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 techniques.
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.
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).
  - LQ with linear policy: Fazel, Ge, Kakade, Mesbahi (2018); Hambly, Xu, Yang (2021).

- Trapped at local minimum:

Continuous-time:

Open: convergence behaviour of PGMs for general models/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).
  - LQ with linear policy: Fazel, Ge, Kakade, Mesbahi (2018); Hambly, Xu, Yang (2021).

- Trapped at local minimum:
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:

Continuous-time:


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

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

over all square integrable, adapted processes \( \alpha \), where \( X^\alpha \) satisfies

\[ dX_t = b_t(X_t, \alpha_t) \, dt + \sigma_t(X_t) \, dW_t, \quad X_0 = \xi_0. \]

- \( f, g, b, \sigma \) are differentiable.
- \( \ell \) is possibly non-smooth and infinite.

\( \ell \) represents control constraints, \( \ell_1 \)-norm or entropy regularisers.
A “naive” gradient direction
Special case with $\ell = \rho = 0$, $\sigma_t(x) := \sigma$

Minimise

$$J(\phi) = \mathbb{E} \left[ \int_0^T f_t(X^\phi_t, \phi_t(X^\phi_t)) \, dt + g(X^\phi_T) \right]$$

over all feedback controls $\phi$, where $X^\phi$ satisfies

$$dX_t = b_t(X_t, \phi_t(X_t)) \, dt + \sigma \, dW_t, \quad X_0 = \xi_0.$$
A “naive” gradient direction

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

Minimise

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

over all feedback controls $\phi$, where $X^\phi$ satisfies

$$dX_t = b_t(X_t, \phi_t(X_t)) \, dt + \sigma \, dW_t, \quad X_0 = \xi_0.$$

For each policy $\phi$ and test policy $\psi$,

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

where $H_t^{re}(x, a, y) := \langle b_t(x, a), y \rangle + f_t(x, a)$, and $u^\phi$ satisfies

$$\partial_t u_t(x) + \frac{1}{2} \sigma^2 \partial_{xx} u_t(x) + H_t^{re}(x, \phi_t(x), \partial_x u_t(x)) = 0, \quad u_T(x) = g(x).$$
A "naive" PGM

\[ H^\text{re}_t(x, a, y) := \langle b_t(x, a), y \rangle + f_t(x, a) \]

- Given \( \phi^0 \), perform gradient descent steps

\[ \phi^{m+1}_t(x) = \phi^m_t(x) - \tau \partial_a H^\text{re}_t(x, \phi^m_t(x), \partial_x u^{\phi^m}_t(x)), \]

where for each \( \phi \), \( u^\phi \) satisfies

\[ \partial_t u_t(x) + \frac{1}{2} \sigma^2 \partial_{xx} u_t(x) + H^\text{re}_t(x, \phi_t(x), \partial_x u_t(x)) = 0, \quad u_T(x) = g(x). \]
A “naive” PGM

\[ H^\text{re}_t(x, a, y) := \langle b_t(x, a), y \rangle + f_t(x, a) \]

- Given \( \phi^0 \), perform gradient descent steps

\[ \phi_t^{m+1}(x) = \phi_t^m(x) - \tau \partial_a H^\text{re}_t(x, \phi_t^m(x), \partial_x u_t^{\phi^m}(x)), \]

where for each \( \phi \), \( u^\phi \) satisfies

\[ \partial_t u_t(x) + \frac{1}{2} \sigma^2 \partial_{xx} u_t(x) + H^\text{re}_t(x, \phi_t(x), \partial_x u_t(x)) = 0, \quad u_T(x) = g(x). \]

- If \( \partial_a H^\text{re}_t(x, \phi^*_t(x), \partial_x u_t^{\phi^*}(x)) = 0 \), then \( \phi^* \) is optimal.
A “naive” PGM

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

Given \( \phi^0 \), perform gradient descent steps

\[ \phi_{m+1}^t(x) = \phi_m^t(x) - \tau \partial_a H_t^{re}(x, \phi_m^t(x), \partial_x u_{m}^t(x)) \]

where for each \( \phi \), \( u^\phi \) satisfies

\[ \partial_t u_t(x) + \frac{1}{2} \sigma^2 \partial_{xx} u_t(x) + H_t^{re}(x, \phi_t(x), \partial_x u_t(x)) = 0, \quad u_T(x) = g(x). \]

- If \( \partial_a H_t^{re}(x, \phi^*_t(x), \partial_x u_{\phi^*}^t(x)) = 0 \), then \( \phi^* \) is optimal.
- \( \phi^{m+1} \) has lower regularity than \( \phi^m \), as \( \partial_x u_{\phi^m} \) has the same regularity as \( \partial_x \phi^m \).
A “naive” PGM

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

- Given \( \phi^0 \), perform gradient descent steps

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

where for each \( \phi \), \( u^\phi \) satisfies

\[
\partial_t u_t(x) + \frac{1}{2} \sigma^2 \partial_{xx} u_t(x) + H_t^{re}(x, \phi_t(x), \partial_x u_t(x)) = 0, \quad u_T(x) = g(x).
\]

- If \( \partial_a H_t^{re}(x, \phi^*_t(x), \partial_x u_t^{\phi^*}(x)) = 0 \), then \( \phi^* \) is optimal.

- \( \phi^{m+1} \) has lower regularity than \( \phi^m \), 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

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

where \( \mathcal{L}^{\phi^m} \) is the generator of \( X^{\phi^m} \).
Minimise

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

over all admissible control processes \( \alpha \), where \( X^\alpha \) satisfies

\[ dX_t = b_t(X_t, \alpha_t) \, dt + \sigma_t(X_t) \, dW_t, \quad X_0 = \xi_0. \]

where

\[ \begin{align*}
& \triangleright f, g, b, \sigma \text{ are differentiable,} \\
& \triangleright \ell \text{ is convex, possibly non-smooth and infinite.}
\end{align*} \]
Stochastic maximum principle

Smooth case: \( \ell = 0 \)

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

\[
dY_t = - \partial_x H_t(X^\alpha_t, \alpha_t, Y_t, Z_t) \, dt + Z_t \, dW_t, \quad Y_T = e^{-\rho T} \partial_x g(X^\alpha_T),
\]

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 \( \alpha \):

\[
\nabla J(\alpha)_t = \partial_a H_t(X^\alpha_t, \alpha_t, Y^\alpha_t, Z^\alpha_t).
\]
Adjoint processes \((Y^\alpha, Z^\alpha)\) for a control \(\alpha\):

\[
dY_t = - \partial_x H_t(X^\alpha_t, \alpha_t, Y_t, Z_t) \, dt + Z_t \, dW_t, \quad Y_T = e^{-\rho T} \partial_x g(X^\alpha_T),
\]

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 \(\alpha\):

\[
\nabla J(\alpha)_t = \partial_a H_t(X^\alpha_t, \alpha_t, Y^\alpha_t, Z^\alpha_t).
\]

\(\alpha\) is a **stationary point** of \(J\) if \(\partial_a H^\text{re}_t(X^\alpha_t, \alpha_t, Y^\alpha_t) = 0\), with

\[
H^\text{re}_t(x, a, y) := \langle b_t(x, a), y \rangle + e^{-\rho t} f_t(x, a).
\]
Open-loop (proximal) gradient descent

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

$$\alpha_{t}^{m+1} = \alpha_{t}^{m} - \tau e^{\rho t} \partial_a H_t^{re}(X^\xi_0, \alpha^m, \alpha_t^m, Y^\xi_0, \alpha^m).$$
Open-loop (proximal) gradient descent

- Smooth case with $\ell = 0$: perform gradient steps
  \[ \alpha_{t+1}^m = \alpha_t^m - \tau e^{\rho t} \partial_a H_t^{re}(X_t^{\xi_0,\alpha_t^m}, \alpha_t^m, Y_t^{\xi_0,\alpha_t^m}). \]

- Nonsmooth case: define the proximal map
  \[ \text{prox}_{\tau \ell}(a) = \arg \min_{z \in \mathbb{R}^k} \left( \frac{1}{2} |z - a|^2 + \tau \ell(z) \right), \quad a \in \mathbb{R}^k, \]
  and perform proximal gradient steps:
  \[ \alpha_{t+1}^m = \text{prox}_{\tau \ell}(\alpha_t^m - \tau e^{\rho t} \partial_a H_t^{re}(X_t^{\xi_0,\alpha_t^m}, \alpha_t^m, Y_t^{\xi_0,\alpha_t^m})). \]
**Proximal policy gradient method**

\[
\prox_{\tau \ell}(a) = \arg \min_{z \in \mathbb{R}^k} \left( \frac{1}{2} |z - a|^2 + \tau \ell(z) \right)
\]

Given \( \phi^0 \), perform proximal gradient steps

\[
\phi^{m+1}_t(x) = \prox_{\tau \ell} \left( \phi^m_t(x) - \tau e^{\rho_t} \partial_a H^r_t(x, \phi^m_t(x), Y^t, x, \phi^m) \right),
\]

where for each \( \phi \),

\[
\begin{align*}
    dX^{t, x, \phi}_s &= b_s(X^{t, x, \phi}_s, \phi_s(X^{t, x, \phi}_s)) \, ds + \sigma_s(X^{t, x, \phi}_s) \, dW_s, \\
    dY^{t, x, \phi}_s &= -\partial_x H_s(X^{t, x, \phi}_s, \phi_s(X^{t, x, \phi}_s), Y^{t, x, \phi}_s, Z^{t, x, \phi}_s) \, ds + Z^{t, x, \phi}_s \, dW_s, \\
    X^{t, x, \phi}_t &= x, \quad Y^{t, x, \phi}_T = e^{-\rho T} \partial_x g(X^{t, x, \phi}_T).
\end{align*}
\]

\( (Y^{t, x, \phi^m}_t)_{t, x} \in [0, T] \times \mathbb{R}^n \) is the Markovian representation of adjoint process \( Y^{\alpha^m} \).
Proximal policy gradient method

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

Given \( \phi^0 \), perform proximal gradient steps

\[ \phi_{t}^{m+1}(x) = \text{prox}_{\tau \ell}(\phi_{t}^{m}(x) - \tau e^{\rho t} \partial_a H_{t}^\text{re}(x, \phi_{t}^{m}(x), Y_{t}^{t,x,\phi^m})) , \]

where for each \( \phi \),

\[
\begin{align*}
\text{d}X_{s}^{t,x,\phi} &= b_s(X_{s}^{t,x,\phi}, \phi_s(X_{s}^{t,x,\phi})) \, \text{d}s + \sigma_s(X_{s}^{t,x,\phi}) \, \text{d}W_s , \\
\text{d}Y_{s}^{t,x,\phi} &= -\partial_x H_s(X_{s}^{t,x,\phi}, \phi_s(X_{s}^{t,x,\phi}), Y_{s}^{t,x,\phi}, Z_{s}^{t,x,\phi}) \, \text{d}s + Z_{s}^{t,x,\phi} \, \text{d}W_s , \\
X_{t}^{t,x,\phi} &= x , \quad Y_{T}^{t,x,\phi} = e^{-\rho T} \partial_x g(X_{T}^{t,x,\phi}) .
\end{align*}
\]

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

Lipschitz regularity of \( x \mapsto Y_{t}^{t,x,\phi^m} \) depends on the Lipschitz regularity of \( \phi^m \), but not \( \partial_x \phi^m \).
Technical conditions

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

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

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

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

**Theorem**

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

Assume one of the following

- Time horizon $T$ is small.
- Discount factor $\rho$ 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 $\phi^0 \in \mathcal{V}_A$ and small $\tau > 0$, there exists $\phi^* \in \mathcal{V}_A$ and $c \in [0, 1)$ such that

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

29 June 2022 From stochastic control to continuous-time RL
Convergence

Assume one of the following:

- Time horizon $T$ is small.
- Discount factor $\rho$ 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., $\kappa_b$ is sufficiently negative.

Theorem

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

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

29 June 2022 From stochastic control to continuous-time RL
Convergence

Practical implication

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),

Reisinger, Stockinger, Zhang (2022),
Stability under numerical approximation of gradient

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

$$
\tilde{\phi}^{m+1}_t(x) = \text{prox}_{\tau \ell}(\tilde{\phi}^m_t(x) - \tau e^{\rho t} \partial a H^re_t(x, \tilde{\phi}^m_t(x), \tilde{Y}^m_t(x))).
$$

Theorem

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

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