

What structures make model-free value learning possible?

An elliptic theory for controlled Markov diffusions

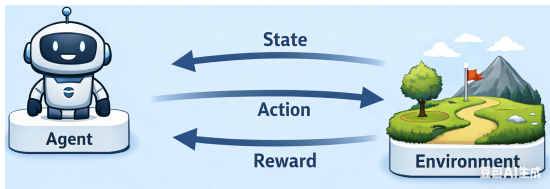
Wenlong Mou

Department of Statistical Sciences
University of Toronto

April 23, 2026

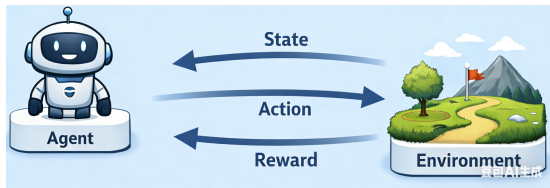
From dynamic programming to function approximation

Reinforcement learning



From dynamic programming to function approximation

Reinforcement learning

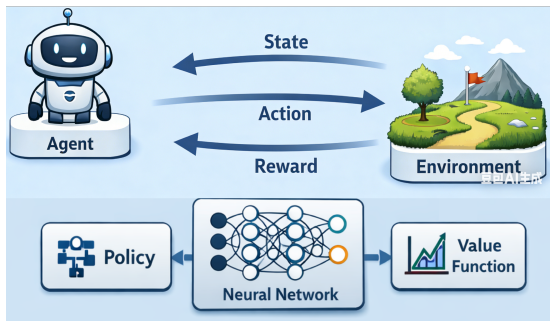


Optimal Q -function

$$Q^*(s, a) = \max_{\pi} \mathbb{E}_{\pi} \left[\sum_t r(s_t, a_t) \mid s_0 = s, a_0 = a \right],$$

From dynamic programming to function approximation

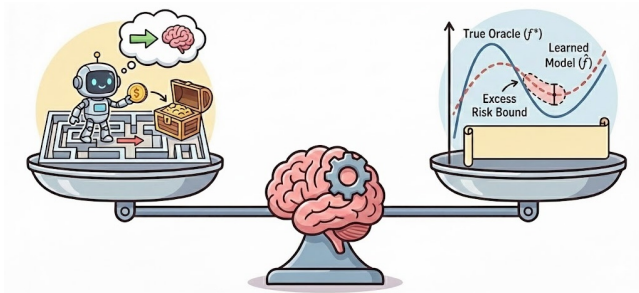
Reinforcement learning with function approximation



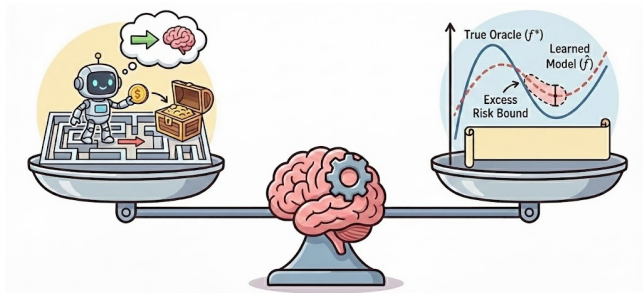
Approximate Optimal Q -function with function class $\mathcal{F}$

$$Q^*(s, a) = \max_{\pi} \mathbb{E}_{\pi} \left[\sum_t r(s_t, a_t) \mid s_0 = s, a_0 = a \right], \quad Q^* \approx \bar{Q} \in \mathcal{F}$$

My dream: make offline RL as easy as supervised learning

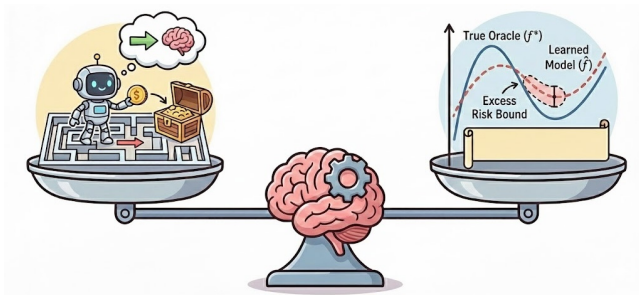


My dream: make offline RL as easy as supervised learning



Regression: given $(X_i, Y_i)_{i=1}^n$, fit $f^*(x) := \mathbb{E}[Y \mid X(x)]$ with function class $\mathcal{F}$.

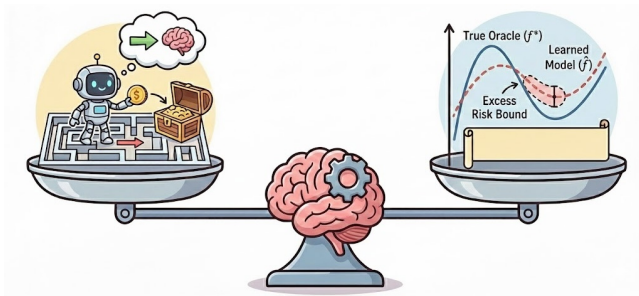
My dream: make offline RL as easy as supervised learning



Regression: given $(X_i, Y_i)_{i=1}^n$, fit $f^*(x) := \mathbb{E}[Y \mid X(x)]$ with function class $\mathcal{F}$.

- ▶ **Approximation:** oracle error $\inf_{f \in \mathcal{F}} \|f - f^*\|$.
- ▶ **Estimation:** critical radius $r_n^2 = \frac{1}{n} \mathcal{C}(\mathcal{F}(r_n))$, for complexity measure $\mathcal{C}$.
- ▶ **Optimization:** under convexity, e^{-cN} convergence with N iterations.

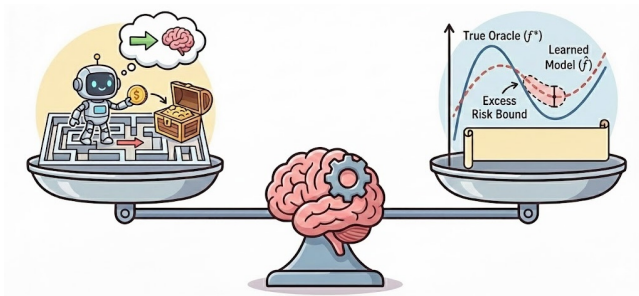
My dream: make offline RL as easy as supervised learning



Regression: given $(X_i, Y_i)_{i=1}^n$, fit $f^*(x) := \mathbb{E}[Y \mid X(x)]$ with function class $\mathcal{F}$.

- **Approximation:** oracle error $\inf_{f \in \mathcal{F}} \|f - f^*\|$.
- **Estimation:** critical radius $r_n^2 = \frac{1}{\sqrt{n}} \mathcal{C}(\mathcal{F}(r_n))$, for complexity measure $\mathcal{C}$.
- **Optimization:** under convexity, e^{-cN} convergence with N iterations.

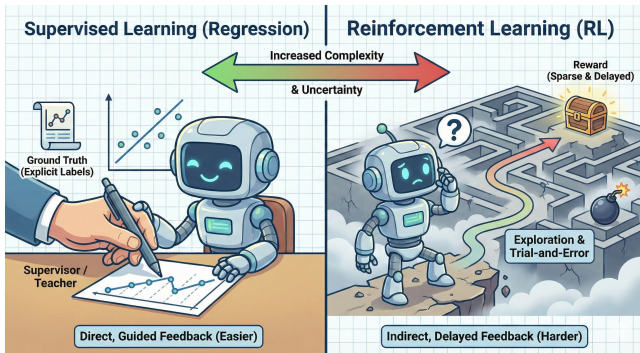
My dream: make offline RL as easy as supervised learning



Regression: given $(X_i, Y_i)_{i=1}^n$, fit $f^*(x) := \mathbb{E}[Y \mid X(x)]$ with function class $\mathcal{F}$.

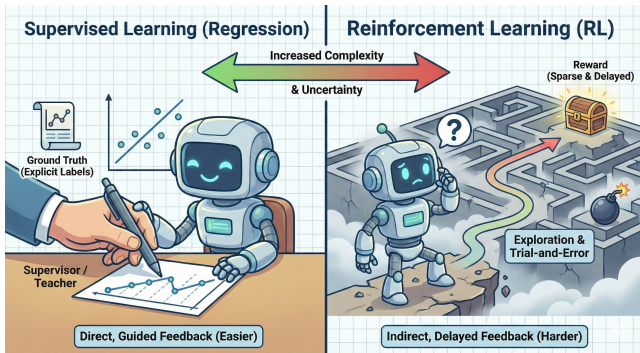
- ▶ **Approximation:** oracle error $\inf_{f \in \mathcal{F}} \|f - f^*\|$.
- ▶ **Estimation:** critical radius $r_n^2 = \frac{1}{\sqrt{n}} \mathcal{C}(\mathcal{F}(r_n))$, for complexity measure $\mathcal{C}$.
- ▶ **Optimization:** under convexity, e^{-cN} convergence with N iterations.

Broken dream: offline RL is harder



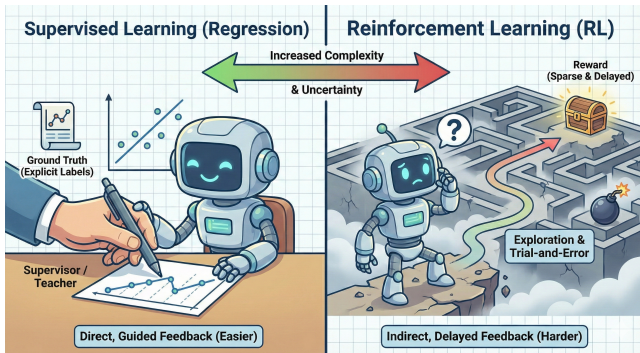
- ▶ In practice, training instability, high variance and slower convergence.
- ▶ In theory, function approximation needs unrealistic structural assumptions on the MDP (Jin et al., 2020, Wang et al., 2021, etc), or unimplementable algorithms (Xie et al., 2021, etc).

Broken dream: offline RL is harder



- ▶ In practice, training instability, high variance and slower convergence.
- ▶ In theory, function approximation needs unrealistic structural assumptions on the MDP (Jin et al., 2020, Wang et al., 2021, etc), or unimplementable algorithms (Xie et al., 2021, etc).

Broken dream: offline RL is harder



- ▶ In practice, training instability, high variance and slower convergence.
- ▶ In theory, function approximation needs unrealistic structural assumptions on the MDP (Jin et al., 2020, Wang et al., 2021, etc), or unimplementable algorithms (Xie et al., 2021, etc).

Root cause of the problem

Value learning solves fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\max_{a'} Q(s', a')]$$

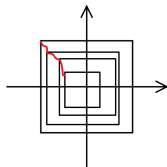
Root cause of the problem

Value learning solves fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\max_{a'} Q(s', a')]$$

Key property: ℓ^∞ -contraction

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty, \quad \text{for all } Q_1, Q_2.$$



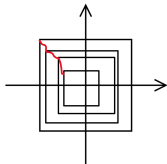
Root cause of the problem

Value learning solves fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\max_{a'} Q(s', a')]$$

Key property: ℓ^∞ -contraction

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty, \quad \text{for all } Q_1, Q_2.$$



- Convergence of value iteration/policy iteration.
- Stochastic approximation under tabular setting (Chen et al., 2021, Mou et al., 2022, etc).

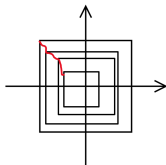
Root cause of the problem

Value learning solves fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [\max_{a'} Q(s', a')]$$

Key property: ℓ^∞ -contraction

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty, \quad \text{for all } Q_1, Q_2.$$



- ▶ Convergence of value iteration/policy iteration.
- ▶ Stochastic approximation under tabular setting (Chen et al., 2021, Mou et al., 2022, etc).

Root cause of the problem

With function approximation, value learning solves **projected** fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\max_{a'} Q(s', a')]$$

$$\bar{Q} = \Pi_{\mathcal{F}} \circ \mathcal{T}(\bar{Q}), \quad \Pi_{\mathcal{F}}(Q) = \operatorname{argmin}_{f \in \mathcal{F}} \|f - Q\|_{\mathbb{L}^2(\mu)}.$$

Root cause of the problem

With function approximation, value learning solves **projected** fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\max_{a'} Q(s', a')]$$

$$\bar{Q} = \Pi_{\mathcal{F}} \circ \mathcal{T}(\bar{Q}), \quad \Pi_{\mathcal{F}}(Q) = \operatorname{argmin}_{f \in \mathcal{F}} \|f - Q\|_{\mathbb{L}^2(\mu)}.$$

- ▶ Composition operator no longer contractive (under either ℓ^∞ or ℓ^2 norm).
- ▶ We can still run the algorithm – fitted Q -iteration and Q -learning

$$\hat{Q}_{t+1} = \operatorname{argmin}_{f \in \mathcal{F}} \|(1 - \alpha)\hat{Q}_t + \alpha\hat{\mathcal{T}}(\hat{Q}_t) - f\|_{\mathbb{L}^2(\mu)}^2$$

No convergence guarantees without strong assumptions.

Root cause of the problem

With function approximation, value learning solves **projected** fixed-point equation

$$Q^* = \mathcal{T}(Q^*), \quad \mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [\max_{a'} Q(s', a')]$$

$$\bar{Q} = \Pi_{\mathcal{F}} \circ \mathcal{T}(\bar{Q}), \quad \Pi_{\mathcal{F}}(Q) = \operatorname{argmin}_{f \in \mathcal{F}} \|f - Q\|_{\mathbb{L}^2(\mu)}.$$

- ▶ Composition operator no longer contractive (under either ℓ^∞ or ℓ^2 norm).
- ▶ We can still run the algorithm – fitted Q -iteration and Q -learning

$$\hat{Q}_{t+1} = \operatorname{argmin}_{f \in \mathcal{F}} \|(1 - \alpha)\hat{Q}_t + \alpha\hat{\mathcal{T}}(\hat{Q}_t) - f\|_{\mathbb{L}^2(\mu)}^2$$

No convergence guarantees without strong assumptions.

Square peg and round hole



Square peg and round hole



- ▶ But we still use value learning in practice.
- ▶ Perhaps some structure helps?
- ▶ With structures, can we design better algorithms?

Square peg and round hole



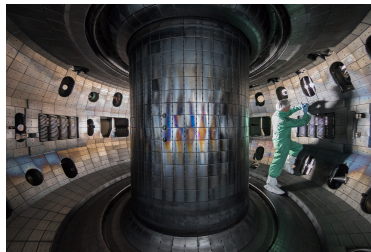
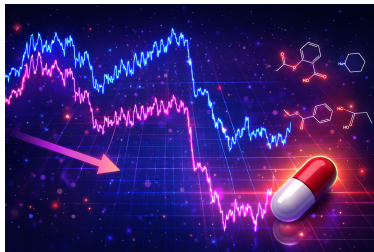
- ▶ But we still use value learning in practice.
- ▶ Perhaps some structure helps?
- ▶ With structures, can we design better algorithms?

Square peg and round hole



- ▶ But we still use value learning in practice.
- ▶ Perhaps some structure helps?
- ▶ With structures, can we design better algorithms?

Moving to continuous-time RL



Continuous-time RL: problem setup

(Controlled) diffusion process in $\mathbb{R}^d$

$$dX_t = b(X_t, A_t)dt + \Lambda^{1/2}(X_t)dB_t.$$

- ▶ B : d -dimensional standard BM.
- ▶ Reward function $r(x, a)$.
- ▶ discount rate $\beta > 0$.

Value function definitions

$$v^\pi(x) = \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t} r(X_t^\pi, A_t^\pi) dt \mid X_0 = x\right],$$
$$v^*(x) = \max_{\pi: \mathbb{X} \rightarrow \Delta^A} v^\pi(x).$$

Discrete-time observations from behaviour policy π_0 .

$$X_0, A_0, R_0, X_\eta, A_\eta, R_\eta, \dots, X_{k\eta}, A_{k\eta}, R_{k\eta}, \dots$$

Continuous-time RL: problem setup

(Controlled) diffusion process in $\mathbb{R}^d$

$$dX_t = b(X_t, A_t)dt + \Lambda^{1/2}(X_t)dB_t.$$

- ▶ B : d -dimensional standard BM.
- ▶ Reward function $r(x, a)$.
- ▶ discount rate $\beta > 0$.

Value function definitions

$$v^\pi(x) = \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t} r(X_t^\pi, A_t^\pi) dt \mid X_0 = x\right],$$
$$v^*(x) = \max_{\pi: \mathbb{X} \rightarrow \Delta^A} v^\pi(x).$$

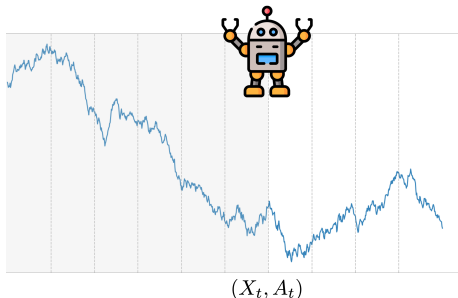
Discrete-time observations from behaviour policy π_0 .

$$X_0, A_0, R_0, X_\eta, A_\eta, R_\eta, \dots, X_{k\eta}, A_{k\eta}, R_{k\eta}, \dots$$

Continuous-time RL as discrete-time RL

Equivalent MDP model for continuous-time policy evaluation.

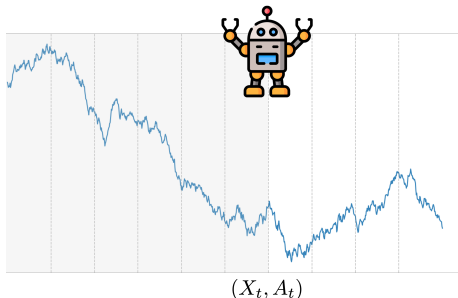
- ▶ Transition kernel $\mathcal{P}_\eta^a$.
- ▶ Discount factor $\gamma = e^{-\beta\eta}$.
- ▶ Reward $\tilde{r}(x, a) \approx \int_0^\eta e^{-\beta t} \mathbb{E}^a[r(X_t, a) | X_0 = x] dt$, numerical approx.



Continuous-time RL as discrete-time RL

Equivalent MDP model for continuous-time policy evaluation.

- ▶ Transition kernel $\mathcal{P}_\eta^a$.
- ▶ Discount factor $\gamma = e^{-\beta\eta}$.
- ▶ Reward $\tilde{r}(x, a) \approx \int_0^\eta e^{-\beta t} \mathbb{E}^a[r(X_t, a) | X_0 = x] dt$, numerical approx.



Additional challenge: time discretization leads to long horizon $\gamma \rightarrow 1$ as $\eta \rightarrow 0$.

Warmup: policy evaluation with linear function approximation

(projected) Bellman equation for policy evaluation

Linear fixed-point equation and its projected version

$$\begin{aligned}v^\pi &= e^{-\beta\eta}\mathcal{P}_\eta^\pi v^\pi + \eta r^\pi, \\ \bar{v}^\pi &= \Pi_{\mathcal{F}_v}(e^{-\beta\eta}\mathcal{P}_\eta^\pi \bar{v}^\pi + \eta r^\pi),\end{aligned}$$

and empirical equations to solve $\hat{v}^\pi$.

(projected) Bellman equation for policy evaluation

Linear fixed-point equation and its projected version

$$\begin{aligned}v^\pi &= e^{-\beta\eta}\mathcal{P}_\eta^\pi v^\pi + \eta r^\pi, \\ \bar{v}^\pi &= \Pi_{\mathcal{F}_v}(e^{-\beta\eta}\mathcal{P}_\eta^\pi \bar{v}^\pi + \eta r^\pi),\end{aligned}$$

and empirical equations to solve $\hat{v}^\pi$.

- ▶ Discrete-time policy evaluation yields ℓ^2 contraction, with factor $e^{-\beta\eta}$.
- ▶ Large effective horizon leads to bad bounds.

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{L}^2(\mu)}^2 \leq O(1/\eta) \cdot \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{L}^2(\mu)}^2 + O\left(\frac{\text{complexity}}{\eta n}\right).$$

(projected) Bellman equation for policy evaluation

Linear fixed-point equation and its projected version

$$\begin{aligned}v^\pi &= e^{-\beta\eta}\mathcal{P}_\eta^\pi v^\pi + \eta r^\pi, \\ \bar{v}^\pi &= \Pi_{\mathcal{F}_v}(e^{-\beta\eta}\mathcal{P}_\eta^\pi \bar{v}^\pi + \eta r^\pi),\end{aligned}$$

and empirical equations to solve $\hat{v}^\pi$.

- ▶ Discrete-time policy evaluation yields ℓ^2 contraction, with factor $e^{-\beta\eta}$.
- ▶ Large effective horizon leads to bad bounds.

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{L}^2(\mu)}^2 \leq O(1/\eta) \cdot \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{L}^2(\mu)}^2 + O\left(\frac{\text{complexity}}{\eta n}\right).$$

(projected) Bellman equation for policy evaluation

Linear fixed-point equation and its projected version

$$\begin{aligned}v^\pi &= e^{-\beta\eta}\mathcal{P}_\eta^\pi v^\pi + \eta r^\pi, \\ \bar{v}^\pi &= \Pi_{\mathcal{F}_v}(e^{-\beta\eta}\mathcal{P}_\eta^\pi \bar{v}^\pi + \eta r^\pi),\end{aligned}$$

and empirical equations to solve $\hat{v}^\pi$.

- ▶ Discrete-time policy evaluation yields ℓ^2 contraction, with factor $e^{-\beta\eta}$.
- ▶ Large effective horizon leads to bad bounds.

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{L}^2(\mu)}^2 \leq O(1/\eta) \cdot \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{L}^2(\mu)}^2 + O\left(\frac{\text{complexity}}{\eta n}\right).$$

Does PDE structure help?

(projected) Bellman equation for policy evaluation

Linear fixed-point equation and its projected version

$$\begin{aligned}v^\pi &= e^{-\beta\eta}\mathcal{P}_\eta^\pi v^\pi + \eta r^\pi, \\ \bar{v}^\pi &= \Pi_{\mathcal{F}_v}(e^{-\beta\eta}\mathcal{P}_\eta^\pi \bar{v}^\pi + \eta r^\pi),\end{aligned}$$

and empirical equations to solve $\hat{v}^\pi$.

- ▶ Discrete-time policy evaluation yields ℓ^2 contraction, with factor $e^{-\beta\eta}$.
- ▶ Large effective horizon leads to bad bounds.

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{L}^2(\mu)}^2 \leq O(1/\eta) \cdot \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{L}^2(\mu)}^2 + O\left(\frac{\text{complexity}}{\eta n}\right).$$

Does PDE structure help? Key insight: for $\eta \rightarrow 0$,

$$\frac{1}{\eta}(v - e^{-\beta\eta}\mathcal{P}_\eta^\pi v) \rightarrow r^\pi + (\beta - \mathcal{L}^\pi)v, \quad \text{where } \mathcal{L}^\pi f = b^\pi \cdot \nabla f + \frac{1}{2}\text{tr}(\Lambda^\pi \nabla^2 f).$$

Blessings of ellipticity

Ellipticity assumption:

$$\lambda_{\min} I \leq \Lambda(x) \leq \lambda_{\max} I, \quad \text{for all } x.$$

Blessings of ellipticity

Ellipticity assumption:

$$\lambda_{\min} I \leq \Lambda(x) \leq \lambda_{\max} I, \quad \text{for all } x.$$

Ellipticity yields positive definite operator

$$\langle (\beta - \mathcal{L}^{\pi_0})f, f \rangle_{\mathbb{L}^2(\mu)} \geq \lambda_{\min} \|f\|_{\mathbb{H}^1(\mu)}^2, \quad \text{for any } f.$$

(Sobolev norm $\|f\|_{\mathbb{H}^1(\mu)} := \mathbb{E}_\mu[|\nabla f|^2] + \mathbb{E}_\mu[f^2]$.)

Blessings of ellipticity

Ellipticity assumption:

$$\lambda_{\min} I \leq \Lambda(x) \leq \lambda_{\max} I, \quad \text{for all } x.$$

Ellipticity yields positive definite operator

$$\langle (\beta - \mathcal{L}^{\pi_0})f, f \rangle_{\mathbb{L}^2(\mu)} \geq \lambda_{\min} \|f\|_{\mathbb{H}^1(\mu)}^2, \quad \text{for any } f.$$

(Sobolev norm $\|f\|_{\mathbb{H}^1(\mu)} := \mathbb{E}_{\mu}[|\nabla f|^2] + \mathbb{E}_{\mu}[f^2]$.)



Lax–Milgram theorem
(Existence and uniqueness in Sobolev space $\mathbb{H}^1$).

Blessings of ellipticity

Ellipticity assumption:

$$\lambda_{\min} I \leq \Lambda(x) \leq \lambda_{\max} I, \quad \text{for all } x.$$

Ellipticity yields positive definite operator

$$\langle (\beta - \mathcal{L}^{\pi_0})f, f \rangle_{\mathbb{L}^2(\mu)} \geq \lambda_{\min} \|f\|_{\mathbb{H}^1(\mu)}^2, \quad \text{for any } f.$$

(Sobolev norm $\|f\|_{\mathbb{H}^1(\mu)} := \mathbb{E}_{\mu}[|\nabla f|^2] + \mathbb{E}_{\mu}[f^2]$.)



Galerkin method

- ▶ Solve the PDE (with known coefficients) by projection onto linear subspace.
- ▶ Bounded approx. factor under uniform ellipticity. (Céa, 1964)

Results for policy evaluation

For diffusions, Bellman operators are “well-conditioned” under Sobolev norms.

Lemma

For v_1, v_2 , we have

$$\begin{aligned}\frac{1}{\eta} \langle (v_1 - e^{-\beta\eta} \mathcal{P}_\eta^{\pi_0} v_1), v_1 \rangle_{\mathbb{L}^2(\mu)} &\geq \min(\beta, \lambda_{\min}) \|v_1\|_{\mathbb{H}^1(\mu)}^2, \\ \left| \frac{1}{\eta} \langle (v_1 - e^{-\beta\eta} \mathcal{P}_\eta^{\pi_0} v_1), v_2 \rangle_{\mathbb{L}^2(\mu)} \right| &\leq c \|v_1\|_{\mathbb{H}^1(\mu)} \cdot \|v_2\|_{\mathbb{H}^1(\mu)}\end{aligned}$$

Results for policy evaluation

For diffusions, Bellman operators are “well-conditioned” under Sobolev norms.

Lemma

For v_1, v_2 , we have

$$\begin{aligned}\frac{1}{\eta} \langle (v_1 - e^{-\beta\eta} \mathcal{P}_\eta^{\pi_0} v_1), v_1 \rangle_{\mathbb{L}^2(\mu)} &\geq \min(\beta, \lambda_{\min}) \|v_1\|_{\mathbb{H}^1(\mu)}^2, \\ \left| \frac{1}{\eta} \langle (v_1 - e^{-\beta\eta} \mathcal{P}_\eta^{\pi_0} v_1), v_2 \rangle_{\mathbb{L}^2(\mu)} \right| &\leq c \|v_1\|_{\mathbb{H}^1(\mu)} \cdot \|v_2\|_{\mathbb{H}^1(\mu)}\end{aligned}$$

Results for policy evaluation

For diffusions, Bellman operators are “well-conditioned” under Sobolev norms.

Lemma

For v_1, v_2 , we have

$$\begin{aligned}\frac{1}{\eta} \langle (v_1 - e^{-\beta\eta} \mathcal{P}_\eta^{\pi_0} v_1), v_1 \rangle_{\mathbb{L}^2(\mu)} &\geq \min(\beta, \lambda_{\min}) \|v_1\|_{\mathbb{H}^1(\mu)}^2, \\ \left| \frac{1}{\eta} \langle (v_1 - e^{-\beta\eta} \mathcal{P}_\eta^{\pi_0} v_1), v_2 \rangle_{\mathbb{L}^2(\mu)} \right| &\leq c \|v_1\|_{\mathbb{H}^1(\mu)} \cdot \|v_2\|_{\mathbb{H}^1(\mu)}\end{aligned}$$

Consequence to learning:

Theorem (Mou and Zhu, 2024, Mou, 2025)

With n samples, the empirical projected Bellman equation yields

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{H}^1}^2 \lesssim \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{H}^1}^2 + \frac{\dim(\mathcal{F}_v)}{n} + O(\eta^{2p})$$

Results for policy evaluation

For diffusions, Bellman operators are “well-conditioned” under Sobolev norms.

Theorem (Mou and Zhu, 2024, Mou, 2025)

With n samples, the empirical projected Bellman equation yields

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{H}^1}^2 \lesssim \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{H}^1}^2 + \frac{\dim(\mathcal{F}_v)}{n} + O(\eta^{2p})$$

- ▶ No blow-up as $\eta \rightarrow 0$.
- ▶ Statistical error indeed enjoys favorable self-mitigating structure.
- ▶ High-order numerical scheme for the reward term yields high-order discretization error $O(\eta^{2p})$.

Results for policy evaluation

For diffusions, Bellman operators are “well-conditioned” under Sobolev norms.

Theorem (Mou and Zhu, 2024, Mou, 2025)

With n samples, the empirical projected Bellman equation yields

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{H}^1}^2 \lesssim \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{H}^1}^2 + \frac{\dim(\mathcal{F}_v)}{n} + O(\eta^{2p})$$

- ▶ No blow-up as $\eta \rightarrow 0$.
- ▶ Statistical error indeed enjoys favorable self-mitigating structure.
- ▶ High-order numerical scheme for the reward term yields high-order discretization error $O(\eta^{2p})$.

Results for policy evaluation

For diffusions, Bellman operators are “well-conditioned” under Sobolev norms.

Theorem (Mou and Zhu, 2024, Mou, 2025)

With n samples, the empirical projected Bellman equation yields

$$\|\hat{v}^\pi - v^\pi\|_{\mathbb{H}^1}^2 \lesssim \inf_{v \in \mathcal{F}_v} \|v - v^\pi\|_{\mathbb{H}^1}^2 + \frac{\dim(\mathcal{F}_v)}{n} + O(\eta^{2p})$$

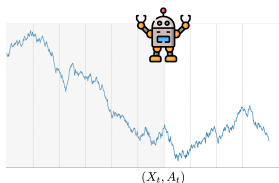
- ▶ No blow-up as $\eta \rightarrow 0$.
- ▶ Statistical error indeed enjoys favorable self-mitigating structure.
- ▶ High-order numerical scheme for the reward term yields high-order discretization error $O(\eta^{2p})$.

Moving to controlled diffusions

Basic setup for control

Offline data from behaviour policy π_0 , applied in discrete time.

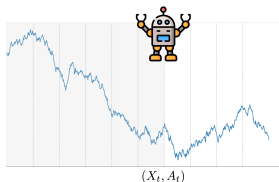
- ▶ $A_{k\eta} \mid X_{k\eta} \sim \pi_0(\cdot \mid X_{k\eta})$ for $k = 0, 1, \dots$.
- ▶ Piecewise constant interpolation, $A_t = A_{k\eta}$ for $t \in [k\eta, (k+1)\eta)$.



Basic setup for control

Offline data from behaviour policy π_0 , applied in discrete time.

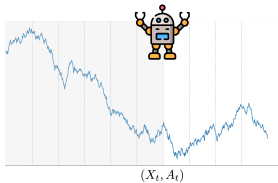
- ▶ $A_{k\eta} \mid X_{k\eta} \sim \pi_0(\cdot \mid X_{k\eta})$ for $k = 0, 1, \dots$.
- ▶ Piecewise constant interpolation, $A_t = A_{k\eta}$ for $t \in [k\eta, (k+1)\eta)$.



Basic setup for control

Offline data from behaviour policy π_0 , applied in discrete time.

- ▶ $A_{k\eta} \mid X_{k\eta} \sim \pi_0(\cdot \mid X_{k\eta})$ for $k = 0, 1, \dots$.
- ▶ Piecewise constant interpolation, $A_t = A_{k\eta}$ for $t \in [k\eta, (k+1)\eta)$.



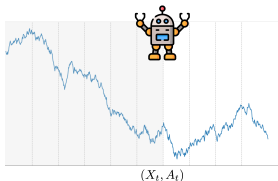
For small η , Q^* is not informative of action values.

$$|Q^*(x, a) - v^*(x)| \leq O(\eta), \quad \text{for all } a \in \mathbb{A}.$$

Basic setup for control

Offline data from behaviour policy π_0 , applied in discrete time.

- ▶ $A_{k\eta} \mid X_{k\eta} \sim \pi_0(\cdot \mid X_{k\eta})$ for $k = 0, 1, \dots$.
- ▶ Piecewise constant interpolation, $A_t = A_{k\eta}$ for $t \in [k\eta, (k+1)\eta)$.



Small-time advantage function (e.g. Jia and Zhou, 2022)

$$q^*(x, a) := \frac{1}{\eta} (Q^*(x, a) - v^*(x)).$$

Use $\mathcal{F}_v$ and $\mathcal{F}_q$ to approximate v^* and q^* , respectively.

Assumptions

- ▶ Regularity of (b, Λ, r) .

- ▶ Uniform ellipticity

$$\lambda_{\min} I \preceq \Lambda(x) \preceq \lambda_{\max} I, \quad \text{for all } x.$$

- ▶ Coverage: for any x , we have

$$\max_{a \in \mathbb{A}} |f(x, a)|^2 \leq c_{\mathbb{A}} \cdot \mathbb{E}_{A \sim \pi_0(\cdot|x)} [|f(x, A)|^2], \quad \text{for all } f \in \mathcal{F}_q - \mathcal{F}_q.$$

Imposed on action space, generally unavoidable for off-policy.

Assumptions

- ▶ Regularity of (b, Λ, r) .
- ▶ Uniform ellipticity

$$\lambda_{\min} I \preceq \Lambda(x) \preceq \lambda_{\max} I, \quad \text{for all } x.$$

- ▶ Coverage: for any x , we have

$$\max_{a \in \mathbb{A}} |f(x, a)|^2 \leq c_{\mathbb{A}} \cdot \mathbb{E}_{A \sim \pi_0(\cdot|x)} [|f(x, A)|^2], \quad \text{for all } f \in \mathcal{F}_q - \mathcal{F}_q.$$

Imposed on action space, generally unavoidable for off-policy.

Assumptions

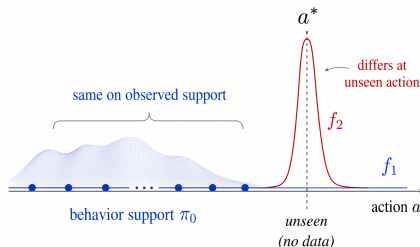
- ▶ Regularity of (b, Λ, r) .
- ▶ Uniform ellipticity

$$\lambda_{\min} I \leq \Lambda(x) \leq \lambda_{\max} I, \quad \text{for all } x.$$

- ▶ Coverage: for any x , we have

$$\max_{a \in \mathbb{A}} |f(x, a)|^2 \leq c_{\mathbb{A}} \cdot \mathbb{E}_{A \sim \pi_0(\cdot|x)} [|f(x, A)|^2], \quad \text{for all } f \in \mathcal{F}_q - \mathcal{F}_q.$$

Imposed on action space, generally unavoidable for off-policy.



Assumptions

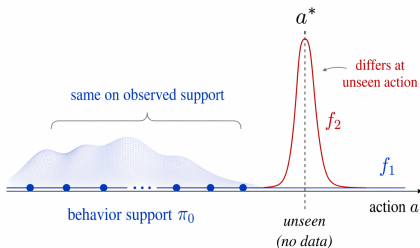
- ▶ Regularity of (b, Λ, r) .
- ▶ Uniform ellipticity

$$\lambda_{\min} I \preceq \Lambda(x) \preceq \lambda_{\max} I, \quad \text{for all } x.$$

- ▶ Coverage: for any x , we have

$$\max_{a \in \mathbb{A}} |f(x, a)|^2 \leq c_{\mathbb{A}} \cdot \mathbb{E}_{A \sim \pi_0(\cdot|x)} [|f(x, A)|^2], \quad \text{for all } f \in \mathcal{F}_q - \mathcal{F}_q.$$

Imposed on action space, generally unavoidable for off-policy.



NO {linear MDP, completeness, state-space coverage}

(projected) Bellman equations

Fixed-point equations with function approximation.

$$\begin{cases} v^* = \mathcal{T}_v[v^*, q^*], \\ q^* = \mathcal{T}_q[v^*, q^*]. \end{cases}$$

Bellman operators (simplified)

$$\mathcal{T}_v[v, q](x) = \eta r^{\pi_0}(x) + e^{-\beta\eta} \mathbb{E}^{\pi_0} [v(X_\eta) + \eta \max_{a \in \mathbb{A}} q(X_\eta, a) | X_0 = x],$$

$$\mathcal{T}_q[v, q](x, a) = \frac{e^{-\beta\eta}}{\eta} (\mathbb{E}^a - \mathbb{E}^{\pi_0}) [v(X_\eta) | X_0 = x].$$

(ignored $O(\eta)$ numerical error in the reward)

(projected) Bellman equations

Fixed-point equations with function approximation.

$$\begin{cases} v^* = \mathcal{T}_v[v^*, q^*], \\ q^* = \mathcal{T}_q[v^*, q^*]. \end{cases} \quad \begin{cases} \bar{v} = \Pi_{\mathcal{F}_v} \circ \mathcal{T}_v[\bar{v}, \bar{q}], \\ \bar{q} = \Pi_{\mathcal{F}_q} \circ \mathcal{T}_q[\bar{v}, \bar{q}]. \end{cases}$$

Bellman operators (simplified)

$$\mathcal{T}_v[v, q](x) = \eta r^{\pi_0}(x) + e^{-\beta\eta} \mathbb{E}^{\pi_0} [v(X_\eta) + \eta \max_{a \in \mathbb{A}} q(X_\eta, a) | X_0 = x],$$

$$\mathcal{T}_q[v, q](x, a) = \frac{e^{-\beta\eta}}{\eta} (\mathbb{E}^a - \mathbb{E}^{\pi_0}) [v(X_\eta) | X_0 = x].$$

(ignored $O(\eta)$ numerical error in the reward)

(projected) Bellman equations

Fixed-point equations with function approximation.

$$\begin{cases} v^* = \mathcal{T}_v[v^*, q^*], \\ q^* = \mathcal{T}_q[v^*, q^*]. \end{cases} \quad \begin{cases} \bar{v} = \Pi_{\mathcal{F}_v} \circ \mathcal{T}_v[\bar{v}, \bar{q}], \\ \bar{q} = \Pi_{\mathcal{F}_q} \circ \mathcal{T}_q[\bar{v}, \bar{q}]. \end{cases} \quad \begin{cases} \hat{v} = \hat{\Pi}_{\mathcal{F}_v} \circ \hat{\mathcal{T}}_v[\hat{v}, \hat{q}], \\ \hat{q} = \hat{\Pi}_{\mathcal{F}_q} \circ \hat{\mathcal{T}}_q[\hat{v}, \hat{q}]. \end{cases}$$

Bellman operators (simplified)

$$\mathcal{T}_v[v, q](x) = \eta r^{\pi_0}(x) + e^{-\beta\eta} \mathbb{E}^{\pi_0} [v(X_\eta) + \eta \max_{a \in \mathbb{A}} q(X_\eta, a) | X_0 = x],$$

$$\mathcal{T}_q[v, q](x, a) = \frac{e^{-\beta\eta}}{\eta} (\mathbb{E}^a - \mathbb{E}^{\pi_0}) [v(X_\eta) | X_0 = x].$$

(ignored $O(\eta)$ numerical error in the reward)

Structures in the Bellman operators

Bellman operators (simplified)

$$\mathcal{T}_v[v, q](x) = \eta r^{\pi_0}(x) + e^{-\beta\eta} \mathbb{E}^{\pi_0} [v(X_\eta) + \eta \max_{a \in \mathbb{A}} q(X_\eta, a) | X_0 = x],$$

$$\mathcal{T}_q[v, q](x, a) = \frac{e^{-\beta\eta}}{\eta} (\mathbb{E}^a - \mathbb{E}^{\pi_0}) [v(X_\eta) | X_0 = x].$$

Structures in the Bellman operators

Bellman operators (simplified)

$$\mathcal{T}_v[v, q](x) = \eta r^{\pi_0}(x) + e^{-\beta\eta} \mathbb{E}^{\pi_0} [v(X_\eta) + \eta \max_{a \in \mathbb{A}} q(X_\eta, a) | X_0 = x],$$

$$\mathcal{T}_q[v, q](x, a) = \frac{e^{-\beta\eta}}{\eta} (\mathbb{E}^a - \mathbb{E}^{\pi_0}) [v(X_\eta) | X_0 = x].$$

- ▶ $\mathcal{T}_v$ is contraction under ℓ^∞ norm.
- ▶ $\mathcal{T}_q$ relates q to v via a linear operator.
- ▶ Anything else?

Structures in the Bellman operators

Bellman operators (simplified)

$$\mathcal{T}_v[v, q](x) = \eta r^{\pi_0}(x) + e^{-\beta\eta} \mathbb{E}^{\pi_0} [v(X_\eta) + \eta \max_{a \in \mathbb{A}} q(X_\eta, a) | X_0 = x],$$

$$\mathcal{T}_q[v, q](x, a) = \frac{e^{-\beta\eta}}{\eta} (\mathbb{E}^a - \mathbb{E}^{\pi_0}) [v(X_\eta) | X_0 = x].$$

- ▶ $\mathcal{T}_v$ is contraction under ℓ^∞ norm.
- ▶ $\mathcal{T}_q$ relates q to v via a linear operator.
- ▶ Anything else?

Key observation: for $\eta \rightarrow 0$

$$\frac{1}{\eta} (v - \mathcal{T}_v[v, q]) \approx r^{\pi_0} + (\beta - \mathcal{L}^{\pi_0})v - \max_{a \in \mathbb{A}} q(\cdot, a),$$

$$\text{where } \mathcal{L}^{\pi_0} f = b^{\pi_0} \cdot \nabla f + \frac{1}{2} \text{tr}(\Lambda^{\pi_0} \nabla^2 f).$$

Diffusion generator $\mathcal{L}^{\pi_0}$ is a second-order elliptic operator.

Blessing of ellipticity for control

Ellipticity-induced ℓ^2 structure again!

Blessing of ellipticity for control

Ellipticity-induced ℓ^2 structure again!

Lemma (“strong monotonicity”)

For v_1, v_2, q_1, q_2 , we have

$$\begin{aligned} \frac{1}{\eta} \langle (v_1 - \mathcal{T}_v[v_1, q_1]) - (v_2 - \mathcal{T}_v[v_2, q_2]), v_1 - v_2 \rangle_{\mathbb{L}^2(\mu)} \\ \geq \lambda_{\min} \|v_1 - v_2\|_{\mathbb{H}^1(\mu)}^2 - c \|q_1 - q_2\|_{\mathbb{L}^2(\mu)}^2 \cdot \|v_1 - v_2\|_{\mathbb{L}^2(\mu)} - O(\eta). \end{aligned}$$

Blessing of ellipticity for control

Ellipticity-induced ℓ^2 structure again!

Lemma (“strong monotonicity”)

For v_1, v_2, q_1, q_2 , we have

$$\begin{aligned} \frac{1}{\eta} \langle (v_1 - \mathcal{T}_v[v_1, q_1]) - (v_2 - \mathcal{T}_v[v_2, q_2]), v_1 - v_2 \rangle_{\mathbb{L}^2(\mu)} \\ \geq \lambda_{\min} \|v_1 - v_2\|_{\mathbb{H}^1(\mu)}^2 - c \|q_1 - q_2\|_{\mathbb{L}^2(\mu)}^2 \cdot \|v_1 - v_2\|_{\mathbb{L}^2(\mu)} - O(\eta). \end{aligned}$$

Lemma (“Lipschitzness”)

For v_1, v_2, q_1, q_2, v , we have

$$\begin{aligned} \left| \frac{1}{\eta} \langle (v_1 - \mathcal{T}_v[v_1, q_1]) - (v_2 - \mathcal{T}_v[v_2, q_2]), v \rangle_{\mathbb{L}^2(\mu)} \right| \\ \leq \|v_1 - v_2\|_{\mathbb{H}^1(\mu)} \cdot \|v\|_{\mathbb{H}^1(\mu)} + c \|q_1 - q_2\|_{\mathbb{L}^2(\mu)} \cdot \|v_1 - v_2\|_{\mathbb{L}^2(\mu)} + O(\eta). \end{aligned}$$

Sobolev-prox fitted Q -iteration

Population-level iterates

$$q^{(t+1)} = \arg \min_{q \in \mathcal{F}_q} \|q - \mathcal{T}_q[v^{(t)}, q^{(t)}]\|_{\mathbb{L}^2(\mu)}^2,$$

$$v^{(t+1)} = \arg \min_{v \in \mathcal{F}_v} \|v - v^{(t)}\|_{\mathbb{H}^1(\mu)}^2 + \frac{2\alpha}{\eta} \langle v^{(t)} - \mathcal{T}_v[v^{(t)}, q^{(t+1)}], v - v^{(t)} \rangle_{\mathbb{L}^2(\mu)}.$$

Sobolev-prox fitted Q -iteration

Population-level iterates

$$q^{(t+1)} = \arg \min_{q \in \mathcal{F}_q} \|q - \mathcal{T}_q[v^{(t)}, q^{(t)}]\|_{\mathbb{L}^2(\mu)}^2,$$

$$v^{(t+1)} = \arg \min_{v \in \mathcal{F}_v} \|v - v^{(t)}\|_{\mathbb{H}^1(\mu)}^2 + \frac{2\alpha}{\eta} \langle v^{(t)} - \mathcal{T}_v[v^{(t)}, q^{(t+1)}], v - v^{(t)} \rangle_{\mathbb{L}^2(\mu)}.$$

- Replace $\mathbb{L}^2$ norm in fitted Q -learning with Sobolev norm for v -update.
- $\alpha \ll 1$: faster time scale for q -update, slower for v -update.

Sobolev-prox fitted Q -iteration

Population-level iterates

$$q^{(t+1)} = \arg \min_{q \in \mathcal{F}_q} \|q - \mathcal{T}_q[v^{(t)}, q^{(t)}]\|_{\mathbb{L}^2(\mu)}^2,$$

$$v^{(t+1)} = \arg \min_{v \in \mathcal{F}_v} \|v - v^{(t)}\|_{\mathbb{H}^1(\mu)}^2 + \frac{2\alpha}{\eta} \langle v^{(t)} - \mathcal{T}_v[v^{(t)}, q^{(t+1)}], v - v^{(t)} \rangle_{\mathbb{L}^2(\mu)}.$$

- ▶ Replace $\mathbb{L}^2$ norm in fitted Q -learning with Sobolev norm for v -update.
- ▶ $\alpha \ll 1$: faster time scale for q -update, slower for v -update.

Sobolev-prox fitted Q -iteration

Population-level iterates

$$q^{(t+1)} = \arg \min_{q \in \mathcal{F}_q} \|q - \mathcal{T}_q[v^{(t)}, q^{(t)}]\|_{\mathbb{L}^2(\mu)}^2,$$

$$v^{(t+1)} = \arg \min_{v \in \mathcal{F}_v} \|v - v^{(t)}\|_{\mathbb{H}^1(\mu)}^2 + \frac{2\alpha}{\eta} \langle v^{(t)} - \mathcal{T}_v[v^{(t)}, q^{(t+1)}], v - v^{(t)} \rangle_{\mathbb{L}^2(\mu)}.$$

- ▶ Replace $\mathbb{L}^2$ norm in fitted Q -learning with Sobolev norm for v -update.
- ▶ $\alpha \ll 1$: faster time scale for q -update, slower for v -update.

Empirical data version

$$\hat{q}^{(t+1)} = \arg \min_{q \in \mathcal{F}_q} \|q - \hat{\mathcal{T}}_q[\hat{v}^{(t)}, \hat{q}^{(t)}]\|_{\mathbb{L}^2(\hat{\mu})}^2,$$

$$\hat{v}^{(t+1)} = \arg \min_{v \in \mathcal{F}_v} \|v - \hat{v}^{(t)}\|_{\mathbb{H}^1(\hat{\mu})}^2 + \frac{2\alpha}{\eta} \langle \hat{v}^{(t)} - \hat{\mathcal{T}}_v[\hat{v}^{(t)}, \hat{q}^{(t+1)}], v - \hat{v}^{(t)} \rangle_{\mathbb{L}^2(\hat{\mu})}.$$

Theoretical guarantees

Localized complexities and critical radii

$$\begin{aligned}(r_v^*)^2 &= \frac{1}{\sqrt{n}} \mathcal{J}_2(\mathcal{F}_v \cap \mathbb{B}_{\mathbb{H}^1}(r_v^*); \|\cdot\|_{\mathbb{H}^1}), \\ (r_q^*)^2 &= \frac{1}{\sqrt{n}} \mathcal{J}_2(\mathcal{F}_q \cap \mathbb{B}_{\mathbb{L}^2}(r_q^*); \|\cdot\|_{\mathbb{L}^2}),\end{aligned}$$

where $\mathcal{J}_2$ is the Dudley entropy integral – canonical in statistical learning theory.

Theoretical guarantees

Localized complexities and critical radii

$$(r_v^*)^2 = \frac{1}{\sqrt{n}} \mathcal{J}_2(\mathcal{F}_v \cap \mathbb{B}_{\mathbb{H}^1}(r_v^*); \|\cdot\|_{\mathbb{H}^1}),$$

$$(r_q^*)^2 = \frac{1}{\sqrt{n}} \mathcal{J}_2(\mathcal{F}_q \cap \mathbb{B}_{\mathbb{L}^2}(r_q^*); \|\cdot\|_{\mathbb{L}^2}),$$

where $\mathcal{J}_2$ is the Dudley entropy integral – canonical in statistical learning theory.

Theorem (Mou, 2026)

For $\beta \geq c_0$, given n samples, N steps of Sobolev-prox fitted Q -iterations yields

$$\begin{aligned} \|\hat{v}_n - v^*\|_{\mathbb{H}^1}^2 + \|\hat{q}_n - q^*\|_{\mathbb{L}^2}^2 &\lesssim \underbrace{\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2}_{\text{Approx. Err.}} \\ &+ \underbrace{e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2}_{\text{Opt. Err.}} + \underbrace{(r_v^*)^2 + (r_q^*)^2}_{\text{Stat. Err.}} + \underbrace{\eta}_{\text{Num. Err.}}. \end{aligned}$$

Theoretical guarantees

Theorem (Mou, 2026)

For $\beta \geq c_0$, given n samples, N steps of Sobolev-prox fitted Q -iterations yields

$$\begin{aligned} \|\hat{v}_n - v^*\|_{\mathbb{H}^1}^2 + \|\hat{q}_n - q^*\|_{\mathbb{L}^2}^2 &\lesssim \underbrace{\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2}_{\text{Approx. Err.}} \\ &+ \underbrace{e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2}_{\text{Opt. Err.}} + \underbrace{(r_v^*)^2 + (r_q^*)^2}_{\text{Stat. Err.}} + \underbrace{\eta}_{\text{Num. Err.}}. \end{aligned}$$

- ▶ **Approximation:** oracle error $\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2$.
- ▶ **Estimation:** critical radii $(r_v^*)^2 + (r_q^*)^2$.
- ▶ **Optimization:** exponential convergence $e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2$.
- ▶ Dream comes true!

Theoretical guarantees

Theorem (Mou, 2026)

For $\beta \geq c_0$, given n samples, N steps of Sobolev-prox fitted Q -iterations yields

$$\begin{aligned} \|\hat{v}_n - v^*\|_{\mathbb{H}^1}^2 + \|\hat{q}_n - q^*\|_{\mathbb{L}^2}^2 &\lesssim \underbrace{\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2}_{\text{Approx. Err.}} \\ &\quad + \underbrace{e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2}_{\text{Opt. Err.}} + \underbrace{(r_v^*)^2 + (r_q^*)^2}_{\text{Stat. Err.}} + \underbrace{\eta}_{\text{Num. Err.}}. \end{aligned}$$

- ▶ **Approximation:** oracle error $\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2$.
- ▶ **Estimation:** critical radii $(r_v^*)^2 + (r_q^*)^2$.
- ▶ **Optimization:** exponential convergence $e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2$.
- ▶ Dream comes true!

Theoretical guarantees

Theorem (Mou, 2026)

For $\beta \geq c_0$, given n samples, N steps of Sobolev-prox fitted Q -iterations yields

$$\begin{aligned} \|\hat{v}_n - v^*\|_{\mathbb{H}^1}^2 + \|\hat{q}_n - q^*\|_{\mathbb{L}^2}^2 &\lesssim \underbrace{\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2}_{\text{Approx. Err.}} \\ &+ \underbrace{e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2}_{\text{Opt. Err.}} + \underbrace{(r_v^*)^2 + (r_q^*)^2}_{\text{Stat. Err.}} + \underbrace{\eta}_{\text{Num. Err.}}. \end{aligned}$$

- ▶ **Approximation:** oracle error $\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2$.
- ▶ **Estimation:** critical radii $(r_v^*)^2 + (r_q^*)^2$.
- ▶ **Optimization:** exponential convergence $e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2$.
- ▶ Dream comes true!

Theoretical guarantees

Theorem (Mou, 2026)

For $\beta \geq c_0$, given n samples, N steps of Sobolev-prox fitted Q -iterations yields

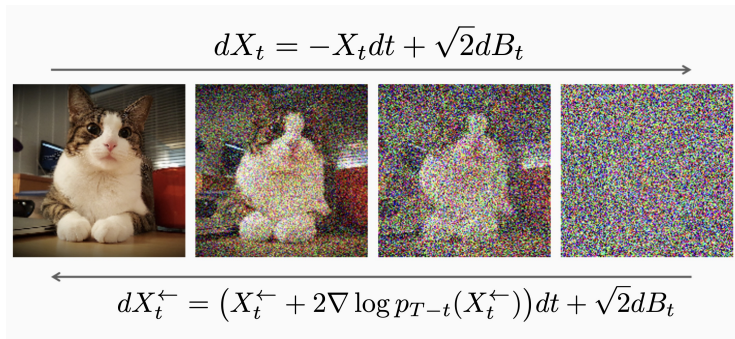
$$\begin{aligned} \|\hat{v}_n - v^*\|_{\mathbb{H}^1}^2 + \|\hat{q}_n - q^*\|_{\mathbb{L}^2}^2 &\lesssim \underbrace{\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2}_{\text{Approx. Err.}} \\ &\quad + \underbrace{e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2}_{\text{Opt. Err.}} + \underbrace{(r_v^*)^2 + (r_q^*)^2}_{\text{Stat. Err.}} + \underbrace{\eta}_{\text{Num. Err.}}. \end{aligned}$$

- ▶ **Approximation:** oracle error $\inf_{v \in \mathcal{F}_v} \|v - v^*\|_{\mathbb{H}^1}^2 + \inf_{q \in \mathcal{F}_q} \|q - q^*\|_{\mathbb{L}^2}^2$.
- ▶ **Estimation:** critical radii $(r_v^*)^2 + (r_q^*)^2$.
- ▶ **Optimization:** exponential convergence $e^{-\lambda_{\min} \alpha N} \|v_0 - v^*\|_{\mathbb{H}^1}^2$.
- ▶ Dream comes true!

Application to diffusion model fine-tuning

Diffusion model fine-tuning as continuous-time RL

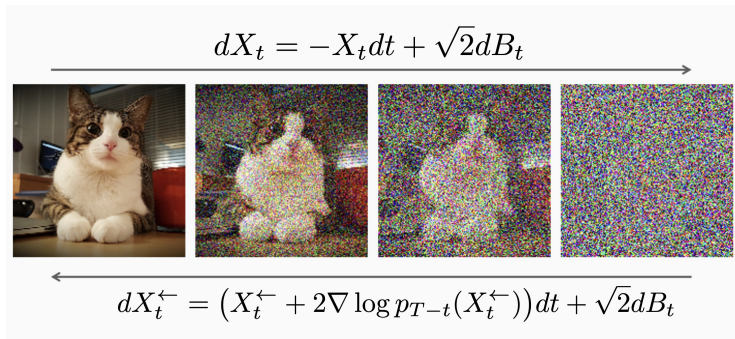
Diffusion generative models



- ▶ [SSDK⁺20, HJA20]: learn the score function to simulate the backward process.
- ▶ State-of-the-art model for vision tasks.

Diffusion model fine-tuning as continuous-time RL

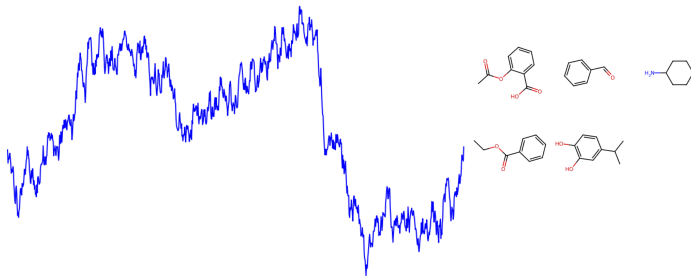
Diffusion generative models



- ▶ [SSDK⁺20, HJA20]: learn the score function to simulate the backward process.
- ▶ State-of-the-art model for vision tasks.

Fine-tuning of diffusion models

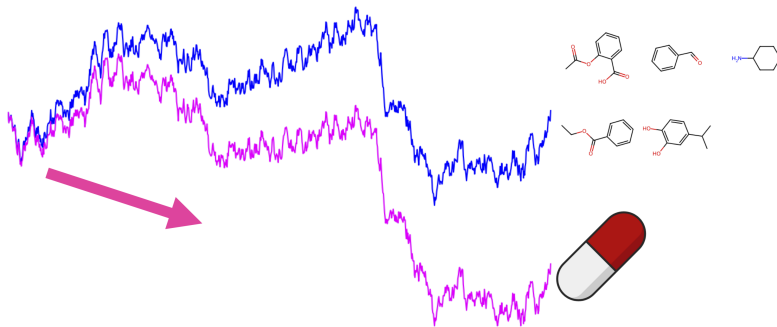
e.g. generative models for drug discovery



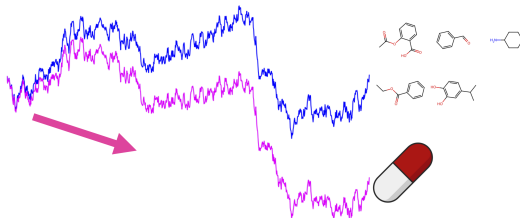
$$dX_t = b_t(X_t)dt + \sqrt{2}dB_t, \quad X_0 \sim \mu_0.$$

Fine-tuning of diffusion models

e.g. generative models for drug discovery



Fine-tuning of diffusion models



Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\min_{\pi} J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}},$$

$$\text{s.t. } dX_t^\pi = (b_t(X_t^\pi) + A_t)dt + \sqrt{2}dB_t, \quad X_0^\pi \sim \mu_0.$$

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\min_{\pi} J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}},$$

$$\text{s.t. } dX_t^\pi = (b_t(X_t^\pi) + A_t)dt + \sqrt{2}dB_t, \quad X_0^\pi \sim \mu_0.$$

- ▶ Entropy regularized framework [SWD⁺17, UZB⁺24, Tan24].
- ▶ Solution via existing RL methods [GZZ24, Tan24, HRX24].
- ▶ Connection to classifier guidance [DN21], learning score vs. log likelihood ratio.

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\min_{\pi} J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}},$$

$$\text{s.t. } dX_t^\pi = (b_t(X_t^\pi) + A_t)dt + \sqrt{2}dB_t, \quad X_0^\pi \sim \mu_0.$$

- ▶ Entropy regularized framework [SWD⁺17, UZB⁺24, Tan24].
- ▶ Solution via existing RL methods [GZZ24, Tan24, HRX24].
- ▶ Connection to classifier guidance [DN21], learning score vs. log likelihood ratio.

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\min_{\pi} J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}},$$

$$\text{s.t. } dX_t^\pi = (b_t(X_t^\pi) + A_t)dt + \sqrt{2}dB_t, \quad X_0^\pi \sim \mu_0.$$

- ▶ Entropy regularized framework [SWD⁺17, UZB⁺24, Tan24].
- ▶ Solution via existing RL methods [GZZ24, Tan24, HRX24].
- ▶ Connection to classifier guidance [DN21], learning score vs. log likelihood ratio.

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\begin{aligned} \min_{\pi} \quad & J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}}, \\ \text{s.t.} \quad & dX_t^\pi = (b_t(X_t^\pi) + A_t) dt + \sqrt{2} dB_t, \quad X_0^\pi \sim \mu_0. \end{aligned}$$

Structures in the fine-tuning problem

- ▶ Control-affine system with quadratic action cost – Cole–Hopf transformation on the value function.
- ▶ Known pre-trained model (dynamics) and unknown reward.
- ▶ Sobolev-prox algorithm for the HJB equation.

$$\hat{f}^{(k+1)} = \arg \min_{f \in \mathcal{F}} \left\{ \|f - \hat{f}^{(k)}\|_{\hat{\mathbb{H}}^1}^2 - 2\eta \hat{B}[f^* - \hat{f}^{(k)}, f - \hat{f}^{(k)}] \right\}.$$

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\begin{aligned} \min_{\pi} \quad & J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}}, \\ \text{s.t.} \quad & dX_t^\pi = (b_t(X_t^\pi) + A_t) dt + \sqrt{2} dB_t, \quad X_0^\pi \sim \mu_0. \end{aligned}$$

Structures in the fine-tuning problem

- ▶ Control-affine system with quadratic action cost – Cole–Hopf transformation on the value function.
- ▶ Known pre-trained model (dynamics) and unknown reward.
- ▶ Sobolev-prox algorithm for the HJB equation.

$$\hat{f}^{(k+1)} = \arg \min_{f \in \mathcal{F}} \left\{ \|f - \hat{f}^{(k)}\|_{\hat{\mathbb{H}}^1}^2 - 2\eta \hat{B}[f^* - \hat{f}^{(k)}, f - \hat{f}^{(k)}] \right\}.$$

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\begin{aligned} \min_{\pi} \quad & J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}}, \\ \text{s.t.} \quad & dX_t^\pi = (b_t(X_t^\pi) + A_t) dt + \sqrt{2} dB_t, \quad X_0^\pi \sim \mu_0. \end{aligned}$$

Structures in the fine-tuning problem

- ▶ Control-affine system with quadratic action cost – Cole–Hopf transformation on the value function.
- ▶ Known pre-trained model (dynamics) and unknown reward.
- ▶ Sobolev-prox algorithm for the HJB equation.

$$\hat{f}^{(k+1)} = \arg \min_{f \in \mathcal{F}} \left\{ \|f - \hat{f}^{(k)}\|_{\hat{\mathbb{H}}^1}^2 - 2\eta \hat{B}[f^* - \hat{f}^{(k)}, f - \hat{f}^{(k)}] \right\}.$$

Fine-tuning of diffusion models

Fine-tuning as control $A_t = \pi_t(X_t^\pi)$.

$$\begin{aligned} \min_{\pi} \quad & J(\pi) := \mathbb{E} \left[y(X_T^\pi) + \int_0^T r_t(X_t^\pi) dt \right] + \underbrace{\alpha D_{\text{KL}}(\mathbb{P}_{[0,T]}^\pi \parallel \mathbb{P}_{[0,T]}^0)}_{\text{regularization}}, \\ \text{s.t.} \quad & dX_t^\pi = (b_t(X_t^\pi) + A_t) dt + \sqrt{2} dB_t, \quad X_0^\pi \sim \mu_0. \end{aligned}$$

Structures in the fine-tuning problem

- ▶ Control-affine system with quadratic action cost – Cole–Hopf transformation on the value function.
- ▶ Known pre-trained model (dynamics) and unknown reward.
- ▶ Sobolev-prox algorithm for the HJB equation.

$$\hat{f}^{(k+1)} = \arg \min_{f \in \mathcal{F}} \left\{ \|f - \hat{f}^{(k)}\|_{\hat{\mathbb{H}}^1}^2 - 2\eta \hat{B}[f^* - \hat{f}^{(k)}, f - \hat{f}^{(k)}] \right\}.$$

Results for diffusion model fine-tuning

Closer look at critical radii in statistical learning theory

$$\begin{aligned}\varepsilon_n^*(\sigma) : \quad & \sigma \cdot \mathcal{J}_2(\mathcal{F} \cap \mathbf{B}(\varepsilon)) = \varepsilon^2, \\ \tilde{\varepsilon}_n : \quad & \mathcal{J}_2(\mathcal{F} \cap \mathbf{B}(\varepsilon)) = \tilde{\varepsilon},\end{aligned}$$

where σ is the noise level.

- ▶ **Low SNR regime:** $\sigma \asymp 1$, effect of noise dominates, “standard rate”.
- ▶ **High SNR regime:** $\sigma \ll 1$, interpolation error may dominate. “fast rate”.
- ▶ Bias-variance trade-off via function class $\mathcal{F}$.

Results for diffusion model fine-tuning

Closer look at critical radii in statistical learning theory

$$\begin{aligned}\varepsilon_n^*(\sigma) : \quad & \sigma \cdot \mathcal{J}_2(\mathcal{F} \cap \mathbf{B}(\varepsilon)) = \varepsilon^2, \\ \tilde{\varepsilon}_n : \quad & \mathcal{J}_2(\mathcal{F} \cap \mathbf{B}(\varepsilon)) = \tilde{\varepsilon},\end{aligned}$$

where σ is the noise level.

- ▶ **Low SNR regime:** $\sigma \asymp 1$, effect of noise dominates, “standard rate”.
- ▶ **High SNR regime:** $\sigma \ll 1$, interpolation error may dominate. “fast rate”.
- ▶ Bias-variance trade-off via function class $\mathcal{F}$.

Theorem (Mou, 2025)

Under suitable assumptions, N steps of Sobolev-prox with n samples yields

$$\begin{aligned}\|\hat{f}^{(N)} - f^*\|_{\mathbb{H}^1} \lesssim & \underbrace{\inf_{f \in \mathcal{F}} \|f - f^*\|_{\mathbb{H}^1}}_{\text{Approx. Err.}} + \underbrace{e^{-c\eta N} \|f^{(0)} - f^*\|_{\mathbb{H}^1}}_{\text{Optim. Err.}} \\ & + \underbrace{\varepsilon_n^*(\|\bar{f} - f^*\|_{\mathbb{W}^{1,p}})}_{\text{Stat. Err.}} + \tilde{\varepsilon}_n (+ \text{high-order terms}).\end{aligned}$$

Results for diffusion model fine-tuning

Closer look at critical radii in statistical learning theory

$$\begin{aligned}\varepsilon_n^*(\sigma) : \quad & \sigma \cdot \mathcal{J}_2(\mathcal{F} \cap \mathbf{B}(\varepsilon)) = \varepsilon^2, \\ \tilde{\varepsilon}_n : \quad & \mathcal{J}_2(\mathcal{F} \cap \mathbf{B}(\varepsilon)) = \tilde{\varepsilon},\end{aligned}$$

where σ is the noise level.

Theorem (Mou, 2025)

Under suitable assumptions, N steps of Sobolev-prox with n samples yields

$$\begin{aligned}\|\hat{f}^{(N)} - f^*\|_{\mathbb{H}^1} \lesssim & \underbrace{\inf_{f \in \mathcal{F}} \|f - f^*\|_{\mathbb{H}^1}}_{\text{Approx. Err.}} + \underbrace{e^{-c\eta N} \|f^{(0)} - f^*\|_{\mathbb{H}^1}}_{\text{Optim.Err.}} \\ & + \underbrace{\varepsilon_n^*(\|\bar{f} - f^*\|_{\mathbb{W}^{1,p}})}_{\text{Stat.Err.}} + \tilde{\varepsilon}_n (+ \text{high-order terms}).\end{aligned}$$

- ▶ Self-mitigating error: effective noise level $\|\bar{f} - f^*\|_{\mathbb{W}^{1,p}}$ depends on the approximation error.
- ▶ Non-standard bias-variance tradeoff: approx. and stat. errors may decrease together as $\mathcal{F}$ grows.

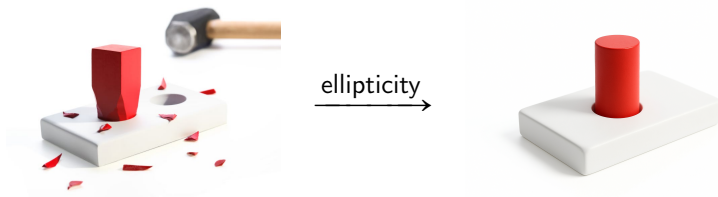
Discussion



ellipticity
→



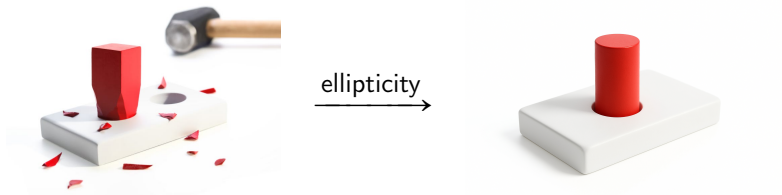
Discussion



More algorithms?

- ▶ Policy-based? Actor-critic?
- ▶ From offline estimation to online learning?

Discussion



More algorithms?

- ▶ Policy-based? Actor-critic?
- ▶ From offline estimation to online learning?

More structures?

- ▶ Continuous spaces: beyond uniform ellipticity.
- ▶ Discrete spaces: elliptic-like structures?

Thank you!