Stochastic Control, Gradient Flows and Reinforcement learning

Lukasz Szpruch
University of Edinburgh, The Alan Turing Institute, London

Continuous time and space Reinforcement Learning

To understand RL in continuous time and space setting we need to answer the following questions:

Continuous time and space Reinforcement Learning

To understand RL in continuous time and space setting we need to answer the following questions:

- How to strike optimal trade-off between model based and model free approaches?

Continuous time and space Reinforcement Learning

To understand RL in continuous time and space setting we need to answer the following questions:

- How to strike optimal trade-off between model based and model free approaches?
- How to strike the optimal balance between exploration (learning) and exploitation (optimal control) ?

Continuous time and space Reinforcement Learning

To understand RL in continuous time and space setting we need to answer the following questions:

- How to strike optimal trade-off between model based and model free approaches?
- How to strike the optimal balance between exploration (learning) and exploitation (optimal control) ?
- Why entropy regularised policy gradient algorithms work so well?

Continuous time and space Reinforcement Learning

To understand RL in continuous time and space setting we need to answer the following questions:

- How to strike optimal trade-off between model based and model free approaches?
- How to strike the optimal balance between exploration (learning) and exploitation (optimal control) ?
- Why entropy regularised policy gradient algorithms work so well?

In this talk I'll discuss 3rd and 2nd questions through the lens of stochastic control theory.

Gradient Flows for Regularised stochastic Control

joint work with David Siska (Edinburgh)

Stochastic Control

For $\xi \in \mathbb{R}^d$ and $\mu \in \mathcal{V}_q^W$, consider the controlled process

$$X_t(\mu) = \xi + \int_0^t \Phi_r(X_r(\mu), \mu_r)\, dr + \int_0^t \Gamma_r(X_r(\mu), \mu_r)\, dW_r\,, \quad t \in [0, T]\,,$$

where

$$\mathcal{V}_q^W := \left\{ \nu : \Omega^W \to \mathcal{M}_q : \mathbb{E}^W \int_0^T \int |a|^q\, \nu_t(da, dt) < \infty \text{ and } \nu_t \in \mathcal{F}_t^W,\ \forall t \in [0, T] \right\}$$

Stochastic Control

For $\xi \in \mathbb{R}^d$ and $\mu \in \mathcal{V}_q^W$, consider the controlled process

$$X_t(\mu) = \xi + \int_0^t \Phi_r(X_r(\mu), \mu_r)\, dr + \int_0^t \Gamma_r(X_r(\mu), \mu_r)\, dW_r\,, \quad t \in [0, T]\,,$$

where

$$\mathcal{V}_q^W := \left\{\nu : \Omega^W \to \mathcal{M}_q : \mathbb{E}^W \int_0^T \int |a|^q\, \nu_t(da, dt) < \infty \text{ and } \nu_t \in \mathcal{F}_t^W,\, \forall t \in [0, T]\right\}$$

Example 1

Relaxed Control

$$\Phi_t(x, m) = \int \phi_t(x, a) m(da)\,, \text{ and } \Gamma_t(x, m)(\Gamma_t(x, m))^\top = \int \gamma_t(x, a)\gamma_t(x, a)^\top m(da)$$

Stochastic Control

For $\xi \in \mathbb{R}^d$ and $\mu \in \mathcal{V}_q^W$, consider the controlled process

$$X_t(\mu) = \xi + \int_0^t \Phi_r(X_r(\mu), \mu_r)\, dr + \int_0^t \Gamma_r(X_r(\mu), \mu_r)\, dW_r\,, \quad t \in [0, T]\,,$$

where

$$\mathcal{V}_q^W := \left\{ \nu : \Omega^W \to \mathcal{M}_q : \mathbb{E}^W \int_0^T \int |a|^q \, \nu_t(da, dt) < \infty \text{ and } \nu_t \in \mathcal{F}_t^W, \forall t \in [0, T] \right\}$$

Example 1

Relaxed Control

$$\Phi_t(x, m) = \int \phi_t(x, a) m(da)\,, \text{ and } \Gamma_t(x, m)(\Gamma_t(x, m))^\top = \int \gamma_t(x, a) \gamma_t(x, a)^\top m(da)$$

Building on [Hu et al., 2021, Hu et al., 2019, Jabir et al., 2019].

Stochastic Control

Given F and g we define the objective functional

$$J^{\sigma}(\nu,\xi) := \mathbb{E}^{W}\left[\int_0^T \left[F_t(X_t(\nu),\nu_t) + \frac{\sigma^2}{2}\mathrm{Ent}(\nu_t)\right] dt + g(X_T(\nu))\middle| X_0(\nu) = \xi\right].$$

$$\mathrm{Ent}(m) := \begin{cases} \int_{\mathbb{R}^d} m(x)\log\left(\frac{m(x)}{\gamma(x)}\right) dx & \text{if } m \text{ is a.c. w.r.t. Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$

and Gibbs measure γ:

$$\gamma(x) = e^{-U(x)} \text{ with } U \text{ s.t. } \int_{\mathbb{R}^d} e^{-U(x)}\, dx = 1\,.$$

Stochastic Control

Given F and g we define the objective functional

$$J^\sigma(\nu,\xi) := \mathbb{E}^W\left[\int_0^T \left[F_t(X_t(\nu),\nu_t) + \frac{\sigma^2}{2}\mathrm{Ent}(\nu_t)\right] dt + g(X_T(\nu)) \middle| X_0(\nu) = \xi\right].$$

$$\mathrm{Ent}(m) := \begin{cases} \int_{\mathbb{R}^d} m(x)\log\left(\frac{m(x)}{\gamma(x)}\right) dx & \text{if } m \text{ is a.c. w.r.t. Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$

and Gibbs measure γ:

$$\gamma(x) = e^{-U(x)} \text{ with } U \text{ s.t. } \int_{\mathbb{R}^d} e^{-U(x)}\, dx = 1\,.$$

Why regularise with Entropy?

- Bridging the gap between stochastic control and entropy regularised Reinforcement Learning (MaxEntRL), [Wang et al., 2020]
- Regularity of Markovian controls [Reisinger and Zhang, 2020]
- Useful when studying inverse RL problems [Cao et al., 2021]

Example 1: Policy gradient with neural network

- Consider a SC problem with the space of actions $A \subseteq \mathbb{R}^a$ given by

$$dX_t^\alpha = b(X_t^\alpha, \alpha_t)\, dt + \sigma(X_t^\alpha, \alpha_t)\, dW_t\,, \quad t \in [0, T]\,, \quad X_0 = x$$

and the objective

$$J(\alpha, x) = \mathbb{E}^W \left[\int_0^T f(X_t^\alpha, \alpha_t)\, dt + g(X_T^\alpha) \right]$$

Example 1: Policy gradient with neural network

- Consider a SC problem with the space of actions $A \subseteq \mathbb{R}^a$ given by

$$dX_t^\alpha = b(X_t^\alpha, \alpha_t)\, dt + \sigma(X_t^\alpha, \alpha_t)\, dW_t\,, \quad t \in [0, T]\,, \quad X_0 = x$$

  and the objective

$$J(\alpha, x) = \mathbb{E}^W \left[\int_0^T f(X_t^\alpha, \alpha_t)\, dt + g(X_T^\alpha) \right]$$

- Aim is to minimize J over all Markov controls $\alpha_t = a(t, X_t)$.

Example 1: Policy gradient with neural network

- Consider a SC problem with the space of actions $A \subseteq \mathbb{R}^a$ given by

$$dX_t^\alpha = b(X_t^\alpha, \alpha_t)\, dt + \sigma(X_t^\alpha, \alpha_t)\, dW_t\,, \quad t \in [0, T]\,, \quad X_0 = x$$

  and the objective

$$J(\alpha, x) = \mathbb{E}^W \left[\int_0^T f(X_t^\alpha, \alpha_t)\, dt + g(X_T^\alpha) \right]$$

- Aim is to minimize J over all Markov controls $\alpha_t = a(t, X_t)$.
- Take $a(t, x) \approx \int \varphi(x; \theta)\, \mu_t(d\theta)$ with φ being the activation function and $\mu_t \in \mathcal{P}_q(\mathbb{R}^p)$ the law of the parameters at time $t \in [0, T]$.

Example 1: Policy gradient with neural network

- Consider a SC problem with the space of actions $A \subseteq \mathbb{R}^a$ given by

$$dX_t^\alpha = b(X_t^\alpha, \alpha_t)\, dt + \sigma(X_t^\alpha, \alpha_t)\, dW_t\,, \quad t \in [0, T]\,, \quad X_0 = x$$

and the objective

$$J(\alpha, x) = \mathbb{E}^W \left[\int_0^T f(X_t^\alpha, \alpha_t)\, dt + g(X_T^\alpha) \right]$$

- Aim is to minimize J over all Markov controls $\alpha_t = a(t, X_t)$.
- Take $a(t,x) \approx \int \varphi(x;\theta)\, \mu_t(d\theta)$ with φ being the activation function and $\mu_t \in \mathcal{P}_q(\mathbb{R}^p)$ the law of the parameters at time $t \in [0, T]$.
- Take

$$\Phi(x, \mu_t) := b\left(x, \int \varphi(x;\theta)\, \mu_t(d\theta)\right), \quad \Gamma(x, \mu_t) := \sigma\left(x, \int \varphi(x;\theta)\, \mu_t(d\theta)\right)$$

$$F(x, \mu_t) := f\left(x, \int \varphi(x;\theta)\, \mu_t(d\theta)\right)$$

Example 1: Generative modelling with causal transport

- Fix $m^\star \in \mathcal{P}([0, T] \times \mathbb{R}^d)$ to be a target distribution. (e.g $\mathbb{Q}$-measure induced by liquid derivatives)

Example 1: Generative modelling with causal transport

- Fix $m^\star \in \mathcal{P}([0,T] \times \mathbb{R}^d)$ to be a target distribution. (e.g $\mathbb{Q}$-measure induced by liquid derivatives)
- The aim of the generative model is to map some basic distribution, in our case $m^0 := \mathcal{L}(\xi) \otimes \mathcal{L}(W)$, into $m^\star$.

Example 1: Generative modelling with causal transport

- Fix $m^\star \in \mathcal{P}([0,T] \times \mathbb{R}^d)$ to be a target distribution. (e.g $\mathbb{Q}$-measure induced by liquid derivatives)
- The aim of the generative model is to map some basic distribution, in our case $m^0 := \mathcal{L}(\xi) \otimes \mathcal{L}(W)$, into $m^\star$.
- The solution to the SDE is given by a measurable map $G^\mu : \mathbb{R}^d \times C[0,T]^d \to C[0,T]^d$ such that $X_t(\mu) := G_t^\mu(\xi, (W_{s \wedge t})_{s\in[0,T]})$

Example 1: Generative modelling with causal transport

- Fix $m^\star \in \mathcal{P}([0,T] \times \mathbb{R}^d)$ to be a target distribution. (e.g $\mathbb{Q}$-measure induced by liquid derivatives)
- The aim of the generative model is to map some basic distribution, in our case $m^0 := \mathcal{L}(\xi) \otimes \mathcal{L}(W)$, into $m^\star$.
- The solution to the SDE is given by a measurable map $G^\mu : \mathbb{R}^d \times C[0,T]^d \to C[0,T]^d$ such that $X_t(\mu) := G_t^\mu(\xi, (W_{s\wedge t})_{s\in[0,T]})$
- One then seeks $\mu^\star$ such that $G_{\#}^{\mu^\star} m^0$ is a good approximation of $m^\star$

Example 1: Generative modelling with causal transport

- Fix $m^\star \in \mathcal{P}([0,T] \times \mathbb{R}^d)$ to be a target distribution. (e.g $\mathbb{Q}$-measure induced by liquid derivatives)
- The aim of the generative model is to map some basic distribution, in our case $m^0 := \mathcal{L}(\xi) \otimes \mathcal{L}(W)$, into $m^\star$.
- The solution to the SDE is given by a measurable map $G^\mu : \mathbb{R}^d \times C[0,T]^d \to C[0,T]^d$ such that $X_t(\mu) := G^\mu_t(\xi, (W_{s\wedge t})_{s\in[0,T]})$
- One then seeks $\mu^\star$ such that $G^{\mu^\star}_{\#} m^0$ is a good approximation of $m^\star$
- optimisation problem on the space of measures: for some $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to R_+$

$$J^\sigma(\nu, \xi) := \mathbb{E}^W \left[\int_0^T \left(D(\mathcal{L}(X_t(\nu)), m^\star_t) + \frac{\sigma^2}{2} \mathrm{Ent}(\nu_t) \right) dt \middle| X_0(\nu) = \xi \right].$$

Example 1: Generative modelling with causal transport

- Fix $m^\star \in \mathcal{P}([0,T] \times \mathbb{R}^d)$ to be a target distribution. (e.g $\mathbb{Q}$-measure induced by liquid derivatives)
- The aim of the generative model is to map some basic distribution, in our case $m^0 := \mathcal{L}(\xi) \otimes \mathcal{L}(W)$, into $m^\star$.
- The solution to the SDE is given by a measurable map $G^\mu : \mathbb{R}^d \times C[0,T]^d \to C[0,T]^d$ such that $X_t(\mu) := G^\mu_t(\xi, (W_{s\wedge t})_{s\in[0,T]})$
- One then seeks $\mu^\star$ such that $G^{\mu^\star}_\# m^0$ is a good approximation of $m^\star$
- optimisation problem on the space of measures: for some $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to R_+$

$$J^\sigma(\nu,\xi) := \mathbb{E}^W\left[\int_0^T \left(D(\mathcal{L}(X_t(\nu)), m^\star_t) + \frac{\sigma^2}{2}\mathrm{Ent}(\nu_t)\right) dt \middle| X_0(\nu) = \xi\right].$$

- See related work on neural SDEs [Cuchiero et al., 2020], [Gierjatowicz et al., 2020],[Cohen et al., 2021] and casual optimal transport [Acciaio et al., 2020, Backhoff-Veraguas et al., 2020]

Stochastic Control

$$X_t(\mu) = \xi + \int_0^t \Phi_r(X_r(\mu), \mu_r)\, dr + \int_0^t \Gamma_r(X_r(\mu), \mu_r)\, dW_r\,, \quad t \in [0, T]\,,$$

$$J^\sigma(\nu, \xi) := \mathbb{E}^W \left[\int_0^T \left[F_t(X_t(\nu), \nu_t) + \frac{\sigma^2}{2} \mathrm{Ent}(\nu_t) \right] dt + g(X_T(\nu)) \middle| X_0(\nu) = \xi \right].$$

Hamiltonian:

$$H_t^\sigma(x, y, z, m) := \Phi_t(x, m) y + \mathrm{tr}(\Gamma_t^\top(x, m) z) + F_t(x, m) + \frac{\sigma^2}{2} \mathrm{Ent}(m)\,.$$

Stochastic Control

$$X_t(\mu) = \xi + \int_0^t \Phi_r(X_r(\mu), \mu_r)\, dr + \int_0^t \Gamma_r(X_r(\mu), \mu_r)\, dW_r\,, \quad t \in [0, T]\,,$$

$$J^\sigma(\nu, \xi) := \mathbb{E}^W \left[\int_0^T \left[F_t(X_t(\nu), \nu_t) + \frac{\sigma^2}{2} \mathrm{Ent}(\nu_t) \right] dt + g(X_T(\nu)) \middle| X_0(\nu) = \xi \right].$$

Hamiltonian:

$$H_t^\sigma(x, y, z, m) := \Phi_t(x, m) y + \mathrm{tr}(\Gamma_t^\top(x, m) z) + F_t(x, m) + \frac{\sigma^2}{2} \mathrm{Ent}(m)\,.$$

Adjoint process with control μ

$$dY_t(\mu) = -(\nabla_x H_t^0)(X_t(\mu), Y_t(\mu), Z_t(\mu), \mu_t)\, dt + Z_t(\mu)\, dW_t\,, \quad t \in [0, T]\,,$$
$$Y_T(\mu) = (\nabla_x g)(X_T(\mu))$$

Theorem 2 (Necessary condition for optimality)

Fix $\sigma > 0$. Fix $q > 2$. If $\nu \in \mathcal{V}_q^W$ is (locally) optimal for $J^\sigma(\cdot, \xi)$, $X(\nu)$ and $Y(\nu)$, $Z(\nu)$ are the associated optimally controlled state and adjoint processes respectively, then for a.a. $(\omega, t) \in \Omega^W \times (0, T)$

$$\nu_t \quad \textit{locally minimizes} \quad H^\sigma(X_t(\nu), Y_t(\nu), Z_t(\nu), \ \nu)\,.$$

Gradient flow

- From work of Benamou-Brenier we know that

$$\mathcal{W}_2(\mu_0, \mu_1) = \inf \left\{ \int |x - y|^2 \pi(dx, dy) \ : \ \pi \in \mathrm{Plan}(\mu_0, \mu_1) \right\}$$

$$= \inf \left\{ \int_0^1 \int |\nu_s|^2 \mu_s(dx) ds \ : \ \mathrm{s.t}\, \partial_s \mu_s + \nabla(\nu_s \mu_s) = 0\,,\ \mu_{t=i} = \mu_i \right\}$$

Gradient flow

- From work of Benamou-Brenier we know that

$$\mathcal{W}_2(\mu_0,\mu_1) = \inf\left\{\int |x-y|^2 \pi(dx,dy) \,:\, \pi \in \mathrm{Plan}(\mu_0,\mu_1)\right\}$$

$$= \inf\left\{\int_0^1 \int |\nu_s|^2 \mu_s(dx)ds \,:\, \text{s.t}\, \partial_s \mu_s + \nabla(\nu_s \mu_s) = 0\,,\ \mu_{t=i} = \mu_i\right\}$$

- Let $b_t : \Omega^W \times [0,\infty] \times \mathbb{R}^p \to \mathbb{R}^p$ time dependent vector field.

Gradient flow

- From work of Benamou-Brenier we know that

$$\mathcal{W}_2(\mu_0,\mu_1) = \inf\left\{\int |x-y|^2\pi(dx,dy)\ :\ \pi\in \mathrm{Plan}(\mu_0,\mu_1)\right\}$$
$$= \inf\left\{\int_0^1\int |\nu_s|^2\mu_s(dx)ds\ :\ \text{s.t}\,\partial_s\mu_s+\nabla(\nu_s\mu_s)=0\,,\ \mu_{t=i}=\mu_i\right\}$$

- Let $b_t : \Omega^W\times[0,\infty]\times\mathbb{R}^p\to\mathbb{R}^p$ time dependent vector field.
- For each $t\in[0,T]$ and $\omega^W\in\Omega^W$ consider

$$\partial_s\nu_{s,t} = \nabla_a\cdot\left(b_{s,t}\nu_{s,t}+\frac{\sigma^2}{2}\nabla_a\nu_{s,t}\right),\ \ s\in[0,\infty)\,,\ \ \nu_{0,t}\in\mathcal{P}_2(\mathbb{R}^p)$$

Gradient flow

- From work of Benamou-Brenier we know that

$$\mathcal{W}_2(\mu_0,\mu_1) = \inf\left\{\int |x-y|^2\pi(dx,dy) \;:\; \pi\in \mathrm{Plan}(\mu_0,\mu_1)\right\}$$
$$= \inf\left\{\int_0^1\int |\nu_s|^2\mu_s(dx)ds \;:\; \text{s.t}\,\partial_s\mu_s+\nabla(\nu_s\mu_s)=0\,,\ \mu_{t=i}=\mu_i\right\}$$

- Let $b_t:\Omega^W\times[0,\infty]\times\mathbb{R}^p\to\mathbb{R}^p$ time dependent vector field.
- For each $t\in[0,T]$ and $\omega^W\in\Omega^W$ consider

$$\partial_s\nu_{s,t}=\nabla_a\cdot\left(b_{s,t}\nu_{s,t}+\frac{\sigma^2}{2}\nabla_a\nu_{s,t}\right),\ \ s\in[0,\infty)\,,\ \ \nu_{0,t}\in\mathcal{P}_2(\mathbb{R}^p)$$

- 'Stochastic gradient flow' X s.t $\mathcal{L}(X_{s,t})=\nu_{s,t}$ for all s is given

$$dX_{t,s}=b_{t,s}ds+\sigma dW_s\,,\ \ s\in[0,\infty)\,.$$

Gradient flow

- From work of Benamou-Brenier we know that

$$\mathcal{W}_2(\mu_0,\mu_1) = \inf\left\{\int |x-y|^2\pi(dx,dy) \,:\, \pi\in \mathrm{Plan}(\mu_0,\mu_1)\right\}$$
$$= \inf\left\{\int_0^1\int |\nu_s|^2\mu_s(dx)ds \,:\, \text{s.t}\, \partial_s\mu_s + \nabla(\nu_s\mu_s) = 0\,,\ \mu_{t=i} = \mu_i\right\}$$

- Let $b_t : \Omega^W \times [0,\infty] \times \mathbb{R}^p \to \mathbb{R}^p$ time dependent vector field.
- For each $t\in[0,T]$ and $\omega^W \in \Omega^W$ consider

$$\partial_s\nu_{s,t} = \nabla_a\cdot\left(b_{s,t}\nu_{s,t} + \frac{\sigma^2}{2}\nabla_a\nu_{s,t}\right),\ \ s\in[0,\infty)\,,\ \ \nu_{0,t}\in\mathcal{P}_2(\mathbb{R}^p)$$

- 'Stochastic gradient flow' X s.t $\mathcal{L}(X_{s,t}) = \nu_{s,t}$ for all s is given

$$dX_{t,s} = b_{t,s}ds + \sigma dW_s\,,\ \ s\in[0,\infty)\,.$$

- Aim: Find b such that $J^\sigma(\nu_{s,\cdot},\xi) \searrow$

GF derivation in the spirit of Otto calculus

- For $\epsilon, \lambda > 0$ let $\nu_t^{\lambda,\epsilon} := \nu_t + \lambda(\nu_{t+\epsilon} - \nu_t)$ we have

$$\partial_s J^\sigma(\nu_{s,\cdot}) = \lim_{\epsilon \to 0} \epsilon^{-1} \left(J^\sigma(\nu_{s+\epsilon,\cdot}) - J^\sigma(\nu_{s,\cdot}) \right)$$

$$= \lim_{\epsilon \to 0} \epsilon^{-1} \left(\int_0^1 \int \frac{\delta J^\sigma}{\delta \nu}(\nu_{s,\cdot}^{\lambda,\epsilon}, y)(\nu_{s+\epsilon,\cdot} - \nu_{s,\cdot})(dy) d\lambda \right)$$

GF derivation in the spirit of Otto calculus

- For $\epsilon, \lambda > 0$ let $\nu_t^{\lambda,\epsilon} := \nu_t + \lambda(\nu_{t+\epsilon} - \nu_t)$ we have

$$\partial_s J^\sigma(\nu_{s,\cdot}) = \lim_{\epsilon \to 0} \epsilon^{-1} \left(J^\sigma(\nu_{s+\epsilon,\cdot}) - J^\sigma(\nu_{s,\cdot})\right)$$

$$= \lim_{\epsilon \to 0} \epsilon^{-1} \left(\int_0^1 \int \frac{\delta J^\sigma}{\delta \nu}(\nu_{s,\cdot}^{\lambda,\epsilon}, y)(\nu_{s+\epsilon,\cdot} - \nu_{s,\cdot})(dy) d\lambda \right)$$

Lemma 3

$$\int_0^1 \int \frac{\delta J^0}{\delta \nu}(\nu_{s,\cdot}^{\lambda,\varepsilon}, a)(\nu_{s+\varepsilon,\cdot} - \nu_{s,\cdot})(da) d\lambda$$

$$= \mathbb{E}^W \left[\int_0^T \left[\int \frac{\delta \mathbf{H}^0}{\delta m}(, \nu_{s,t}^{\lambda,\varepsilon}, a)(\nu_{s+\varepsilon,t} - \nu_{s,t})(da) \right] dt \right].$$

GF derivation in the spirit of Otto calculus

- For $\epsilon, \lambda > 0$ let $\nu_t^{\lambda,\epsilon} := \nu_t + \lambda(\nu_{t+\epsilon} - \nu_t)$ we have

$$\partial_s J^\sigma(\nu_{s,\cdot}) = \lim_{\epsilon\to 0} \epsilon^{-1} \left(J^\sigma(\nu_{s+\epsilon,\cdot}) - J^\sigma(\nu_{s,\cdot})\right)$$
$$= \lim_{\epsilon\to 0} \epsilon^{-1} \left(\int_0^1 \int \frac{\delta J^\sigma}{\delta \nu}(\nu_{s,\cdot}^{\lambda,\epsilon}, y)(\nu_{s+\epsilon,\cdot} - \nu_{s,\cdot})(dy) d\lambda\right)$$

Lemma 3

$$\int_0^1 \int \frac{\delta J^0}{\delta \nu}(\nu_{s,\cdot}^{\lambda,\varepsilon}, a)(\nu_{s+\varepsilon,\cdot} - \nu_{s,\cdot})(da) d\lambda$$
$$= \mathbb{E}^W \left[\int_0^T \left[\int \frac{\delta \mathbf{H}^0}{\delta m}(, \nu_{s,t}^{\lambda,\varepsilon}, a)(\nu_{s+\varepsilon,t} - \nu_{s,t})(da)\right] dt\right].$$

- Hence, formally differentiating entropy,

$$\partial_s J^\sigma(\nu_{s,\cdot})$$
$$= \lim_{\epsilon\to 0} \epsilon^{-1} \left(\int_0^1 \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}^{\lambda,\epsilon}, y)(\nu_{s+\epsilon,t} - \nu_{s,t})(dy)\right] dt d\lambda\right)$$

GF derivation in the spirit of Otto calculus

- Assuming continuity of the hamiltonian and noting that $\nu_t^{\lambda,\epsilon} \to \nu_t$ as $\epsilon \to 0$

GF derivation in the spirit of Otto calculus

- Assuming continuity of the hamiltonian and noting that $\nu_t^{\lambda,\epsilon} \to \nu_t$ as $\epsilon \to 0$

$$\partial_s J^\sigma(\nu_{s,\cdot}) = \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}, y) \partial_s \nu_{s,t}(dy) \right] dt$$

$$= \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}, y)\ \nabla_a \cdot \left(b_{s,t}\nu_{s,t} + \frac{\sigma^2}{2} \nabla_a \nu_{s,t} \right)(dy) \right] dt\,.$$

GF derivation in the spirit of Otto calculus

- Assuming continuity of the hamiltonian and noting that $\nu_t^{\lambda,\epsilon} \to \nu_t$ as $\epsilon \to 0$

$$\partial_s J^\sigma(\nu_{s,\cdot}) = \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}, y) \partial_s \nu_{s,t}(dy) \right] dt$$
$$= \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}, y) \; \nabla_a \cdot \left(b_{s,t}\nu_{s,t} + \frac{\sigma^2}{2} \nabla_a \nu_{s,t} \right)(dy) \right] dt \,.$$

- Integration by parts yield

$$\partial_s J^\sigma(\nu_{s,\cdot}) = -\mathbb{E}^W \int_0^T \left[\int (\nabla_a \frac{\delta \mathbf{H}^\sigma}{\delta \nu})(\nu_{s,\cdot}, y) \left(b_{s,t}\nu_{s,t} + \frac{\sigma^2}{2} \nabla_a \nu_{s,t} \right)(dy) \right] dt \,.$$

GF derivation in the spirit of Otto calculus

- Assuming continuity of the hamiltonian and noting that $\nu_t^{\lambda,\epsilon} \to \nu_t$ as $\epsilon \to 0$

$$\partial_s J^\sigma(\nu_{s,\cdot}) = \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}, y) \partial_s \nu_{s,t}(dy) \right] dt$$

$$= \mathbb{E}^W \int_0^T \left[\int \frac{\delta \mathbf{H}^\sigma}{\delta \nu}(\nu_{s,\cdot}, y) \; \nabla_a \cdot \left(b_{s,t} \nu_{s,t} + \frac{\sigma^2}{2} \nabla_a \nu_{s,t} \right)(dy) \right] dt \,.$$

- Integration by parts yield

$$\partial_s J^\sigma(\nu_{s,\cdot}) = -\mathbb{E}^W \int_0^T \left[\int (\nabla_a \frac{\delta \mathbf{H}^\sigma}{\delta \nu})(\nu_{s,\cdot}, y) \left(b_{s,t} \nu_{s,t} + \frac{\sigma^2}{2} \nabla_a \nu_{s,t} \right)(dy) \right] dt \,.$$

- Hence take

$$b_{s,t} := (\nabla_a \frac{\delta \mathbf{H}^0}{\delta \nu})(\nu_{s,\cdot}, y) + \frac{\sigma^2}{2} (\nabla_a U)(a)$$

Energy dissipation

Theorem 4

Assume that $X_{s,\cdot}, Y_{s,\cdot}, Z_{s,\cdot}$ are the forward and backward processes arising from control $\nu_{s,\cdot} \in \mathcal{V}_2^W$ and data $\xi \in \mathbb{R}^d$. Then

$$\frac{d}{ds} J^\sigma(\nu_{s,\cdot}) = \; -\mathbb{E}^W \int_0^T \int \left(\left(\nabla_a \frac{\delta \mathbf{H}_t^\sigma}{\delta m} \right) (a, \nu_{s,t}) \right)^2 \nu_{s,t}\,(da)\,dt\,.$$

Energy dissipation

Theorem 4

Assume that $X_{s,\cdot}, Y_{s,\cdot}, Z_{s,\cdot}$ are the forward and backward processes arising from control $\nu_{s,\cdot} \in \mathcal{V}_2^W$ and data $\xi \in \mathbb{R}^d$. Then

$$\frac{d}{ds} J^\sigma(\nu_{s,\cdot}) = -\mathbb{E}^W \int_0^T \int \left(\left(\nabla_a \frac{\delta \mathbf{H}_t^\sigma}{\delta m} \right) (a, \nu_{s,t}) \right)^2 \nu_{s,t}(da)\, dt\,.$$

- Proof relies on Itô formula for measures and PDE estimates
- See related work [Karatzas et al., 2018]

SDE / BSDE System Representation for Gradient Flow

Consider with $\theta_{t,0} = \theta_t^0$ and $s \geq 0$:

$$d\theta_{s,t} = -\Big((\nabla_a \frac{\delta H_t^0}{\delta m})(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}, \theta_{s,t}) + \frac{\sigma^2}{2} (\nabla_a U)(\theta_{s,t}) \Big) \, ds + \sigma \, dB_s \,, \tag{1}$$

coupled with

$$\begin{cases} \nu_{s,t} & = \mathcal{L}(\theta_{s,t} | \mathcal{F}_t^W) \,, \\ X_{s,t} & = \xi + \int_0^t \Phi_r(X_{s,r}, \nu_{s,r}) \, dr + \int_0^t \Gamma_r(X_{s,r}, \nu_{s,r}(da)) \, dW_r \,, \quad t \in [0, T] \,, \\ dY_{s,t} & = -(\nabla_x H_t^0)(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}) \, dt + Z_{s,t} \, dW_t \,, \\ Y_{s,T} & = (\nabla_x g)(X_T) \,. \end{cases} \tag{2}$$

SDE / BSDE System Representation for Gradient Flow

Consider with $\theta_{t,0} = \theta_t^0$ and $s \geq 0$:

$$d\theta_{s,t} = -\Big((\nabla_a \frac{\delta H_t^0}{\delta m})(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}, \theta_{s,t}) + \frac{\sigma^2}{2}(\nabla_a U)(\theta_{s,t})\Big)\, ds + \sigma\, dB_s\,, \tag{1}$$

coupled with

$$\begin{cases} \nu_{s,t} &= \mathcal{L}(\theta_{s,t}|\mathcal{F}_t^W)\,, \\ X_{s,t} &= \xi + \int_0^t \Phi_r(X_{s,r}, \nu_{s,r})\, dr + \int_0^t \Gamma_r(X_{s,r}, \nu_{s,r}(da))\, dW_r\,, \quad t \in [0,T]\,, \\ dY_{s,t} &= -(\nabla_x H_t^0)(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t})\, dt + Z_{s,t}\, dW_t\,, \\ Y_{s,T} &= (\nabla_x g)(X_T)\,. \end{cases} \tag{2}$$

- After discretising time take a step of gradient descent $\theta^i_{s_k,t}$ for $i = 1, \dots, N$ and compute $\nu^N_{s_k,t} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_{s_k,t}}$

SDE / BSDE System Representation for Gradient Flow

Consider with $\theta_{t,0} = \theta_t^0$ and $s \geq 0$:

$$d\theta_{s,t} = -\Big((\nabla_a \frac{\delta H_t^0}{\delta m})(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}, \theta_{s,t}) + \frac{\sigma^2}{2} (\nabla_a U)(\theta_{s,t}) \Big)\, ds + \sigma\, dB_s\,, \tag{1}$$

coupled with

$$\begin{cases} \nu_{s,t} & = \mathcal{L}(\theta_{s,t} | \mathcal{F}_t^W)\,, \\ X_{s,t} & = \xi + \int_0^t \Phi_r(X_{s,r}, \nu_{s,r})\, dr + \int_0^t \Gamma_r(X_{s,r}, \nu_{s,r}(da))\, dW_r\,, \quad t \in [0,T]\,, \\ dY_{s,t} & = -(\nabla_x H_t^0)(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t})\, dt + Z_{s,t}\, dW_t\,, \\ Y_{s,T} & = (\nabla_x g)(X_T)\,. \end{cases} \tag{2}$$

- After discretising time take a step of gradient descent $\theta^i_{s_k,t}$ for $i = 1, \dots, N$ and compute $\nu^N_{s_k,t} = \frac{1}{N} \sum_{i=1}^N \delta_{\theta^i_{s_k,t}}$
- Solve forward process $X_{s_k,t}(\nu^N_{s_k,t})$ on $[0, T]$

SDE / BSDE System Representation for Gradient Flow

Consider with $\theta_{t,0} = \theta^0_t$ and $s \geq 0$:

$$d\theta_{s,t} = -\Big((\nabla_a \frac{\delta H^0_t}{\delta m})(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}, \theta_{s,t}) + \frac{\sigma^2}{2}(\nabla_a U)(\theta_{s,t}) \Big)\, ds + \sigma\, dB_s \,, \tag{1}$$

coupled with

$$\begin{cases} \nu_{s,t} & = \mathcal{L}(\theta_{s,t}|\mathcal{F}^W_t)\,, \\ X_{s,t} & = \xi + \int_0^t \Phi_r(X_{s,r}, \nu_{s,r})\, dr + \int_0^t \Gamma_r(X_{s,r}, \nu_{s,r}(da))\, dW_r\,, \quad t \in [0,T]\,, \\ dY_{s,t} & = -(\nabla_x H^0_t)(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t})\, dt + Z_{s,t}\, dW_t\,, \\ Y_{s,T} & = (\nabla_x g)(X_T)\,. \end{cases} \tag{2}$$

- After discretising time take a step of gradient descent $\theta^i_{s_k,t}$ for $i = 1, \dots, N$ and compute $\nu^N_{s_k,t} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_{s_k,t}}$
- Solve forward process $X_{s_k,t}(\nu^N_{s_k,t})$ on $[0,T]$
- 'Back-propagate' $(Y_{s_k,t}(\nu^N_{s_k,t}), Z_{s_k,t}(\nu^N_{s_k,t}))$ on $[0,T]$

SDE / BSDE System Representation for Gradient Flow

Consider with $\theta_{t,0} = \theta_t^0$ and $s \geq 0$:

$$d\theta_{s,t} = -\Big((\nabla_a \frac{\delta H_t^0}{\delta m})(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}, \theta_{s,t}) + \frac{\sigma^2}{2}(\nabla_a U)(\theta_{s,t}) \Big)\, ds + \sigma\, dB_s \,, \tag{1}$$

coupled with

$$\begin{cases} \nu_{s,t} & = \mathcal{L}(\theta_{s,t} | \mathcal{F}_t^W)\,, \\ X_{s,t} & = \xi + \int_0^t \Phi_r(X_{s,r}, \nu_{s,r})\, dr + \int_0^t \Gamma_r(X_{s,r}, \nu_{s,r}(da))\, dW_r\,, \quad t \in [0,T]\,, \\ dY_{s,t} & = -(\nabla_x H_t^0)(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t})\, dt + Z_{s,t}\, dW_t\,, \\ Y_{s,T} & = (\nabla_x g)(X_T)\,. \end{cases} \tag{2}$$

- After discretising time take a step of gradient descent $\theta^i_{s_k,t}$ for $i = 1, \dots, N$ and compute $\nu^N_{s_k,t} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_{s_k,t}}$
- Solve forward process $X_{s_k,t}(\nu^N_{s_k,t})$ on $[0,T]$
- 'Back-propagate' $(Y_{s_k,t}(\nu^N_{s_k,t}), Z_{s_k,t}(\nu^N_{s_k,t}))$ on $[0,T]$
- Update the 'gradient' and produce the next step $\theta^i_{s_{k+1},t}$

SDE / BSDE System Representation for Gradient Flow

Consider with $\theta_{t,0} = \theta_t^0$ and $s \geq 0$:

$$d\theta_{s,t} = -\Big((\nabla_a \frac{\delta H_t^0}{\delta m})(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t}, \theta_{s,t}) + \frac{\sigma^2}{2}(\nabla_a U)(\theta_{s,t}) \Big)\, ds + \sigma dB_s\,, \tag{1}$$

coupled with

$$\begin{cases} \nu_{s,t} & = \mathcal{L}(\theta_{s,t}|\mathcal{F}_t^W)\,, \\ X_{s,t} & = \xi + \int_0^t \Phi_r(X_{s,r}, \nu_{s,r})\, dr + \int_0^t \Gamma_r(X_{s,r}, \nu_{s,r}(da))\, dW_r\,, \quad t \in [0,T]\,, \\ dY_{s,t} & = -(\nabla_x H_t^0)(X_{s,t}, Y_{s,t}, Z_{s,t}, \nu_{s,t})\, dt + Z_{s,t}\, dW_t\,, \\ Y_{s,T} & = (\nabla_x g)(X_T)\,. \end{cases} \tag{2}$$

- After discretising time take a step of gradient descent $\theta^i_{s_k,t}$ for $i = 1, \dots, N$ and compute $\nu^N_{s_k,t} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_{s_k,t}}$
- Solve forward process $X_{s_k,t}(\nu^N_{s_k,t})$ on $[0,T]$
- 'Back-propagate' $(Y_{s_k,t}(\nu^N_{s_k,t}), Z_{s_k,t}(\nu^N_{s_k,t}))$ on $[0,T]$
- Update the 'gradient' and produce the next step $\theta^i_{s_{k+1},t}$
- Probabilistic numerical analysis plus propagation of chaos reuslts yield precise error rates in terms of N, learning rate etc.

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \ \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a, \nu) \text{ is constant } \text{ for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t, \omega^W) \in (0, T) \times \Omega^W \right\}.$$

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a, \nu) \text{ is constant} \text{ for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t, \omega^W) \in (0, T) \times \Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \ \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a,\nu) \text{ is constant } \text{ for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t,\omega^W) \in (0,T)\times\Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$ and that it admits unique invariant measure $\mu^* \in \mathcal{V}_q^W$ such that for any $\mu^0 \in \mathcal{V}_q^W$, $\lim_{s\to\infty} \rho_q(P_s\mu^0, \mu^*) = 0$.

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a,\nu) \text{ is constant for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t,\omega^W) \in (0,T) \times \Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$ and that it admits unique invariant measure $\mu^* \in \mathcal{V}_q^W$ such that for any $\mu^0 \in \mathcal{V}_q^W$, $\lim_{s\to\infty} \rho_q(P_s\mu^0, \mu^*) = 0$. Then

i) $\mathcal{I}^\sigma = \{\mu^*\}$. *In other words, μ^* is the only control which satisfies the first order condition.*

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a,\nu) \text{ is constant for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t,\omega^W) \in (0,T) \times \Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$ and that it admits unique invariant measure $\mu^* \in \mathcal{V}_q^W$ such that for any $\mu^0 \in \mathcal{V}_q^W$, $\lim_{s\to\infty} \rho_q(P_s\mu^0, \mu^*) = 0$. Then

i) $\mathcal{I}^\sigma = \{\mu^*\}$. *In other words, μ^* is the only control which satisfies the first order condition.*

ii) *The unique minimizer of J^σ is μ^*.*

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a,\nu) \text{ is constant } \text{ for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t,\omega^W) \in (0,T) \times \Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$ and that it admits unique invariant measure $\mu^* \in \mathcal{V}_q^W$ such that for any $\mu^0 \in \mathcal{V}_q^W$, $\lim_{s\to\infty} \rho_q(P_s\mu^0, \mu^*) = 0$. Then

- i) $\mathcal{I}^\sigma = \{\mu^*\}$. *In other words, μ^* is the only control which satisfies the first order condition.*
- ii) *The unique minimizer of J^σ is μ^*.*

From the first order condition we have that for a.a. $(\omega^W, t) \in \Omega^W \times (0,T)$

$$\mu_t^*(a) = \mathcal{Z}_t^{-1} e^{-\frac{2}{\sigma^2}\frac{\delta \mathbf{H}_t^0}{\delta m}(a,\mu^*)}\gamma(a), \quad \mathcal{Z}_t := \int e^{-\frac{2}{\sigma^2}\frac{\delta \mathbf{H}_t^0}{\delta m}(a,\mu^*)}\gamma(a)da.$$

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a,\nu) \text{ is constant } \text{ for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t,\omega^W) \in (0,T)\times\Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$ and that it admits unique invariant measure $\mu^* \in \mathcal{V}_q^W$ such that for any $\mu^0 \in \mathcal{V}_q^W$, $\lim_{s\to\infty} \rho_q(P_s\mu^0,\mu^*) = 0$. Then

i) $\mathcal{I}^\sigma = \{\mu^*\}$. *In other words, μ^* is the only control which satisfies the first order condition.*

ii) *The unique minimizer of J^σ is μ^*.*

From the first order condition we have that for a.a. $(\omega^W, t) \in \Omega^W \times (0,T)$

$$\mu_t^*(a) = \mathcal{Z}_t^{-1} e^{-\frac{2}{\sigma^2}\frac{\delta \mathbf{H}_t^0}{\delta m}(a,\mu^*)}\gamma(a)\,, \quad \mathcal{Z}_t := \int e^{-\frac{2}{\sigma^2}\frac{\delta \mathbf{H}_t^0}{\delta m}(a,\mu^*)}\gamma(a)da\,.$$

Define a set of local minimisers

$$\mathcal{I}^\sigma := \left\{ \nu \in \mathcal{V}_q^W : \frac{\delta \mathbf{H}_t^\sigma}{\delta m}(a,\nu) \text{ is constant } \text{ for a.a. } a \in \mathbb{R}^p, \text{ a.a. } (t,\omega^W) \in (0,T) \times \Omega^W \right\}.$$

Theorem 5 (Siska, Szpruch 2020)

Assume that for any $\mu^0 \in \mathcal{V}_q^W$ the MFLD xhas unique solution $P_s\mu^0$ and that it admits unique invariant measure $\mu^* \in \mathcal{V}_q^W$ such that for any $\mu^0 \in \mathcal{V}_q^W$, $\lim_{s\to\infty} \rho_q(P_s\mu^0, \mu^*) = 0$. Then

- i) $\mathcal{I}^\sigma = \{\mu^*\}$. *In other words, μ^* is the only control which satisfies the first order condition.*
- ii) *The unique minimizer of J^σ is μ^*.*

From the first order condition we have that for a.a. $(\omega^W, t) \in \Omega^W \times (0,T)$

$$\mu_t^*(a) = \mathcal{Z}_t^{-1} e^{-\frac{2}{\sigma^2}\frac{\delta \mathbf{H}_t^0}{\delta m}(a,\mu^*)}\gamma(a), \quad \mathcal{Z}_t := \int e^{-\frac{2}{\sigma^2}\frac{\delta \mathbf{H}_t^0}{\delta m}(a,\mu^*)}\gamma(a)da.$$

- $\mu_t^*(a)$ is related to Boltzmann exploration it entropy regularised RL with $\frac{\delta \mathbf{H}_t^0}{\delta m}(\cdot,\nu)$ in place of Q-function

Theorem 6 (Exponential convergence to invariant measure)

Assume that $\lambda = \frac{q}{2}\left(\sigma^2\kappa + \eta_1 - \eta_2\right) > 0$. *Then there is* $\mu^* \in \mathcal{V}_q^W$ *such that for any* $s \geq 0$ *we have* $P_s\mu^* = \mu^*$ *and* μ^* *is unique. For any* $\mu^0 \in \mathcal{V}_q^W$ *we have that*

$$\rho_q(P_s\mu^0, \mu^*) \leq e^{-\frac{1}{q}\lambda s}\rho_q(\mu^0, \mu^*)\,.$$

where for $\mu, \mu' : \Omega^W \to \mathcal{V}_2^W$ we have

$$\mathcal{W}_q^T(\mu, \nu) := \left(\int_0^T \mathcal{W}_q(\mu_t, \nu_t)^q \, dt\right)^{1/q}$$

$$\rho_q(\mu, \mu') = \left(\mathbb{E}^W\left[|\mathcal{W}_q^T(\mu, \mu')|^q\right]\right)^{1/q}$$

Extensions

- Reproduce all the results for Feedback Markov Controls (possibly via HJB route)

Extensions

- Reproduce all the results for Feedback Markov Controls (possibly via HJB route)
- Study the setting from the perspective of the agent that 'samples' optimal control

Extensions

- Reproduce all the results for Feedback Markov Controls (possibly via HJB route)
- Study the setting from the perspective of the agent that 'samples' optimal control
- Study the regret in the setting when the coefficients are unknown.

Linear-Convex RL problems

joint work with Tanut (Nash) Treetanthiploet (Turing) and Yufei Zhang (LSE)

Linear-convex control problem with known parameter

Fix $\theta = (A, B) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times p}$ and consider

$$V(\theta) = \inf_{\alpha\in\mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)} J(\alpha;\theta), \quad J(\alpha;\theta) = \mathbb{E}\left[\int_0^T f(t, X_t^{\theta,\alpha}, \alpha_t)\,\mathrm{d}t + g(X_T^{\theta,\alpha})\right],$$

where $X^{\theta,\alpha} \in \mathcal{S}^2_{\mathbb{F}}(\Omega;\mathbb{R}^d)$ is the strong solution to the following dynamics:

$$\mathrm{d}X_t = (AX_t + B\alpha_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0, T], \quad X_0 = x_0,$$

Linear-convex control problem with known parameter

Fix $\theta = (A, B) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times p}$ and consider

$$V(\theta) = \inf_{\alpha\in\mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)} J(\alpha;\theta), \quad J(\alpha;\theta) = \mathbb{E}\left[\int_0^T f(t, X_t^{\theta,\alpha}, \alpha_t)\,\mathrm{d}t + g(X_T^{\theta,\alpha})\right],$$

where $X^{\theta,\alpha} \in \mathcal{S}^2_{\mathbb{F}}(\Omega;\mathbb{R}^d)$ is the strong solution to the following dynamics:

$$\mathrm{d}X_t = (AX_t + B\alpha_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0, T], \quad X_0 = x_0,$$

Assumption 1

(i) *There exist measurable functions f_0 and h such that*

$$f(t, x, a) = f_0(t, x, a) + h(a), \quad \forall (t, x, a) \in [0, T] \times \mathbb{R}^d \times \mathbb{R}^p.$$

Furthermore $f_0(t, x, \cdot)$ is convex, $f_0(t, \cdot, \cdot)$ has Lipschitz continuous derivative and h is lower semicontinuous and convex.

(ii) *There exists $\lambda > 0$ s.t for all t, $(x, a), (x', a')$, and $\eta \in [0, 1]$,*

$$\eta f(t, x, a) + (1-\eta) f(t, x', a') \geq f(t, \eta x + (1-\eta)x', \eta a + (1-\eta)a') + \eta(1-\eta)\frac{\lambda}{2}|a - a'|^2.$$

(iii) *g is convex and differentiable with a Lipschitz continuous derivative.*

Feedback controls

Proposition 2 (M Basei, X Guo, A Hu, Y Zhang, 2021)

For any given $\theta = (A, B)$ *the LC control admits a unique optimal control* α^θ *which satisfies*

$$\alpha_t^\theta = \psi_\theta(t, X_t^\theta), \quad \mathrm{d}\mathbb{P} \otimes dt \text{ a.e.}$$

Furthermore $\forall\ (t, x) \in [0, T] \times \mathbb{R}^d$ *and* $\theta, \theta' \in \Theta$,

$$|\psi_\theta(t, x) - \psi_{\theta'}(t, x)| \leq C(1 + |x|)|\theta - \theta'|$$

Feedback controls

Proposition 2 (M Basei, X Guo, A Hu, Y Zhang, 2021)

For any given $\theta = (A, B)$ the LC control admits a unique optimal control α^θ which satisfies

$$\alpha_t^\theta = \psi_\theta(t, X_t^\theta), \quad \mathrm{d}\mathbb{P} \otimes dt \text{ a.e.}$$

Furthermore $\forall\ (t, x) \in [0, T] \times \mathbb{R}^d$ and $\theta, \theta' \in \Theta$,

$$|\psi_\theta(t, x) - \psi_{\theta'}(t, x)| \leq C(1 + |x|)|\theta - \theta'|$$

- When θ is known, this is the classical LC stochastic control problem.

Feedback controls

Proposition 2 (M Basei, X Guo, A Hu, Y Zhang, 2021)

For any given $\theta = (A, B)$ the LC control admits a unique optimal control α^θ which satisfies

$$\alpha_t^\theta = \psi_\theta(t, X_t^\theta), \quad \mathrm{d}\mathbb{P} \otimes dt \text{ a.e.}$$

Furthermore $\forall\ (t, x) \in [0, T] \times \mathbb{R}^d$ *and* $\theta, \theta' \in \Theta$,

$$|\psi_\theta(t, x) - \psi_{\theta'}(t, x)| \leq C(1 + |x|)|\theta - \theta'|$$

- When θ is known, this is the classical LC stochastic control problem.
- When θ is unknown, one needs to balance exploitation (optimal control), and exploration (learning via interactions with the random environment).

Exploration-Exploitation tradeoff in Episodic Learning

- After $(m-1)$ learning episodes, let $\hat{\theta}^{(m-1)}$ be the estimated value of an unknown parameter

Exploration-Exploitation tradeoff in Episodic Learning

- After $(m-1)$ learning episodes, let $\hat{\theta}^{(m-1)}$ be the estimated value of an unknown parameter
- Given $\hat{\theta}^{(m-1)}$ agent exercises a feedback control $\psi^{(m)}$ (which may depend on $\hat{\theta}^{(m)}$ or not) and observes

$$\mathrm{d}X^m = (AX_t^m + B\psi^m(t, X_t^m))\mathrm{d}t + \mathrm{d}W_t^m, \quad t \in [0, T], \quad X_0 = x_0, \, .$$

Exploration-Exploitation tradeoff in Episodic Learning

- After $(m-1)$ learning episodes, let $\hat{\theta}^{(m-1)}$ be the estimated value of an unknown parameter
- Given $\hat{\theta}^{(m-1)}$ agent exercises a feedback control $\psi^{(m)}$ (which may depend on $\hat{\theta}^{(m)}$ or not) and observes

$$\mathrm{d}X^m = (AX_t^m + B\psi^m(t, X_t^m))\mathrm{d}t + \mathrm{d}W_t^m, \quad t \in [0, T], \quad X_0 = x_0, .$$

- The expected cost for each episode is

$$J(\psi^{(m)}; \theta) = \mathbb{E}\left[\int_0^T f(t, X_t^{\theta,\psi^{(m)}}, \psi^{(m)}(t, X_t^{\theta,\psi^{(m)}}))\,\mathrm{d}t + g(X_T^{\theta,\psi^{(m)}})\right]$$

Exploration-Exploitation tradeoff in Episodic Learning

- After $(m-1)$ learning episodes, let $\hat{\theta}^{(m-1)}$ be the estimated value of an unknown parameter
- Given $\hat{\theta}^{(m-1)}$ agent exercises a feedback control $\psi^{(m)}$ (which may depend on $\hat{\theta}^{(m)}$ or not) and observes

$$\mathrm{d}X^m = (AX_t^m + B\psi^m(t, X_t^m))\mathrm{d}t + \mathrm{d}W_t^m, \quad t \in [0, T], \quad X_0 = x_0, \,.$$

- The expected cost for each episode is

$$J(\psi^{(m)}; \theta) = \mathbb{E}\left[\int_0^T f(t, X_t^{\theta,\psi^{(m)}}, \psi^{(m)}(t, X_t^{\theta,\psi^{(m)}}))\,\mathrm{d}t + g(X_T^{\theta,\psi^{(m)}})\right]$$

- Using $(X^i)_{i=1}^m$ agent constructs $\hat{\theta}^{(m)}$

Exploration-Exploitation tradeoff in Episodic Learning

- After $(m-1)$ learning episodes, let $\hat{\theta}^{(m-1)}$ be the estimated value of an unknown parameter
- Given $\hat{\theta}^{(m-1)}$ agent exercises a feedback control $\psi^{(m)}$ (which may depend on $\hat{\theta}^{(m)}$ or not) and observes

$$\mathrm{d}X^m = (AX_t^m + B\psi^m(t, X_t^m))\mathrm{d}t + \mathrm{d}W_t^m, \quad t \in [0, T], \quad X_0 = x_0, \,.$$

- The expected cost for each episode is

$$J(\psi^{(m)}; \theta) = \mathbb{E}\left[\int_0^T f(t, X_t^{\theta,\psi^{(m)}}, \psi^{(m)}(t, X_t^{\theta,\psi^{(m)}}))\,\mathrm{d}t + g(X_T^{\theta,\psi^{(m)}})\right]$$

- Using $(X^i)_{i=1}^m$ agent constructs $\hat{\theta}^{(m)}$
- How to design optimal algorithm $\boldsymbol{\Psi} = (\psi^{(1)}, \ldots, \psi^{(N)})$ that strikes optimal balance between exploration and exploitation ?

Regret Analysis

- Let $\boldsymbol{\Psi} = (\psi^{(1)}, \dots, \psi^{(N)})$ be a learning algorithm

Regret Analysis

- Let $\boldsymbol{\Psi} = (\psi^{(1)}, \ldots, \psi^{(N)})$ be a learning algorithm
- The expected regret of learning with $N \in \mathbb{N}$ episodes is

$$\mathcal{R}(N, \boldsymbol{\Psi}) = \sum_{m=1}^{N} \left(J(\psi^{(m)}; \theta) - J(\psi; \theta) \right)$$

where

  - $J(\psi; \theta)$ is the optimal cost as if the parameters were known
  - $J(\psi^{(m)}; \theta)$ is the cost for the m-th episode

Regret Analysis

- Let $\mathbf{\Psi} = (\psi^{(1)}, \ldots, \psi^{(N)})$ be a learning algorithm
- The expected regret of learning with $N \in \mathbb{N}$ episodes is

$$\mathcal{R}(N, \mathbf{\Psi}) = \sum_{m=1}^{N} \left(J(\psi^{(m)}; \theta) - J(\psi; \theta) \right)$$

  where
  - $J(\psi; \theta)$ is the optimal cost as if the parameters were known
  - $J(\psi^{(m)}; \theta)$ is the cost for the m-th episode

Optimal results from literature:

- For LQ-RL with self-exploration property one can construct $\mathbf{\Psi}$ s.t $\mathcal{R}(N, \mathbf{\Psi}) = \mathcal{O}((\ln N)(\ln \ln N))$, [Basei et al., 2020].

Regret Analysis

- Let $\boldsymbol{\Psi} = (\psi^{(1)}, \ldots, \psi^{(N)})$ be a learning algorithm
- The expected regret of learning with $N \in \mathbb{N}$ episodes is

$$\mathcal{R}(N, \boldsymbol{\Psi}) = \sum_{m=1}^{N} \left(J(\psi^{(m)}; \theta) - J(\psi; \theta) \right)$$

  where
  - $J(\psi; \theta)$ is the optimal cost as if the parameters were known
  - $J(\psi^{(m)}; \theta)$ is the cost for the m-th episode

Optimal results from literature:

- For LQ-RL with self-exploration property one can construct $\boldsymbol{\Psi}$ s.t $\mathcal{R}(N, \boldsymbol{\Psi}) = \mathcal{O}((\ln N)(\ln \ln N))$, [Basei et al., 2020].
- For LC-RL with irregular cost function and with self-exploration property one can construct $\boldsymbol{\Psi}$ s.t $\mathcal{R}(N, \boldsymbol{\Psi}) = \mathcal{O}(\sqrt{N \ln N})$, [Guo et al., 2021].

Do we need to explore?

Do we need to explore?

- Consider the 1d controlled SDE

$$\mathrm{d}X_t = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\mathrm{d}t + \mathrm{d}W_t$$

with $(B_1, B_2) \neq (0,0)$ and the cost

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\mathrm{d}t + X_T^2\right],$$

Do we need to explore?

- Consider the 1d controlled SDE

$$\mathrm{d}X_t = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\mathrm{d}t + \mathrm{d}W_t$$

with $(B_1, B_2) \neq (0,0)$ and the cost

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\mathrm{d}t + X_T^2\right],$$

- The optimal policy is given by $\psi_\theta(t,x) = -p_t B^\top x$, where $(p_t)_t$ satisfies corresponding Riccati equation.

Do we need to explore?

- Consider the 1d controlled SDE

$$\mathrm{d}X_t = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\mathrm{d}t + \mathrm{d}W_t$$

with $(B_1, B_2) \neq (0, 0)$ and the cost

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\mathrm{d}t + X_T^2\right],$$

- The optimal policy is given by $\psi_\theta(t,x) = -p_t B^\top x$, where $(p_t)_t$ satisfies corresponding Riccati equation.
- Assume agent learns by only executing optimal (greedy) policy

Do we need to explore?

- Consider the 1d controlled SDE

$$\mathrm{d}X_t = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\mathrm{d}t + \mathrm{d}W_t$$

with $(B_1, B_2) \neq (0,0)$ and the cost

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\mathrm{d}t + X_T^2\right],$$

- The optimal policy is given by $\psi_\theta(t,x) = -p_t B^\top x$, where $(p_t)_t$ satisfies corresponding Riccati equation.
- Assume agent learns by only executing optimal (greedy) policy
- Assume that after the first episode we have $(B_1^{(1)}, 0)$, $B_1^{(1)} \neq 0$ and consequently agent executes $(\alpha_{1,t}, 0)$ and only learns about B_1 in the next episode

Do we need to explore?

- Consider the 1d controlled SDE

$$\mathrm{d}X_t = (B_1\alpha_{1,t} + B_2\alpha_{2,t})\mathrm{d}t + \mathrm{d}W_t$$

with $(B_1, B_2) \neq (0,0)$ and the cost

$$J(\alpha;\theta) = \mathbb{E}\left[\int_0^T (\alpha_{1,t}^2 + \alpha_{2,t}^2)\mathrm{d}t + X_T^2\right],$$

- The optimal policy is given by $\psi_\theta(t,x) = -p_t B^\top x$, where $(p_t)_t$ satisfies corresponding Riccati equation.
- Assume agent learns by only executing optimal (greedy) policy
- Assume that after the first episode we have $(B_1^{(1)}, 0)$, $B_1^{(1)} \neq 0$ and consequently agent executes $(\alpha_{1,t}, 0)$ and only learns about B_1 in the next episode
- and if it happens that for all $m \in \mathbb{N}$, $(B_1^{(m)}, 0)$, $B_1^{(m)} \neq 0$, the optimal model and the optimal policy will never be learned.

Bayesian Perspective

- One should view the unknown parameter as a (hidden or unobserved) random variable $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$.

Bayesian Perspective

- One should view the unknown parameter as a (hidden or unobserved) random variable $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$.
- Given fixed policy our aim is to estimate $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$ where

$$\mathrm{d}X_t = \boldsymbol{\theta} Z_t^{\alpha} \mathrm{d}t + \mathrm{d}W_t$$

Bayesian Perspective

- One should view the unknown parameter as a (hidden or unobserved) random variable $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$.
- Given fixed policy our aim is to estimate $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$ where

$$\mathrm{d}X_t = \boldsymbol{\theta} Z_t^{\alpha} \mathrm{d}t + \mathrm{d}W_t$$

- Given prior $\pi_0(\boldsymbol{\theta}) = N(\hat{\theta}_0, v_0)$ the posterior is given by

$$\pi\left(\boldsymbol{\theta} | \mathcal{F}_t^{X,\alpha}\right) = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}(t, X^{\alpha})\pi_0(\theta)$$
$$\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}\left(v_0^{-1} + \int_0^t (Z_s^{\alpha})(Z_s^{\alpha})^\top \mathrm{d}s\right)\boldsymbol{\theta}^\top + \boldsymbol{\theta}\left(v_0^{-1}\hat{\theta}_0^\top + \int_0^t (Z_s^{\alpha})\mathrm{d}X_s\right)\right).$$

Bayesian Perspective

- One should view the unknown parameter as a (hidden or unobserved) random variable $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$.
- Given fixed policy our aim is to estimate $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{B})$ where

$$\mathrm{d}X_t = \boldsymbol{\theta} Z_t^{\alpha} \mathrm{d}t + \mathrm{d}W_t$$

- Given prior $\pi_0(\boldsymbol{\theta}) = N(\hat{\theta}_0, v_0)$ the posterior is given by

$$\pi\left(\boldsymbol{\theta} | \mathcal{F}_t^{X,\alpha}\right) = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}(t, X^\alpha)\pi_0(\theta)$$
$$\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}\left(v_0^{-1} + \int_0^t (Z_s^\alpha)(Z_s^\alpha)^\top \mathrm{d}s\right)\boldsymbol{\theta}^\top + \boldsymbol{\theta}\left(v_0^{-1}\hat{\theta}_0^\top + \int_0^t (Z_s^\alpha)\mathrm{d}X_s\right)\right).$$

- We see that the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{F}_t^X) = N(\hat{\theta}_t, V_t)$ where

$$\hat{\boldsymbol{\theta}}_t = \mathbb{E}[\boldsymbol{\theta}|\mathcal{F}_t^{X,\alpha}] = \left(v_0^{-1}\hat{\theta}_0^\top + \int_0^t (Z_s^\alpha)\mathrm{d}X_s\right)^\top \left(v_0^{-1} + \int_0^t (Z_s^\alpha)(Z_s^\alpha)^\top \mathrm{d}s\right)^{-1}$$
$$V_t = \mathrm{Var}[\boldsymbol{\theta}|\mathcal{F}_t^{X,\alpha}] = \left(v_0^{-1} + \int_0^t (Z_s^\alpha)(Z_s^\alpha)^\top \mathrm{d}s\right)^{-1}.$$

Phased Exploration with Greedy Exploitation

Algorithm 1: PEGE Algorithm

Input: $\mathfrak{m} : \mathbb{N} \to \mathbb{N}$.

Initialize $m = 0$.
for $k = 1, 2, \dots$ **do**
  Execute the exploration policy ψ^e for one episode, and $m \leftarrow m + 1$.
  Update the estimate $\hat{\boldsymbol{\theta}}_m$ and set $\overline{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_m$.
  for $l = 1, 2, \dots, \mathfrak{m}(k)$ **do**
    Execute the greedy policy $\psi_{\overline{\boldsymbol{\theta}}}$ for one episode, and $m \leftarrow m + 1$.
  end
end

- Here greedy policy is given by

$$\Psi_m(\omega, t, x) = \psi_m(\hat{\boldsymbol{\theta}}^{\boldsymbol{\Psi}, m-1}(\omega), V^{\boldsymbol{\theta}, \boldsymbol{\Psi}, m-1}(\omega), t, x)$$

- Sufficient statistics are updates as at the episodes $j = n+1, \dots, m$:

$$V^{\boldsymbol{\theta}, \boldsymbol{\Psi}, m} = \left((V^{\boldsymbol{\theta}, \boldsymbol{\Psi}, n})^{-1} + \sum_{j=n+1}^{m} \int_0^T Z_s^{\boldsymbol{\theta}, \boldsymbol{\Psi}, j} (Z_s^{\boldsymbol{\theta}, \boldsymbol{\Psi}, j})^\top \mathrm{d}s \right)^{-1},$$

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{\Psi}, m} = \left(\hat{\boldsymbol{\theta}}^{\boldsymbol{\Psi}, n} (V^{\boldsymbol{\theta}, \boldsymbol{\Psi}, n})^{-1} + \sum_{j=n+1}^{m} \left(\int_0^T Z_s^{\boldsymbol{\theta}, \boldsymbol{\Psi}, j} (\mathrm{d}X_s^{\boldsymbol{\theta}, \boldsymbol{\Psi}, j})^\top \right)^\top \right) V^{\boldsymbol{\theta}, \boldsymbol{\Psi}, m}.$$

Regret Analysis

Let $\mathcal{E}^{\boldsymbol{\Psi}} = \{m \in \mathbb{N} | \boldsymbol{\Psi}_m = \psi^e\}$ and consider

$$\mathcal{R}(N, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \sum_{m=1}^{N} \left(J(\psi^{(m)}; \theta) - J(\psi_\theta; \theta) \right)$$

$$= \sum_{m \in [1,N] \cap \mathcal{E}^{\boldsymbol{\Psi}}} \left(J(\psi^e, \boldsymbol{\theta}) - J(\psi_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \right) + \sum_{m \in [1,N] \cap (\mathcal{E}^{\boldsymbol{\Psi}})^c}^{N} \left(J(\psi_{\hat{\boldsymbol{\theta}}_{m-1}}, \boldsymbol{\theta}) - J(\psi_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \right)$$

Regret Analysis

Let $\mathcal{E}^{\boldsymbol{\Psi}} = \{m \in \mathbb{N} | \boldsymbol{\Psi}_m = \psi^e\}$ and consider

$$\mathcal{R}(N, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \sum_{m=1}^{N} \left(J(\psi^{(m)}; \theta) - J(\psi_\theta; \theta) \right)$$

$$= \sum_{m \in [1,N] \cap \mathcal{E}^{\boldsymbol{\Psi}}} \left(J(\psi^e, \boldsymbol{\theta}) - J(\psi_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \right) + \sum_{m \in [1,N] \cap (\mathcal{E}^{\boldsymbol{\Psi}})^c}^{N} \left(J(\psi_{\hat{\boldsymbol{\theta}}_{m-1}}, \boldsymbol{\theta}) - J(\psi_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \right)$$

Assumption 3 (Performance Gap)

There exist constants $L_\Theta, \beta > 0$, $r \in (0, 1]$ *such that for all* $\theta_0 \in \Theta$,

$$|J(\psi_\theta; \theta_0) - J(\psi_{\theta_0}; \theta_0)| \leq L_\Theta |\theta - \theta_0|^{2r}, \quad \forall \theta \in \mathbb{B}_\beta(\theta_0),$$

where ψ_θ *is an optimal feedback control with parameter* θ.

Regret Analysis

Let $\mathcal{E}^{\boldsymbol{\Psi}} = \{m \in \mathbb{N} | \boldsymbol{\Psi}_m = \psi^e\}$ and consider

$$\mathcal{R}(N, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \sum_{m=1}^{N} \Big(J(\psi^{(m)}; \theta) - J(\psi_\theta; \theta) \Big)$$

$$= \sum_{m \in [1,N] \cap \mathcal{E}^{\boldsymbol{\Psi}}} \left(J(\psi^e, \boldsymbol{\theta}) - J(\psi_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \right) + \sum_{m \in [1,N] \cap (\mathcal{E}^{\boldsymbol{\Psi}})^c}^{N} \Big(J(\psi_{\hat{\boldsymbol{\theta}}_{m-1}}, \boldsymbol{\theta}) - J(\psi_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \Big)$$

Assumption 3 (Performance Gap)

There exist constants $L_\Theta, \beta > 0$, $r \in (0, 1]$ *such that for all* $\theta_0 \in \Theta$,

$$|J(\psi_\theta; \theta_0) - J(\psi_{\theta_0}; \theta_0)| \leq L_\Theta |\theta - \theta_0|^{2r}, \quad \forall \theta \in \mathbb{B}_\beta(\theta_0),$$

where ψ_θ *is an optimal feedback control with parameter* θ.

We then have

$$\mathcal{R}(N, \boldsymbol{\Psi}, \boldsymbol{\theta}) \lesssim \big(J(\psi^e, \boldsymbol{\theta}) + V(\boldsymbol{\theta}) \big) \kappa^{\boldsymbol{\Psi}}(N) + \sum_{m \in [1,N] \cap \mathcal{E}^{(\boldsymbol{\Psi})^c}}^{N} L_\Theta |\hat{\boldsymbol{\theta}}_{m-1} - \boldsymbol{\theta}|^{2r}$$

Performance Gap analysis

Theorem 7

Let $h \equiv 0$*. Then for any* $\beta > 0$*, there exists* $L_\Theta > 0$ *such that performance gap assumption holds with* $r = 1$*.*

Performance Gap analysis

Theorem 7

Let $h \equiv 0$. Then for any $\beta > 0$, there exists $L_\Theta > 0$ such that performance gap assumption holds with $r = 1$.

- First show $J(\cdot;\theta_0) : \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p) \to \mathbb{R} \cup \{\infty\}$ is convex and has a Lipschitz continuous derivative
- Conclude that $J(\alpha;\theta_0) - J(\alpha^{\theta_0};\theta_0) \leq C\|\alpha - \alpha^{\theta_0}\|^2_{\mathcal{H}^2}$ for all $\alpha \in \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)$

Performance Gap analysis

Theorem 7

Let $h \equiv 0$. Then for any $\beta > 0$, there exists $L_\Theta > 0$ such that performance gap assumption holds with $r = 1$.

- First show $J(\cdot;\theta_0) : \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p) \to \mathbb{R} \cup \{\infty\}$ is convex and has a Lipschitz continuous derivative
- Conclude that $J(\alpha;\theta_0) - J(\alpha^{\theta_0};\theta_0) \leq C\|\alpha - \alpha^{\theta_0}\|^2_{\mathcal{H}^2}$ for all $\alpha \in \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)$

Theorem 8

Let the cost function be given form

$$f(t,x,a) := f_0(t,x)^\top a + h_{en}(a), \quad h_{en}(a) = \sum_{i=1}^{p} a_i \ln(a_i),$$

Assume further that $f_0(t,\cdot) \in C^4_b(\mathbb{R}^d)$ and $g \in C^4_b(\mathbb{R}^d)$ uniformly in t.
Then the performance gap assumption holds with $r = 1$.

Performance Gap analysis

Theorem 7

Let $h \equiv 0$. Then for any $\beta > 0$, there exists $L_\Theta > 0$ such that performance gap assumption holds with $r = 1$.

- First show $J(\cdot;\theta_0) : \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p) \to \mathbb{R} \cup \{\infty\}$ is convex and has a Lipschitz continuous derivative
- Conclude that $J(\alpha;\theta_0) - J(\alpha^{\theta_0};\theta_0) \leq C\|\alpha - \alpha^{\theta_0}\|^2_{\mathcal{H}^2}$ for all $\alpha \in \mathcal{H}^2_{\mathbb{F}}(\Omega;\mathbb{R}^p)$

Theorem 8

Let the cost function be given form

$$f(t,x,a) := f_0(t,x)^\top a + h_{en}(a), \quad h_{en}(a) = \sum_{i=1}^{p} a_i \ln(a_i),$$

Assume further that $f_0(t,\cdot) \in C^4_b(\mathbb{R}^d)$ and $g \in C^4_b(\mathbb{R}^d)$ uniformly in t.
Then the performance gap assumption holds with $r = 1$.

- Expand cost function into 2nd order Taylor series around the minimiser.

Optimal Regret

Theorem 9

Consider PEGE algorithm. We have

- *For* $\mathfrak{m}(k) = \lfloor k^r \rfloor$ *for all* $k \in \mathbb{N}$

$$\mathcal{R}(N, \boldsymbol{\Psi}^{PEGE}, \boldsymbol{\theta})] \leq CN^{\frac{1}{1+r}} (\log N)^r \quad \text{for all } N \in \mathbb{N} \cap [2, \infty).$$

- *Assume self-exploration property holds. Then for* $\mathfrak{m}(k) = 2^k$

$$\mathbb{E}^{\mathbb{P}}[\mathcal{R}(N, \boldsymbol{\Psi}^{PEGE}, \boldsymbol{\theta})] \leq \begin{cases} CN^{1-r}\big(\log N\big)^r, & r \in (0,1), \\ C\big(\log N\big)^2, & r = 1, \end{cases}$$

Optimal Regret

Theorem 9

Consider PEGE algorithm. We have

- *For* $\mathfrak{m}(k) = \lfloor k^r \rfloor$ *for all* $k \in \mathbb{N}$

$$\mathcal{R}(N, \mathbf{\Psi}^{PEGE}, \boldsymbol{\theta})] \leq CN^{\frac{1}{1+r}} (\log N)^r \quad \text{for all } N \in \mathbb{N} \cap [2, \infty).$$

- *Assume self-exploration property holds. Then for* $\mathfrak{m}(k) = 2^k$

$$\mathbb{E}^{\mathbb{P}}[\mathcal{R}(N, \mathbf{\Psi}^{PEGE}, \boldsymbol{\theta})] \leq \begin{cases} CN^{1-r} \big(\log N \big)^r, & r \in (0,1), \\ C \big(\log N \big)^2, & r = 1, \end{cases}$$

- Proof requires concentration inequalities for conditional sub-exponential random variables
- One can also obtain high probability bounds for pathwsie regret
- Easy extension to ϵ-greedy algorithms.

References I

[Acciaio et al., 2020] Acciaio, B., Backhoff-Veraguas, J., and Zalashko, A. (2020). Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *Stochastic Processes and their Applications*, 130(5):2918–2953.

[Backhoff-Veraguas et al., 2020] Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., and Eder, M. (2020). All adapted topologies are equal. *Probability Theory and Related Fields*, 178(3):1125–1172.

[Basei et al., 2020] Basei, M., Guo, X., Hu, A., and Zhang, Y. (2020). Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *arXiv preprint arXiv:2006.15316*.

[Cao et al., 2021] Cao, H., Cohen, S. N., and Szpruch, L. (2021). Identifiability in inverse reinforcement learning. *arXiv preprint arXiv:2106.03498*.

[Cohen et al., 2021] Cohen, S. N., Reisinger, C., and Wang, S. (2021). Arbitrage-free neural-sde market models.

[Cuchiero et al., 2020] Cuchiero, C., Khosrawi, W., and Teichmann, J. (2020). A generative adversarial network approach to calibration of local stochastic volatility models. *arXiv preprint arXiv:2005.02505*.

[Gierjatowicz et al., 2020] Gierjatowicz, P., Sabate-Vidales, M., Siska, D., Szpruch, L., and Zuric, Z. (2020). Robust pricing and hedging via neural sdes. *Available at SSRN 3646241*.

[Guo et al., 2021] Guo, X., Hu, A., and Zhang, Y. (2021). Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls. *arXiv preprint arXiv:2104.09311*.

[Hu et al., 2019] Hu, K., Kazeykina, A., and Ren, Z. (2019). Mean-field langevin system, optimal control and deep neural networks. *arXiv preprint arXiv:1909.07278*.

[Hu et al., 2021] Hu, K., Ren, Z., Šiška, D., and Szpruch, Ł. (2021). Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré.

[Jabir et al., 2019] Jabir, J.-F., Šiška, D., and Szpruch, Ł. (2019). Mean-field neural odes via relaxed optimal control. *arXiv preprint arXiv:1912.05475*.

References II

[Karatzas et al., 2018] Karatzas, I., Schachermayer, W., and Tschiderer, B. (2018). Trajectorial otto calculus. *arXiv preprint arXiv:1811.08686.*

[Reisinger and Zhang, 2020] Reisinger, C. and Zhang, Y. (2020). Regularity and stability of feedback relaxed controls. *arXiv preprint arXiv:2001.03148.*

[Wang et al., 2020] Wang, H., Zariphopoulou, T., and Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 21:198–1.