Deep Bellman Hedging

Hans Buehler*, Phillip Murray†, Ben Wood‡

London June 30th, 2022

Abstract

We present an actor-critic-type reinforcement learning algorithm for solving the problem of hedging a portfolio of financial instruments such as securities and over-the-counter derivatives using purely historic data.

The key characteristics of our approach are: the ability to hedge with derivatives such as forwards, swaps, futures, options; incorporation of trading frictions such as trading cost and liquidity constraints; applicability for any reasonable portfolio of financial instruments; realistic, continuous state and action spaces; and formal risk-adjusted return objectives.

Most importantly, the trained model provides an optimal hedge for arbitrary initial portfolios and market states without the need for re-training.

We also prove existence of finite solutions to our Bellman equation, and show the relation to our vanilla Deep Hedging approach [BGTW19]

Contents

1 Introduction

This note discusses a model-free, data-driven method of managing a portfolio of financial instruments such as stock, FX, securities and derivatives with *reinforcement learing* "AI" methods. It is a dynamic programming "Bellman" version of our Deep Hedging approach [BGTW19]. The

*Technical University München
†JP Morgan London, Imperial College London
‡JP Morgan London

key characteristics of our Deep Hedging framework above several other proposed methods are the ability to hedge with derivatives such as forwards, swaps, futures, options; incorporation of trading frictions such as trading cost and liquidity constraints; applicability for any reasonable portfolio of financial instruments; realistic, continuous state and action spaces; and formal risk-adjusted return objectives. However, our original approach solved this problem for a given initial portfolio and market state. That means that it needs to be re-trained, say, daily to reflect changes in our trading universe or the market. The method proposed here, on the other hand, attempts to solve the optimal hedging problem for any portfolio and market state, as long as they are reasonably close to the data used to train the model.

The work presented here is an extension of our patent application [BMW20]. The main contribution is to provide a numerical implementation method for the practical problem of being able to represent arbitrary portfolios of derivatives as states using purely historic data. We also clarify conditions under which the corresponding Bellman equation is well-posed and admits finite solutions.

Quant Finance 2.0

The motivation for the work presented in this article – and of the Deep Hedging framework in general – is to build financial risk management models which "learn" to trade from historic data and experiences. Today, portfolios of derivatives, securities and other instruments are managed using the traditional quantitative finance engineering paradigm borne out of the seminal work by Black, Scholes & Morten. However, practical experience on any trading desk is that such models do not perform sufficiently well to be automated directly. To start with, they do not take into account trading frictions such as cost and liquidity constraints. Even beyond that they suffer from the underlying engineering approach which prioritizes a focus on interpolating hedging instruments such as forwards, options, swaps over realistic market dynamics. It is an indication of the state of affairs that standard text books on financial engineering in quantitative finance do not discuss real data out-of-sample performance of the models proposed.

As a result, the prices and risk management signals ("greeks") are overly simplistic and do not capture important real-life dynamics. A typical trader will therefore need to overwrite the prices and trading signals from such as standard models with their own heuristics.

Our *Deep Hedging* framework takes a different approach and focuses on robust performance under real-life dynamics, enabed by the use of modern machine learning techniques. The current article is the closest attempting to mimicing a trader's real life behaviour in that here we will give an AI the same historic "experience" a real trader would have. Of course, our model will still be limited by the coverage of historic scenarios used to train it. Hence, human oversight is still required to cater for abrupt changes in market scenarios or starkly adverse risk scenarios.

The website `http://deep-hedging.com` gives an overview over available material on the wider topic.

Related Works

There a few related works concerning the use of machine learning methods for managing portfolios of financial instruments which include derivatives, starting with our own [BGTW19]. However, there we solved the optimal trading problem for a fixed initial portfolio and a fixed initial market state using *periodic policy search*, a method akin to "American Monte Carlo".

In [DJK+20] the authors discuss the use of Bellman methods for this task, namely using DQN and a number of similar methods. They also use risk-adjusted returns in the form of a mean-variance objective. However, in their work the state and action spaces are finite which are not realistic in practise. Moreover, their parametrization of the derivative portfolio is limited to single vanilla options. They also do not cover derivatives as hedging instruments.

In [Hal17] the authors also develop a discrete state approach, where the problem is solved for each derivative position separately. The authors focus in their first work on vanilla options and minimize the terminal variance of the delta-hedged position. In their later [Hal19] the authors present methods to smooth the state space. In neither account are derivatives as hedging instruments supported.

There is a larger literature on the application of AI methods for managing portfolio risks in the context of perpetual assets such as stock and FX portfolios which might reasonably be approximated by normal assets. See the summary [KR19] for an overview, where they also cover the related topic of trade execution with AI methods.

Underlying our work is the use of *dynamic risk measures*, a topic with a wide literature. We refer the interested reader to [DS05] and [**?**] among many others.

2 Deep Bellman Hedging

In this note we will use a notation much more similar to standard reinforcement learning literature, chiefly [SB18]. That means in particular that we will formulate our approach essentially as a continuous state Markov Decision Process (MDP) problem. We will make a decision from some point in time to another. That would typically be intraday or from day to day. To simplify our discussion we will assume we are making a decision "today" and then again "tomorrow". Variables which are valid tomorrow will be denoted by $'$. We will strive to use bold letters for vectors. A product of two vectors is element wise, while "$\cdot$" represents the dot product. We will strive to use small letters for instances of data, and capital letters for random variables.

We denote by $\mathbf{m}$ the **market state** today. The market contains all information available to us today such as current market prices, time, past prices, bid/asks, social media feeds and the like. The set of all market states is denoted by $\mathcal{M} \subset \mathbb{R}^N$. All quantities observed today are a function of the market state.[1] The market tomorrow is a random variable $\mathbf{M}'$ whose distribution is assumed to only depend on $\mathbf{m}$, and not on our trading activity.[2] In terms of notation, think $\mathbf{m} \equiv \mathbf{m}_t$ and $\mathbf{M}' \equiv \mathbf{M}_{t+1}$. The expectation operator of a function f of $\mathbf{M}'$ conditional on $\mathbf{m}$ is written as $\mathbb{E}[f(\mathbf{M}')|m] := \int f(\mathbf{m}')\mathbb{P}[d\mathbf{m}'|\mathbf{m}]$.

We will trade financial instruments such as securities, OTC derivatives or currencies. We will loosely refer to them as "derivatives" as the most general term, even if we explicitly include primary asset such as stocks and currencies. We use $\mathcal{X}$ to refer to the space of these instruments. For $x \in \mathcal{X}$ we denote by $r(x, \mathbf{m}) \in \mathbb{R}$ the cashflows arising from holding x today, aggregated into our accounting currency.[3] Cashflows here cover everything from expiry settlements, coupons, dividends, to payments arising from borrowing or lending an asset. For a vector $\mathbf{x}$ of instruments we use $\mathbf{r}(\mathbf{x}; \mathbf{m})$ to denote the vector of their cashflows.

[1] Mathematically, we say that $\mathbf{m}$ generates today's σ-algebra.

[2] See the lecture notes [**?**] for an example of incorporating market impact.

[3] This implies implies that spot-FX transactions are frictionless.

An instrument changes with the passage of time: an instrument $x \in \mathcal{X}$ today becomes $x' \in \mathcal{X}$ tomorrow, representing only cashflows from tomorrow onwards. If the expiry of the instrument is today, then $x' = 0$.

Every instrument x we may trade has a **book value** which we denote by $B(x, \mathbf{m})$. The book value of a financial instrument is its official mark-to-market, computed using the prevailing market data $\mathbf{m}$. This could be a simple closing price, a weighted mid-price, or the result of computing more complex standard derivative model. Following our notation $B(x', \mathbf{M}')$ denotes the book value of the instrument tomorrow. We use $\mathbf{B}(\mathbf{x}, \mathbf{m})$ for the vector of book values if $\mathbf{x}$ is a vector of instruments. We like to stress that contrary to [BMPW22] here the book value is with respect to only to today's and future cashflows, not past cashflows.

In order to take into account the value of money across time, we will also assume are given a bank account – usually called the *numeraire* - which charges the same overnight interest rate for credits and deposits. The respective one-day discount factor from tomorrow to today is denoted by $\beta(\mathbf{m})$ and we will assume that there is some β^* such that $\beta(\mathbf{m}) \leq \beta^* < 1$. Contrary to [BGTW19] we do not assume that cashflows are implicitly discounted using our bank account.

The discounted profit-and-loss (P&L) for a given instrument $x \in \mathcal{X}$ is the random variable

$$dB(x, \mathbf{m}, \mathbf{M}') := \underbrace{\beta(\mathbf{s})\, B(x', \mathbf{M}') - B(x, \mathbf{m})}_{\text{Change in book value}} + \underbrace{r(x, \mathbf{m})}_{\text{Cashflows}} \quad . \tag{1}$$

If $\mathbf{x} \in \mathcal{X}^n$ is a vector, then $d\mathbf{B}(\mathbf{x}, \mathbf{m}, \mathbf{M}')$ denotes the vector of P&Ls.

Trading

A trader is in charge of a **portfolio** $z \in \mathcal{X}$ – also called "book" – of financial instruments such as currencies, securities and over-the-counter (OTC) derivatives. We call the combined $\mathbf{s} := (z, \mathbf{m})$ our **state** today which takes values in $\mathbf{s} \in \mathcal{X} \times \mathcal{S}$. We will switch in our notation between writing functions in both variables $(z, \mathbf{m})$ and only in $(\mathbf{s})$ depending on context.

In order to risk manage her portfolio, the trader has access to n liquid hedging instruments $\mathbf{h} \equiv \mathbf{h}(\mathbf{s}) \in \mathcal{X}^n$ in each time step. These are any liquid instruments such such as forwards, options, swaps etc. Across different market states they will usually not be the contractually same fixed-strike fixed-maturity instruments: instead, they will usually be defined relative the prevailing market in terms of time-to-maturities and strikes relative to at-the-money. See [BGTW19] for details.

The **action** of buying[4] $\mathbf{a} \in \mathbb{R}^n$ units of our hedging instruments will incur transaction cost $c(\mathbf{a}; z, \mathbf{m})$ on top of the book value. Making cost dependent on both the current portfolio and the market allows modelling trading restrictions based on our current position such as short-sell constraints, or restrictions based on risk exposure. Transaction cost as function of $\mathbf{a}$ is assumed to be normalized to $c(0; \mathbf{s}) = 0$, non-negative, and convex.[5] The convex set of admissible actions is given as $\mathcal{A}(\mathbf{s}) := \{\mathbf{a} \in \mathbb{R}^n : c(\mathbf{a}; \mathbf{s}) < \infty\}$.

A **trading policy** π is a function $\pi(\mathbf{s}) \equiv \pi(z, \mathbf{s})$ which determines the next action based on our current state, i.e. simply $\mathbf{a} := \pi(\mathbf{s})$.

A trader will usually manage her book by referring to the change in book values plus any other cashflows, most notably cashflows and the cost of hedging. The associated **reward** for

[4] Purchasing a negative quantity is a sell.

[5] Convexity excludes fixed transaction cost.

taking an action $\mathbf{a}$ per time step is given as

$$R(\mathbf{a}; z, \mathbf{m}) := dB(z, \mathbf{m}, \mathbf{M}') + \mathbf{a} \cdot d\mathbf{B}(\mathbf{h}, \mathbf{m}, \mathbf{M}') - c(\mathbf{a}, z, \mathbf{m}) \ . \tag{2}$$

The new joint portfolio tomorrow is given by

$$z'_{\mathbf{a}} := z' + \mathbf{a} \cdot \mathbf{h}' \ . \tag{3}$$

The new state tomorrow is a random variable depending on our action which we write as

$$\mathbf{S}'_{\mathbf{a}} := (z'_{\mathbf{a}}, \mathbf{M}') \ .$$

2.1 The Bellman Equation for Monetary Utilities

Standard reinforcement learning as discussed for example in [SB18] usually aims to maximize the "discounted" expected future rewards of running a given policy. Essentially, the optimal value function V^* is stipulated to satisfy a Bellman equation

$$\begin{cases} V^*\big(z; \mathbf{m}\big) & \stackrel{!}{=} & TV^*(z, \mathbf{m}) \\ \\ Tf(z, \mathbf{m}) & := & \sup_{\mathbf{a} \in \mathcal{A}(z, \mathbf{m})} : \ \mathbb{E}\big[\ \beta(\mathbf{m}) \ f\big(z' + \mathbf{a} \cdot \mathbf{h}'; \mathbf{M}' \big) + R(\mathbf{a}; z, \mathbf{m}) \ \big| \ \mathbf{m} \ \big] \ . \end{cases} \tag{4}$$

Instead of using the expectation it is more natural in finance to choose an operator U which takes into account risk aversion: this roughly means that if two events have the same expected outcome, then we prefer the one with the lower uncertainty. The respective Bellman equation becomes

$$\begin{cases} V^*\big(z; \mathbf{m}\big) & \stackrel{!}{=} & TV^*(z, \mathbf{m}) \\ \\ Tf(z, \mathbf{m}) & := & \sup_{\mathbf{a} \in \mathcal{A}(z, \mathbf{m})} : \ U\big[\ \beta(\mathbf{m}) \ f\big(z' + \mathbf{a} \cdot \mathbf{h}'; \mathbf{M}' \big) + R(\mathbf{a}; z, \mathbf{m}) \ \big| \ \mathbf{m} \ \big] \ . \end{cases} \tag{5}$$

The action $\mathbf{a}$ in above operator has to be found per state $\mathbf{s} = (z, \mathbf{m})$.

We would like to stress that the "value function" here represents the "excess value" of a portfolio over its book value. If V^* were zero, that would mean the optimal risk-adjuste value for a portfolio is given as the book value. Remark 2 makes this statement explicit.

There are many different reasonable risk-adjusted return metrics U used in finance, most notably mean-volatility, mean-variance and their downside versions which where all first discussed in the seminal [Mar52]. Mean-volatility in particular remains a popular choice for many practical applications. However, it is well known that mean-volatility, mean-variance and their downside variants are not monotone, which means that even if $f(\mathbf{s}) \geq g(\mathbf{s})$ for all states $\mathbf{s}$ it is not guaranteed that $U[f(\mathbf{S}') \geq U[g(\mathbf{S}')]$, c.f. [Bue17]. The lack of monotonicity means that standard convergence proofs for the Bellman equation do not apply; see section 5.

We will here take a somewhat more formal route and focus on *monetary utilities*. A functional U is a **monetary utility** if it is monotone increasing (more is better),[6] concave (we are risk averse)[7] and *cash-invariant*. The latter means that for any function $y(\mathbf{s})$ then $U[f(\mathbf{S}', \mathbf{s}) + y(\mathbf{s})|\mathbf{s}] = U[f(\mathbf{S}', \mathbf{s})|\mathbf{s}] + y(\mathbf{s})$. The intuition behind this property is if we add a cash amount y

[6] If $f \geq g$ then $U[f(\mathbf{S}')] \geq U[g(\mathbf{S}')]$.

[7] For $X = f(\mathbf{S}'), Y = g(\mathbf{S}')$ and $\alpha \in [0, 1]$ we have $U[\alpha X + (1 - \alpha)Y] \geq \alpha U[X] + (1 - \alpha)U[Y]$.

to our portfolio, then its monetary utility increases by this amount.[8] An important implication of cash-invariance is that our optimal actions do not depend on our current wealth.

The negative of a monetary utility is also called a *convex risk measure*, c.f. [FS16]. See also [DS05] on the topic of dynamic and time-consistent risk measures.

As in [BMPW22] we will focus on monetary utilities given as *optimized certainty equivalents* (**OCE**) of a utility function, introduced by [BTT07]. A utility function $u : \mathbb{R} \to \mathbb{R}$ here is assumed to be C^1, monotone increasing, and concave. We also normalize it to $u(0) = 0$ and $u'(0) = 1$. The respective OCE monetary utility is then defined by

$$U\big[\ f(\mathbf{S}') \ \big|\ \mathbf{s}\ \big] := \sup_{y(\mathbf{s})\in\mathbb{R}} \mathbb{E}\big[\ u\big(\ f(\mathbf{S}') + y(\mathbf{s})\ \big)\ \big|\ \mathbf{s}\ \big] - y(\mathbf{s})$$

The function y will be modelled as a neural network.

Examples of COE utility functions are

- **Expectation** (risk-neutral): $u(x) := x$.
- **Worst Case**: $u(x) := \inf x$.
- **CVaR** or **Expected Short Fall**: $u(x) := (1+\lambda)\ \min\{0, X\}$.
- **Entropy**: $u(x) := (1 - e^{-\lambda x})/\lambda$ in which case $U[f(\mathbf{S}')|\mathbf{s}] = -\frac{1}{\lambda}\mathbb{E}[\exp(-\lambda f(\mathbf{S}'))|\mathbf{s}]$. The entropy reduces to mean-variance if the variables concerned are normal. It has many other desirable properties, but it also penalizes losses rather harshly: an unhedgable short position in a Black&Scholes stock has negative infinite entropy.
- **Truncated Entropy**: to avoid the harsh penalaties for short positions imposed by the exponential utility we might instead use $u(x) := (1 - e^{-\lambda x})/\lambda 1_{x>0} + (x - \frac{1}{2}\lambda x^2)1_{x<0}$.
- **Vicky**: the following functional was proposed in [HH09]: $u(x) := \frac{1}{\lambda}\left(1 + \lambda x - \sqrt{1 + \lambda^2 x^2}\right)$.
- **Normalized quadratic utility**: $u(x) := -\frac{1}{2}\lambda(x - \frac{1}{\lambda})^2 1_{x<\frac{1}{\lambda}} + \frac{1}{2\lambda}$.

We call a monetary utility *coherent* if $U[n(\mathbf{s})\, f(\mathbf{S}')|\mathbf{s}] = n(\mathbf{s})U[f(\mathbf{S}')|\mathbf{s}]$. An OCE monetary utility is coherent if $u(n\,x) = n\,u(nx)$. Coherence is not the most natural property: it says that the value of risk of a position growths linearly with position size. Usually, we would assume that it increases superlinearly. The practical relevance of this property for us is that if U is coherent, then we can move the discount factor β in and out of our monetary utility: $\beta(\mathbf{s})\, U[f(\mathbf{S}')|\mathbf{s}] = U[\beta(\mathbf{s})\, f(\mathbf{S}')|\mathbf{s}]$.

We say U is *time-consistent* if iterative application lead to the same monetary utility in the sense that $U[U[f(\mathbf{S}'')|\mathbf{S}']|\mathbf{s}] = U(f(\mathbf{S}''|\mathbf{s})$. The only time-consistent OCE montary utilities are the entropy and the expectation, c.f. [KS09].

We may now present the first key result of this article: we say that any statistical arbitrage is finite if

$$\sup_{\mathbf{a}\in\mathcal{A}(\mathbf{s}),\mathbf{s}\in\mathcal{S}} R(\mathbf{a},\mathbf{s}) < \infty\ .$$

This can be achieved for example if $\mathcal{A}(\mathbf{s})$ is bounded.

[8] We have shown in [Bue17] that cash-invariance is equivalent to being able to write-off parts of our portfolio for the worst possible outcome.

THEOREM 1 *Assume any statistical arbitrage is finite.*

Then the Bellman equation (5) has a unique finite solution.

The proof can be found in section 5. It relies on monotonicity and cash-invariance of the monetary utility.

REMARK 2 (Using only Cashflows as Rewards) *Our definition of our rewards (2) as the full mark to market of the hedged portfolio is in so far unusual as the reward term contains future variables, namely the book value of the hedged book tomorrow.*

A more classic approach would be to let the rewards represent only actual cashflows, e.g.

$$\tilde{R}(\mathbf{a}, z, \mathbf{m}) := \underbrace{r(z, \mathbf{m})}_{\substack{\text{Cashflows from} \\ \text{our portfolio}}} \quad \underbrace{-\mathbf{a} \cdot \mathbf{B}(\mathbf{h}, \mathbf{m}) - c(\mathbf{a}, z, \mathbf{m})}_{\substack{\text{Proceeds from} \\ \text{trading } \mathbf{a}}} , \tag{6}$$

The numerical challenge with this formulation is that cashflows are relatively rare for most hedging instruments: if we trade a 1M vanilla option, then it only has one cashflow at maturity – if it ends up in the money that is. That means that learning the value of future cashflows is harder when we train with only daily cashflows. Hence, it is numerically more efficient to solve for the difference between the optimal value function and the book value of a portfolio.

Theoretically, though, the two are equivalent: let $\tilde{V}^*(z, \mathbf{m}) := V^*(z, \mathbf{m}) + B(z, \mathbf{m})$. *Then* $\tilde{V}^*$ *solves the Bellman equation*

$$\tilde{V}^*(z, \mathbf{m}) \stackrel{!}{=} \sup_{\mathbf{a} \in \mathcal{A}(\mathbf{s})} : \; U\Big[\; \beta(\mathbf{m}) \; \tilde{V}^*(\, z' + \mathbf{a} \cdot \mathbf{h}', \mathbf{M}\,) + \tilde{R}(\mathbf{a}, z, \mathbf{m}) \; \Big| \; \mathbf{s} \; \Big] \; . \tag{7}$$

REMARK 3 (Multiple time steps) *It is straight forward to formally extend (5) to multiple time steps. Let* $\mathbf{S}^{(1)} := \mathbf{s}$ *and* $\mathbf{S}^{(i+1)} = \mathbf{S}^{(i)\prime}$. *We use the same numbering for other variables. Define the discount factors* $\beta_i := \prod_{e=1}^{i} \beta(\mathbf{M}^{(e-1)})$ *and set*

$$T_n f(z, \mathbf{m}) := \sup_{\pi} : \; U\left[\; \beta_n \; f\big(\, z^{(n)} + \mathbf{A}^{(n)} \cdot \mathbf{H}^{(n)}; \, \mathbf{M}^{(n+1)} \, \big) + \sum_{i=1}^{n} \beta_{i-1} \, R(\mathbf{A}^{(i)}, \mathbf{S}^{(i)}) \; \Big| \; \mathbf{m} \; \right] \tag{8}$$

where we used $\mathbf{A}^{(n)} := \pi(z^{(n)}, \mathbf{M}^{(n)})$. *Our indexing scheme means that* $T_1 = T$.

It is straight forward to amend the proof of theorem 1 to show that if any statistical arbitrage is finite, then the associated equation $T_n f = f$ *also has a unique finite solution* V_n^*.

However, for $n > 1$ *in general* $V_n^* \neq T^n V^*$ *unless* U *is the expectation.*[9]

[9] We show the claim for $n = 2$. Let $R^{(n)} := R(\mathbf{A}^{(n)}, \mathbf{S}^{(n)})$

$$\begin{aligned}
T^2 f(\mathbf{s}) \quad &= \quad \sup_{\pi} : U\left[\beta(\mathbf{S}^{(1)}) \; \left\{ \sup_{\pi} U\left[\beta(\mathbf{S}^{(2)}) \; f(\pi' \cdots, \mathbf{M}^{(3)}) + R^{(2)} \,|\, \mathbf{S}^{(1)} \right] \right\} + R^{(1)} \,|\, \mathbf{s} \right] \\
&= \quad \sup_{\pi} : U\left[\beta(\mathbf{S}^{(1)}) \; \left\{ U\left[\beta(\mathbf{S}^{(2)}) \; f(\cdots, \mathbf{M}^{(3)}) + R^{(2)} \,|\, \mathbf{S}^{(1)} \right] \right\} + R^{(1)} \,|\, \mathbf{s} \right] \\
&\stackrel{(*)}{\geq / =} \quad \sup_{\pi} : U\Big[\; U\left[\beta(\mathbf{S}^{(1)}) \beta(\mathbf{S}^{(2)}) \; f(\cdots, \mathbf{M}^{(3)}) + R^{(2)} + \beta(\mathbf{S}^{(1)}) R^{(1)} \,|\, \mathbf{S}^{(1)} \right] \,|\, \mathbf{s} \Big] \\
&\stackrel{(**)}{=} \quad \sup_{\pi} U\,[\cdots |\, \mathbf{s}\,] \; = \; T_2 f(\mathbf{s}) \; .
\end{aligned}$$

Here, $(*)$ is an equality for the expectation or any other coherent monetary utility. For all others convexity and $U(0) = 0$ imply the stated inequality. The final equality $(*)$ is only true of U is time-consistent which means either the entropy or the expectation. □

3 Numerical Implementation

We now present an algorithm which will iteratively approach an optimal solution of our Bellman equation (5). This is an extension over the entropy case presented in [BMW20].

We first initialize $V^{(0)}(\,\mathbf{s}\,) := 0$ for all portfolios and states $\mathbf{s}$.[10]

Then we solve iteratively for each n the following scheme $(n-1) \to n$:

1. **Actor**: given $V^{(n-1)}$ we wish to find an optimal neural network policy $\pi^{(n)}$ which solves for all states $\mathbf{s} = (z, \mathbf{m}) \in \mathcal{S}$

$$\sup_{\pi(\mathbf{s})} : \; U\Big[\, \beta(\mathbf{s})\, V^{(n-1)}\big(\, z'_\pi;\, \mathbf{M}'\,\big) + R(\pi; \mathbf{s}) \,\Big|\, \mathbf{s}\,\Big] \qquad (z_\pi := z' + \pi(\mathbf{s}) \cdot \mathbf{h}') \;. \tag{9}$$

   (We recall that $c(\mathbf{a}, \mathbf{s}) = \infty$ whenever $\mathbf{s} \notin \mathcal{A}(\mathbf{s})$.) In the case of our OCE monetary utility, we will need to find both a network $\pi^{(n)}$ and a network $y^{(n)}$ to satisfy for all states $\mathbf{s}$

$$\sup_{\pi, y} : \; \mathbb{E}\Big[\, \beta(\mathbf{s})\, u\Big(\, V^{(n-1)}(\, z'_\pi;\, \mathbf{M}'\,) + y(\mathbf{s})\Big) - y(\mathbf{s}) + R(\pi; \mathbf{s}) \,\Big|\, \mathbf{s}\,\Big] \;. \tag{10}$$

   We will approach this by stipulating that we have a density $\mathbb{Q}$ over all sample $\mathcal{S}$, for example a uniform distribution if $\mathcal{S}$ is a finite set. This allows us defining the unconditional expectation operator $\mathbb{E}[\cdot] = \int \mathbb{Q}[d\mathbf{s}]\, \mathbb{E}[\cdot|\mathbf{s}]$.

   We then solve

$$\sup_{\pi, y} : \; \mathbb{E}\Big[\, u\Big(\, \beta(\mathbf{S})\; V^{(n-1)}(\, Z'_\pi;\, \mathbf{M}'\,) + y(\mathbf{S})\Big) - y(\mathbf{S}) + R(\pi; \mathbf{S}) \,\Big] \quad (Z'_\pi := Z + \pi(\mathbf{S}) \cdot \mathbf{H}') \;. \tag{11}$$

   Under $\mathbb{Q}$ the current market state, the portfolio and the hedging instrument representation are random variables, hence we have referred to them with capital letters.

   The existence of $\mathbb{Q}$ is not trivial: it is meant to represent the probability of possible portfolio and market state conditions. We will discuss this later when we comment on implementation.

2. **Critic (Interpolation)**: as next step, we estimate a new value function $V^{(n)}$ given $\pi^{(n)}$ and $y^{(n)}$. This means fitting a neural network $V^{(n)}$ such that

$$V^{(n)}(z, \mathbf{m}) \equiv T V^{(n-1)}(z, \mathbf{m}) \tag{12}$$

   We note that solving (11) numerically with packages like TensorFlow or PyTorch will also yield samples $TV^{(n-1)}(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{S}$. Assuming this is the case we may find network weights for $V^{(n)}$ by solving the interpolation problem

$$\inf_{V} : \mathbb{E}\Big[\, \Big(-V^{(n)}(Z, \mathbf{M}) + T V^{(n-1)}(Z, \mathbf{M})\Big)^2 \,\Big]$$

   over our discrete sample space.

   Instead of using neural networks for the last step we may also consider classic interpolation techniques such as kernel interpolators.

[10] It is not a good idea to initialize a network with zero to achieve this as all gradients will look rather the same. Assume $\mathcal{N}(\theta; x)$ is a neural network initialized by random weights θ_0. Then use the *Buehler-zero* network $N(\theta; x) := \mathcal{N}(\theta; x) - \mathcal{N}(\theta_0; x)$.

This scheme is reasonably intuitive as it iteratively improves the estimation of the monetary utility $V^{(n)}$ and the optimal action $a^{(n)}$. There is a question on how many training epochs to use when solving, in each step, for the action and the value function. In [SB18] there is a suggestion that using just *one* step is sufficient. The authors call this the **actor-critic** method. There are several other discussions on the viability of such methdos, see also [MBM+16] and the references therein.

REMARK 4 *In some applications we may not be able to use samples of $TV^{(n-1)}$ to solve (12), but make use of trained $a^{(n)}$ and $y^{(n)}$ directly. We therefore may solve*

$$\inf_V : \mathbb{E}\left[\left(-V(\,Z;\,\mathbf{M}\,) + \mathbb{E}\Big[\,\beta(\mathbf{S})\;u\Big(V^{(n-1)}(\,Z'_{\pi^{(n)}},\mathbf{M}'\,) + y^{(n)}(\mathbf{S})\Big) - y^{(n)}(\mathbf{S}) + R(\pi^{(n)},\mathbf{S})\;\Big|\;\mathbf{S}\,\Big]\right)^2\right]. \tag{13}$$

The nested expectation is numerically suboptimal. In order to address this, we solve instead the unconditional

$$\inf_V : \mathbb{E}\left[\left(-V(\,Z;\,\mathbf{M}\,) + \beta(\mathbf{S})\;u\Big(V^{(n-1)}(\,Z'_{\pi^{(n)}},\mathbf{M}'\,) + y^{(n)}(\mathbf{S})\Big) - y^{(n)}(\mathbf{S}) + R(\pi^{(n)},\mathbf{S})\Big)^2\right). \tag{14}$$

which has the same gradient in V, and therefore the same optimal solution.[11]

3.1 Representing Portfolios

The most obvious challenge when applying the approach presented in section 3 is the need to represent our portfolio is some numerically efficient way. The following is an extension of the patent [BMW20] where we proposed using a more cumbersome signature representation of our trader instruments a'la [LNA19].

Assume that we are given historic market data $\mathbf{m}_t$ at time points $\tau_0,\ldots,\tau_N$. Further assume that at each point τ_j we had in our book instruments $\mathbf{x}^t = (x^{t,1},\ldots,x^{t,m_t})$ with $x^{t,i}\in\mathcal{X}$.

As $\mathbf{x}$ were actual historic instruments, we have for each $x^{t,i}$ a vector $\mathbf{f}_t^{t,i}\in\mathbb{R}^F$ of historic risk metrics computed in t, such as the book value, a range of greeks, scenarios and other calculations made in τ_t to assist humans in their risk management decisions. We assume that those metrics $\mathbf{f}_t^t = (\mathbf{f}_t^{t,1},\ldots,\mathbf{f}_t^{t,m_t})$ are also available for *the same instruments* at the next time step τ_{t+1}, denoted by $\mathbf{f}_{t+1}^t$. Instrument which expire between τ_t and τ_{t+1} will have their book value and all greeks and scenario values set to zero.

It is a reasonable assumption that those metrics $\mathbf{f}$ have decent predictive power for the behaviour of our instruments; after all this is what human traders use to do drive their risk management decisions. Hence we will use them as **instrument features**. We will here only consider linear features such that for any weight vector $\mathbf{w}\in\mathbb{R}^{m_t}$ the feature vector (the greeks, scenarios etc) of the weighted instrument $\mathbf{w}\cdot\mathbf{x}^t$ is correctly given as $\mathbf{w}\cdot\mathbf{f}^t$, so there is no need to

[11] *Proof* – Assume that $V(\mathbf{s})\equiv V(\theta;\mathbf{s})$ where θ are our network parameters. Denote by ∂_i the derivative with respect to the ith parameter. Our equation then has the form $\inf_\theta f(\theta)$ where

$$f(\theta) := \mathbb{E}\big[(V(\theta;\mathbf{S}) + \mathbb{E}[h(\mathbf{S}')|\mathbf{S}] + g(\mathbf{S}))^2\big]$$

The gradient is

$$\partial_{\theta_i} f'(\theta) = 2\,\mathbb{E}\big[\partial_i V(\theta;\mathbf{S})(V(\theta;\mathbf{S}) + \mathbb{E}[h(\mathbf{S}')|\mathbf{S}] + g(\mathbf{S}))\big] = 2\,\mathbb{E}\big[\partial_i V(\theta;\mathbf{S})(V(\theta;\mathbf{S}) + h(\mathbf{S}') + g(\mathbf{S}))\big]$$

Therefore f has the same gradient as $\theta\mapsto\mathbb{E}\big[(V(\theta;\mathbf{S}) + h(\mathbf{S}') + g(\mathbf{S}))^2\big]$. □

recompute it later.[12] We have referred to such a representation in [BMW20] as *Finite Markov Representation*, or short *FMR*.

We further denote by r_t^i the historic aggregated cashflows of $x^{t,i}$ over the period $[\tau_t, \tau_t)$, all in our accounting currency. We set $\mathbf{r}_t := (r_t^1, \dots, r_t^{m_t})$. The aggregated cashflows of a weighted instrument $\mathbf{w} \cdot \mathbf{x}^{(t)}$ are $\mathbf{w} \cdot \mathbf{r}_t$. Similarly, we use $\mathbf{B}_u^t = (\mathbf{B}_u^{t:1}, \dots, \mathbf{B}_u^{i,m_t})$ to refer to the book values of our instruments in $u \in \{t, t+1\}$, respectively.

We also assume that we have for all our hedging instruments access to their respective feature vectors $\mathbf{f}_t^{h:t}$ for both τ_t and τ_{t+1}. It is important to recall that the greeks $\mathbf{f}_{t+1}^{h:t,i}$ refer to the features of the ith hedging instrument traded at τ_t, but computed at τ_{t+1}. That means in particular $\mathbf{f}_{t+1}^{h:t,i} \neq \mathbf{f}_{t+1}^{h:t+1,i}$ as the instrument definition changes between time steps. We also denote by $\mathbf{b}_u^{h:t}$ the book values of our hedging instruments for $u \in \{t, t+1\}$.

In addition to our instrument features, we also assume that we chose a reasonable subset of **market features** at each time step τ_t. We continue to use the symbol $\mathbf{m}$ for those features even though in practise we will not use the entire available state vector.

We will now generate random scenarios as follows

1. Randomly choose $t \in \{0, \dots, N-1\}$, which determines the market states $\mathbf{m} := \mathbf{m}_t$ and $\mathbf{m}' := \mathbf{m}_{t+1}$.

2. Identify the hedging instruments $\mathbf{h}$ with their finite Markov representation

Terminal FMR of hedging instruments	$\mathbf{h}'$	:=	$\mathbf{f}_{t+1}^{h:t}$
Book values for our hedging instruments	$\mathbf{B}(\mathbf{h}, \mathbf{s})$	:=	$\mathbf{b}_t^{h:t}$
	$\mathbf{B}(\mathbf{h}, \mathbf{s}')$	:=	$\mathbf{b}_{t+1}^{h:t}$
Cashflows of the hedging instruments	$\mathbf{r}(\mathbf{h}, \mathbf{m})$	:=	$r_t^{h:t}$
Cost	$c(\mathbf{a}; \mathbf{s})$	←	$\mathbf{s}_t, \mathbf{f}_t^h$

   The concrete implementation of the last line depends on the specifics of the cost function. For example, proportional transaction cost on net traded feature exposure are implemented using a weight vector $\boldsymbol{\gamma} \in \mathbb{R}^F$ by setting $c(\mathbf{a}; \mathbf{s}) := |\mathbf{a} \cdot (\boldsymbol{\gamma}\, \mathbf{f}_t^t)|$.

3. Choose a random weight vector $\mathbf{w} \in \mathbb{R}^{m_t}$ and define a sample portfolio as $z := \mathbf{w} \cdot \mathbf{x}$ with

Initial and terminal FMR of the portfolio	z	:=	$\mathbf{w} \cdot \mathbf{f}_t^t$
	z'	:=	$\mathbf{w} \cdot \mathbf{f}_{t+1}^t$
Book value of our portfolio	$B(z, \mathbf{s})$	:=	$\mathbf{w} \cdot \mathbf{b}_t^t$
	$B(z, \mathbf{s}')$	:=	$\mathbf{w} \cdot \mathbf{b}_{t+1}^t$
Cashflows of the portfolio	$r(z, \mathbf{m})$	:=	$\mathbf{w} \cdot \mathbf{x}_t$.

   The construction of a reasonable randomization of the weight vector is important: if the samples are too different from likely portfolios, then the resulting model will underperform. However, if only historic portfolios are used, then the model is less able to learn handling deviations. More importantly, though, generating portfolios increases sample size.

[12] We note that this linearity is satisfied for all common risk metric calculations except VaR and counterparty credit calculations.

4 Relation to Vanilla Deep Hedging

We will now discuss the relation of equation (5) to the solution of a corresponding vanilla Deep Hedging Problem.

We start by stating our original Deep Hedging problem [BGTW19] adapting the notation used here so far. We fix some initial time $t = 0$ with state $\mathbf{s} \equiv \mathbf{s}_0 \equiv \mathbf{S}_0$. Subsequent states are denoted by $\mathbf{S}_{t-1} := \mathbf{S}'_t$. We use a similar notation for all other variables. We also define the stochastic discount factor to zero as $\beta_t := \beta(\mathbf{S}_t)\beta_{t-1}$ starting with $\beta_0 := 1$.

For this part we will need to assume that every hedging instrument has a time-to-maturity less than τ^* in the sense that if we by $\mathbf{a} \cdot \mathbf{h}^t$ at time t, then all the book value and all cashflows from the portfolio are zero beyond $t + \tau^*$. This assumption excludes perpetual assets such as shares or currencies. We will need to trade those with their respective forwards in the current setup.

Assume we are starting with an initial portfolio z and follow a trading policy π. Let $\mathbf{A}_t := \pi(Z_t, \mathbf{M}_t)$. Assume the portfolio has maturity T^*. Then

$$\sum_{t=0}^{\infty} \beta_t R(\mathbf{A}_t, \mathbf{S}_t) = \underbrace{-\mathbf{B}(z, \mathbf{s}_0) + \sum_{t=0}^{T^*} \beta_t r(z, \mathbf{M}_t)}_{\text{P\&L from } z} + \sum_{t=0}^{\infty} \beta_t \underbrace{\left(-\mathbf{A}_t \cdot \mathbf{B}(\mathbf{h}^t, \mathbf{M}_t) - c(\mathbf{A}_t, \mathbf{S}_t) + \mathbf{A}_t \cdot \sum_{u=t}^{t+\tau^*} \frac{\beta_u}{\beta_t} r(\mathbf{h}^t, \mathbf{M}_u) \right)}_{\text{P\&L from trading } \mathbf{h}^t \text{ in } t} \tag{15}$$

If the market is free of statistical arbitrage, then $\mathbb{E}[\sum_{t=0}^{\infty} \beta_t R(\mathbf{A}_t, \mathbf{S}_t)] < \infty$,[13] and therefore $U(\cdot) \leq \mathbb{E}[\cdots] < \infty$.

The **Vanilla Deep Hedging** problem for an infinite trading horizon is then defined as

$$U^*(\mathbf{s}_0) := \sup_{\pi} : \; U\left[\sum_{t=0}^{\infty} \beta_t R(\pi, \mathbf{S}_t) \;\middle|\; \mathbf{s}_0 \right] \tag{16}$$

This formulation is justified if the market is free of statistical arbitrage since then $U^*(\mathbf{s}_0) < \infty$. That means that if U is time-consistent – which means it is the entropy of the expectation –, then U^* satisfies the dynamic programming equation

$$\begin{cases} \hat{U}^*\big(z;\, \mathbf{m}\big) & \stackrel{!}{=} \quad \hat{T}\hat{U}^*(z, \mathbf{m}) \\ \hat{T}f(z, \mathbf{m}) & := \quad \sup_{\mathbf{a} \in \mathcal{A}(z,\mathbf{m})} : \; U\big[\, f\big(z' + \mathbf{a} \cdot \mathbf{h}'; \mathbf{M}'\big) \,\big|\, \mathbf{m} \,\big] + \hat{R}(\mathbf{a}; z, \mathbf{m}) \;. \end{cases} \tag{17}$$

with discounted cashflow rewards

$$\hat{R}(\mathbf{a}; z, \mathbf{m}) := \beta_{t(\mathbf{s})} \left(\, r(z + \mathbf{a} \cdot \mathbf{h}, \mathbf{m}) - \mathbf{a} \cdot B(\mathbf{h}, \mathbf{m}) - c(\mathbf{a}, z, \mathbf{m}) \, \right) \;.$$

(We used $t(\mathbf{s})$ to extract calendar time from the state $\mathbf{s}$.) This is structurally similar to the cashflow rewards (6). Since we have discounted all cashflows in (17) we must interpret $\hat{U}^*$ as units of numeraire. In other words, the actual cash value is $U^*(\mathbf{s}) := \hat{U}^*(\mathbf{s})/\beta_{t(\mathbf{s})}$. Inserting this into our operator yields

$$U^*(z, \mathbf{m}) = \sup_{\mathbf{a} \in \mathcal{A}(z,\mathbf{m})} : \; \frac{1}{\beta_{t(\mathbf{m})}} \, U\big[\, \beta_{t(\mathbf{M}')} U^*\big(z' + \mathbf{a} \cdot \mathbf{h}'; \mathbf{M}'\big) \,\big|\, \mathbf{m} \,\big] + \tilde{R}(\mathbf{a}; z, \mathbf{m}) \;. \tag{18}$$

[13] In (15) the P&L from z is finite since all $x \in \mathcal{X}$ are integrable. If we have no statistical arbitrage it means that each expected P&L from trading $\mathbf{h}^t$ in t has non-positive expectation. Dominance convergence yields the claim. □

If U were coherent, we would get

$$U*(z,\mathbf{m}) = \sup_{\mathbf{a}\in\mathcal{A}(z,\mathbf{m})} : \; U\big[\; \beta(\mathbf{m})\; U^*\big(\, z' + \mathbf{a}\cdot\mathbf{h}';\, \mathbf{M}'\,\big)\,\big|\,\mathbf{m}\;\big] + \tilde{R}(\mathbf{a};z,\mathbf{m})\;.$$

which via the discussion in remark 2 is equivalent to our original Bellman equation (5). However, since the entropy is not coherent it means that this equivalence only holds for the expectation operator.

To summarize:

PROPOSITION 5 *If the market is free of arbitrage, if all hedging instruments have a common finite time-to-maturity, and if $U = \mathbb{E}$, then the value function of the vanilla Deep Hedging problem satisfies our Deep Bellman Hedging equation (5).*

5 Existence of Finite Solutions for Deep Bellman Hedging

We will now prove with theorem 1 convergence of our Deep Bellman Hedging equations. This is easiest understood when the space $\mathcal{Z}$ of future cashflows is parameterized in $\mathbb{R}^{|\mathcal{Z}|}$ with a finite Markov representation. However, in more generality we may assume that $\mathcal{Z}$ represents the set of suitably integrable adapted stochastic processes with values in $\mathbb{R}$. Therefore, we may just assume that $(\mathcal{S},\mathbb{Q}), \mathcal{S} = \mathcal{Z}\times\mathcal{M}$ is a measure space. In the following we will consider the function space F by the $\mathbb{Q}$-equivalence classes of functions $f : \mathcal{S}\to\mathbb{R}$.

Let as before

$$(Tf)(z,\mathbf{m}) := \sup_{\mathbf{a}\in\mathcal{A}(z,\mathbf{m})} : \; U\big[\beta(\mathbf{m})f\big(z' + \mathbf{a}\cdot\mathbf{h}', \mathbf{M}'\big)\big|\mathbf{m}\big] + R(\mathbf{a},z,\mathbf{m}) \tag{19}$$

for $\beta(\mathbf{m}) \le \beta^* < 1$. Then the Bellman equation $f = Tf$ has a unique, finite solution.

We will demonstrate the proof for bounded value functions. See [RZRP03] on how to extend results of convergence of Bellman operators to the unbounded case. The below mimicd the spirit of the proof of the classic Banach contraction theorem.

Proof of theorem 1– **Step 1:** equip F with the supremum norm. We wish to show that for $\|f\|_\infty < \infty$ we have $\|Tf\| < \infty$ We have

$$Tf \le T\|f\| = T0 + \|f\|$$

because of monotonicity and cash-invariance. Since we assumed $\sup_{\mathbf{a}\in\mathcal{A}(\mathbf{s}),\mathbf{s}} : R(\mathbf{a},\mathbf{s}) < \infty$ we find that $T0 < \infty$.

REMARK 6 *It is clear from (19) the gains we can make from the rewards in any time step must be bounded to ensure convergence to a finite optimal point.*

Step 2: We wish to show that Tf is a contraction for bounded f, i.e. $\|Tf - Tg\| \le \beta^*\|f-g\|$ for our $\beta^* < 1$. Note that $f(x) - g(x) \le \|f-g\|$. Monotonicity and cash invariance of the operator T yield

$$(Tf)(x) \le T\left(g + \|f-g\|\right)(x) \le Tg(x) + \beta^*\|f-g\|$$

and, similarly,

$$(Tg)(x) \le T\left(f + \|f-g\|\right)(x) \le Tf(x) + \beta^*\|f-g\|\;.$$

Jointly this gives

$$\|Tf - Tg\| \leq \beta^* \|f - g\| \ .$$

Step 3 Chose f_0 and let $f_n := Tf_{n-1}$ such that $f_n = T^n f_0$. We know that $\|Tf_1 - Tf_0\| \leq \beta^* \|f_1 - f_0\|$ and therefore iteratively $\|Tf_n - Tf_{n-1}\| \leq \beta^{*n} \|f_n - f_{n-1}\|$. Trianlge inequality implies $\|Tf_n - Tf_m\| \leq \sum_{i=m+1}^{n} \|Tf_i - Tf_{i-1}\| \leq \|f_1 - f_0\| \sum_{i=m+1}^{n} \beta^{*i} \downarrow 0$. This means Tf_n is a Cauchy sequence and therefore converges to a unique point $f_n \to f$.

Step 4 To show that f is a fixed point note that $\|Tf - f| \leq \|Tf - f_n\| + \|f_n - f\| \leq \beta^* \|f - f_{n-1}\| + \|f_n - f\| \downarrow 0$. □

Disclaimer

References

[BGTW19] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. Deep hedging. *Quantitative Finance*, 0(0):1–21, 2019.

[BMPW22] Hans Buehler, Phillip Murray, Mikko S. Pakkanen, and Ben Wood. Deep hedging: Learning to remove the drift. *Risk*, March 2022.

[BMW20] Hans Buehler, Louis Moussu, and Ben Wood. Method for optimizing a hedging strategy for portfolio management us patent 16/744514, 2020.

[BTT07] Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, July 2007.

[Bue17] H. Buehler. Statistical hedging. *SSRN*, 2017.

[DJK+20] Jiayi Du, Muyang Jin, Petter N. Kolm, Gordon Ritter, Yixuan Wang, and Bofei Zhang. Deep reinforcement learning for option replication and hedging. *The Journal of Financial Data Science*, 2(4):44–57, 2020.

[DS05] Kai Detlefsen and Giacomo Scandolo. Conditional and dynamic convex risk measures. *Finance and Stochastics*, 9:539–561, 02 2005.

[FS16] Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter, 2016.

[Hal17] Igor Halperin. Qlbs: Q-learner in the black-scholes(-merton) worlds. *SSRN Electronic Journal*, 12 2017.

[Hal19] Igor Halperin. The qlbs q-learner goes nuqlear: fitted q iteration, inverse rl, and option portfolios. *Quantitative Finance*, 19:1543 – 1553, 2019.

[HH09] V. Henderson and D. Hobson. Utility indifference pricing: An overview, 01 2009.

[KR19] Petter N. Kolm and Gordon Ritter. Modern perspectives on reinforcement learning in finance. *Econometrics: Mathematical Methods & Programming eJournal*, 2019.

[KS09] Michael Kupper and Walter Schachermayer. Representation results for law invariant time consistent functions. *Mathematics and Financial Economics*, 2(3):189–210, 2009.

[LNA19] Terry Lyons, Sina Nejad, and Imanol Perez Arribas. Non-parametric pricing and hedging of exotic derivatives. *Applied Mathematical Finance*, 27:457 – 494, 2019.

[Mar52] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[MBM+16] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016.

[RZRP03] Juan Pablo Rincón-Zapatero and Carlos Rodríguez-Palmero. Existence and uniqueness of solutions to the bellman equation in the unbounded case. *Econometrica*, 71(5):1519–1555, 2003.

[SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.