Provably convergent policy gradient methods for continuous-time stochastic control

Yufei Zhang

Department of Statistics, London School of Economics

with Christoph Reisinger and Wolfgang Stockinger

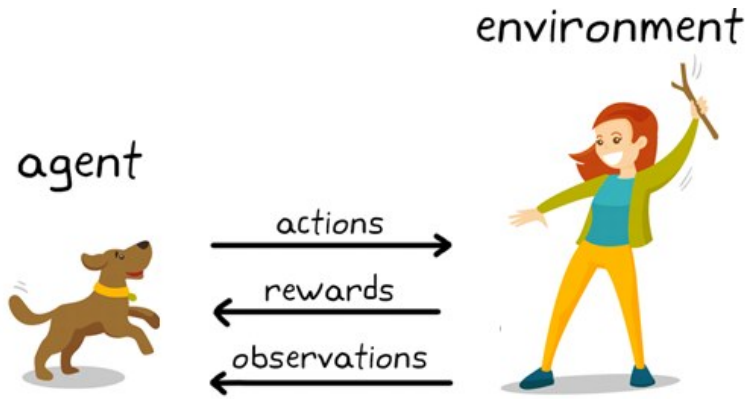
BSDE2022, 29 June 2022

Motivation and overview

- Stochastic control problems are ubiquitous.
- Continuous-time models well understood in this community.
- Reinforcement learning (RL) methods increasingly popular.
- Analysis restricted to discrete-time models.

This talk:

- Analysis of policy gradient methods for continuous-time models using control techiniques.



Stochastic control v.s. RL

Classical control theory focuses on:

- existence and uniqueness of optimal control processes.
- characterisation and regularity of value function.
- little attention has been on feedback control, i.e., a function mapping system states to actions.

Stochastic control v.s. RL

Classical control theory focuses on:

- existence and uniqueness of optimal control processes.
- characterisation and regularity of value function.
- little attention has been on feedback control, i.e., a function mapping system states to actions.

Learning algorithm naturally of feedback form:

- regularity of feedback control (a.k.a. policy).
- convergence/regret rate analysis:
  - critical for understanding algorithm efficiency.

Policy gradient method

- Approximate a policy in a parametric form, and update the policy parametrization iteratively based on gradients of the objective function.

Policy gradient method

- Approximate a policy in a parametric form, and update the policy parametrization iteratively based on gradients of the objective function.
- Analysing the convergence of PGMs is technically challenging, as the objective of a control problem (even for LQ problems) is typically nonconvex with respect to the policies.

Related works on PGMs

Analysis restricted to discrete-time models and specific policy parameterisation.

- Linear convergence to optimality:
  - tabular MDP with softmax policy: Mei, Xiao, Szpesvari, Schuurmans (2020).
  - entropy-regularised MDP with one-layer neural network policy: Kerimkulov, Leahy, Siska, Szpruch (2022).
  - LQ with linear policy: Fazel, Ge, Kakade, Mesbahi (2018); Hambly, Xu, Yang (2021).

Related works on PGMs

Analysis restricted to discrete-time models and specific policy parameterisation.

- Linear convergence to optimality:
  - tabular MDP with softmax policy: Mei, Xiao, Szpesvari, Schuurmans (2020).
  - entropy-regularised MDP with one-layer neural network policy: Kerimkulov, Leahy, Siska, Szpruch (2022).
  - LQ with linear policy: Fazel, Ge, Kakade, Mesbahi (2018); Hambly, Xu, Yang (2021).
- Trapped at local minimum:
  - LQ with piecewise linear policy: Chen, Agazzi (2021).

Related works on PGMs

Analysis restricted to discrete-time models and specific policy parameterisation.

- Linear convergence to optimality:
  - tabular MDP with softmax policy: Mei, Xiao, Szpesvari, Schuurmans (2020).
  - entropy-regularised MDP with one-layer neural network policy: Kerimkulov, Leahy, Siska, Szpruch (2022).
  - LQ with linear policy: Fazel, Ge, Kakade, Mesbahi (2018); Hambly, Xu, Yang (2021).
- Trapped at local minimum:
  - LQ with piecewise linear policy: Chen, Agazzi (2021).

Continuous-time:

- Algorithm design: Jia, Zhou (2000, 2021).

Open: convergence behaviour of PGMs for general models/policies.

Our work

Reisinger, Stockinger, Zhang (2022)

- General stochastic control problem: nonlinear state dynamics and nonconvex, nonsmooth costs.
- Non-parametric time-dependent policies.
- Linear convergence to stationary points.

Minimise

$$J(\alpha) = \mathbb{E}\left[\int_0^T e^{-\rho t}\left(f_t(X_t^\alpha, \alpha_t) + \ell(\alpha_t)\right)\,\mathrm{d}t + e^{-\rho T} g(X_T^\alpha)\right]$$

over all square integrable, adapted processes α, where X^α satisfies

$$\mathrm{d}X_t = b_t(X_t, \alpha_t)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}W_t, \quad X_0 = \xi_0.$$

- f, g, b, σ are differentiable.
- ℓ is possibly non-smooth and infinite.

ℓ represents control constraints, ℓ_1-norm or entropy regularisers.

A "naive" gradient direction

Special case with $\ell = \rho = 0$, $\sigma_t(x) \coloneqq \sigma$

Minimise

$$J(\phi) = \mathbb{E}\left[\int_0^T f_t(X_t^\phi, \phi_t(X_t^\phi))\,\mathrm{d}t + g(X_T^\phi)\right]$$

over all feedback controls ϕ, where X^ϕ satisfies

$$\mathrm{d}X_t = b_t(X_t, \phi_t(X_t))\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad X_0 = \xi_0.$$

A "naive" gradient direction

Special case with $\ell = \rho = 0$, $\sigma_t(x) \coloneqq \sigma$

Minimise

$$J(\phi) = \mathbb{E}\left[\int_0^T f_t(X_t^\phi, \phi_t(X_t^\phi))\,\mathrm{d}t + g(X_T^\phi)\right]$$

over all feedback controls ϕ, where X^ϕ satisfies

$$\mathrm{d}X_t = b_t(X_t, \phi_t(X_t))\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad X_0 = \xi_0.$$

For each policy ϕ and test policy ψ,

$$\left.\frac{\mathrm{d}J(\phi + \epsilon\psi)}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = \mathbb{E}\left[\int_0^T \langle \partial_a H_t^{\mathrm{re}}(X_t^\phi, \phi_t(X_t^\phi), \partial_x u_t^\phi(X_t^\phi)), \psi_t(X_t^\phi)\rangle\,\mathrm{d}t\right],$$

where $H_t^{\mathrm{re}}(x, a, y) \coloneqq \langle b_t(x, a), y\rangle + f_t(x, a)$, and u^ϕ satisfies

$$\partial_t u_t(x) + \tfrac{1}{2}\sigma^2 \partial_{xx} u_t(x) + H_t^{\mathrm{re}}\big(x, \phi_t(x), \partial_x u_t(x)\big) = 0, \quad u_T(x) = g(x).$$

A "naive" PGM

$H_t^{\mathrm{re}}(x, a, y) \coloneqq \langle b_t(x, a), y \rangle + f_t(x, a)$

- Given ϕ^0, perform gradient descent steps

$$\phi_t^{m+1}(x) = \phi_t^m(x) - \tau \partial_a H_t^{\mathrm{re}}(x, \phi_t^m(x), \partial_x u_t^{\phi^m}(x)),$$

where for each ϕ, u^ϕ satisfies

$$\partial_t u_t(x) + \tfrac{1}{2}\sigma^2 \partial_{xx} u_t(x) + H_t^{re}\big(x, \phi_t(x), \partial_x u_t(x)\big) = 0, \; u_T(x) = g(x).$$

A "naive" PGM

$H_t^{\mathrm{re}}(x, a, y) \coloneqq \langle b_t(x, a), y \rangle + f_t(x, a)$

- Given ϕ^0, perform gradient descent steps

$$\phi_t^{m+1}(x) = \phi_t^m(x) - \tau \partial_a H_t^{\mathrm{re}}(x, \phi_t^m(x), \partial_x u_t^{\phi^m}(x)),$$

where for each ϕ, u^ϕ satisfies

$$\partial_t u_t(x) + \tfrac{1}{2}\sigma^2 \partial_{xx} u_t(x) + H_t^{re}\big(x, \phi_t(x), \partial_x u_t(x)\big) = 0, \; u_T(x) = g(x).$$

- If $\partial_a H_t^{\mathrm{re}}(x, \phi_t^\star(x), \partial_x u_t^{\phi^\star}(x)) = 0$, then $\phi^\star$ is optimal.

A "naive" PGM

$H_t^{\mathrm{re}}(x, a, y) \coloneqq \langle b_t(x, a), y \rangle + f_t(x, a)$

- Given ϕ^0, perform gradient descent steps

$$\phi_t^{m+1}(x) = \phi_t^m(x) - \tau \partial_a H_t^{\mathrm{re}}(x, \phi_t^m(x), \partial_x u_t^{\phi^m}(x)),$$

where for each ϕ, u^ϕ satisfies

$$\partial_t u_t(x) + \tfrac{1}{2}\sigma^2 \partial_{xx} u_t(x) + H_t^{re}\big(x, \phi_t(x), \partial_x u_t(x)\big) = 0, \; u_T(x) = g(x).$$

- If $\partial_a H_t^{\mathrm{re}}(x, \phi_t^\star(x), \partial_x u_t^{\phi^\star}(x)) = 0$, then $\phi^\star$ is optimal.
- ϕ^{m+1} has lower regularity than ϕ^m, as $\partial_x u^{\phi^m}$ has the same regularity as $\partial_x \phi^m$.

A "naive" PGM

$H_t^{\mathrm{re}}(x, a, y) := \langle b_t(x, a), y \rangle + f_t(x, a)$

- Given ϕ^0, perform gradient descent steps

$$\phi_t^{m+1}(x) = \phi_t^m(x) - \tau \partial_a H_t^{\mathrm{re}}(x, \phi_t^m(x), \partial_x u_t^{\phi^m}(x)),$$

where for each ϕ, u^ϕ satisfies

$$\partial_t u_t(x) + \tfrac{1}{2}\sigma^2 \partial_{xx} u_t(x) + H_t^{re}\big(x, \phi_t(x), \partial_x u_t(x)\big) = 0, \; u_T(x) = g(x).$$

- If $\partial_a H_t^{\mathrm{re}}(x, \phi_t^\star(x), \partial_x u_t^{\phi^\star}(x)) = 0$, then $\phi^\star$ is optimal.
- ϕ^{m+1} has lower regularity than ϕ^m, as $\partial_x u^{\phi^m}$ has the same regularity as $\partial_x \phi^m$. To see it, observe $v := \partial_x u^{\phi^m}$ solves

$$\begin{aligned}\partial_t v_t(x) + \mathcal{L}^{\phi^m} v_t(x) = -[&\partial_x H_t^{re}\big(x, \phi_t^m(x), v_t(x)\big) \\ &+ \partial_a H_t^{re}\big(x, \phi_t^m(x), v_t(x)\big) \partial_x \phi_t^m(x)], \quad v_T(x) = \partial_x g(x),\end{aligned}$$

where $\mathcal{L}^{\phi^m}$ is the generator of X^{ϕ^m}.

Minimise

$$J(\alpha) = \mathbb{E}\left[\int_0^T e^{-\rho t}\left(f_t(X_t^\alpha, \alpha_t) + \ell(\alpha_t)\right)\,\mathrm{d}t + e^{-\rho T} g(X_T^\alpha)\right]$$

over all admissible control processes α, where X^α satisfies

$$\mathrm{d}X_t = b_t(X_t, \alpha_t)\,\mathrm{d}t + \sigma_t(X_t)\,\mathrm{d}W_t, \quad X_0 = \xi_0.$$

where

- f, g, b, σ are differentiable,
- ℓ is convex, possibly non-smooth and infinite.

Stochastic maximum principle

Smooth case: $\ell = 0$

- Adjoint processes (Y^α, Z^α) for a control α:

$$\mathrm{d}Y_t = -\partial_x H_t(X_t^\alpha, \alpha_t, Y_t, Z_t)\,\mathrm{d}t + Z_t\,\mathrm{d}W_t,\ Y_T = e^{-\rho T}\partial_x g(X_T^\alpha),$$

where H is the Hamiltonian:

$$H_t(x, a, y, z) = \langle b_t(x, a), y\rangle + \langle \sigma_t(x), z\rangle + e^{-\rho t} f_t(x, a).$$

- Gradient of $J(\cdot)$ at α:

$$\nabla J(\alpha)_t = \partial_a H_t(X_t^\alpha, \alpha_t, Y_t^\alpha, Z_t^\alpha).$$

Stochastic maximum principle

Smooth case: $\ell = 0$

- Adjoint processes (Y^α, Z^α) for a control α:

$$\mathrm{d}Y_t = -\partial_x H_t(X_t^\alpha, \alpha_t, Y_t, Z_t)\,\mathrm{d}t + Z_t\,\mathrm{d}W_t,\ Y_T = e^{-\rho T}\partial_x g(X_T^\alpha),$$

where H is the Hamiltonian:

$$H_t(x, a, y, z) = \langle b_t(x, a), y\rangle + \langle \sigma_t(x), z\rangle + e^{-\rho t} f_t(x, a).$$

- Gradient of $J(\cdot)$ at α:

$$\nabla J(\alpha)_t = \partial_a H_t(X_t^\alpha, \alpha_t, Y_t^\alpha, Z_t^\alpha).$$

- α is a stationary point of J if $\partial_a H_t^{\mathrm{re}}(X_t^\alpha, \alpha_t, Y_t^\alpha) = 0$, with

$$H_t^{\mathrm{re}}(x, a, y) := \langle b_t(x, a), y\rangle + e^{-\rho t} f_t(x, a).$$

Open-loop (proximal) gradient descent

- Smooth case with $\ell = 0$: perform gradient steps

$$\alpha_t^{m+1} = \alpha_t^m - \tau e^{\rho t} \partial_a H_t^{\mathrm{re}}(X_t^{\xi_0, \alpha^m}, \alpha_t^m, Y_t^{\xi_0, \alpha^m}).$$

Open-loop (proximal) gradient descent

- Smooth case with $\ell = 0$: perform gradient steps

$$\alpha_t^{m+1} = \alpha_t^m - \tau e^{\rho t} \partial_a H_t^{\mathrm{re}}(X_t^{\xi_0, \alpha^m}, \alpha_t^m, Y_t^{\xi_0, \alpha^m}).$$

- Nonsmooth case: define the proximal map

$$\mathrm{prox}_{\tau\ell}(a) = \arg\min_{z \in \mathbb{R}^k} \left(\frac{1}{2}|z - a|^2 + \tau \ell(z) \right), \quad a \in \mathbb{R}^k,$$

and perform proximal gradient steps:

$$\alpha_t^{m+1} = \mathrm{prox}_{\tau\ell}\left(\alpha_t^m - \tau e^{\rho t} \partial_a H_t^{\mathrm{re}}(X_t^{\xi_0, \alpha^m}, \alpha_t^m, Y_t^{\xi_0, \alpha^m})\right).$$

Proximal policy gradient method

$\mathrm{prox}_{\tau\ell}(a) = \arg\min_{z\in\mathbb{R}^k}\left(\frac{1}{2}|z-a|^2 + \tau\ell(z)\right)$

Given ϕ^0, perform proximal gradient steps

$$\phi_t^{m+1}(x) = \mathrm{prox}_{\tau\ell}\big(\phi_t^m(x) - \tau e^{\rho t}\partial_a H_t^{\mathrm{re}}(x, \phi_t^m(x), Y_t^{t,x,\phi^m})\big),$$

where for each ϕ,

$$\begin{aligned}
\mathrm{d}X_s^{t,x,\phi} &= b_s(X_s^{t,x,\phi}, \phi_s(X_s^{t,x,\phi}))\,\mathrm{d}s + \sigma_s(X_s^{t,x,\phi})\,\mathrm{d}W_s,\\
\mathrm{d}Y_s^{t,x,\phi} &= -\partial_x H_s(X_s^{t,x,\phi}, \phi_s(X_s^{t,x,\phi}), Y_s^{t,x,\phi}, Z_s^{t,x,\phi})\,\mathrm{d}s + Z_s^{t,x,\phi}\,\mathrm{d}W_s,\\
X_t^{t,x,\phi} &= x, \quad Y_T^{t,x,\phi} = e^{-\rho T}\partial_x g(X_T^{t,x,\phi}).
\end{aligned}$$

- $(Y_t^{t,x,\phi^m})_{(t,x)\in[0,T]\times\mathbb{R}^n}$ is the Markovian representation of adjoint process Y^{α^m}.

Proximal policy gradient method

$\mathrm{prox}_{\tau\ell}(a) = \arg\min_{z\in\mathbb{R}^k}\left(\frac{1}{2}|z-a|^2 + \tau\ell(z)\right)$

Given ϕ^0, perform proximal gradient steps

$$\phi_t^{m+1}(x) = \mathrm{prox}_{\tau\ell}\big(\phi_t^m(x) - \tau e^{\rho t}\partial_a H_t^{\mathrm{re}}(x, \phi_t^m(x), Y_t^{t,x,\phi^m})\big),$$

where for each ϕ,

$$\begin{aligned}
\mathrm{d}X_s^{t,x,\phi} &= b_s(X_s^{t,x,\phi}, \phi_s(X_s^{t,x,\phi}))\,\mathrm{d}s + \sigma_s(X_s^{t,x,\phi})\,\mathrm{d}W_s,\\
\mathrm{d}Y_s^{t,x,\phi} &= -\partial_x H_s(X_s^{t,x,\phi}, \phi_s(X_s^{t,x,\phi}), Y_s^{t,x,\phi}, Z_s^{t,x,\phi})\,\mathrm{d}s + Z_s^{t,x,\phi}\,\mathrm{d}W_s,\\
X_t^{t,x,\phi} &= x, \quad Y_T^{t,x,\phi} = e^{-\rho T}\partial_x g(X_T^{t,x,\phi}).
\end{aligned}$$

- $(Y_t^{t,x,\phi^m})_{(t,x)\in[0,T]\times\mathbb{R}^n}$ is the Markovian representation of adjoint process Y^{α^m}.
- Lipschitz regularity of $x \mapsto Y_t^{t,x,\phi^m}$ depends on the Lipschitz regularity of ϕ^m, but not $\partial_x\phi^m$.

Technical conditions

- ℓ is lower semicontinuous and the action set $\mathbf{A} := \{z \in \mathbb{R}^k \,|\, \ell(z) < \infty\}$ is nonempty,

Technical conditions

- ℓ is lower semicontinuous and the action set $\mathbf{A} := \{z \in \mathbb{R}^k \,|\, \ell(z) < \infty\}$ is nonempty,
- $\exists \mu, \nu \geq 0$ s.t. $\mu + \nu > 0$ and f is (μ-strongly) convex in control, ℓ is (ν-strongly) convex in control,

Technical conditions

- ℓ is lower semicontinuous and the action set $\mathbf{A} := \{z \in \mathbb{R}^k \,|\, \ell(z) < \infty\}$ is nonempty,
- $\exists \mu, \nu \geq 0$ s.t. $\mu + \nu > 0$ and f is (μ-strongly) convex in control, ℓ is (ν-strongly) convex in control,
- $b_t(x, a) = \hat{b}_t(x) + \bar{b}_t(x)a$.
- and several regularity conditions on σ, f, g, $\hat{b}$ and $\bar{b}$, e.g.

$$\langle x - x', \hat{b}_t(x) - \hat{b}_t(x') \rangle \leq \kappa_{\hat{b}} |x - x'|^2.$$

Policy space $\mathcal{V}_{\mathbf{A}}$

- $\mathcal{V}_{\mathbf{A}}$ contains all Borel functions $\phi : [0, T] \times \mathbb{R}^n \to \mathbf{A}$ that are Lipschitz continuous and linearly growth in x.

Theorem

For all $\phi^0 \in \mathcal{V}_{\mathbf{A}}$ and $\tau > 0$, the iterates $(\phi^m)_{m \in \mathbb{N}}$ are well-defined and in $\mathcal{V}_{\mathbf{A}}$.

Convergence

Assume *one* of the following

Theorem

If [....], then for all $\phi^0 \in \mathcal{V}_{\boldsymbol{A}}$ *and small* $\tau > 0$, *there exists* $\phi^\star \in \mathcal{V}_{\boldsymbol{A}}$ *and* $c \in [0, 1)$ *such that*

- $\alpha^{\phi^\star}$ *is a stationary point of* J;
- $|\phi^{m+1} - \phi^\star|_0 \leq c|\phi^m - \phi^\star|_0$ *and* $\|\alpha^{\phi^m} - \alpha^{\phi^\star}\|_{\mathcal{H}^2} \leq \mathcal{O}(c^m)$ *for all* m.

Convergence

Assume *one* of the following

- Time horizon T is small.
- Discount factor ρ is large.
- Running cost is sufficiently convex in control, i.e., $\mu + \nu$ is sufficiently large.
- Costs depend weakly on state.
- Control affects state dynamics weakly.
- State dynamics is strongly dissipative, i.e., $\kappa_{\hat{b}}$ is sufficiently negative.

Theorem

If [....], then for all $\phi^0 \in \mathcal{V}_{\boldsymbol{A}}$ and small $\tau > 0$, there exists $\phi^\star \in \mathcal{V}_{\boldsymbol{A}}$ and $c \in [0, 1)$ such that

- $\alpha^{\phi^\star}$ *is a stationary point of* J*;*
- $|\phi^{m+1} - \phi^\star|_0 \leq c|\phi^m - \phi^\star|_0$ *and* $\|\alpha^{\phi^m} - \alpha^{\phi^\star}\|_{\mathcal{H}^2} \leq \mathcal{O}(c^m)$ *for all* m*.*

- Apply PGM for subintervals: Coache, Jaimungal (2021).
- Fictitious discount factor regularisation: Guo, Hu, Zhang (2021).
- Convexify by entropy: Siska, Szpruch (2020), Kerimkulov, Leahy, Siska, Szpruch (2022).
- Dissipativity and controllability: Hu (2019).

Conclusions

Theoretically, gradient iterations over feedback controls are

- linearly convergent for nonconvex, nonsmooth running cost;
- stable to numerical perturbations.

In practice,

- hybrid method using
  - PDEs for adjoint variables (value function and gradient),
  - and particle simulation for mean-field problems.
- Improved interpretability and robust to perturbations.

Reisinger, Stockinger, Zhang (2021),
A fast iterative PDE-based algorithm for feedback controls of nonsmooth mean-field control problems, *arXiv:2108.06740.*

Reisinger, Stockinger, Zhang (2022),
Linear convergence of a policy gradient method for finite horizon continuous time stochastic control problems, *arXiv:2203.11758.*

- Given the feedback control $\widetilde{\phi}^m$ at the m-th iteration;
- a function $\widetilde{\mathcal{Y}}^{\widetilde{\phi}^m} : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ approximating the solution map $[0,T] \times \mathbb{R}^n \ni (t,x) \mapsto \mathcal{Y}_t^{\widetilde{\phi}^m}(x) := Y_t^{t,x,\widetilde{\phi}^m} \in \mathbb{R}^n$.
- Then perform an approximate proximal gradient update

$$\widetilde{\phi}_t^{m+1}(x) = \mathrm{prox}_{\tau\ell}(\widetilde{\phi}_t^m(x) - \tau e^{\rho t} \partial_a H_t^{\mathrm{re}}(x, \widetilde{\phi}_t^m(x), \widetilde{\mathcal{Y}}_t^{\widetilde{\phi}^m}(x))).$$

Theorem

In the set-up from earlier, there exist $c \in [0,1)$ and $C \geq 0$ s.t.

$$|\widetilde{\phi}^m - \phi^\star|_0 \leq c^m |\phi^0 - \phi^\star|_0 + C \sum_{j=0}^{m-1} c^{m-1-j} |\mathcal{Y}^{\widetilde{\phi}^j} - \widetilde{\mathcal{Y}}^{\widetilde{\phi}^j}|_0.$$