q-Learning in Continuous Time

Xunyu Zhou

Columbia University

(Based on Joint Work with Yanwei Jia)

October 2022, In Memory of Tomas Björk

Tomas Björk

Background and Motivation

Gibbs Sampler and Boltzmann Exploration

q-Learning

Conclusions

Outline

Intra-Personal Equilibrium for Mean–Variance Portfolio Selection

- Basak and Chabakauri (*RFS* 2010) considers mean–variance problem in a Black–Scholes market

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma}{2}\text{Var}[X_{t,x}]$$

Intra-Personal Equilibrium for Mean–Variance Portfolio Selection

- Basak and Chabakauri (*RFS* 2010) considers mean–variance problem in a Black–Scholes market

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma}{2}\text{Var}[X_{t,x}]$$

- Inherently time inconsistent and intra-personal game framework

Intra-Personal Equilibrium for Mean–Variance Portfolio Selection

- Basak and Chabakauri (*RFS* 2010) considers mean–variance problem in a Black–Scholes market

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma}{2}\text{Var}[X_{t,x}]$$

- Inherently time inconsistent and intra-personal game framework
- Optimal dollar amount $h(t)$

Intra-Personal Equilibrium for Mean–Variance Portfolio Selection

- Basak and Chabakauri (*RFS* 2010) considers mean–variance problem in a Black–Scholes market

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma}{2}\text{Var}[X_{t,x}]$$

- Inherently time inconsistent and intra-personal game framework
- Optimal dollar amount $h(t)$
- Reproduced by Björk and Murgoci (2009) by extended HJB equation

Intra-Personal Equilibrium for Mean–Variance Portfolio Selection

- Basak and Chabakauri (*RFS* 2010) considers mean–variance problem in a Black–Scholes market

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma}{2}\text{Var}[X_{t,x}]$$

- Inherently time inconsistent and intra-personal game framework
- Optimal dollar amount $h(t)$
- Reproduced by Björk and Murgoci (2009) by extended HJB equation
- Consider instead *state-dependent* risk aversion

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma(x)}{2}\text{Var}[X_{t,x}]$$

Intra-Personal Equilibrium for Mean–Variance Portfolio Selection

- Basak and Chabakauri (*RFS* 2010) considers mean–variance problem in a Black–Scholes market

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma}{2}\text{Var}[X_{t,x}]$$

- Inherently time inconsistent and intra-personal game framework
- Optimal dollar amount $h(t)$
- Reproduced by Björk and Murgoci (2009) by extended HJB equation
- Consider instead *state-dependent* risk aversion

$$\text{Maximize } \mathbb{E}[X_{t,x}] - \frac{\gamma(x)}{2}\text{Var}[X_{t,x}]$$

- When $\gamma(x) = 1/x$, optimal dollar amount as a feedback policy is $c(t)x$

MATHEMATICAL FINANCE

An International Journal of Mathematics, Statistics and Financial Economics

Mathematical Finance, Vol. 24, No. 1 (January 2014), 1–24

MEAN–VARIANCE PORTFOLIO OPTIMIZATION WITH STATE-DEPENDENT RISK AVERSION

TOMAS BJÖRK

Stockholm School of Economics

AGATHA MURGOCI

*Copenhagen Business School**

XUN YU ZHOU

University of Oxford

The objective of this paper is to study the mean–variance portfolio optimization in continuous time. Since this problem is time inconsistent we attack it by placing the problem within a game theoretic framework and look for subgame perfect Nash equilibrium strategies. This particular problem has already been studied in Basak and Chabakauri where the authors assumed a constant risk aversion parameter. This assumption leads to an equilibrium control where the dollar amount invested in the risky asset is independent of current wealth, and we argue that this result is unrealistic from an economic point of view. In order to have a more realistic model we instead study the case when the risk aversion depends dynamically on current wealth. This is a substantially more complicated problem than the one with constant risk aversion but, using the general theory of time-inconsistent control developed in Björk and Murgoci, we provide a fairly detailed analysis on the general case. In particular, when the risk aversion is inversely proportional to wealth, we provide an analytical solution where the equilibrium dollar amount invested in the risky asset is proportional to current wealth. The equilibrium for this model thus appears more reasonable than the one for the model with constant risk aversion.

KEY WORDS: mean–variance, time inconsistency, time-inconsistent control, dynamic programming, stochastic control, Hamilton–Jacobi–Bellman equation.

1. INTRODUCTION

Mean–variance (MV) analysis for optimal asset allocation is one of the classical results of financial economics. After the original publication in Markowitz (1952), a vast number of papers have been published on this topic. Most of these papers deal with the single period case, and there is a very good reason for this: It is very easy to see that an MV optimal

The authors are greatly indebted to Ivar Ekeland, Ali Lazrak, Traian Pirvu, and Suleyman Basak for very helpful discussions. We are also very grateful to two anonymous referees for a number of comments, which have improved the paper considerably.

*Agatha Murgoci's affiliation was wrongly published online as Stockholm School of Economics on 3 Feb 2012.

Manuscript received January 2011; final revision received October 2011.

Address correspondence to Tomas Björk, Department of Finance, Stockholm School of Economics, PO Box 6501, Stockholm SE 11383, Sweden; e-mail: Tomas.Bjork@hhs.se.

DOI: 10.1111/j.1467-9965.2011.00515.x

Tomas Björk

Stockholm School of Economics

	All	Since 2017
Citations	7828	2407
h-index	26	14
i10-index	39	17

TITLE	CITED BY	YEAR
Arbitrage theory in continuous time T Björk Oxford university press	3551	2009
Bond market structure in the presence of marked point processes T Björk, Y Kabanov, W Runggaldier Mathematical Finance 7 (2), 211-239	457	1997
Mean–variance portfolio optimization with state-dependent risk aversion T Björk, A Murgoci, XY Zhou Mathematical Finance: An International Journal of Mathematics, Statistics …	447	2014
Interest rate dynamics and consistent forward rate curves T Björk, BJ Christensen Mathematical Finance 9 (4), 323-348	438	1999
A general theory of Markovian time inconsistent stochastic control problems T Bjork, A Murgoci Available at SSRN 1694759	350	2010
Towards a general theory of bond markets T Björk, G Di Masi, Y Kabanov, W Runggaldier Finance and Stochastics 1 (2), 141-174	281	1997
A note on Wick products and the fractional Black-Scholes model T Björk, H Hult Finance and Stochastics 9 (2), 197-209	244	2005
On time-inconsistent stochastic control in continuous time T Björk, M Khapko, A Murgoci Finance and Stochastics 21 (2), 331-360	220	2017
A theory of Markovian time-inconsistent stochastic control in discrete time T Björk, A Murgoci Finance and Stochastics 18 (3), 545-592	173	2014
On the existence of finite-dimensional realizations for nonlinear forward rate models T Björk, L Svensson Mathematical Finance 11 (2), 205-243	172	2001

Outline

Reinforcement Learning

- Reinforcement learning (RL): an active and fast developing subareas in machine learning

Reinforcement Learning

- Reinforcement learning (RL): an active and fast developing subareas in machine learning
- RL mimics humans' – especially children's – learning process

Reinforcement Learning

- Reinforcement learning (RL): an active and fast developing subareas in machine learning
- RL mimics humans' – especially children's – learning process
- An RL agent learns the best strategies based on *trial and error*, through interactions with the black box environment (e.g. the market)

Reinforcement Learning

- Reinforcement learning (RL): an active and fast developing subareas in machine learning
- RL mimics humans' – especially children's – learning process
- An RL agent learns the best strategies based on *trial and error*, through interactions with the black box environment (e.g. the market)
- RL learns strategies directly, *not* a model

Reinforcement Learning

- Reinforcement learning (RL): an active and fast developing subareas in machine learning
- RL mimics humans' – especially children's – learning process
- An RL agent learns the best strategies based on *trial and error*, through interactions with the black box environment (e.g. the market)
- RL learns strategies directly, *not* a model
- This is in sharp contrast with classical model-based methods

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem
- Key components of model-free RL

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem
- Key components of model-free RL
  - *Exploration* (trial and error): broaden search space via randomization (stochastic policies)

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem
- Key components of model-free RL
  - *Exploration* (trial and error): broaden search space via randomization (stochastic policies)
  - *Policy evaluation* (PE): estimate value function of a given policy using samples only

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem
- Key components of model-free RL
  - *Exploration* (trial and error): broaden search space via randomization (stochastic policies)
  - *Policy evaluation* (PE): estimate value function of a given policy using samples only
  - *Policy improvement* (PI): improve and update current policy based on learned value function

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem
- Key components of model-free RL
  - *Exploration* (trial and error): broaden search space via randomization (stochastic policies)
  - *Policy evaluation* (PE): estimate value function of a given policy using samples only
  - *Policy improvement* (PI): improve and update current policy based on learned value function
    - *Policy gradient* (PG): update current policy along gradient of value function in policy

Key Elements of Reinforcement Learning

- RL: stochastic control (dynamic optimization) without knowledge about environment
- Key elements of model-based stochastic control: Bellman's principle, HJB equation, verification theorem
- Key components of model-free RL
  - *Exploration* (trial and error): broaden search space via randomization (stochastic policies)
  - *Policy evaluation* (PE): estimate value function of a given policy using samples only
  - *Policy improvement* (PI): improve and update current policy based on learned value function
    - *Policy gradient* (PG): update current policy along gradient of value function in policy
    - *Q-learning*: learn the Q-function to generate an improved policy

Pitfalls of Current RL Study

- Two major limitations in existing study on RL

Pitfalls of Current RL Study

- Two major limitations in existing study on RL
- Most algorithms developed for discrete-time Markov Decision Processes (MDPs), and little attention paid to problems with *continuous* time and spaces

Pitfalls of Current RL Study

- Two major limitations in existing study on RL
- Most algorithms developed for discrete-time Markov Decision Processes (MDPs), and little attention paid to problems with *continuous* time and spaces
  - World is inherently continuous in time

Pitfalls of Current RL Study

- Two major limitations in existing study on RL
- Most algorithms developed for discrete-time Markov Decision Processes (MDPs), and little attention paid to problems with *continuous* time and spaces
  - World is inherently continuous in time
  - Abundant real-life examples in which an agent can or need to interact with a *random* environment at ultra-high frequency

Pitfalls of Current RL Study

- Two major limitations in existing study on RL
- Most algorithms developed for discrete-time Markov Decision Processes (MDPs), and little attention paid to problems with *continuous* time and spaces
  - World is inherently continuous in time
  - Abundant real-life examples in which an agent can or need to interact with a *random* environment at ultra-high frequency
  - Few existing studies in continuous setting restricted to *deterministic* systems (Baird 1993, Doya 2000, Frémaux et al. 2013, Lee and Sutton 2021)

Pitfalls of Current RL Study

- Two major limitations in existing study on RL
- Most algorithms developed for discrete-time Markov Decision Processes (MDPs), and little attention paid to problems with *continuous* time and spaces
  - World is inherently continuous in time
  - Abundant real-life examples in which an agent can or need to interact with a *random* environment at ultra-high frequency
  - Few existing studies in continuous setting restricted to *deterministic* systems (Baird 1993, Doya 2000, Frémaux et al. 2013, Lee and Sutton 2021)
- Many RL algorithms were devised in heuristic and *ad hoc* manners with underlying objectives not always clearly stated

Pitfalls of Current RL Study

- Two major limitations in existing study on RL
- Most algorithms developed for discrete-time Markov Decision Processes (MDPs), and little attention paid to problems with *continuous* time and spaces
  - World is inherently continuous in time
  - Abundant real-life examples in which an agent can or need to interact with a *random* environment at ultra-high frequency
  - Few existing studies in continuous setting restricted to *deterministic* systems (Baird 1993, Doya 2000, Frémaux et al. 2013, Lee and Sutton 2021)
- Many RL algorithms were devised in heuristic and *ad hoc* manners with underlying objectives not always clearly stated
- In short, there seems a lack of an overarching theoretical understanding and a *unified* framework for RL methods

RL in Continuous Time and Spaces

- Bridge these gaps by providing a unified theoretical underpinning of RL in continuous time with possibly continuous state and action spaces

RL in Continuous Time and Spaces

- Bridge these gaps by providing a unified theoretical underpinning of RL in continuous time with possibly continuous state and action spaces
- Carry out all theoretical analysis for the continuous setting and take discrete *observations* at the final, algorithmic stage

RL in Continuous Time and Spaces

- Bridge these gaps by providing a unified theoretical underpinning of RL in continuous time with possibly continuous state and action spaces
- Carry out all theoretical analysis for the continuous setting and take discrete *observations* at the final, algorithmic stage
- Rule out sensitivity in time step size

RL in Continuous Time and Spaces

- Bridge these gaps by providing a unified theoretical underpinning of RL in continuous time with possibly continuous state and action spaces
- Carry out all theoretical analysis for the continuous setting and take discrete *observations* at the final, algorithmic stage
- Rule out sensitivity in time step size
- Make use of well-developed tools in stochastic calculus, differential equations, and stochastic control, which enables better interpretability/explainability to underlying learning technologies

RL in Continuous Time and Spaces

- Bridge these gaps by providing a unified theoretical underpinning of RL in continuous time with possibly continuous state and action spaces
- Carry out all theoretical analysis for the continuous setting and take discrete *observations* at the final, algorithmic stage
- Rule out sensitivity in time step size
- Make use of well-developed tools in stochastic calculus, differential equations, and stochastic control, which enables better interpretability/explainability to underlying learning technologies
- Provide new perspectives on RL overall

Research Questions

- How to explore strategically?
- How to do PE?
- How to do PI generally?
- How to do PG specifically?
- Financial applications?

A Pentalogy

- H. Wang, T. Zariphopoulou and X. Zhou, "Reinforcement learning in continuous time and space: A stochastic control approach", *Journal of Machine Learning Research*, 2020.
- Y. Jia and X. Zhou, "Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach", *Journal of Machine Learning Research*, 2022a.
- Y. Jia and X. Zhou, "Policy gradient and actor–critic learning in continuous time and space: Theory and algorithms", *Journal of Machine Learning Research*, 2022b.
- Y. Jia and X. Zhou, "q-Learning in continuous time", arXiv:2207.00713, 2022c.
- Y. Huang, Y. Jia and X. Zhou, "Data-driven mean-variance portfolio selection", work in progress.

Outline

Problem Formulation

- $\left(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{F}_t^W\}_{t\geq 0}\right)$, Brownian motion $W = \{W_t,\, t \geq 0\}$
- Action space $\mathcal{A}$: representing constraints on an agent's actions (or "controls")
- Admissible action $a = \{a_t,\, t \geq 0\}$: an $\{\mathcal{F}_t^W\}_{t\geq 0}$-adapted measurable process taking value in $\mathcal{A}$
- State (or "feature") dynamics in $\mathbb{R}^d$

$$\mathrm{d}X_t = b(t, X_t, a_t)\mathrm{d}t + \sigma(t, X_t, a_t)\mathrm{d}W_t,\ t > 0$$

- Objective: to achieve maximum expected total reward represented by *optimal value function*

$$w\,(t,x) := \sup \mathbb{E}\left[\int_t^T r\,(s, X_s, a_s)\,\mathrm{d}s + h(X_T)\middle| X_t = x\right],$$

where $(t,x) \in [0,T] \times \mathbb{R}^d$

Classical Model-Based Approach

- Dynamic programming (Fleming and Soner 1992, Yong and Z. 1998)
- HJB equation: optimal value function w satisfies

$$\frac{\partial v}{\partial t}(t,x) + \sup_{a \in \mathcal{A}} H(t,x,a,\frac{\partial v}{\partial x}(t,x),\frac{\partial^2 v}{\partial x^2}(t,x)) = 0; \quad v(T,x) = h(x)$$

- ... where (generalized) *Hamiltonian* (Yong and Z. 1998)

$$H(t,x,a,p,P) = \frac{1}{2}\mathrm{tr}\left[\sigma(t,x,a)' P \sigma(t,x,a)\right] + p \cdot b(t,x,a) + r(t,x,a)$$

- Verification theorem: optimal (feedback) control *policy* is

$$\mathbf{a}(t,x) = \mathrm{argmax}_{a \in \mathcal{A}} H\left(t,x,a,\frac{\partial v}{\partial x}(t,x),\frac{\partial^2 v}{\partial x^2}(t,x)\right)$$

- *Deterministic* policy, devised at $t = 0$
- This approach requires the knowledge of environment (functional forms of b, σ, r, h)

Trial and Error: Exploration through Randomization

- Absence of knowledge of environment

Trial and Error: Exploration through Randomization

- Absence of knowledge of environment
- Exploration is modelled by a *distribution* (*randomization*) of policies

Trial and Error: Exploration through Randomization

- Absence of knowledge of environment
- Exploration is modelled by a *distribution* (*randomization*) of policies
- *Stochastic* policies

Trial and Error: Exploration through Randomization

- Absence of knowledge of environment
- Exploration is modelled by a *distribution* (*randomization*) of policies
- *Stochastic* policies
- Actions are sampled from a policy to be *actually* executed

Trial and Error: Exploration through Randomization

- Absence of knowledge of environment
- Exploration is modelled by a *distribution* (*randomization*) of policies
- *Stochastic* policies
- Actions are sampled from a policy to be *actually* executed
- Notion of controls extended to distributions/measures

Trial and Error: Exploration through Randomization

- Absence of knowledge of environment
- Exploration is modelled by a *distribution* (*randomization*) of policies
- *Stochastic* policies
- Actions are sampled from a policy to be *actually* executed
- Notion of controls extended to distributions/measures
- Randomization itself is *independent* of Brownian motion W

Stochastic Policies

- Probability space is rich enough to support $Z \sim U(0, 1)$ independent of W

Stochastic Policies

- Probability space is rich enough to support $Z \sim U(0,1)$ independent of W
- $\mathcal{F}_s = \mathcal{F}_s^W \vee \sigma(Z)$

Stochastic Policies

- Probability space is rich enough to support $Z \sim U(0,1)$ independent of W
- $\mathcal{F}_s = \mathcal{F}_s^W \vee \sigma(Z)$
- $\mathcal{P}(\mathcal{A})$: collection of probability density functions (pdfs) on $\mathcal{A}$

Stochastic Policies

- Probability space is rich enough to support $Z \sim U(0,1)$ independent of W
- $\mathcal{F}_s = \mathcal{F}_s^W \vee \sigma(Z)$
- $\mathcal{P}(\mathcal{A})$: collection of probability density functions (pdfs) on $\mathcal{A}$
- Let $\boldsymbol{\pi} : (t,x) \in [0,T] \times \mathbb{R}^d \mapsto \boldsymbol{\pi}(\cdot|t,x) \in \mathcal{P}(\mathcal{A})$ be a given (stochastic) policy

Stochastic Policies

- Probability space is rich enough to support $Z \sim U(0,1)$ independent of W
- $\mathcal{F}_s = \mathcal{F}_s^W \vee \sigma(Z)$
- $\mathcal{P}(\mathcal{A})$: collection of probability density functions (pdfs) on $\mathcal{A}$
- Let $\boldsymbol{\pi} : (t,x) \in [0,T] \times \mathbb{R}^d \mapsto \boldsymbol{\pi}(\cdot|t,x) \in \mathcal{P}(\mathcal{A})$ be a given (stochastic) policy
- At each time s, an action a_s is sampled from distribution $\boldsymbol{\pi}(\cdot|s,X_s)$

An Exploratory Formulation

- Let $\{\mathcal{F}_s\}_{s\geq 0}$-progressively measurable action process $a^{\boldsymbol{\pi}} = \{a_s^{\boldsymbol{\pi}}, t \leq s \leq T\}$ be generated from $\boldsymbol{\pi}$

An Exploratory Formulation

- Let $\{\mathcal{F}_s\}_{s\geq 0}$-progressively measurable action process $a^{\boldsymbol{\pi}} = \{a_s^{\boldsymbol{\pi}}, t \leq s \leq T\}$ be generated from $\boldsymbol{\pi}$
- Corresponding state process follows

$$\mathrm{d}X_s^{\boldsymbol{\pi}} = b(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}})\mathrm{d}s + \sigma(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}})\mathrm{d}W_s,\ s \in [t, T];\ \ X_t^{\boldsymbol{\pi}} = x$$

An Exploratory Formulation

- Let $\{\mathcal{F}_s\}_{s\geq 0}$-progressively measurable action process $a^{\boldsymbol{\pi}} = \{a_s^{\boldsymbol{\pi}}, t \leq s \leq T\}$ be generated from $\boldsymbol{\pi}$
- Corresponding state process follows

$$\mathrm{d}X_s^{\boldsymbol{\pi}} = b(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}})\mathrm{d}s + \sigma(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}})\mathrm{d}W_s,\ s \in [t, T];\quad X_t^{\boldsymbol{\pi}} = x$$

- A *regularizer* is included to encourage exploration

$$J(t, x; \boldsymbol{\pi}) = \mathbb{E}^{\mathbb{P}}\left(\int_t^T \left[r(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}) + \gamma p\big(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}, \boldsymbol{\pi}(\cdot|s, X_s^{\boldsymbol{\pi}})\big)\right] \mathrm{d}s + h(X_T^{\boldsymbol{\pi}}) \Big| X_t^{\boldsymbol{\pi}} = x \right)$$

An Exploratory Formulation

- Let $\{\mathcal{F}_s\}_{s\geq 0}$-progressively measurable action process $a^{\boldsymbol{\pi}} = \{a_s^{\boldsymbol{\pi}}, t \leq s \leq T\}$ be generated from $\boldsymbol{\pi}$
- Corresponding state process follows

$$\mathrm{d}X_s^{\boldsymbol{\pi}} = b(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}})\mathrm{d}s + \sigma(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}})\mathrm{d}W_s,\ s \in [t, T];\quad X_t^{\boldsymbol{\pi}} = x$$

- A *regularizer* is included to encourage exploration

$$J(t, x; \boldsymbol{\pi}) = \mathbb{E}^{\mathbb{P}}\Bigg(\int_t^T \left[r(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}) + \gamma p\big(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}, \boldsymbol{\pi}(\cdot|s, X_s^{\boldsymbol{\pi}})\big)\right] \mathrm{d}s + h(X_T^{\boldsymbol{\pi}}) \Big| X_t^{\boldsymbol{\pi}} = x\Bigg)$$

- Entropy regularizer (Wang, Zariphoupoulou, Z. 2020)

$$p\big(t, x, a, \pi(\cdot)\big) = -\log \pi(a)$$

Entropy Regularization and Gibbs Measure

- With entropy regularization, optimal stochastic policy (Wang, Zariphoupoulou, Z. 2020)

$$\boldsymbol{\pi}^*(a|t,x) = \frac{1}{Z(\gamma)} \exp\left(\frac{1}{\gamma} H(t,x,a,v_x(t,x),v_{xx}(t,x))\right)$$

where

$$Z(\gamma) = \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} H(t,x,a,v_x(t,x),v_{xx}(t,x))\right) \mathrm{d}a$$

- *Gibbs measure* or *Boltzmann exploration*
- Gaussian in LQ case
- Mean–variance (Wang and Z. 2020)

Outline

A Policy Improvement Theorem

Theorem (Wang and Z. 2020, Jia and Z. 2022c)

Given $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, *define*

$$\boldsymbol{\pi}'(\cdot|t,x) \propto \exp\left\{\frac{1}{\gamma}H\Big(t,x,\cdot,\frac{\partial J}{\partial x}(t,x;\boldsymbol{\pi}),\frac{\partial J}{\partial x^2}(t,x;\boldsymbol{\pi})\Big)\right\}.$$

If $\boldsymbol{\pi}' \in \boldsymbol{\Pi}$, *then*

$$J(t,x;\boldsymbol{\pi}') \geq J(t,x;\boldsymbol{\pi}).$$

Moreover, if the following map

$$\mathcal{I}(\boldsymbol{\pi}) = \frac{\exp\{\frac{1}{\gamma}H\big(t,x,\cdot,\frac{\partial J}{\partial x}(t,x;\boldsymbol{\pi}),\frac{\partial J}{\partial x^2}(t,x;\boldsymbol{\pi})\big)\}}{\int_{\mathcal{A}}\exp\{\frac{1}{\gamma}H\big(t,x,a,\frac{\partial J}{\partial x}(t,x;\boldsymbol{\pi}),\frac{\partial J}{\partial x^2}(t,x;\boldsymbol{\pi})\big)\}\mathrm{d}a}, \quad \boldsymbol{\pi} \in \boldsymbol{\Pi}$$

has a fixed point $\boldsymbol{\pi}^*$ *on* $\boldsymbol{\Pi}$, *then* $\boldsymbol{\pi}^*$ *is the optimal policy.*

Q-Learning

- The previous theorem is not implementable for learning because H is unknown

Q-Learning

- The previous theorem is not implementable for learning because H is unknown
- Recall classical stochastic control

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^T r\left(s,X_s,a_s\right)\mathrm{d}s+h(X_T)\middle| X_t=x\right]$$

Q-Learning

- The previous theorem is not implementable for learning because H is unknown
- Recall classical stochastic control

$$w\left(t,x\right) = \sup \mathbb{E}\left[\int_t^T r\left(s, X_s, a_s\right) \mathrm{d}s + h(X_T) \middle| X_t = x\right]$$

- Bellman's principle of optimality

$$w\left(t,x\right) = \sup \mathbb{E}\left[\int_t^{t+\Delta t} r\left(s, X_s, a_s\right) \mathrm{d}s + w(t+\Delta t, X_{t+\Delta t}) \middle| X_t = x\right]$$

Q-Learning

- The previous theorem is not implementable for learning because H is unknown
- Recall classical stochastic control

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^T r\left(s,X_s,a_s\right)\mathrm{d}s+h(X_T)\middle|\,X_t=x\right]$$

- Bellman's principle of optimality

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^{t+\Delta t} r\left(s,X_s,a_s\right)\mathrm{d}s+w(t+\Delta t,X_{t+\Delta t})\middle|\,X_t=x\right]$$

- Q-function

$$Q_{\Delta t}\left(t,x,a\right)=\mathbb{E}\left[\int_t^{t+\Delta t} r\left(s,X_s,a\right)\mathrm{d}s+w(t+\Delta t,X_{t+\Delta t})\middle|\,X_t=x\right]$$

Q-Learning

- The previous theorem is not implementable for learning because H is unknown
- Recall classical stochastic control

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^T r\left(s,X_s,a_s\right)\mathrm{d}s+h(X_T)\middle| X_t=x\right]$$

- Bellman's principle of optimality

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^{t+\Delta t} r\left(s,X_s,a_s\right)\mathrm{d}s+w(t+\Delta t,X_{t+\Delta t})\middle| X_t=x\right]$$

- Q-function

$$Q_{\Delta t}\left(t,x,a\right)=\mathbb{E}\left[\int_t^{t+\Delta t} r\left(s,X_s,a\right)\mathrm{d}s+w(t+\Delta t,X_{t+\Delta t})\middle| X_t=x\right]$$

- $Q^*_{\Delta t}\left(t,x\right)=\sup_a Q_{\Delta t}\left(t,x,a\right)$

Q-Learning

- The previous theorem is not implementable for learning because H is unknown
- Recall classical stochastic control

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^T r\left(s,X_s,a_s\right)\mathrm{d}s+h(X_T)\middle|\, X_t=x\right]$$

- Bellman's principle of optimality

$$w\left(t,x\right)=\sup\mathbb{E}\left[\int_t^{t+\Delta t} r\left(s,X_s,a_s\right)\mathrm{d}s+w(t+\Delta t,X_{t+\Delta t})\middle|\, X_t=x\right]$$

- Q-function

$$Q_{\Delta t}\left(t,x,a\right)=\mathbb{E}\left[\int_t^{t+\Delta t} r\left(s,X_s,a\right)\mathrm{d}s+w(t+\Delta t,X_{t+\Delta t})\middle|\, X_t=x\right]$$

- $Q^*_{\Delta t}\left(t,x\right)=\sup_a Q_{\Delta t}\left(t,x,a\right)$
- In chess, "what should be the current best move, assuming I will always follow the best moves afterwards"?

No Q-Function in Continuous Time!

- Q-learning works inherently for discrete-time only: Δt is fixed

No Q-Function in Continuous Time!

- Q-learning works inherently for discrete-time only: Δt is fixed
- Q-function collapses in continuous time when $\Delta t \to 0$ (Tallec et al. 2019)

No Q-Function in Continuous Time!

- Q-learning works inherently for discrete-time only: Δt is fixed
- Q-function collapses in continuous time when $\Delta t \to 0$ (Tallec et al. 2019)
- Impact of any action a is negligible on $[t, t + \Delta t]$ when $\Delta t \to 0$

No Q-Function in Continuous Time!

- Q-learning works inherently for discrete-time only: Δt is fixed
- Q-function collapses in continuous time when $\Delta t \to 0$ (Tallec et al. 2019)
- Impact of any action a is negligible on $[t, t + \Delta t]$ when $\Delta t \to 0$
- What should be a proper continuous-time counterpart of Q-function?

Continuous Time

- Given a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, define

$$
\begin{aligned}
& Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) \\
:= & \mathbb{E}^{\mathbb{P}}\Big[\int_t^{t+\Delta t} r(s, X_s^a, a)\mathrm{d}s \\
& + \mathbb{E}^{\mathbb{P}}[\int_{t+\Delta t}^{T} [r(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}) - \gamma \log \boldsymbol{\pi}(a_s^{\boldsymbol{\pi}} | s, X_s^{\boldsymbol{\pi}})]\mathrm{d}s + h(X_T^{\boldsymbol{\pi}}) | X_{t+\Delta t}^a] \Big| X_t^{\boldsymbol{\pi}} = x\Big] \\
= & J(t, x; \boldsymbol{\pi}) + \left[\frac{\partial J}{\partial t}(t, x; \boldsymbol{\pi}) + H\left(t, x, a, \frac{\partial J}{\partial x}(t, x; \boldsymbol{\pi}), \frac{\partial^2 J}{\partial x^2}(t, x; \boldsymbol{\pi})\right)\right] \Delta t + o(\Delta t)
\end{aligned}
$$

Continuous Time

- Given a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, define

$$
\begin{aligned}
& Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) \\
:= & \mathbb{E}^{\mathbb{P}}\Big[\int_t^{t+\Delta t} r(s, X_s^a, a) \mathrm{d}s \\
& + \mathbb{E}^{\mathbb{P}}[\int_{t+\Delta t}^{T} [r(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}) - \gamma \log \boldsymbol{\pi}(a_s^{\boldsymbol{\pi}} | s, X_s^{\boldsymbol{\pi}})] \mathrm{d}s + h(X_T^{\boldsymbol{\pi}}) | X_{t+\Delta t}^a] \Big| X_t^{\boldsymbol{\pi}} = x \Big] \\
= & J(t, x; \boldsymbol{\pi}) + \left[\frac{\partial J}{\partial t}(t, x; \boldsymbol{\pi}) + H\left(t, x, a, \frac{\partial J}{\partial x}(t, x; \boldsymbol{\pi}), \frac{\partial^2 J}{\partial x^2}(t, x; \boldsymbol{\pi})\right)\right] \Delta t + o(\Delta t)
\end{aligned}
$$

- Leading term J is independent of a, as expected

Continuous Time

- Given a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, define

$$
\begin{aligned}
&Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) \\
:=& \mathbb{E}^{\mathbb{P}}\Big[\int_t^{t+\Delta t} r(s, X_s^a, a) \mathrm{d}s \\
&+ \mathbb{E}^{\mathbb{P}}\big[\int_{t+\Delta t}^T [r(s, X_s^{\boldsymbol{\pi}}, a_s^{\boldsymbol{\pi}}) - \gamma \log \boldsymbol{\pi}(a_s^{\boldsymbol{\pi}} | s, X_s^{\boldsymbol{\pi}})] \mathrm{d}s + h(X_T^{\boldsymbol{\pi}}) | X_{t+\Delta t}^a \big] \Big| X_t^{\boldsymbol{\pi}} = x \Big] \\
=& J(t, x; \boldsymbol{\pi}) + \left[\frac{\partial J}{\partial t}(t, x; \boldsymbol{\pi}) + H\left(t, x, a, \frac{\partial J}{\partial x}(t, x; \boldsymbol{\pi}), \frac{\partial^2 J}{\partial x^2}(t, x; \boldsymbol{\pi}) \right) \right] \Delta t + o(\Delta t)
\end{aligned}
$$

- Leading term J is independent of a, as expected
- Consider the first-order term instead!

q-Function

Definition (Jia and Z. 2022c)

The q-function associated with a given stochastic policy $\boldsymbol{\pi} \in \mathbf{\Pi}$ is defined as

$$q(t, x, a; \boldsymbol{\pi}) = \frac{\partial J}{\partial t}(t, x; \boldsymbol{\pi}) + H\left(t, x, a, \frac{\partial J}{\partial x}(t, x; \boldsymbol{\pi}), \frac{\partial^2 J}{\partial x^2}(t, x; \boldsymbol{\pi})\right).$$

Discussions

- q-Function is first-order *derivative* of conventional Q-function in time:

$$q(t, x, a; \boldsymbol{\pi}) = \lim_{\Delta t \to 0} \frac{Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) - J(t, x; \boldsymbol{\pi})}{\Delta t}$$

Discussions

- q-Function is first-order *derivative* of conventional Q-function in time:

$$q(t, x, a; \boldsymbol{\pi}) = \lim_{\Delta t \to 0} \frac{Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) - J(t, x; \boldsymbol{\pi})}{\Delta t}$$

- A continuous-time notion because *it does not depend on any time-discretization*

Discussions

- q-Function is first-order *derivative* of conventional Q-function in time:

$$q(t, x, a; \boldsymbol{\pi}) = \lim_{\Delta t \to 0} \frac{Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) - J(t, x; \boldsymbol{\pi})}{\Delta t}$$

- A continuous-time notion because *it does not depend on any time-discretization*
- Vital advantage for learning algorithm design as performance of RL algorithms is very sensitive wrt time discretization step (Tallec et al. 2019)

Discussions

- q-Function is first-order *derivative* of conventional Q-function in time:

$$q(t, x, a; \boldsymbol{\pi}) = \lim_{\Delta t \to 0} \frac{Q_{\Delta t}(t, x, a; \boldsymbol{\pi}) - J(t, x; \boldsymbol{\pi})}{\Delta t}$$

- A continuous-time notion because *it does not depend on any time-discretization*
- Vital advantage for learning algorithm design as performance of RL algorithms is very sensitive wrt time discretization step (Tallec et al. 2019)
- Policy improvement theorem can now be expressed in terms of q-function:

$$\boldsymbol{\pi}'(\cdot|t, x) \propto \exp\left\{\frac{1}{\gamma} q(t, x, \cdot; \boldsymbol{\pi})\right\}$$

Discussions

- q-Function is first-order *derivative* of conventional Q-function in time:

$$q(t,x,a;\boldsymbol{\pi}) = \lim_{\Delta t \to 0} \frac{Q_{\Delta t}(t,x,a;\boldsymbol{\pi}) - J(t,x;\boldsymbol{\pi})}{\Delta t}$$

- A continuous-time notion because *it does not depend on any time-discretization*
- Vital advantage for learning algorithm design as performance of RL algorithms is very sensitive wrt time discretization step (Tallec et al. 2019)
- Policy improvement theorem can now be expressed in terms of q-function:

$$\boldsymbol{\pi}'(\cdot|t,x) \propto \exp\left\{\frac{1}{\gamma} q(t,x,\cdot;\boldsymbol{\pi})\right\}$$

- Only need to learn q-function $q(\cdot,\cdot,\cdot;\boldsymbol{\pi})$ under any policy $\boldsymbol{\pi}$

Martingale Characterization

Theorem (Jia and Z. 2022c)

Let a policy $\boldsymbol{\pi} \in \mathbf{\Pi}$, a function $\hat{J} \in C^{1,2}\big([0,T) \times \mathbb{R}^d\big) \cap C\big([0,T] \times \mathbb{R}^d\big)$ and a continuous function $\hat{q} : [0,T] \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ be given satisfying

$$\hat{J}(T,x) = h(x), \int_{\mathcal{A}} \big[\hat{q}(t,x,a) - \gamma \log \boldsymbol{\pi}(a|t,x)\big] \boldsymbol{\pi}(a|t,x)\mathrm{d}a = 0, \quad \forall (t,x).$$

Then $\hat{J}$ and $\hat{q}$ are respectively the value function and the q-function associated with $\boldsymbol{\pi}$ if and only if for all $(t,x) \in [0,T] \times \mathbb{R}^d$, the following process

$$\hat{J}(s, X_s^{\boldsymbol{\pi}}; \boldsymbol{\pi}) + \int_t^s [r(t', X_{t'}^{\boldsymbol{\pi}}, a_{t'}^{\boldsymbol{\pi}}) - \hat{q}(t', X_{t'}^{\boldsymbol{\pi}}, a_{t'}^{\boldsymbol{\pi}})]\mathrm{d}t'$$

is an $(\{\mathcal{F}_s\}_{s\geq 0}, \mathbb{P})$-martingale, where $\{X_s^{\boldsymbol{\pi}}, t \leq s \leq T\}$ is the state process with $X_t^{\boldsymbol{\pi}} = x$. If it holds further that $\boldsymbol{\pi}(a|t,x) = \frac{\exp\{\frac{1}{\gamma}\hat{q}(t,x,a)\}}{\int_{\mathcal{A}} \exp\{\frac{1}{\gamma}\hat{q}(t,x,a)\}\mathrm{d}a}$, then $\boldsymbol{\pi}$ is the optimal policy and $\hat{J}$ is the optimal value function.

Function Approximation and Martingality

- *Function approximation*: approximates function f to be learned by a parametric family of functions f^θ where $\theta \in \mathbb{R}^L$ (finite dimensional approximation)

Function Approximation and Martingality

- *Function approximation*: approximates function f to be learned by a parametric family of functions f^θ where $\theta \in \mathbb{R}^L$ (finite dimensional approximation)
- Parametric form may be inspired by problem structure or neural networks

Function Approximation and Martingality

- *Function approximation*: approximates function f to be learned by a parametric family of functions f^θ where $\theta \in \mathbb{R}^L$ (finite dimensional approximation)
- Parametric form may be inspired by problem structure or neural networks
- Martingality of M leads to two loss functions for learning algorithms (Jia and Z. 2022a)

Function Approximation and Martingality

- *Function approximation*: approximates function f to be learned by a parametric family of functions f^θ where $\theta \in \mathbb{R}^L$ (finite dimensional approximation)
- Parametric form may be inspired by problem structure or neural networks
- Martingality of M leads to two loss functions for learning algorithms (Jia and Z. 2022a)
  - Martingale loss function (to be solved by stochastic gradient descent):

$$\mathbb{E}\int_0^T |M_T^\theta - M_t^\theta|^2 \mathrm{d}t \to \text{ min.}$$

Function Approximation and Martingality

- *Function approximation*: approximates function f to be learned by a parametric family of functions f^θ where $\theta \in \mathbb{R}^L$ (finite dimensional approximation)
- Parametric form may be inspired by problem structure or neural networks
- Martingality of M leads to two loss functions for learning algorithms (Jia and Z. 2022a)
  - Martingale loss function (to be solved by stochastic gradient descent):
$$\mathbb{E}\int_0^T |M_T^\theta - M_t^\theta|^2 \mathrm{d}t \to \text{ min.}$$
  - Martingale Orthogonality Conditions (to be solved by stochastic approximation or least square):
$$\mathbb{E}\int_0^T \xi_t \mathrm{d}M_t^\theta = 0$$
for any $\xi \in L^2_{\mathcal{F}}([0,T]; M)$ (*test function*)

Help with Function Approximation

- Let J^θ and q^ψ be function approximators satisfying

$$J^\theta(T,x) = h(x), \ \int_{\mathcal{A}} \left[q^\psi(t,x,a) - \gamma \log \boldsymbol{\pi}^\psi(a|t,x) \right] \boldsymbol{\pi}^\psi(a|t,x) \mathrm{d}a = 0,$$

Help with Function Approximation

- Let J^θ and q^ψ be function approximators satisfying

$$J^\theta(T,x) = h(x), \ \int_{\mathcal{A}} \left[q^\psi(t,x,a) - \gamma \log \boldsymbol{\pi}^\psi(a|t,x)\right] \boldsymbol{\pi}^\psi(a|t,x)\mathrm{d}a = 0,$$

- ... and

$$\boldsymbol{\pi}^\psi(a|t,x) = \frac{\exp\{\frac{1}{\gamma}q^\psi(t,x,a)\}}{\int_{\mathcal{A}}\exp\{\frac{1}{\gamma}q^\psi(t,x,a)\}\mathrm{d}a}$$

Help with Function Approximation

- Let J^θ and q^ψ be function approximators satisfying

$$J^{\theta}(T,x)=h(x),\ \int_{\mathcal{A}}\left[q^{\psi}(t,x,a)-\gamma\log\boldsymbol{\pi}^{\psi}(a|t,x)\right]\boldsymbol{\pi}^{\psi}(a|t,x)\mathrm{d}a=0,$$

- ... and

$$\boldsymbol{\pi}^{\psi}(a|t,x)=\frac{\exp\{\frac{1}{\gamma}q^{\psi}(t,x,a)\}}{\int_{\mathcal{A}}\exp\{\frac{1}{\gamma}q^{\psi}(t,x,a)\}\mathrm{d}a}$$

- Lead to more special parametric form of q-function approximator q^ψ, potentially facilitating more efficient learning

An Example

- When system dynamic is linear in a and reward quadratic in a, Hamiltonian and hence q-function are quadratic in a

An Example

- When system dynamic is linear in a and reward quadratic in a, Hamiltonian and hence q-function are quadratic in a
- So we can parameterize

$$q^{\psi}(t,x,a) = -\frac{1}{2} q_2^{\psi}(t,x) \circ \Big(a - q_1^{\psi}(t,x)\Big)^2 + q_0^{\psi}(t,x)$$

An Example

- When system dynamic is linear in a and reward quadratic in a, Hamiltonian and hence q-function are quadratic in a
- So we can parameterize

$$q^{\psi}(t,x,a) = -\frac{1}{2} q_2^{\psi}(t,x) \circ \left(a - q_1^{\psi}(t,x)\right)^2 + q_0^{\psi}(t,x)$$

- Corresponding policy is normal

$$\boldsymbol{\pi}^{\psi}(\cdot|t,x) = \mathcal{N}\left(q_1^{\psi}(t,x), \gamma\left(q_2^{\psi}(t,x)\right)^{-1}\right)$$

An Example

- When system dynamic is linear in a and reward quadratic in a, Hamiltonian and hence q-function are quadratic in a
- So we can parameterize

$$q^{\psi}(t,x,a) = -\frac{1}{2} q_2^{\psi}(t,x) \circ \left(a - q_1^{\psi}(t,x)\right)^2 + q_0^{\psi}(t,x)$$

- Corresponding policy is normal

$$\boldsymbol{\pi}^{\psi}(\cdot|t,x) = \mathcal{N}\left(q_1^{\psi}(t,x), \gamma\left(q_2^{\psi}(t,x)\right)^{-1}\right)$$

- Its entropy value is $-\frac{1}{2}\log\det q_2^{\psi}(t,x) + \frac{m}{2}\log 2\pi e\gamma$

An Example

- When system dynamic is linear in a and reward quadratic in a, Hamiltonian and hence q-function are quadratic in a
- So we can parameterize

$$q^{\psi}(t,x,a) = -\frac{1}{2} q_2^{\psi}(t,x) \circ \left(a - q_1^{\psi}(t,x)\right)^2 + q_0^{\psi}(t,x)$$

- Corresponding policy is normal

$$\boldsymbol{\pi}^{\psi}(\cdot|t,x) = \mathcal{N}\left(q_1^{\psi}(t,x), \gamma\left(q_2^{\psi}(t,x)\right)^{-1}\right)$$

- Its entropy value is $-\frac{1}{2}\log\det q_2^{\psi}(t,x) + \frac{m}{2}\log 2\pi e\gamma$
- The second constraint on q^{ψ} then yields

$$q_0^{\psi}(t,x) = \frac{\gamma}{2}\log\left(\det q_2^{\psi}(t,x)\right) - \frac{m\gamma}{2}\log 2\pi$$

An Example

- When system dynamic is linear in a and reward quadratic in a, Hamiltonian and hence q-function are quadratic in a
- So we can parameterize

$$q^{\psi}(t,x,a) = -\frac{1}{2} q_2^{\psi}(t,x) \circ \left(a - q_1^{\psi}(t,x)\right)^2 + q_0^{\psi}(t,x)$$

- Corresponding policy is normal

$$\boldsymbol{\pi}^{\psi}(\cdot|t,x) = \mathcal{N}\left(q_1^{\psi}(t,x), \gamma\left(q_2^{\psi}(t,x)\right)^{-1}\right)$$

- Its entropy value is $-\frac{1}{2}\log\det q_2^{\psi}(t,x) + \frac{m}{2}\log 2\pi e\gamma$
- The second constraint on q^{ψ} then yields

$$q_0^{\psi}(t,x) = \frac{\gamma}{2}\log\left(\det q_2^{\psi}(t,x)\right) - \frac{m\gamma}{2}\log 2\pi$$

- ... leading to a more specific parametric form

$$q^{\psi}(t,x,a) = -\frac{1}{2} q_2^{\psi}(t,x) \circ \left(a - q_1^{\psi}(t,x)\right)^2 + \frac{\gamma}{2}\log\left(\det q_2^{\psi}(t,x)\right) - \frac{m\gamma}{2}\log 2\pi$$

Algorithm: Martingale Loss Function

- Minimize martingale loss function:

$$\frac{1}{2}\mathbb{E}^{\mathbb{P}}\left[\int_0^T\left[h(X_T^{\boldsymbol{\pi}^\psi}) - J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + \int_t^T [r(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi}) - q^\psi(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi})]\mathrm{d}s\right]^2 \mathrm{d}t\right]$$

Algorithm: Martingale Loss Function

- Minimize martingale loss function:

$$\frac{1}{2}\mathbb{E}^{\mathbb{P}}\left[\int_0^T \left[h(X_T^{\boldsymbol{\pi}^\psi}) - J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + \int_t^T [r(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi}) - q^\psi(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi})]\mathrm{d}s\right]^2 \mathrm{d}t\right]$$

- Intrinsically offline

Algorithm: Martingale Loss Function

- Minimize martingale loss function:

$$\frac{1}{2}\mathbb{E}^{\mathbb{P}}\left[\int_0^T\left[h(X_T^{\boldsymbol{\pi}^\psi}) - J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + \int_t^T [r(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi}) - q^\psi(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi})]\mathrm{d}s\right]^2 \mathrm{d}t\right]$$

- Intrinsically offline
- SGD to update

$$\theta \leftarrow \theta + \alpha_\theta \int_0^T \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\boldsymbol{\pi}^\psi}) G_{t:T}\mathrm{d}t$$

$$\psi \leftarrow \psi + \alpha_\psi \int_0^T \int_t^T \frac{\partial q^\psi}{\partial \psi}(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi})\mathrm{d}s G_{t:T}\mathrm{d}t$$

where

$$G_{t:T} = h(X_T^{\boldsymbol{\pi}^\psi}) - J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + \int_t^T [r(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi}) - q^\psi(s, X_s^{\boldsymbol{\pi}^\psi}, a_s^{\boldsymbol{\pi}^\psi})]\mathrm{d}s,$$

and α_θ and α_ψ are learning rates

Algorithm: Martingale Orthogonality Conditions

- Apply martingale orthogonality conditions to get following system of equations in (θ, ψ):

$$\mathbb{E}^{\mathbb{P}}\left[\int_0^T \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\mathrm{d}t\right]\right] = 0,$$

and

$$\mathbb{E}^{\mathbb{P}}\left[\int_0^T \frac{\partial q^\psi}{\partial \psi}(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\right]\right] = 0$$

Algorithm: Martingale Orthogonality Conditions

- Apply martingale orthogonality conditions to get following system of equations in (θ, ψ):

$$\mathbb{E}^{\mathbb{P}}\left[\int_0^T \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\mathrm{d}t\right]\right] = 0,$$

  and

$$\mathbb{E}^{\mathbb{P}}\left[\int_0^T \frac{\partial q^\psi}{\partial \psi}(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\right]\right] = 0$$

- Stochastic approximation to update (θ, ψ) either offline by

$$\theta \leftarrow \theta + \alpha_\theta \int_0^T \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\right],$$

$$\psi \leftarrow \psi + \alpha_\psi \int_0^T \frac{\partial q^\psi}{\partial \psi}(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\right],$$

  or online by

$$\theta \leftarrow \theta + \alpha_\theta \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\right],$$

$$\psi \leftarrow \psi + \alpha_\psi \frac{\partial q^\psi}{\partial \psi}(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\left[\mathrm{d}J^\theta(t, X_t^{\boldsymbol{\pi}^\psi}) + r(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t - q^\psi(t, X_t^{\boldsymbol{\pi}^\psi}, a_t^{\boldsymbol{\pi}^\psi})\mathrm{d}t\right].$$

Outline

What Do We Need To Learn About Environment?

- Classical model-based approach: separates “estimation” and “optimization”

What Do We Need To Learn About Environment?

- Classical model-based approach: separates "estimation" and "optimization"
- Model-free RL approach: skips estimating a model and learns optimizing policies directly via PG or Q/q-learning

What Do We Need To Learn About Environment?

- Classical model-based approach: separates "estimation" and "optimization"
- Model-free RL approach: skips estimating a model and learns optimizing policies directly via PG or Q/q-learning
- But RL still learns *something* about the environment: q-function or Hamiltonian

What Do We Need To Learn About Environment?

- Classical model-based approach: separates "estimation" and "optimization"
- Model-free RL approach: skips estimating a model and learns optimizing policies directly via PG or Q/q-learning
- But RL still learns *something* about the environment: q-function or Hamiltonian
- It is the Hamiltonian, rather than each and every individual model coefficient, that needs to be learned/estimated for optimization

What Do We Need To Learn About Environment?

- Classical model-based approach: separates "estimation" and "optimization"
- Model-free RL approach: skips estimating a model and learns optimizing policies directly via PG or Q/q-learning
- But RL still learns *something* about the environment: q-function or Hamiltonian
- It is the Hamiltonian, rather than each and every individual model coefficient, that needs to be learned/estimated for optimization
- From a pure computational standpoint, estimating a single function is much more efficient and robust than estimating multiple functions (b, σ, r, h) in terms of avoiding or reducing over-parameterization, sensitivity to errors and accumulation of errors

What Do We Need To Learn About Environment?

- Classical model-based approach: separates "estimation" and "optimization"
- Model-free RL approach: skips estimating a model and learns optimizing policies directly via PG or Q/q-learning
- But RL still learns *something* about the environment: q-function or Hamiltonian
- It is the Hamiltonian, rather than each and every individual model coefficient, that needs to be learned/estimated for optimization
- From a pure computational standpoint, estimating a single function is much more efficient and robust than estimating multiple functions (b, σ, r, h) in terms of avoiding or reducing over-parameterization, sensitivity to errors and accumulation of errors
- Itô's formula shows q-function can be learned through temporal differences of the value function; so the task of learning and optimizing can be accomplished in a data-driven way

What Do We Need To Learn About Environment?

- Classical model-based approach: separates "estimation" and "optimization"
- Model-free RL approach: skips estimating a model and learns optimizing policies directly via PG or Q/q-learning
- But RL still learns *something* about the environment: q-function or Hamiltonian
- It is the Hamiltonian, rather than each and every individual model coefficient, that needs to be learned/estimated for optimization
- From a pure computational standpoint, estimating a single function is much more efficient and robust than estimating multiple functions (b, σ, r, h) in terms of avoiding or reducing over-parameterization, sensitivity to errors and accumulation of errors
- Itô's formula shows q-function can be learned through temporal differences of the value function; so the task of learning and optimizing can be accomplished in a data-driven way
- This would not be the case if we chose to learn individual model coefficients separately