MF-PhiBE for LQR Mean-Field Control from Discrete-Time Data

Martin Hernandez

UCLA, Department of Statistics & Data Science

Joint work with Yuhua Zhu, Qinxin Yan, and Erhan Bayraktar

martinh@g.ucla.edu

Foundations of Multi-Agent and Mean Field Reinforcement Learning

May 2026

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s^+_\eta, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r^\pi_\lambda(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r^\pi_\lambda(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

LQR benchmark.

$$r(s,\mu,a) = -\left(s^\top Q s + \bar{\mu}^\top \bar{Q} \bar{\mu} + a^\top N a\right)$$

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

LQR benchmark.

$$r(s,\mu,a) = -\big(s^\top Q s + \bar{\mu}^\top \bar{Q} \bar{\mu} + a^\top N a\big)$$

where $\mu_t = \mathrm{Law}(s_t)$ and

$$ds_t = (As_t + \bar{A}\bar{\mu}_t + Ba_t)\, dt + \gamma\, dW_t, \quad \bar{\mu}_t = \mathbb{E}[s_t].$$

Using Riccati

$$\pi^*(\cdot|s,\mu) = \mathcal{N}\Big(-M^{-1}B^\top(Ks + \bar{K}\bar{\mu}),\ \frac{\gamma}{2}M^{-1}\Big)$$

where K and $\bar{K}$ solve a Riccati equation independent of γ.

Question

The deterministic problem ($\gamma = 0$) defines the same optimal policy

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

LQR benchmark.

$$r(s,\mu,a) = -\big(s^\top Q s + \bar\mu^\top \bar Q \bar\mu + a^\top N a\big)$$

where $\mu_t = \mathrm{Law}(s_t)$ and

$$ds_t = (As_t + \bar A \bar\mu_t + Ba_t)\, dt + \gamma\, dW_t, \quad \bar\mu_t = \mathbb{E}[s_t].$$

Using Riccati.

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

LQR benchmark.

$$r(s,\mu,a) = -\big(s^\top Q s + \bar\mu^\top \bar Q \bar\mu + a^\top N a\big)$$

where $\mu_t = \mathrm{Law}(s_t)$ and

$$ds_t = (As_t + \bar A \bar\mu_t + Ba_t)\, dt + \gamma\, dW_t, \quad \bar\mu_t = \mathbb{E}[s_t].$$

Using Riccati.

$$\pi^*(\cdot|s,\mu) = \mathcal{N}\left(-N^{-1}B^\top(K(s-\bar\mu) + \Lambda\bar\mu), \frac{\lambda}{2}N^{-1}\right)$$

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

LQR benchmark.

$$r(s,\mu,a) = -\big(s^\top Q s + \bar\mu^\top \bar Q \bar\mu + a^\top N a\big)$$

where $\mu_t = \mathrm{Law}(s_t)$ and

$$ds_t = (As_t + \bar A \bar\mu_t + Ba_t)\, dt + \gamma\, dW_t, \quad \bar\mu_t = \mathbb{E}[s_t].$$

Using Riccati.

$$\pi^*(\cdot|s,\mu) = \mathcal{N}\left(-N^{-1}B^\top\big(K(s-\bar\mu) + \Lambda\bar\mu\big), \frac{\lambda}{2}N^{-1}\right)$$

where K and Λ solve a Riccati equation independent of γ.

Data and problem setting

Observed data at step Δt.

$$\mathcal{D} = \{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}.$$

Question. Can one use one-step data while keeping the continuous-time control problem, rather than replacing it by a discrete-time Bellman equation?

Consider the value function

$$\mathcal{V}^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[\int_0^\infty e^{-\beta t} r_\lambda^\pi(s, \mu_t)\, dt \,\middle|\, \mu_0 = \mu\right],$$

where $r_\lambda^\pi(s,\mu) = \mathbb{E}_\pi[r(s,\mu,a) - \lambda \log p_\pi]$
and denote by

$$\mathcal{V}^*(\mu) = \sup_\pi \mathcal{V}^\pi(\mu),$$

where $\pi(da \mid s, \mu)$ is a randomized feedback policy.

LQR benchmark.

$$r(s,\mu,a) = -\big(s^\top Q s + \bar{\mu}^\top \bar{Q}\bar{\mu} + a^\top N a\big)$$

where $\mu_t = \mathrm{Law}(s_t)$ and

$$ds_t = (As_t + \bar{A}\bar{\mu}_t + Ba_t)\, dt + \gamma\, dW_t, \quad \bar{\mu}_t = \mathbb{E}[s_t].$$

Using Riccati.

$$\pi^*(\cdot|s,\mu) = \mathcal{N}\left(-N^{-1}B^\top(K(s-\bar{\mu}) + \Lambda\bar{\mu}), \frac{\lambda}{2}N^{-1}\right)$$

where K and Λ solve a Riccati equation independent of γ.

Key Message

The deterministic problem ($\gamma = 0$) defines the same optimal policy.

MF-PhiBE and LQR policy approximation

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s,\mu,a) = \mathbb{E}\left[\frac{s' - s}{\Delta t}\middle| s,\mu,a\right]$$

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s,\mu,a) := \mathbb{E}\left[\frac{s^{+}-s}{\Delta t}\,\middle|\, s,\mu,a\right].$$

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s,\mu,a) := \mathbb{E}\left[\frac{s^{+}-s}{\Delta t}\,\middle|\, s,\mu,a\right].$$

LQR structure. For the LQR benchmark,

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s, \mu, a) := \mathbb{E}\left[\frac{s^{+} - s}{\Delta t} \,\middle|\, s, \mu, a\right].$$

LQR structure. For the LQR benchmark,

$$\hat{b}(s, \mu, a) = \hat{A}s + \hat{\bar{A}}\bar{\mu} + \hat{B}a.$$

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s, \mu, a) := \mathbb{E}\left[\frac{s^{+} - s}{\Delta t} \,\middle|\, s, \mu, a\right].$$

LQR structure. For the LQR benchmark,

$$\hat{b}(s, \mu, a) = \hat{A}s + \hat{\bar{A}}\bar{\mu} + \hat{B}a.$$

Since $\gamma = 0$ defines the same policy, MF-PhiBE

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s, \mu, a) := \mathbb{E}\left[\frac{s^{+} - s}{\Delta t} \,\middle|\, s, \mu, a\right].$$

LQR structure. For the LQR benchmark,

$$\hat{b}(s, \mu, a) = \hat{A}s + \hat{\bar{A}}\bar{\mu} + \hat{B}a.$$

Since $\gamma = 0$ defines the same policy, MF-PhiBE

$$(\mathcal{L}^{\pi}_{\hat{b},0} - \beta)\hat{\mathcal{V}}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0,$$

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s,\mu,a) := \mathbb{E}\left[\frac{s^{+}-s}{\Delta t}\,\middle|\, s,\mu,a\right].$$

LQR structure. For the LQR benchmark,

$$\hat{b}(s,\mu,a) = \hat{A}s + \hat{\bar{A}}\bar{\mu} + \hat{B}a.$$

Since $\gamma = 0$ defines the same policy, MF-PhiBE

$$(\mathcal{L}^{\pi}_{\hat{b},0} - \beta)\hat{\mathcal{V}}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0,$$

and

$$\hat{\mathcal{V}}^{*}(\mu) = \sup_{\pi} \hat{\mathcal{V}}^{\pi}(\mu).$$

MF-PhiBE and LQR policy approximation

Policy evaluation. For a fixed policy, the exact continuous-time value solves

$$(\mathcal{L}^{\pi}_{b,\Sigma} - \beta)\mathcal{V}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0.$$

MF-PhiBE keeps the continuous-time equation and replaces the unknown local drift by

$$\hat{b}(s,\mu,a) := \mathbb{E}\left[\frac{s^{+}-s}{\Delta t}\,\middle|\, s,\mu,a\right].$$

LQR structure. For the LQR benchmark,

$$\hat{b}(s,\mu,a) = \hat{A}s + \hat{\bar{A}}\bar{\mu} + \hat{B}a.$$

Since $\gamma = 0$ defines the same policy, MF-PhiBE

$$(\mathcal{L}^{\pi}_{\hat{b},0} - \beta)\hat{\mathcal{V}}^{\pi}(\mu) + r^{\pi}_{\lambda}(\mu) = 0,$$

and

$$\hat{\mathcal{V}}^{*}(\mu) = \sup_{\pi} \hat{\mathcal{V}}^{\pi}(\mu).$$

MF-PhiBE error in LQR

The optimal MF-PhiBE policy is Gaussian, $\hat{\pi}^{*}(\cdot \mid s,\mu) = \mathcal{N}\left(\hat{m}^{*}(s,\mu), \frac{\lambda}{2}N^{-1}\right)$, with

$$\hat{m}^{*}(s,\mu) = -N^{-1}\hat{B}^{\top}\left(\hat{K}(s-\bar{\mu}) + \hat{\Lambda}\bar{\mu}\right), \quad m^{*}(s,\mu) = -N^{-1}B^{\top}\left(K(s-\bar{\mu}) + \Lambda\bar{\mu}\right),$$

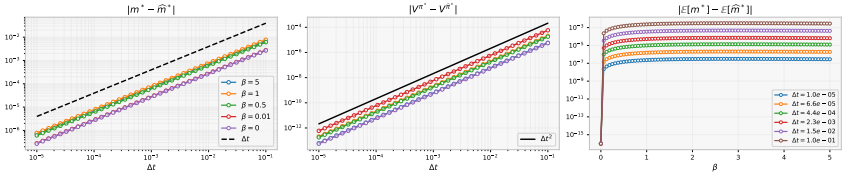
then

$$\|\hat{m}^{*}(s,\mu) - m^{*}(s,\mu)\| \le C_{s,\mu}\Delta t, \qquad |\mathcal{V}^{\pi^{*}} - \mathcal{V}^{\hat{\pi}^{*}}| \le C_{\mu}(\Delta t)^{2}.$$

Moreover, when $\beta = 0$, we have $\mathbb{E}_{s\sim\mu}[m^{*}(s,\mu)] = \mathbb{E}_{s\sim\mu}[\hat{m}^{*}(s,\mu)]$.

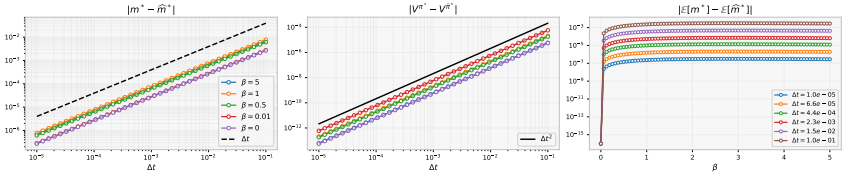
Model-free actor-critic

Convergence order verification:



Model-free actor-critic

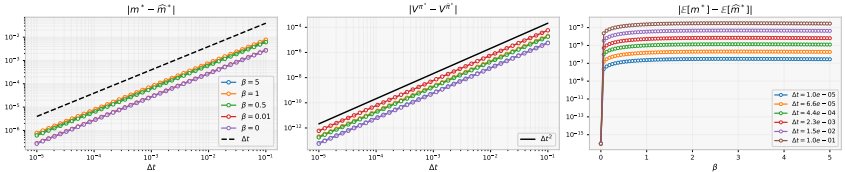
Convergence order verification:



Algorithm for LQR:

Model-free actor-critic

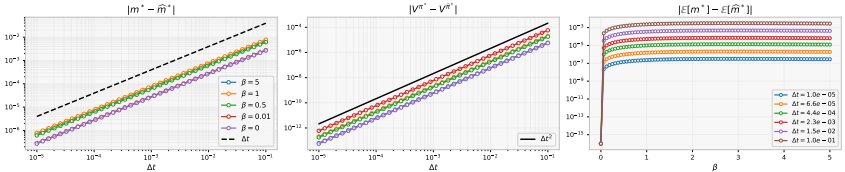
Convergence order verification:



Algorithm for LQR: Let π_ω.

Model-free actor-critic

Convergence order verification:



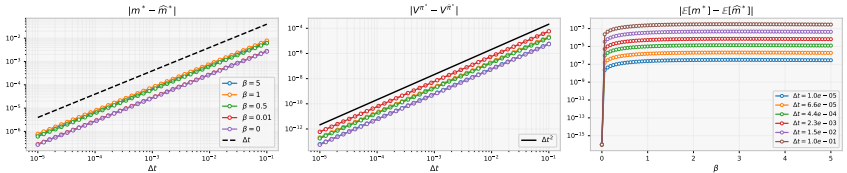
Algorithm for LQR: Let π_ω.

DATA

$$\{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}$$

Model-free actor-critic

Convergence order verification:



Algorithm for LQR: Let π_ω.

DATA

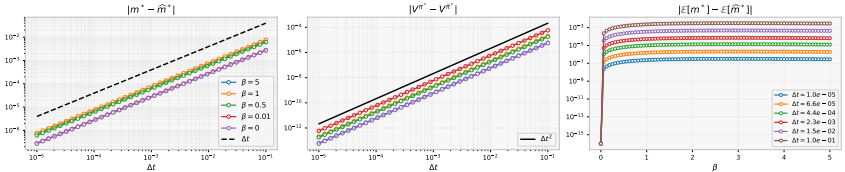
$$\{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}$$

Estimate drift:

$$\hat{b} \sim \tilde{b}_\eta := \frac{s_\eta^+ - s_\eta}{\Delta t}$$

Model-free actor-critic

Convergence order verification:



Algorithm for LQR: Let π_ω.

DATA

$$\{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}$$

CRITIC

Estimate drift:

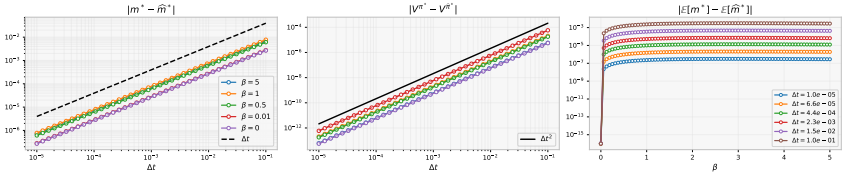
$$\hat{b} \sim \tilde{b}_\eta := \frac{s_\eta^+ - s_\eta}{\Delta t}$$

Solve MF-PhiBE:

$$\left(\tilde{\mathcal{L}}^{\pi_\omega}_{\tilde{b}_\eta, 0} - \beta\right)\hat{V}^{\pi_\omega} + r_{\lambda,\eta} = 0$$

Model-free actor-critic

Convergence order verification:



Algorithm for LQR: Let π_ω.

DATA

$\{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=\mathbf{1}}^{N_D}$

CRITIC

Estimate drift:

$$\hat{b} \sim \tilde{b}_\eta := \frac{s_\eta^+ - s_\eta}{\Delta t}$$

Solve MF-PhiBE:

$$\big(\hat{\mathcal{L}}^{\pi_\omega}_{\tilde{b}_\eta, 0} - \beta\big)\hat{V}^{\pi_\omega} + r_{\lambda,\eta} = 0$$

$\hat{V}^{\pi_\omega}$

ACTOR

Policy gradient:

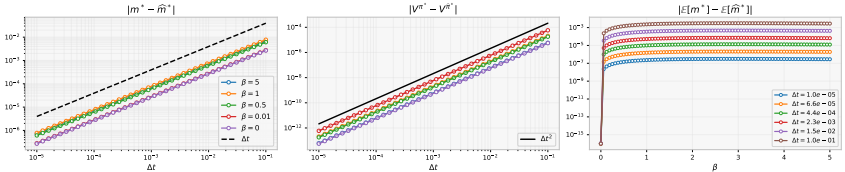
$$\widehat{\nabla J}(\omega) := \textstyle\sum_\eta w_\eta\, \hat{q}_\eta^{\,\omega}\, \nabla_\omega \log p_{\pi_\omega}(a_\eta \mid s_\eta, \mu_\eta)$$

$$\hat{q}_\eta^{\,\omega} := r_{\lambda,\eta} + \tilde{b}_\eta \cdot \partial_\mu \hat{V}^{\pi_\omega}(\mu_\eta)(s_\eta)$$

$$\omega \leftarrow \omega + \alpha\, \widehat{\nabla J}(\omega)$$

Model-free actor-critic

Convergence order verification:



Algorithm for LQR: Let π_ω.

DATA

$\{(s_\eta, \mu_\eta, a_\eta, s_\eta^+, r_\eta)\}_{\eta=1}^{N_D}$

CRITIC

Estimate drift:

$$\hat{b} \sim \tilde{b}_\eta := \frac{s_\eta^+ - s_\eta}{\Delta t}$$

Solve MF-PhiBE:

$$\left(\hat{\mathcal{L}}^{\pi_\omega}_{\tilde{b}_\eta, 0} - \beta\right)\hat{V}^{\pi_\omega} + r_{\lambda,\eta} = 0$$

$\hat{V}^{\pi_\omega}$

ACTOR

Policy gradient:

$$\widehat{\nabla J}(\omega) := \sum_\eta w_\eta \, \hat{q}_\eta^{\,\omega} \, \nabla_\omega \log p_{\pi_\omega}(a_\eta \mid s_\eta, \mu_\eta)$$

$$\hat{q}_\eta^{\,\omega} := r_{\lambda,\eta} + \tilde{b}_\eta \cdot \partial_\mu \hat{V}^{\pi_\omega}(\mu_\eta)(s_\eta)$$

$$\omega \leftarrow \omega + \alpha \, \widehat{\nabla J}(\omega)$$

π_ω

Numerical implementation

- The value function and the policy are parameterized:

$$\pi_\omega(t|s,\mu) = \mathcal{N}(\omega_0(s-\bar{p}) + \omega_1\bar{p}, \tfrac{1}{2}N^{-1})$$

$$V_\theta(\mu) = -(\theta_0 + \theta_1\bar{p} + \theta_2 \mathrm{Var}(\mu))$$

- We solve PHJBE with Galerkin with cylindrical basis functions
- Only the first and second moments of μ are needed.

Beyond vanilla gradient descent: Since we have access to ∇J, we can implement different gradient-based optimization methods.

Numerical implementation

- The value function and the policy are parameterized:

$$\pi_{\theta_\pi}(t|s,\mu) = \mathcal{N}(\omega_0 t(s-\bar{p}) + \omega_1 \bar{p}, \frac{2}{3}\theta^{(2)})$$

$$V_{\theta_V}(\mu) = -(\theta_0 + \theta_1 \bar{p} + \theta_2 \mathrm{Var}(\mu))$$

- We solve PhiBE with Galerkin with cylindrical basis functions
- Only the first and second moments of μ_t are needed.

Beyond vanilla gradient descent: Since we have access to ∇J, we can implement different gradient based optimization methods.

Numerical implementation

- The value function and the policy are parameterized:

$$\pi_\omega(\cdot|s,\mu) = \mathcal{N}\big(\omega_1(s-\bar{\mu}) + \omega_2\bar{\mu},\, \tfrac{\lambda}{2}N^{-1}\big)$$

$$V_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2 \mathrm{Var}(\mu))$$

- We solve PhiBE with Galerkin with cylindrical basis functions
- Only the first and second moments of μ are needed.

Beyond vanilla gradient descent: Since we have access to ∇J, we can implement different gradient based optimization methods.

Numerical implementation

- The value function and the policy are parameterized:

$$\pi_\omega(\cdot|s,\mu) = \mathcal{N}\big(\omega_1(s-\bar{\mu})+\omega_2\bar{\mu},\, \tfrac{\lambda}{2}N^{-1}\big)$$

$$\mathcal{V}_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2\,\mathrm{Var}(\mu))$$

- We solve PhiBE with Galerkin with cylindrical basis functions
- Only the first and second moments of μ are needed.

Beyond vanilla gradient descent: Since we have access to ∇J, we can implement different gradient based optimization methods.

Numerical implementation

- The value function and the policy are parameterized:

  $\pi_\omega(\cdot|s,\mu) = \mathcal{N}\big(\omega_1(s-\bar{\mu})+\omega_2\bar{\mu}, \frac{\lambda}{2}N^{-1}\big)$

  $\mathcal{V}_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2\,\mathrm{Var}(\mu))$

- We solve PhiBE with Galerkin with cylindrical basis functions.
- Only the first and second moments of μ are needed.

Beyond vanilla gradient descent: Since we have access to $\nabla\mathcal{J}$, we can implement different gradient based optimization methods.

Numerical implementation

- The value function and the policy are parameterized:

  $\pi_\omega(\cdot|s,\mu) = \mathcal{N}\big(\omega_1(s-\bar{\mu})+\omega_2\bar{\mu}, \frac{\lambda}{2}N^{-1}\big)$

  $\mathcal{V}_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2\,\mathrm{Var}(\mu))$

- We solve PhiBE with Galerkin with cylindrical basis functions.
- **Only the first and second moments of μ are needed.**

Beyond vanilla gradient descent: Since we have access to $\nabla\mathcal{J}$, we can implement different gradient based optimization methods.

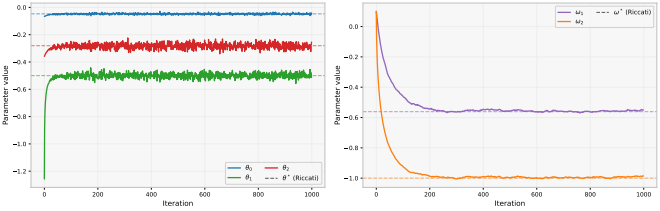
Numerical implementation

- The value function and the policy are parameterized:

$$\pi_\omega(\cdot|s,\mu) = \mathcal{N}\big(\omega_1(s-\bar{\mu})+\omega_2\bar{\mu}, \tfrac{\lambda}{2}N^{-1}\big)$$

$$\mathcal{V}_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2\,\mathrm{Var}(\mu))$$

- We solve PhiBE with Galerkin with cylindrical basis functions.
- **Only the first and second moments of μ are needed.**



Beyond vanilla gradient descent: Since we have access to ∇J, we can implement different gradient based optimization methods.

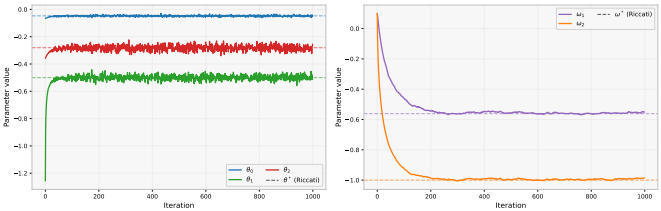
Numerical implementation

- The value function and the policy are parameterized:

$$\pi_\omega(\cdot|s,\mu) = \mathcal{N}\big(\omega_1(s-\bar{\mu})+\omega_2\bar{\mu}, \tfrac{\lambda}{2}N^{-1}\big)$$

$$\mathcal{V}_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2\,\mathrm{Var}(\mu))$$

- We solve PhiBE with Galerkin with cylindrical basis functions.
- **Only the first and second moments of μ are needed.**



Beyond vanilla gradient descent: Since we have access to $\widehat{\nabla J}$, we can implement different gradient-based optimization methods.

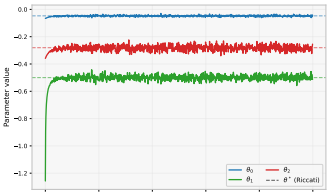
Numerical implementation

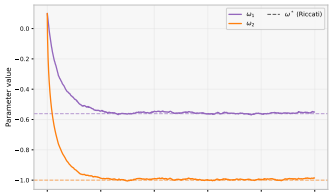
- The value function and the policy are parameterized:

$$\pi_\omega(\cdot|s,\mu) = \mathcal{N}(\omega_1(s-\bar{\mu})+\omega_2\bar{\mu}, \tfrac{\lambda}{2}N^{-1})$$

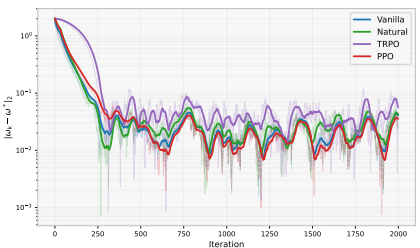
$$\mathcal{V}_\theta(\mu) = -(\theta_0 + \theta_1\bar{\mu} + \theta_2\,\mathrm{Var}(\mu))$$

- We solve PhiBE with Galerkin with cylindrical basis functions.
- **Only the first and second moments of μ are needed.**

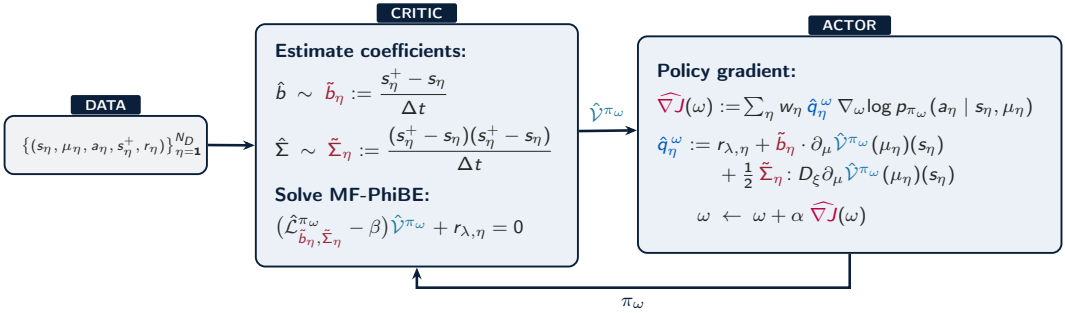




Beyond vanilla gradient descent: Since we have access to $\widehat{\nabla J}$, we can implement different gradient-based optimization methods.

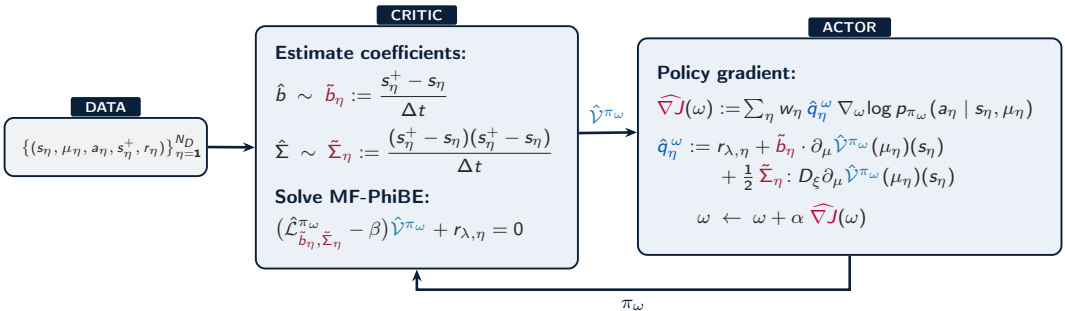


Model-free actor-critic: general setting



Thanks for your attention!

Model-free actor-critic: general setting



Thanks for your attention!

Modeling Stochastic Multi-Agent Interaction in Intraday Battery Energy Storage Dispatch with Market Power

Hezhong Zhang

Department of Statistics and Applied Probability
University of California, Santa Barbara

May 2026

https://arxiv.org/pdf/2605.01178

Joint work with Ruimeng Hu and Mike Ludkovski

Stochastic Differential Game Model (N agents)

Exogenous supply dynamics

$$dQ_t = \kappa(\theta(t) - Q_t)\,dt + \sigma^0 dW_t^0 \tag{1}$$

Battery State of Charge dynamics

$$dS_t^i = (\mu_t^i + \alpha_t^i)\,dt + \sigma^i \rho^i dW_t^0 + \sigma^i \sqrt{1-(\rho^i)^2}\,d\tilde{W}_t^i, \tag{2}$$

where $\mu_t^i = a^i(t) + b^i(t)Q_t$ is hybrid asset self generation, and $\alpha_t^i = \alpha^i(t, Q_t, \mathbf{S}_t)$, charges ($\alpha_t^i > 0$) and discharges ($\alpha_t^i < 0$).

Price

$$P_t^i = \underbrace{\bar{P}^i}_{\text{Base price}} - c_1^i \underbrace{\left(Q_t - \mathbf{w}_i^\top \boldsymbol{\alpha}_t\right)}_{\text{Net Supply}} \tag{3}$$

Optimization problem (Linear Quadratic (LQ) Structure)

$$\hat{\alpha}^i \in \arg\min_{\alpha^i} J^i(\alpha^i, \hat{\boldsymbol{\alpha}}^{-i}),$$

$$J^i(\boldsymbol{\alpha}) = \mathbb{E}\Big[\int_0^T \underbrace{P_t^i \alpha_t^i + c_2^i(\alpha_t^i)^2 + c_3^i(S_t^i - \zeta^i(t))^2}_{f^i(t, Q_t, \mathbf{S}_t, \alpha_t^i, \boldsymbol{\alpha}_t^{-i})}\,dt + \underbrace{c_4^i(S_T^i - \zeta^i(T))^2}_{g^i(Q_t, \mathbf{S}_t)}\Big] \tag{4}$$

Main Technical Contribution

Markovian LQ Differential Game $\Longrightarrow$ Best Response Functions $\Longrightarrow$

Coupled HJB System $\Longrightarrow$ LQ Ansatz $\Longrightarrow$ Coupled Riccati ODEs

- The best response function V^i for each agent i evaluated at Nash Equilibrium $\hat{\boldsymbol{\alpha}}$ satisfies

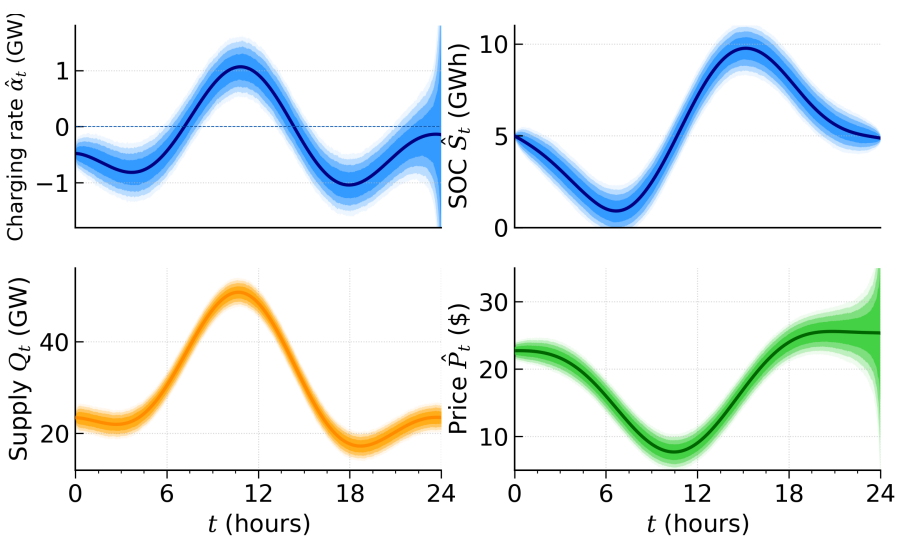
$$\partial_t V^i(t,q,\mathbf{s}) + \inf_{\alpha^i}\left\{ f^i(t,q,\mathbf{s},\alpha^i,\hat{\boldsymbol{\alpha}}^{-i}) + \mathcal{L}^{\alpha^i,\hat{\boldsymbol{\alpha}}^{-i}} V^i(t,q,\mathbf{s}) \right\} = 0, \quad (5)$$

  with $V^i(T,q,\mathbf{s}) = g^i(q,\mathbf{s})$. The LQ ansatz is of the form.

$$V^i(t,q,\mathbf{s}) = \begin{bmatrix} q \\ \mathbf{s} \end{bmatrix}^\top \begin{bmatrix} p_0^i(t) & \mathbf{p}^i(t)^\top \\ \mathbf{p}^i(t) & \mathbf{P}^i(t) \end{bmatrix} \begin{bmatrix} q \\ \mathbf{s} \end{bmatrix} + \begin{bmatrix} r_0^i(t) \\ \mathbf{r}^i(t) \end{bmatrix}^\top \begin{bmatrix} q \\ \mathbf{s} \end{bmatrix} + u^i(t). \quad (6)$$

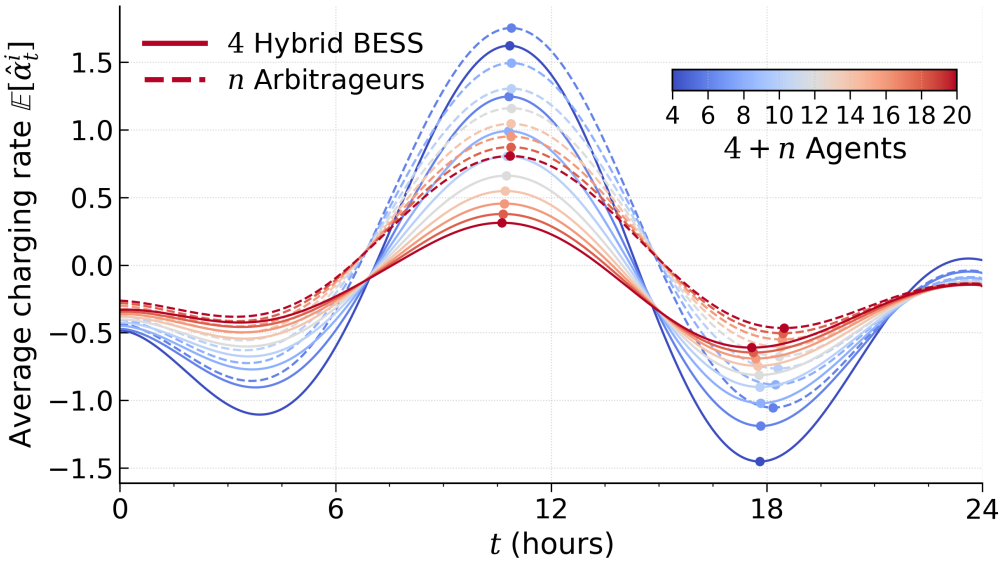
- The heterogeneous setting requires solving $\mathcal{O}(N^3)$ coupled Riccati ODEs, while under homogeneous setting the system collapses to $\mathcal{O}(1)$.

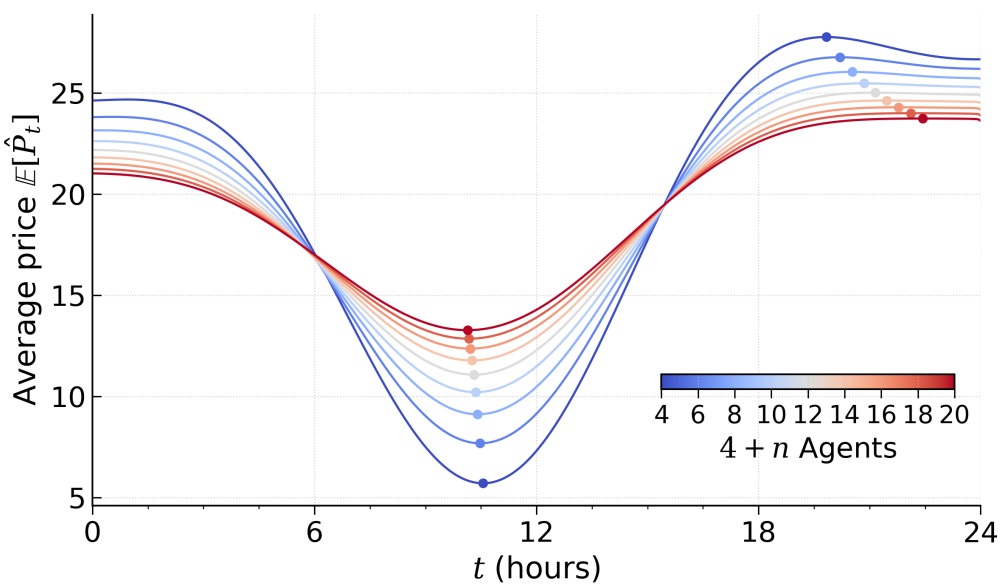
What Does Equilibrium Look Like? (Homogeneous, $N = 8$)



- Batteries charge when prices are low, and discharge during high-price periods
- SOC decreases in the discharging period and increase during charging
- Resulting prices are lowest in the morning and highest during the evening.
- Evening Prices flattened

How Does Competition Reshape Equilibrium?





- More competition reduces individual market power, leading to less charging and discharging
- Gap between the charging rate of Hybrid and Arbitrageurs agents increase as number of agents increase
- Peak shifts down and valley shifts up, meaning price goes up in the charging period and goes down when discharging

Conclusion and Future Work

Main Takeaways

- Strategic BESS interaction creates endogenous price effects
- Increasing competition reduces individual market power
- The LQ game framework enables tractable equilibrium analysis

Future Directions

- Non-LQ structure
- Deep learning / reinforcement learning methods for non-LQ games
- Major–Minor and mean field game formulations
- Heterogeneous and asymmetric equilibrium structures

Thank you!

Population-Aware Imitation Learning in Mean-Field Games with Common Noise

Grégoire Lambrecht

Joint work with Mathieu Laurière

NYU Shanghai

IMSI Workshop, Chicago

May 2026

Why Read Our Paper?

Key Contributions of Lambrecht and Laurière (2026)

- **Theory:** Convergence guarantees for imitation learning proxies in Mean-Field Games with common noise.
- **Algorithm:** *Deep Fictitious Play*, a heuristic extension of *Fictitious Play* to the common noise setting.
- **Robustness:** Empirical evidence showing the advantage of *population-aware* policies over *population-unaware* ones.

arXiv preprint →

What is Imitation Learning in MFGs?

- Observe experts playing at equilibrium.
- Assume that the agents' objective (i.e., the reward function) is unknown.
- Questions:
  - **(Q1) Equilibrium Recovery:** *Is it possible to recover actions to reconstruct an equilibrium ?*
  - **(Q2) Performance Matching:** *Is it possible to recover actions to perform well against the experts?*
- Without common noise: Ramponi et al. (2023).

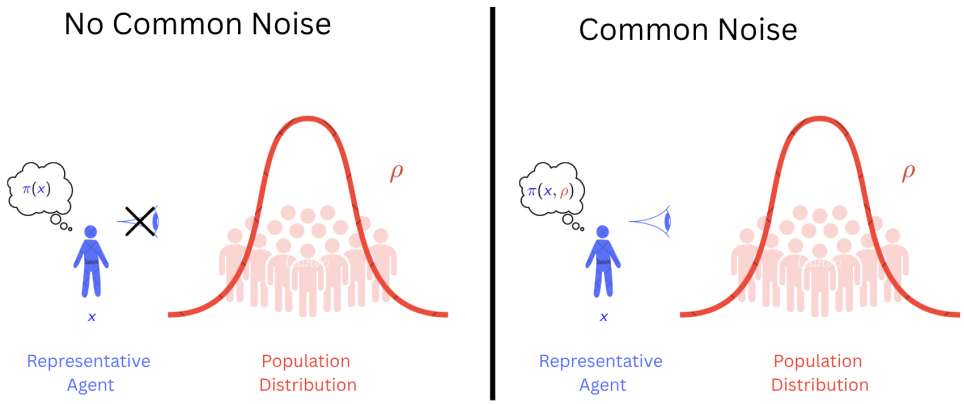
Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Laurière, and Matthieu Geist. On imitation in mean-field games. Advances in Neural Information Processing Systems, 36, 2023.

What is common noise?

- An additional source of randomness affecting all agents simultaneously.

What changes with common noise?

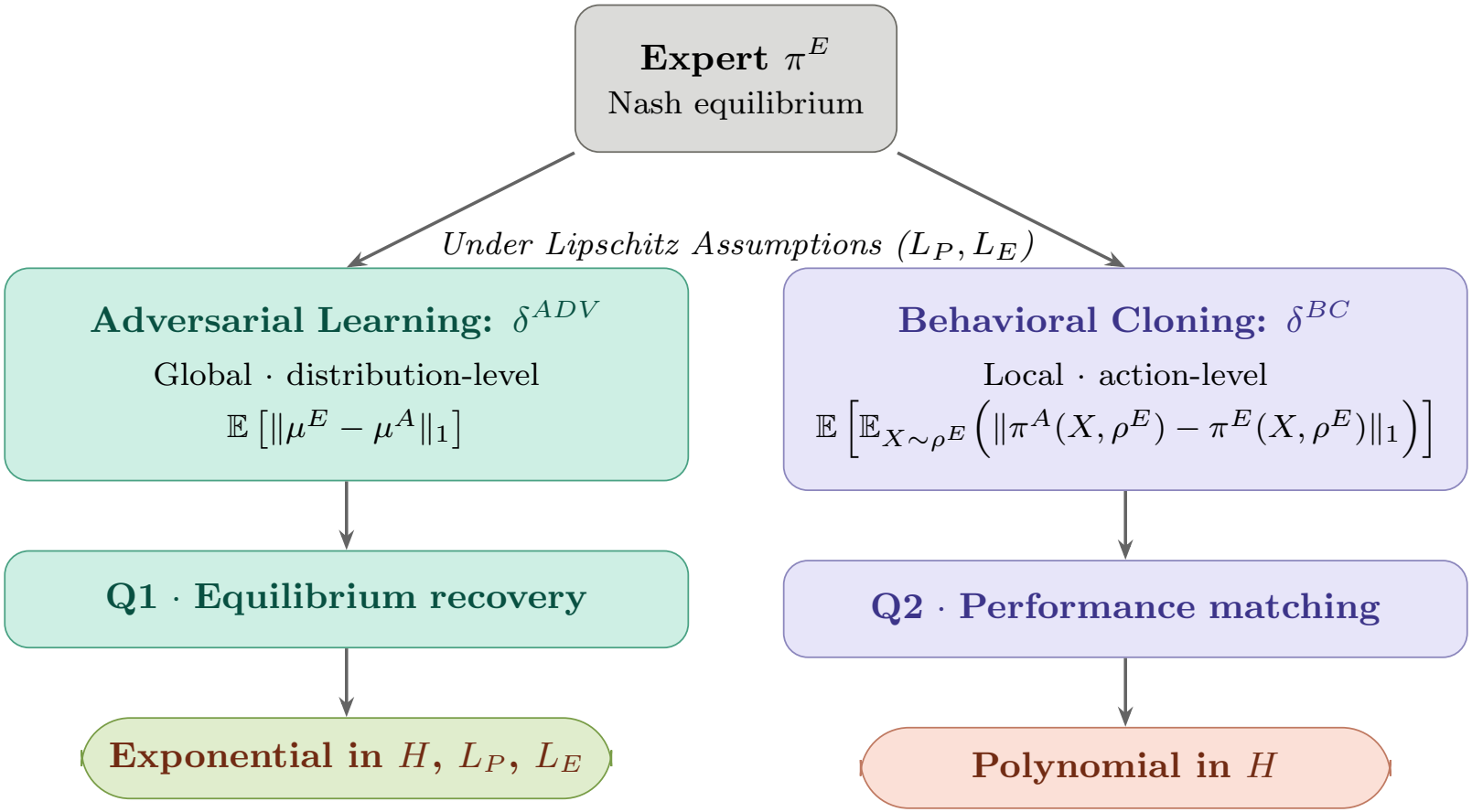
- The population distribution becomes stochastic.
- Policies become population-aware.



Can population-unaware imitation learning methods still perform well under common noise?

Theoretical Results

- π^E: Expert Policy, π^A: Candidate Policy.
- **(Q1) Equilibrium Recovery:** Exploitability$(\pi^A) \leq C\delta(\pi^A, \pi^E)$
- **(Q2) Perf. Matching:** Perf. vs. Expert$(\pi^A) \leq C'\delta(\pi^A, \pi^E)$



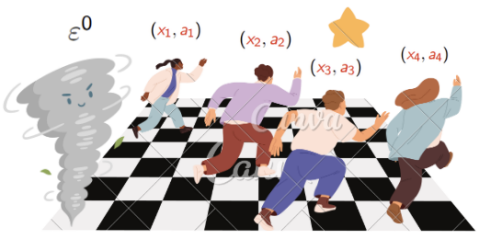
Learning Algorithm, Numerical Results and Key Takeaways

Expert World at Equilibrium

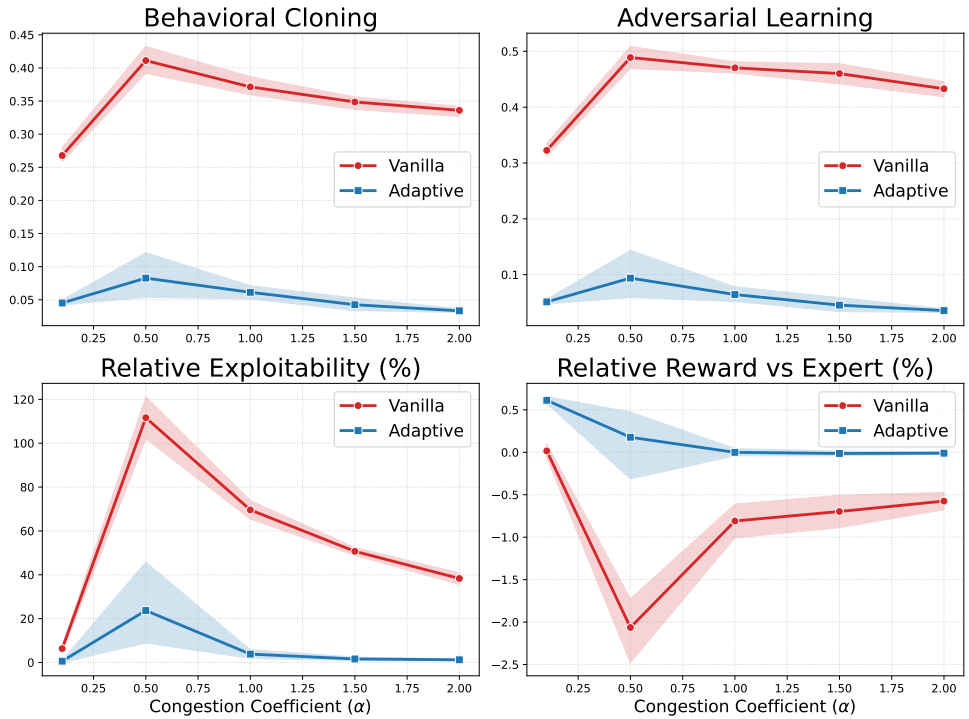
Learner World

Learning $\pi^\theta(a|x,\rho)$

$$L(\theta) \sim \sum_{i=1}^{4} \|\pi^\theta(x_i, \hat{\rho}) - \hat{\pi}(x_i)\|_1$$



Observe (x_i, a_i)



Thank You

Policy Gradient Learning for Distributionally Robust Markov Decision Processes under Wasserstein Ambiguity

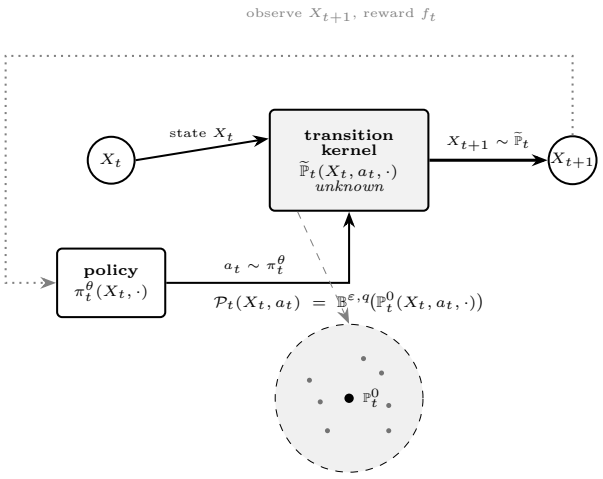
Y. Hafsi S. Mekkaoui H. Pham K. Yan

CMAP, École Polytechnique
School of Mathematical Sciences, Xiamen University

Foundations of Multi-Agent and Mean Field Reinforcement Learning

May 21, 2026

From MDPs to Robust MDPs



Standard MDP. Tuple $(\mathcal{X}, A, \{\mathbb{P}_t\}_{t<T}, f, g, T)$ with $\mathcal{X} \subset \mathbb{R}^d$, $A \subset \mathbb{R}^m$ compact, $T \in \mathbb{N}^*$, $\mathcal{T} := \{0, \dots, T-1\}$, $\mathbb{P}_t : \mathcal{X} \times A \to \mathcal{P}(\mathcal{X})$, $f : \mathcal{T} \times \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$, $g : \mathcal{X} \to \mathbb{R}$, and trajectory reward $R(S) = \sum_t f(t, X_t, a_t, X_{t+1}) + g(X_T)$.

Randomized parametric policy. $\pi_t^\theta : \mathcal{X} \to \mathcal{P}(A)$, $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ compact; for $\mu_0 \in \mathcal{P}(\mathcal{X})$, optimize $J(\theta) := \mathbb{E}_{\mu_0}[V_0^{\pi^\theta}(X_0)]$.

Specialization. For $\varepsilon \geq 0$, $q \geq 1$, $\mathbb{P}_t^0 : \mathcal{X} \times A \to \mathcal{P}(\mathcal{X})$: $\mathcal{P}_t(x, a) = \mathbb{B}^{\varepsilon,q}\big(\mathbb{P}_t^0(x, a, \cdot)\big)$, $\mathcal{W}_q$-ball of radius ε.

Robust criterion:

$$V(x) \;=\; \sup_{\pi} \inf_{\mathbb{P} \in \mathcal{B}_{x,\pi}} \mathbb{E}^{\mathbb{P}}[R(S)].$$

Dynamic programming and the differentiation obstacle

Theorem (Robust DPP, Thm. 2.1)

For all $(t,x) \in \mathcal{T} \times \mathcal{X}$ and all admissible π,

$$V_T^\pi(x) = g(x), \qquad V_t^\pi(x) = \mathbb{E}_{a\sim\pi_t(x,\cdot)}\Big[\inf_{\mathbb{P}_t \in \mathcal{P}_t(x,a)} \mathbb{E}^{\mathbb{P}_t}\big[f(t,x,a,X_{t+1}) + V_{t+1}^\pi(X_{t+1})\big]\Big].$$

A Borel-measurable worst-case selector $\mathbb{P}_t^{*,\pi}$ attains the infimum and $V^\pi(x_0) = \mathbb{E}^{\mathbb{P}^{*,\pi}}[R(S)]$.

Obstacle. The recursion under the parametric policy reads

$$V_t^\theta(x) \;=\; \int_A \int_{\mathcal{X}} \big[\, f(t,x,a,x') + V_{t+1}^\theta(x') \big]\, \mathbb{P}_t^{*,\theta}(x,a,\mathrm{d}x')\, \pi_t^\theta(x,\mathrm{d}a).$$

The kernel $\mathbb{P}_t^{*,\theta}$ *depends implicitly on* θ through V_{t+1}^θ. In general $\theta \mapsto \mathbb{P}_t^{*,\theta}$ admits no closed form and is not amenable to a straightforward induction.

Strategy. Pass to a *dual* representation in which $\mathbb{P}_t^{*,\theta}$ never appears explicitly.

Wasserstein duality and policy gradient

Constrained OT. For $V : \mathcal{X} \to \mathbb{R}$ continuous, $\mathbb{P}^0 \in \mathcal{P}(\mathcal{X})$, $\varepsilon \geq 0$, $c(x,y) = \|x-y\|^q$:

$$\inf_{\mathbb{P}\in\mathbb{B}^{\varepsilon,q}(\mathbb{P}^0)} \mathbb{E}^{\mathbb{P}}[V(Y)] \;=\; \inf_{\substack{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{X}):\\ \mathrm{pr}_{1\#}\gamma=\mathbb{P}^0,\ \mathbb{E}^{\gamma}[c(X,Y)]\leq\varepsilon^q}} \mathbb{E}^{\gamma}[V(Y)].$$

Strong duality (Blanchet–Murthy, 2019). With $\mathcal{F}^{\lambda}(V)(x) := \sup_{y\in\mathcal{X}}\{V(y) - \lambda c(x,y)\}$, $\lambda \geq 0$:

$$\inf_{\mathbb{P}\in\mathbb{B}^{\varepsilon,q}(\mathbb{P}^0)} \mathbb{E}^{\mathbb{P}}[V(Y)] \;=\; \sup_{\lambda\geq 0} \mathbb{E}^{\mathbb{P}^0}\Big[\inf_{y\in\mathcal{X}}\{V(y) + \lambda c(X,y)\}\Big] - \varepsilon^q\lambda.$$

Dual robust Bellman. For $(t,\theta,x,a) \in \mathcal{T}\times\Theta\times\mathcal{X}\times A$:

$$G_t^{\theta}(x,a) := \sup_{\lambda\geq 0}\big\{\mathbb{E}_{X\sim\mathbb{P}_t^0(x,a,\cdot)}\big[-\mathcal{F}^{\lambda}(-f(t,x,a,\cdot) - V_{t+1}^{\theta})(X)\big] - \varepsilon^q\lambda\big\} \in \mathbb{R},$$

$$\Lambda_t^{*,\theta}(x,a) := \arg\max_{\lambda\in[0,\Lambda]}\{\cdots\} \subset [0,\Lambda], \quad Y_{t,\theta,\lambda}^*(x,a;X) := \operatorname*{argmin}_{y\in\mathcal{X}}\{f + V_{t+1}^{\theta} + \lambda c(X,y)\} \subset \mathcal{X}.$$

Theorem (Policy gradient, Thm. 3.1)

For every $r \in \mathbb{R}^{d_\theta}$, $\theta \mapsto G_t^{\theta}(x,a)$ admits one-sided directional derivatives:

$$D_{\theta}^{+}G_t^{\theta}(x,a)[r] \;=\; \sup_{\lambda\in\Lambda_t^{*,\theta}(x,a)} \mathbb{E}_{X\sim\mathbb{P}_t^0(x,a,\cdot)}\Big[\inf_{y\in Y_{t,\theta,\lambda}^*(x,a;X)} D_{\theta}V_{t+1}^{\theta}(y)[r]\Big],$$

$$D_{\theta}^{-}G_t^{\theta}(x,a)[r] \;=\; \inf_{\lambda\in\Lambda_t^{*,\theta}(x,a)} \mathbb{E}_{X\sim\mathbb{P}_t^0(x,a,\cdot)}\Big[\sup_{y\in Y_{t,\theta,\lambda}^*(x,a;X)} D_{\theta}V_{t+1}^{\theta}(y)[r]\Big].$$

Robust actor–critic algorithm

Approximators updated backward in time, $t = T-1, \dots, 0$:

$$V_{\psi,t}(x) \approx V_t^{\theta}(x), \qquad U_{\xi,t}(x) \approx \nabla_\theta V_t^{\theta}(x).$$

Per step: draw $a_t \sim \pi_t^{\theta}(x_t, \cdot)$ and $X_{t+1}^{(i)} \sim \mathbb{P}_t^0(x_t, a_t, \cdot)$, $i = 1, \dots, N$.

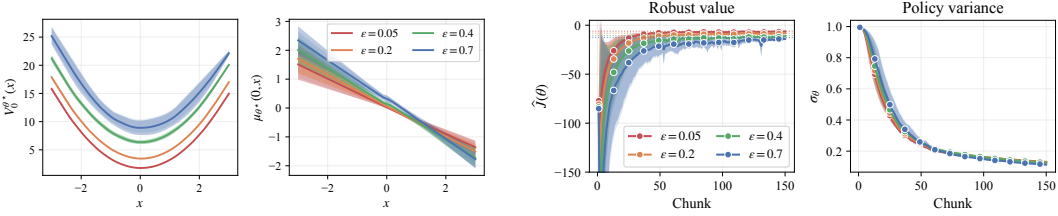
1. **Critic.** Solve the scalar dual on $[0, \Lambda]$, $\widehat{\lambda}_t^* \in \arg\max_{\lambda \in [0,\Lambda]} \widehat{F}_t^{\theta}(\lambda; x_t, a_t)$, then regress $V_{\psi,t}(x_t)$ on $\widehat{G}_t^{\theta}(x_t, a_t)$.
2. **Inner selector.** For each i, $y^*\big(X_{t+1}^{(i)}\big) \in \arg\min_{y \in \mathcal{X}} \big\{ f(t, x_t, a_t, y) + V_{\psi,t+1}(y) + \widehat{\lambda}_t^*\, c(X_{t+1}^{(i)}, y) \big\}$.
3. **Gradient critic.** Form $\widehat{\nabla_\theta G_t^{\theta}}(x_t, a_t) = \frac{1}{N} \sum_{i=1}^{N} U_{\xi,t+1}\big(y^*(X_{t+1}^{(i)})\big)$, then regress $U_{\xi,t}(x_t)$ on $\widehat{z}_t = \widehat{G}_t^{\theta}(x_t, a_t)\, \nabla_\theta \log \pi_t^{\theta}(x_t, a_t) + \widehat{\nabla_\theta G_t^{\theta}}(x_t, a_t)$.
4. **Actor.** $\theta \leftarrow \theta + \eta_\theta \cdot M^{-1} \sum_{j=1}^{M} U_{\xi,0}(X_0^{(j)})$, $X_0^{(j)} \sim \mu_0$.

Remarks. (i) No small-ε restriction. (ii) Tabular case is exact; function approximation is a scalability device. (iii) Framework extends to any convex, weakly-compact ambiguity admitting strong duality.

Numerical evidence — LQ control

Setup. Dynamics $X_{t+1} = AX_t + Bu_t + \Xi w_t$, quadratic costs $x^\top Qx + u^\top Ru$, terminal $x^\top P_T x$. Adversary perturbs the empirical noise distribution within a Wasserstein ball. Policy parametrized as a truncated Gaussian with neural-network mean $\mu_\theta(t,x)$ and learned scalar variance σ_θ^2.

Recovery of the Riccati structure. The learned $V_0^{\theta^\star}$ is quadratic in x and $\mu_{\theta^\star}(0,\cdot)$ is linear, matching the closed-form benchmark of Kim–Yang (2021):



Training is stable for all ε; the policy variance σ_θ collapses toward 0, so the stochastic policy becomes nearly deterministic, consistent with the deterministic Riccati optimizer $K_t x + L_t$.

Take-away. Wasserstein duality + directional Danskin ⇒ tractable robust actor–critic; recovers the Riccati optimum in LQ. `github.com/YadhHafsi/Policy-Gradient-Learning-for-DRO`