known.
- Instead, RL relies on samples to find an optimal policy.
- A RL solution implements Mark Twain's and Vapnik's principles: it focuses on the target problem and crosses out all wrong words!

$$\boxed{\text{RL pricing} = \text{BS} - \text{No Arbitrage} - \text{Model} - \text{PDEs} + \text{Q-Learning}}$$

Batch-mode RL

Data: N trajectories, the information set $\mathcal{F}_t = \left\{ \mathcal{F}_t^{(n)} \right\}_{n=1}^{N}$.

For each trajectory n, $\mathcal{F}_t^{(n)}$ contains:

- The stock price S_t
- The hedge position a_t
- Instantaneous reward R_t
- The next-time value S_{t+1}:

$$\mathcal{F}_t^{(n)} = \left\{ \left(S_t^{(n)}, a_t^{(n)}, R_t^{(n)}, S_{t+1}^{(n)} \right) \right\}_{t=0}^{T-1} \tag{3}$$

Bellman equation and rewards in the MDP option pricing model

The Bellman equation for our model:

$$V_t^{\pi}(S_t) = \mathbb{E}_t^{\pi}\left[R(S_t, a_t, S_{t+1}) + \gamma V_{t+1}^{\pi}(S_{t+1})\right]$$

Here $R(S_t, a_t, S_{t+1})$ is a one-step time-dependent random reward

$$R_t(S_t, a_t, S_{t+1}) = \gamma a_t \Delta S_t - \lambda \mathrm{Var}_t\left[\Pi_t\right]$$

The one-step reward is a risk-adjusted portfolio return of the Markowitz theory!

Action-value function

The action-value function, or Q-function, is defined by an expectation of the same expression as in the definition of the value function, but conditioned on both the current state S_t *and* the initial action $a = a_t$, while following a policy π afterwards:

$$\begin{aligned} Q_t^\pi(x,a) &= \mathbb{E}_t\left[-\Pi_t(S_t)\middle| S_t = x, a_t = a\right] \\ &- \lambda\,\mathbb{E}_t^\pi\left[\sum_{t'=t}^{T} e^{-r(t'-t)}\mathrm{Var}_t\left[\Pi_{t'}(S_{t'})\right]\ \ x,a\right] \end{aligned} \tag{4}$$

The Bellman equation for the Q-function:

$$Q_t^\pi(x,a) = \mathbb{E}_t\left[R_t\middle| x,a\right] + \gamma\mathbb{E}_t^\pi\left[V_{t+1}^\pi\left(S_{t+1}\right)\middle| x\right] \tag{5}$$

An optimal action-value function $Q_T^\star(x,a)$ is obtained when (4) is evaluated with an optimal policy $\pi_t^\star$:

$$\pi_t^\star = \arg\max_\pi Q_t^\pi(x,a) \tag{6}$$

Q-Learning

> Give me a place to stand, and a set of basis functions rich enough, and I will move the world.
>
> Anonymous

- Q-Learning (Watson 1989) is a model-free and off-policy method of solving RL directly from data.
- In its original form (Watkins 1989), applies only for a discrete-state/discrete-action MDP model.
- Q-Learning converges with probability one, given enough data (Watkins 1989).
- Extended to continuous state-action cases in Fitted Q Iteration (FQI) method (Ernest et. al. 2005, Murphy 2005).
- FQI expands optimal action and Q-function in a set of *basis functions*, similar to the HMC method of Potters and Bouchaud (2001).

Q-Learning for the Black-Scholes problem (QLBS) model

- The QBLS model is a MDP model that reduces to the BS model (if stock prices are log-normal!) in the limit $\Delta t \to 0$ and $\lambda \to 0$
- This limit is **degenerate**: risk disappears, nothing to optimize anymore!
- Dynamic Programming solution when the model is *known*
- Reinforcement Learning FQI solution when the model is *unknown*
- As the RL solution relies on Q-Learning and FQI, it is a *model-free* and *off-policy* solution
- Simple semi-analytical solutions for both the DP and RL settings (needs only Linear Algebra)
- Extendable in many ways, e.g. can use an Inverse RL (IRL) formulation

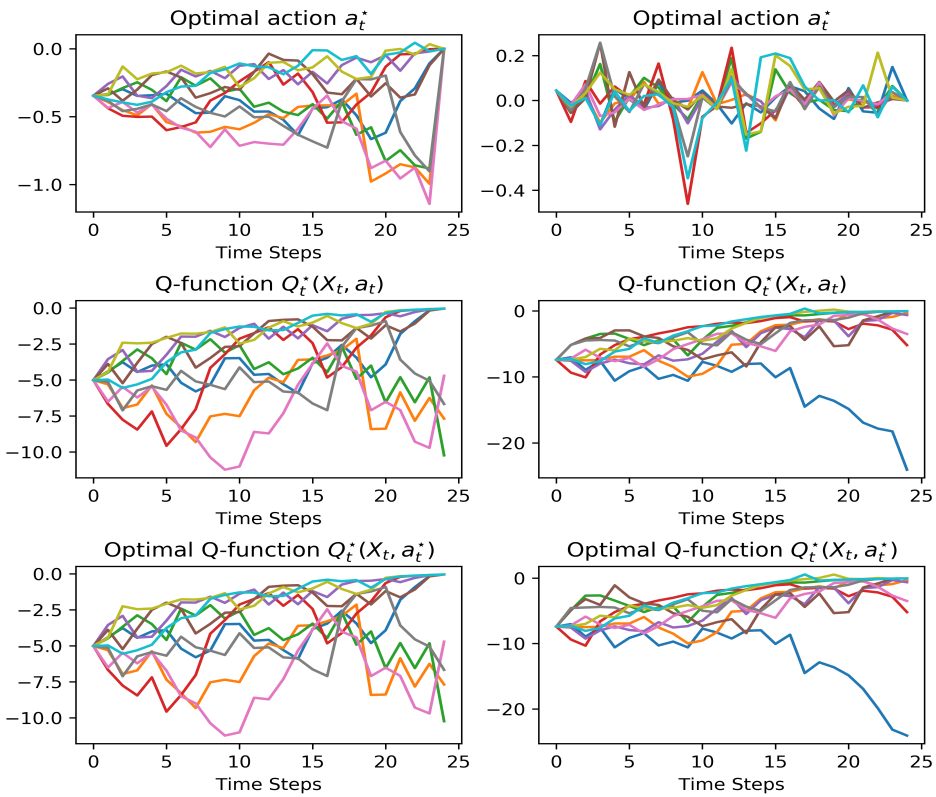
QLBS vs BS: comparison

- The classical BS model and other Mathematical Finance models compute a "fair" "risk-neutral" option prices, ignore risk in options.
- In the QLBS model, residual risk in options is *priced*, consistently with a hedge applied.
- In the BS model, hedging comes *after* pricing.
- In the QLBS model, hedging comes *ahead* of pricing.
- In the QLBS model, the price and the hedge are part of the same formula and are outputs of the same value maximization procedure. In the BS model, they are two different expressions.
- The Black-Scholes model is obtained in the continuous-time limit of such Markowitz model with log-normal dynamics of stock prices.
- The QLBS model involves only finite sums and linear algebra. The BS model involves special functions such as $\mathcal{N}(d_1)$ and $\mathcal{N}(d_2)$.

FQI solution: on-policy learning

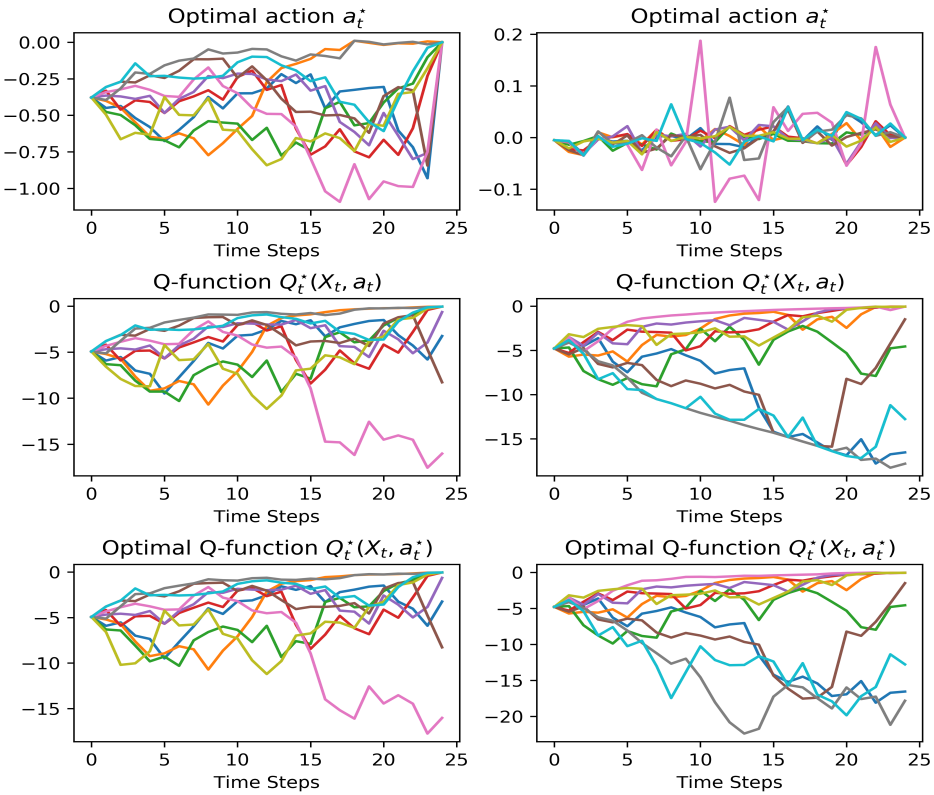
The ATM European put option:

Parameter values: $K = 100$, $T = 1$, $S_0 = 100$, $\mu = 0.05$, $\sigma = 0.15$, $r = 0.03$, $\Delta t = 1/24$, $\lambda = 0.001$. Two sets of MC with $N_{MC} = 50,000$.



FQI solution: off-policy learning

Produce randomized hedges from optimal DP hedges by multiplying by a random uniform number in the interval $[1-\eta, 1+\eta]$ where $0 < \eta < 1$. Take the values of $\eta = [0.15, 0.25, 0.35, 0.5]$ to test the noise tolerance of our algorithms. Results for $\eta = 0.5$:



Noise tolerance for off-policy learning

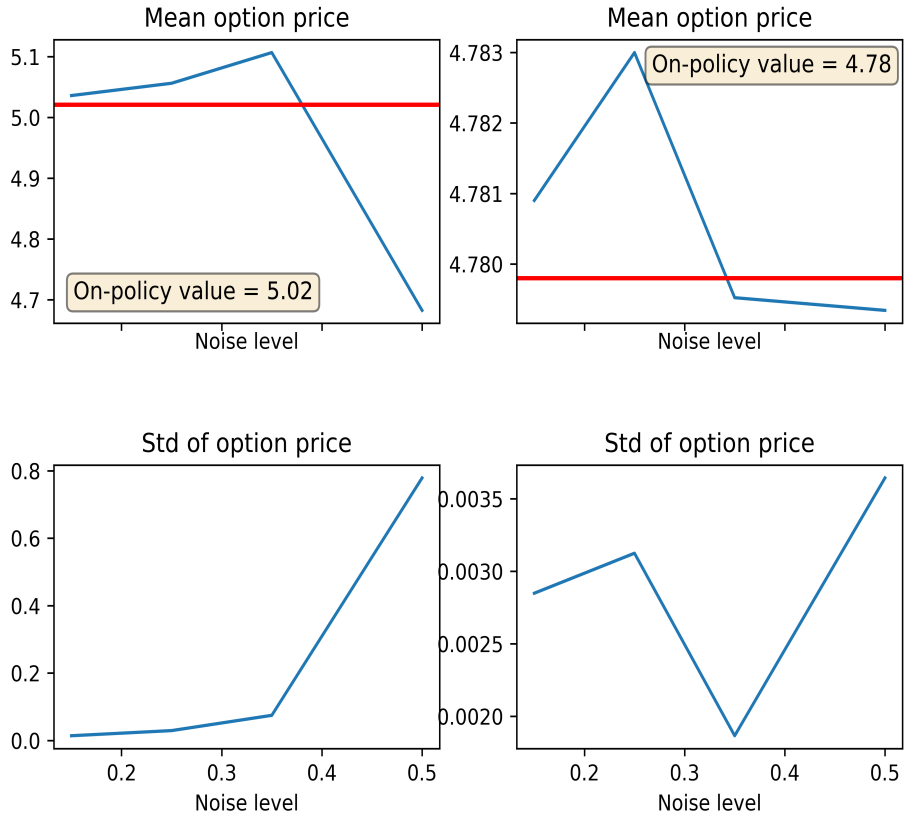


Figure: Means and standard deviations of option prices obtained with *off-policy* FQI learning with data obtained by randomization of DP optimal actions. Horizontal red lines show values obtained with *on-policy* learning corresponding to $\eta = 0$.

Summary

1. The QLBS model shows that we can price and hedge options using only the option replication idea of Black, Scholes and Merton and Q-Learning, and nothing else!
2. The model does **not** need the following: No Arbitrage, a model of stock prices, the BS equation, and PDE's in general, uses instead the trading data and Q-learning.
3. The original BS model is obtained as a *non-physical* (continuous-time and zero risk) limit of a multi-period Markowitz problem for a portfolio of a stock and cash.
3. The QLBS approach is extendable to multiple factors, transaction costs, early exercises etc.
4. Extensions to high-dimensional portfolio settings are non-trivial.
5. The QLBS model suggests that we can do option pricing in Quantitative Finance without no-arbitrage. Can other ideas from Physics and Machine Learning help to build tractable models without competitive market equilibrium?

Thank you!

References:
I. Halperin, "QLBS: Q-Learner in the Black-Scholes (-Merton) Worlds", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3087076 (2017).
I. Halperin, "The QLBS Q-Learner Goes NuQLear: Fitted Q Iteration, Inverse RL, and Option Portfolios", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3102707 (2018).