1. Preliminaries

Bellman Optim. Operator: $(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[\max_{a'} Q(s',a')]$

1.1 Stochastic Processes

DTMC Rems: 1) The transition matrix (stochastic Matrix) satisfies: $P_{ij} \ge 0$, $\sum_j P_{ij} = 1$ $\forall i$. 2) Let $X_0 \sim \mu_0$, then $X_t \sim \mu_t = \mu_0 P^t$. 3) μ is called the stationary distr. of a DTMC with transition Mat P if $\mu = \mu P$

Thm: DTMC is said to be irreducible if all states communicate, i.e. $\forall i,j\ \exists \tau < \infty: P(X_\tau = j | X_0 = i) > 0$

Thm: DTMC said to be aperiodic if $\forall s \in S\ \exists k_0 \in \mathbb{N}$: $\forall k \ge k_0: p^{(k)}(s|s) > 0$ ⟺ transition graph has a closed loop from s to s $\forall$ lengths $k \ge k_0 \in \mathbb{N}$

Thm: Any finite state DTMC has a stationary distr. μ s.t. $\mu = \mu P$. If finite-state DTMC is irreducible, then μ is unique. If DTMC is aperiodic, $\{X_t\}$ converges to it: $\lim_{t\to\infty} P(X_t = j | X_0 = i) = \mu_j$

1.2 Optimization:

Linear Programming: Primal: $\min c^T x$ s.t. $Ax \ge b$, Dual: $\max b^T y$ s.t. $A^T y = c$, $y \ge 0$

Thm: If one of the them (primal &dual) has an optimal sol., then so does the other & their optimal values are same

Convexity: $f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y)$ ⟺ $f(x) \ge f(x_0) + \nabla f(x_0)^T (x - x_0)$ ⟺ $\nabla^2 f(x) \succeq 0$

Descent Methods: Goal: $\min_x f(x)$. Scheme: $x_{t+1} \leftarrow x_t + \gamma_t d_t$. d_t s.t. $\nabla f(x_t)^T d_t < 0$. Eg.: 1) GD: $d_t = -\nabla f(x_t)$, 2) Natural GD $d_t := Q_t \nabla f(x_t)$, Q psd, 3) Newton's Method: $d_t = -[\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$

Stepsize: $\gamma_t > 0$ s.t. $f(x_{t+1}) < f(x_t)$. Eg: 1) const: $\gamma_t = \gamma$, 2) Diminishing Stepsize: $\gamma_t = \frac{1}{t}$, 3) Exact line search: $\gamma_t = \arg\min_\gamma f(x_t + \gamma d_t)$

Convergence & Complexity: Given $\{x_t\}_t$ & Accuracy Measure $\varepsilon(x_t)$: $\lim_{t\to\infty} \frac{\varepsilon(x_{t+1})}{\varepsilon(x_t)^p} = q$ ⟹ 1) Lin. Conv.: $p=1$, $q \in (0,1)$ e.g. $\varepsilon(x_t) = \mathcal{O}(e^{-t})$, 2) Sublinear Conv: $p=1$, $q=1$, e.g. $\varepsilon(x_t) = \mathcal{O}(\frac{1}{t})$, 3) Superlinear conv.: $p=1$, $q=0$, e.g.: $\varepsilon(x_t) = \mathcal{O}(e^{-t^2})$ 4) Conv. of order p: $p > 1$, $q > 0$, when $p=2$ "quadratic conv." e.g. $\varepsilon(x_t) = \mathcal{O}(e^{-a p^t})$

Thm (Conv of GD): $x_{t+1} \leftarrow x_t - \gamma \nabla f(x_t)$

Suppose f is cont. diff'able with L-Lipschitz gradients. Let $\gamma = \frac{1}{L}$

1) In general: $\min_{k=0,\dots,t} \|\nabla f(x_k)\|_2^2 \le \frac{2L}{t+1}(f(x_0) - f(x^*))$
2) If f is convex: $f(x_t) - f(x^*) \le \frac{2L}{t} \|x_0 - x^*\|_2^2$
3) Strongly convex: If f is strongly convex ($f(x) - \frac{\mu}{2}\|x\|_2^2$ convex) ↳ $f(x_t) - f(x^*) \le (1 - \frac{\mu}{L})^t [f(x_0) - f(x^*)]$

Stochastic Optim: Let $f(x) = \mathbb{E}_{\xi \sim p(.)}[F(x,\xi)] \Rightarrow \nabla_x f(x) = \mathbb{E}_\xi[\nabla_x F(x,\xi)]$

↳ unbiased estim: $\hat{\nabla} f(x) = \nabla F(x,\xi)$, $\xi \sim p(.)$, $\hat\nabla_x f(x) = \frac{1}{b}\sum_{j=1}^b \nabla_x F(x,\xi_j)$

SGD: $x_{t+1} \leftarrow x_t - \gamma_t \nabla F(x_t, \xi_t)$, $\xi_t \sim P(.)$ → unbiased estim

↳ Rem: SGD is not a monotonic descent method.

Thm: Assume: (i) $f(x) = \mathbb{E}_\xi[F(x,\xi)]$ is L-smooth, (ii) $\forall x: \mathbb{E}[\|\nabla F(x,\xi) - \nabla f(x)\|_2^2] \le \sigma^2$

↳ Then SGD w/ $\gamma_t = \min\{\frac{1}{L}, \frac{\gamma}{\sigma\sqrt{T}}\}$ achieves: $\min_{t=0,\dots,T-1} \mathbb{E}[\|\nabla f(x_t)\|_2^2] \le \frac{2L(F(x_0) - F(x^*))}{T} + \frac{\sigma}{\sqrt{T}}\left(\frac{2[f(x_0) - f(x^*)]}{\gamma} + L\gamma\right)$

Fixed Point Problem: Let $\mathcal{T}: \mathbb{R}^d \to \mathbb{R}^d$ be a operator. The fixed-point problem aims to find a sol x^* s.t. $\mathcal{T}x^* = x^*$

Def: $\mathcal{T}$ is a contraction wrt. norm $\|.\|$ if for some $0 < \rho < 1$: $\|\mathcal{T}x - \mathcal{T}y\| \le \rho\|x - y\|$ ⟹ Banach fpt: $\exists!$ fp x^*

FP-Algo: $x_{t+1} \leftarrow \mathcal{T}x_t$ yields Linear conv.: $\|x_t - x^*\| \le \rho^t \|x_0 - x^*\|$

Def: an operator $\mathcal{T}$ is non-expansive if: $\|\mathcal{T}x - \mathcal{T}y\| \le \|x - y\|$

Relaxed FP-Algo: $x_{t+1} \leftarrow (1-\gamma)x_t + \gamma \mathcal{T}x_t$ for $\gamma \in (0,1)$ ⟹ Sublin. Conv. $\|\mathcal{T}x_t - x_t\|_2^2 \le \frac{\|x_0 - x^*\|_2^2}{\gamma(1-\gamma)(t+1)}$

2. Dynamic & Linear Programming

2.1 MDPs:

Def: MDP ≈ controlled MC + performance obj.. An MDP consists of S: State space, $\mathcal{A}$: action space, $P(s'|s,a): S \times \mathcal{A} \to \Delta(S)$ transit. model, $r(s,a): S \times \mathcal{A} \to \mathbb{R}$ reward fct., μ_0: init. state distr.: $s_0 \sim \mu_0$

Def: Controlled Markov Property: $P(s_{t+1} = s' | \dots) = P(s'|s,a)$

Performance Criteria: Finite Horizon: Cumul. Rew.: $\mathbb{E}[\sum_{t=0}^{T-1} r(s_t,a_t)]$, Average Reward: $\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t,a_t)]$ $\gamma = 0$ ⟹ immediate reward

∞-horizon: $\gamma \in [0,1]$: Discounted Rew.: $\mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)]$ $\gamma^0 = 1$

Average Rew: $\liminf_{T\to\infty} \mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t,a_t)]$ $|r(\dots)| \le D$

Rem: $\forall \varepsilon > 0, T \ge \mathcal{O}(\frac{1}{1-\gamma}\log(D/\varepsilon)) \Rightarrow \sum_{t=T}^\infty \gamma^t r(s_t,a_t) \le \varepsilon$

State Value Fct: $V^\pi(s) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 = s, \pi]$

State-Action Value Fct.: $Q^\pi(s,a) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 = s, a_0 = a, \pi]$

Rem: $Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^\pi(s')$, $V^\pi(s,a) = \sum_a \pi(a|s) Q^\pi(s,a)$ ⟹ Bellman eq!

Def: $V^*(s) = \max_\pi V^\pi(s)$, $Q^*(s,a) := \max_\pi Q^\pi(s,a)$

↳ $V^*(s) = \max_a Q^*(s,a)$

Thm: For an ∞-horizon MDP $M = (S, \mathcal{A}, P, \mu_0, \gamma)$ ∃ stationary & deterministic policy π s.t. $\forall (s,a) \in S \times \mathcal{A}$: $V^\pi(s) = V^*(s)$, $Q^\pi(s,a) = Q^*$

↳ $\pi^*(s) := \arg\max_{a \in \mathcal{A}} Q^*(s,a)$

2.2 Bellman Eq's & Bellman Optimality

Bellman Consistency Eq: $V^\pi(s) = \mathbb{E}_{a \sim \pi(.|s)}[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^\pi(s')]$

↳ Matrix Form: $V^\pi = R^\pi + \gamma P^\pi V^\pi$, $R^\pi_s = \sum_a \pi(a|s) r(s,a)$

↳ $V^\pi = (\mathbb{1} - \gamma P^\pi)^{-1} R^\pi$, always invertible for $\gamma \in [0,1)$

Bellman Expectation Operator: Def $\mathcal{T}^\pi: \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ s.t. $\mathcal{T}^\pi V := R^\pi + \gamma P^\pi V$. V^π is a fixed point of $\mathcal{T}^\pi$

↳ Contraction ⟹ $\lim_{t\to\infty} (\mathcal{T}^\pi)^t V_0 = V^\pi$

Bellman Optimality Eq.: $V^*(s) = \max_{a \in \mathcal{A}}[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')]$
$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \cdot (\max_{a' \in \mathcal{A}} Q^*(s',a'))$

Bellman Optimality Operator: $(\mathcal{T}V)(s) := \max_{a \in \mathcal{A}}[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')]$
optim val fct V^* is the fixed point: $\mathcal{T}V^* = V^*$

Contraction Property: Bellman Optim. Op. $\mathcal{T}$ defined above is a γ-contraction mapping under ℓ_∞ norm: $\forall V, V' \in \mathbb{R}^{|S|}$: $\|\mathcal{T}V' - \mathcal{T}V\|_\infty \le \gamma \|V' - V\|_\infty$, $\|x\|_\infty := \max_i |x_i|$

Proof: $|(\mathcal{T}V')(s) - (\mathcal{T}V)(s)| = |\max_a[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V'(s')] - \max_a[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')]|$
$\le \max_a \gamma \sum_{s'} P(s'|s,a) |V'(s') - V(s')|$
$\le \|V' - V\|_\infty \max_a \gamma \sum_{s'} P(s'|s,a) = \gamma \|V' - V\|_\infty$ □

2.3 Value Iteration: Goal: find V^* ⟺ find FP of $\mathcal{T}$

Algo: Start with arbitrary guess V_0 (e.g. $V = 0$). For $t \in 1:T$ apply $\mathcal{T}$: $V_{t+1} \leftarrow \mathcal{T}V_t$ ($\mathcal{T}$ Bellman optim op.)

Rem: After obtaining V^* by VI, we can obtain an optimal policy from the greedy policy: $\pi^*(s) = \arg\max_a[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')]$

Equiv: Q-value Iter.: $Q^* \Rightarrow \pi^*(s) = \arg\max_a Q^*(s,a)$

↳ Per Iteration cost of VI: $\mathcal{O}(|S|^2 |\mathcal{A}|)$

Linear Conv.: $\|V_t - V^*\|_\infty = \|\mathcal{T}V_{t-1} - \mathcal{T}V^*\|_\infty \le \gamma \|V_{t-1} - V^*\| \le \dots \le \gamma^t \|V_0 - V^*\|_\infty$

2.4 Policy Iteration: Directly search for π^*. 2 steps:

1) Policy evaluation: Compute V^π of current policy
2) Policy improvement: Update π by greedy policy wrt V^π

Policy Improvement Thm: If a (deterministic) policy π' satisfies $Q^\pi(s, \pi'(s)) \ge V^\pi(s)$ $\forall s \in S$, then $V^{\pi'}(s) \ge V^\pi(s)$ $\forall s \in S$

Algo: $\pi_{t+1}(s) \leftarrow \arg\max_a Q^{\pi_t}(s,a)$

Proof: Let $s \in S$ & $s' \sim P(.|s, \pi'(s))$: $V^\pi(s) \le Q^\pi(s, \pi'(s))$
$= \mathbb{E}_{\pi'}[r(s_0, \pi'(s_0)) + \gamma V^\pi(s_1) | s_0 = s] \le \mathbb{E}_{\pi'}[r_0 + \gamma Q^\pi(s_1, \pi'(s_1)) | s_0 = s]$
$\le \mathbb{E}_{\pi'}[r_0 + \gamma r_1 + \gamma^2 V^\pi(s_2) | s_0 = s] \le \dots \le \mathbb{E}_{\pi'}[r_0 + \gamma r_1 + \gamma^2 r_2 + \dots | s_0 = s] = V^{\pi'}(s)$

PI Algo: Start with arbitrary π_0. For $t = 1:T$ do:
1) Compute V^π: (Option 1) VI: $V_t \leftarrow \mathcal{T}^{\pi_t} V_t$ until conv. (Option 2): Closed form sol: $V_t = (\mathbb{1} - \gamma P^{\pi_t})^{-1} R^{\pi_t}$
2) Policy Improvement: $\pi_{t+1}(s) = \arg\max_a[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi_t}(s')]$

Rem: PI converges in finitely many steps

Thm: PI attains Lin. conv. rate: $\|V^{\pi_t} - V^*\|_\infty \le \gamma^t \|V^{\pi_0} - V^*\|_\infty$

Proof: $V^{\pi_{t+1}}(s) \ge V^{\pi_t}(s)$ $\forall s$, let $a = \pi_{t+1}(s)$ ⟹ $V^{\pi_{t+1}}(s) = r(s,a) + \mathbb{E}_{s'|s,a}[V^{\pi_{t+1}}(s')] \ge r(s,a) + \mathbb{E}_{s'|s,a}[V^{\pi_t}(s')] = \mathcal{T}V^{\pi_t}(s)$
⟹ $V^* \ge V^{\pi_t} \ge \mathcal{T}V^{\pi_{t-1}}$ ⟹ $\|V^{\pi_t} - V^*\|_\infty \le \|\mathcal{T}V^{\pi_{t-1}} - \mathcal{T}V^*\|_\infty \le \gamma \|V^{\pi_{t-1}} - V^*\|_\infty \le \gamma^t \|V_0 - V^*\|_\infty$

Rems: VI vs PI:
VI: One sweep through operator $\mathcal{T}$, per iter. cost: $\mathcal{O}(|S|^2|\mathcal{A}|)$ Linear Convergence rate, cheap cost, more iterations.
PI: Multiple sweeps through operator $\mathcal{T}^\pi$, per iter cost: $\mathcal{O}(\frac{|S|^2|\mathcal{A}|}{1-\gamma})$ or $\mathcal{O}(|S|^3 + |S|^2|\mathcal{A}|)$, linear conv. rate Expensive cost, fewer iterations

2.5 Linear Programming Approach: V^* unique & sol of:

$\min_V \sum_{s \in S} \mu_0(s) V(s)$ s.t. $V(s) \ge r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')$ $\forall s,a$

Dual: $\max_\lambda \sum_{s,a} r(s,a) \lambda(s,a)$ s.t. $\sum_a \lambda(s,a) = \mu_0(s) + \gamma \sum_{s',a'} P(s|s',a') \lambda(s',a')$, $\lambda(s,a) \ge 0$ $\forall s,a$

Occupancy Measure: Given policy π & $s_0 \sim \mu_0$, the state-action occupancy measure of policy π is:
$\lambda^\pi(s,a) := \sum_{t=0}^\infty \gamma^t P(s_t = s, a_t = a | s_0 \sim \mu_0, \pi)$ $d^\pi_{\mu_0}(s,a) \ge 0$ & $\sum_{s,a} d^\pi_{\mu_0}(s,a) = 1$

State-Action distr: $d^\pi_{\mu_0}(s,a) := (1-\gamma)\mathbb{E}[\sum_{t=0}^\infty \gamma^t \mathbb{1}\{s_t = s, a_t = a\} | s_0 \sim \mu_0, \pi]$

Value Obj. Equiv: $\sum_s \mu_0(s) V^\pi(s) = \sum_{s,a} \lambda^\pi(s,a) r(s,a)$

Proof: $\sum_s \dots = \mathbb{E}_{s \sim \mu_0}[V^\pi(s)] = \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 \sim \mu_0]$
$= \sum_{s,a} \sum_{t=0}^\infty \gamma^t P(s_t = s, a_t = a | \pi) r(s,a) = \sum_{s,a} \lambda^\pi(s,a) r(s,a)$

Bellman Eq for state occupancy Measure:
$\lambda^\pi(s,a) = \mu_0(s)\pi(a|s) + \gamma \sum_{s',a'} \pi(a|s) P(s|s',a') \lambda^\pi(s',a')$
↳ $\sum_a \dots$ yields dual constraints in dual LP

$\lambda^\pi(s,a) = P(s_0 = s, a_0 = a) + \sum_{t=1}^\infty \gamma^t P(s_t = s, a_t = a | \pi, s_0 \sim \mu_0)$
$= \mu_0(s)\pi(a|s) + \sum_{t=1}^\infty \gamma^t \sum_{s',a'} P(s_t = s, a_t = a | s_{t-1} = s', a_{t-1} = a', \pi, s_0 \sim \mu) P(s_{t-1} = s', a_{t-1} = a' | \pi, s_0 \sim \mu_0)$
$= \mu_0(s)\pi(a|s) + \sum_{t=1}^\infty \gamma^t \sum_{s',a'} P(s_t = s, a_t = a | s_{t-1} = s', a_{t-1} = a') P(s_{t-1} = s', a_{t-1} = a' | \pi, s_0 \sim \mu_0)$
$= \mu_0(s)\pi(a|s) + \gamma \sum_{s',a'} \pi(a|s) P(s|s',a') \sum_{t=1}^\infty \gamma^{t-1} P(s_{t-1} = s', a_{t-1} = a' | \pi, s_0 \sim \mu_0)$
$= \mu_0(s)\pi(a|s) + \gamma \sum_{s',a'} \pi(a|s) P(s|s',a') \lambda^\pi(s',a')$

Optimal policy: In dual formulation: 1) solve dual LP to obtain λ^* ⟹ 2) $\pi^*(a|s) = \frac{\lambda^*(s,a)}{\sum_a \lambda^*(s,a)}$

3. Value-Based Methods

3.1 From Planning to RL:

Value Based: Learn V^*, Q^*, Policy Based: learn π^*, Model-based: learn P&R

Functional Approx: $V_\theta(s)$, $Q_\theta(s,a)$, $\pi_\theta(a|s)$, $P_\theta(s'|s,a)$

Representation Learning: e.g. Linear: $V_\theta(s) = [\phi_1(s), \dots, \phi_d(s)] \begin{bmatrix}\theta_1 \\ \vdots \\ \theta_d\end{bmatrix}$, non-linear: NNs

3.2 Model-free Prediction:

Goal: Given policy $\pi: S \to \Delta(\mathcal{A})$, estimate $V^\pi(s)$ or $Q^\pi(s,a)$ from episodes of experience under π: $V^\pi(s) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 = s, \pi]$

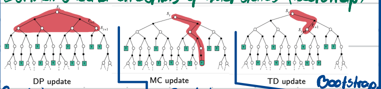
Monte Carlo Method: For each episode do: generate $\tau = (s_0, a_0, \dots)$ for each state s_t do: $G_t = r_{t+1} + \gamma r_{t+2} + \dots$, $n_t(s_t) \leftarrow n_t(s_t) + 1$, $V(s_t) \leftarrow V(s_t) + \frac{1}{n_t}(G_t - V(s_t))$ $V \to V^\pi$ if visit $\forall s_t$ ∞ often (no Bootstrap)

Value estims are indep. & don't build on that of other states

TD-Learning: $V(s_t) \leftarrow V(s_t) + \alpha_t(r_t + \gamma V(s_{t+1}) - V(s_t))$ $=: \delta$ TD error

Algo: For each step do: Observe $(s_t, a_t, r_t, s_{t+1}) \sim \pi$
Update $V(s_t) \leftarrow V(s_t) + \alpha_t(r_t + \gamma V(s_{t+1}))$ Conv. if {$\forall s$ visit ∞-often, α_t RM cond}

Estimates Build on estims of other states (Bootstrap)



DP update: Bootstrap, no sampling exploits Markov prop.
MC update: no Bootstrap, sampling model-free, does not exploit Markov property
TD update: Bootstrap, Sampling, online, model-free, exploits Markov p.

Rems: MC return unbiased, but higher variance. TD target is biased but has low variance.

MC error = sum of TD errors: $G_t - V(s_t) = \sum_{k=t}^\infty \gamma^{k-t} \delta_k$

Multiple-step TD learning: Let T termination time step, $\gamma \in [0,1]$
↳ $G_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n})$, $G_t^\infty = MC$
↳ $V(s_t) \leftarrow V(s_t) + \alpha_t(G_t^{(n)} - V(s_t))$, $\delta^{(n)}$: n-step TD error
↳ $n = 0$: TD, $n = \infty$: MC, $0 < n < \infty$: better balance of B&V

TD(λ): $G_t^\lambda = (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1} G_t^{(n)}$, $\lambda \in [0,1]$
$V(s_t) \leftarrow V(s_t) + \alpha[G_t^\lambda - V(s_t)]$ δ_t = TD Error of TD(0)

Rem: $\lambda = 0$ reduces to TD, $\lambda = 1$ reduces to MC

Implementation: $V(s) \leftarrow V(s) + \alpha \delta_t e_t(s)$ with e_t the eligibility trace: $e_t(s) = \gamma\lambda e_{t-1}(s) + \mathbb{1}\{s_t = s\} = \sum_{k=0}^t \gamma^{t-k} \mathbb{1}\{s_t = s\}$

Rem: conv. faster than TD(0) when λ is approp. chosen

SARSA: Algo: ∀step do: Observe $(s_t, a_t, r_t, s_{t+1}, a_{t+1})$
Update $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t(r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$

3.3 Model-free Control

Methods	DP	MC	TD(0)
Model knowledge	Need	No need	No need
Bootstrap	Yes	No	Yes
When do updates	After next step	After whole episode	After next step
Use Markov property	Yes	No	Yes
Bias	-	Unbiased	Biased
Variance	-	Big	Small
Convergence rate	Linear rate	-	?

Model-free Policy Iter: Algo: 1) Init π_0, 2) ∀ iter t: Estim (instead of compute) V^{π_t} & Q^{π_t}, Update policy: $\pi_{t+1}(s) \leftarrow \arg\max_a Q^{\pi_t}(s,a)$

Exploration Scheme: Issues: (i) policy eval. step requires visiting all state-action pairs ∞-often. (ii) greedy policy lacks exploration

Remedy: ε-greedy: $\pi_Q^\varepsilon(a|s) = \begin{cases}\arg\max_a Q(s,a) & p = 1-\varepsilon \\ \mathcal{U}(\mathcal{A}) & p = \varepsilon\end{cases}$

Boltzmann exploration: $\pi_Q^\beta(a|s) = \frac{\exp(\beta Q(s,a))}{\sum_{a'} \exp(\beta Q(s,a'))}$

GLIE Policy: (Greedy in the Limit with infinite Exploration) policy satisfies:
a) each state-action pair is visited ∞-often (with prob 1)
b) Policy converges to a greedy policy (w. p→1)
↳ e.g.: ε-greedy ε↓0, softmax policy β↑∞, small/large β {exploration/exploitation}
$\varepsilon = \begin{cases}\text{large} \\ \text{small}\end{cases} \Rightarrow \begin{cases}\text{exploration} \\ \text{exploitation}\end{cases}$

State-specific ε-greedy & Boltzmann: $\varepsilon_t(s) = \frac{c}{n_t(s)}$, $0 < c < 1$, $\beta_t(s) = \frac{\log(n_t(s))}{C_t(s)}$, $C_t(s) \ge \max_{a,a'} |Q_t(s,a) - Q_t(s,a')|$

SARSA for Model-free control: GLIE(Q) = {ε-greedy, any other}

Algo: Repeat ∀episodes: {Init s_0, chose $a_0 \sim GLIE(Q, s_0)$
Repeat ∀t≥0 of episode: {Take action a_t, obs. r_t & s_{t+1}, chose $a_{t+1} \sim GLIE(Q, s_{t+1})$, $Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha_t(s_t,a_t) \cdot [r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$} until s_t terminal }

Thm: For MDPs with finite state-actions & bounded rew. if LR satisfies $0 \le \alpha_t(s,a) \le 1$, $\sum_t \alpha_t(s,a) = \infty$, $\sum_t \alpha_t^2(s,a) < \infty$ where $\alpha_t(s,a) = 0$ unless $(s,a) = (s_t, a_t)$ & SARSA $\xrightarrow{conv.} Q^*$

Model-free Value Iteration:

Bellman optim eq: $Q^*(s,a) = \mathcal{T}_Q Q^*(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(.|s,a)}[\max_{a'} Q^*(s',a')]$

Q Learning: Algo: ∀t: 1) Observe (s_t, a_t, r_t, s_{t+1}) ← from any pol.
2) Update: $Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha_t[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$

Thm: If all (s,a) pairs visited ∞-often & LR $\alpha_t \in (0,1)$ satisfies $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$, then Q-learning Algo converges to the optimal state-action-value fct Q^*. slower

SARSA: on-policy & faster conv., Q-Learning: off-policy & V

Rem: Maximization Bias in Q-Learning: $\max_a Q(s_{t+1}, a) = Q(s_{t+1}, \arg\max_a Q(s_{t+1}, a))$ ⟹ Overestim bec same Q for {select, eval.}

Double Learning (DQN): Goal: lower the maxim. bias.

Sol: Decouple choice of maximization action & its estim
↳ Learn 2 indep. estims Q_1 & Q_2. Use Q_1 to select max a. Use Q_2 to estim value at best action.

Algo: ∀t do: {Observe (s_t, a_t, r_t, s_{t+1}). with prob $\beta \in (0,1)$: {
$Q_1(s_t,a_t) \leftarrow Q_1(s_t,a_t) + \alpha_t[r_t + \gamma Q_2(s_{t+1}, \arg\max_a Q_1(s_{t+1},a)) - Q_1(s_t,a_t)]$}
With prob (1-β): {$Q_2(s_t,a_t) \leftarrow Q_2(s_t,a_t) + \alpha_t[r_t + \gamma Q_1(s_{t+1}, \arg\max_a Q_2(s_{t+1},a)) - Q_2(s_t,a_t)]$}}

Rem: Conv to Q^* under same cond as Q-Learning

3.4 Value Fct Approx: If |S| & |A| large → need approx. learn $\phi(s)^T w$

Idea: $V(s) \approx V_w(s)$ & $Q(s,a) \approx Q_w(s,a)$. Options: NNs, Linear: $\phi(s)^T w$

Objective 1: minimize MSE: here: d: stationary distr under π,
$\min_w J_1(w) = \frac{1}{2}\mathbb{E}_{s\sim d}[(V_w(s) - V^\pi(s))^2] := \frac{1}{2}\|V_w - V^\pi\|_D^2$ $D = \text{diag}(d)$, $\|x\|_D^2 = x^T D x$

$J_1(w)$ is convex under Linear fct. approx, but non-convex in gen.

Rem: $\nabla_w J_1(w) = \mathbb{E}_{s\sim d}[(V_w(s) - V^\pi(s))\nabla_w V_w(s)]$, stoch. grad. $\mathbb{E}_{s\sim d} \to s_t \sim d$

Estim $V^\pi(s_t)$: (i) MC return, (ii) TD target, (iii) λ-return

Update steps: MC: $w_{t+1} \leftarrow w_t - \alpha_t(V_{w_t}(s_t) - G_t)\nabla V_{w_t}(s_t)$
TD: $w_{t+1} \leftarrow w_t - \alpha_t(V_{w_t}(s_t) - (r_t + \gamma V_{w_t}(s_{t+1})))\nabla V_{w_t}(s_t)$
TD(λ): $w_{t+1} \leftarrow w_t - \alpha_t(V_{w_t}(s_t) - G_t^\lambda)\nabla V_{w_t}(s_t)$

Rem: MC: =SGD with unbiased gradient, high var, with Linear FA converges to global optimum, with non-Linear FA: local min
TD: regr. with moving target, does not always conv., conv. with lin FA

Linear TD: $V_w(s) = \phi(s)^T w$, def feature mat $\Phi \in \mathbb{R}^{|S| \times d}$ s.t. $V_w = \Phi w$ (full column rank)
↳ $w_{t+1} \leftarrow w_t + \alpha_t(r_t + \gamma \phi(s_{t+1})^T w_t - \phi(s_t)^T w_t)\phi(s_t) \equiv w_t + \alpha_t g(w_t)$
↳ Limit point w^* of linear TD should satisfy $g(w^*) = 0$:
↳ $\mathbb{E}[(r + \gamma\phi(s')^T w - \phi(s)^T w)\phi(s)] = \Phi^T D(\mathcal{T}^\pi \Phi w^* - \Phi w^*) \stackrel{!}{=} 0$

Here $\mathcal{T}^\pi$ is Bellman consistency operator. sol ⟺ FP of:

Projected Bellman Eq: $\Phi w = \Pi_D \mathcal{T}^\pi \Phi w$, Range($\Phi$) = $\{\Phi w: w \in \mathbb{R}^d\}$

Projection Operator onto range(Φ): $\Pi_D(x) := \arg\min_{x' \in \text{Range}(\Phi)} \|x - x'\|_D^2$

Performance Bound:

Denote w^* as the FP to the projected Bellman eq & $V_{w^*} = \Phi w^*$
↳ $\|V_{w^*} - V^\pi\|_D \le \frac{1}{1-\gamma}\|V^\pi - \Pi_D V^\pi\|_D$
↳ TD sol not worse than MC sol. If $V^\pi \in \text{span}(\Phi)$ ⟹ MC&TD conv. to it

Proof: Show that $\Pi_D \mathcal{T}^\pi$ is a γ contraction wrt $\|.\|_D$:
Use (i) Projections are non-expansive, (ii) $\mathcal{T}^\pi$ is a γ contraction
↳ $\mathcal{T}^\pi V = R^\pi + \gamma P^\pi V \Rightarrow \|\Pi_D \mathcal{T}^\pi V_1 - \Pi_D \mathcal{T}^\pi V_2\|_D \le \|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_D \le \gamma\|V_1 - V_2\|_D$
↳ $\|V_{w^*} - V^\pi\|_D \le \|V_{w^*} - \Pi_D V^\pi\|_D + \|\Pi_D V^\pi - V^\pi\|_D$
$= \|\Pi_D \mathcal{T}^\pi V_{w^*} - \Pi_D \mathcal{T}^\pi V^\pi\|_D + \|\Pi_D V^\pi - V^\pi\|_D \le \gamma\|V_{w^*} - V^\pi\|_D + \|\Pi_D V^\pi - V^\pi\|_D$

used: triangle ineq, FP property of V_{w^*} & V^π, γ-contr. of $\Pi_D \mathcal{T}^\pi$

Alternative Objectives: $J_2(w) = \frac{1}{2}\|V_w - \mathcal{T}^\pi V_w\|_D^2$ (TD-λ)
$J_3(w) = \frac{1}{2}\|V_w - \Pi_D \mathcal{T}^\pi V_w\|_D^2$ (GTD2)

Q-Learning with fct. Approx.:
$w_{t+1} \leftarrow w_t + \alpha_t[r_t + \gamma \max_a Q_{w_t}(s_{t+1}, a) - Q_{w_t}(s_t, a_t)]\nabla Q_{w_t}(s_t, a_t)$
$\leftarrow w_t + \alpha_t[r_t + \gamma \max_a \phi(s_{t+1}, a)w - \phi(s_t, a_t)w]\phi(s_t, a_t)$

Deadly Triad: Divergence is likely to happen when we combine bootstrapping, off-policy learning & fct approx!
FA: update for one state might worsen approx for another
Off-Pol.: might not update Bootstrap values sufficiently often
Divergence: even with lin FA or TD-L   large NNs
Sol: can be alleviated by λ-return, ε-greedy or ✓

3.5 Convergence Analysis:

Corresponding ODE to linear TD update is:
$\dot{w} = \Phi^T D(\gamma P^\pi - \mathbb{1})\Phi w + \Phi D R \Leftrightarrow \frac{d}{dt}(w(t) - w^*) = A \cdot (w(t) - w^*)$
$=: A$ (neg. def) ⟹ ODE globally stable FP of $\mathcal{T}^\pi$

Thm: Suppose Φ is full rank & stepsize satisfies RM-cond. Then the iterates of linear TD converges, $w_k \xrightarrow{k\to\infty} w^*$

Rem: synchron. version Q-L & ODE: $\dot{Q} = \mathcal{T}Q - Q := (\gamma P^\pi - \mathbb{1})Q + R$
$Q_{k+1}(s,a) = Q_k(s,a) + \alpha_t[r(s,a) + \gamma \max_a Q_k(s',a) - Q_k(s,a)]$

Because of γ-contraction of $\mathcal{T}$: ODE is globally asympt. stable

Thm: Consider the stepsize $\alpha_k = \alpha$ in linear TD: $\mathbb{E}[\|w_k - w^*\|^2] \le \mathcal{O}((1 - c\alpha)^{k - \tau_{mix}}) + \mathcal{O}(\alpha \cdot \tau_{mix})$
where: τ_{mix}: geometric mixing of underlying MC.

Rem: Requires $\mathcal{O}(\frac{1}{\varepsilon^2})$ sample complexity to conv. to TD FP with ε accur.

4 Policy Gradient Methods:

4.1 Overview:

Value-Based Methods: ⊕: easy to gen. π from V, Bootstrap, easy Bias-Var-Tradeoff
⊖: π deterministic, doesn't scale (|S|&|A|↗), instability with off-policy-Learn. under FA, small value error might lead to big π-err

Policy-Based Methods: Idea: represent policy as $\pi_\theta(.|s)$ & find θ that maximizes long-term rewards.

Policy Optimization:
$\max_\theta J(\pi_\theta) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 \sim \mu, \pi_\theta] = \mathbb{E}_{s\sim\mu}[V^{\pi_\theta}(s)]$

Alternatively: $J_{avg}(\pi_\theta) = \mathbb{E}_{s \sim d^{\pi_\theta}}[V^{\pi_\theta}(s)] = \sum_s d^{\pi_\theta}(s) V^{\pi_\theta}(s)$ where $d^\pi(s)$ is the state stationary distr. under π

Parametrizations: Discrete: $\pi_\theta(a|s) = \theta_{s,a} \ge 0$ & $\sum_a \theta_{s,a} = 1$
Log-Linear: $\pi_\theta(a|s) = \frac{\exp(\theta \cdot \phi(s,a))}{\sum_{a'} \exp(\theta \cdot \phi(s,a'))}$ θ & $\phi \in \mathbb{R}^d$,
Softmax: $\pi_\theta(a|s) = \exp(\theta_{s,a}) / \sum_{a'} \exp(\theta_{s,a'})$ $\theta \in \mathbb{R}^{|S|\times|A|}$
Neural Softmax: $\pi_\theta(a|s) = \exp(f_\theta(s,a)) / \sum_{a'} \exp(f_\theta(s,a'))$ NN

Gaussian (cont.): $\pi_\theta(a|s) = (\sqrt{2\pi}\sigma_\theta(s))^{-1} \exp(-\frac{(a - \mu_\theta(s))^2}{2\sigma_\theta^2(s)})$
$\mu_\theta(s)$ & $\sigma_\theta^2(s)$ are differentiable FA.

Policy Gradient Methods: $\theta_{t+1} \leftarrow \theta_t + \alpha \hat\nabla_\theta J(\pi_{\theta_t})$

4.2 Policy Gradient Estimation:

MC Estim: Let $F(\theta) = \mathbb{E}_{\xi \sim p(.)}[f(\theta,\xi)] \Rightarrow \nabla_\theta F(\theta) = \mathbb{E}_{\xi \sim p(.)}[\nabla_\theta f(\theta,\xi)]$

Unbiased Grad estims: $\hat\nabla_\theta F(\theta) = \nabla_\theta f(\theta,\xi)$, $\hat\nabla_\theta F(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta f(\theta,\xi_i)$ $\xi \sim p(.)$, $\xi_i \sim p(.)$

Score Fct: $F(\theta) = \mathbb{E}_{\xi \sim p_\theta(.)}[f(\xi)] \Rightarrow \nabla_\theta F(\theta) = \mathbb{E}_{\xi \sim p_\theta(.)}[f(\xi)\nabla_\theta \log(p_\theta(\xi))]$

Unbiased estim: as before

Policy Optim.: $\max_\theta J(\pi_\theta) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) | s_0 \sim \mu, \pi] = \mathbb{E}_{\tau \sim p_\theta}[R(\tau)]$

Note: $\tau = (s_0, a_0, s_1, a_1, \dots)$ random trajectory with prob $p_\theta(\tau)$
$p_\theta(\tau) := \mu(s_0)\prod_{t=0}^\infty \pi_\theta(a_t|s_t) P(s_{t+1}|s_t,a_t)$.
$R(\tau)$ = total reward over the random trajectory
↳ $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim p_\theta}[R(\tau) \cdot \nabla_\theta \log(p_\theta(\tau))]$

Policy Gradient Thm (REINFORCE) $= \sum_{t=0}^\infty \nabla_\theta \log(\pi_\theta(a_t|s_t))$
$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim p_\theta}[R(\tau)(\sum_{t=0}^\infty \nabla_\theta \log(\pi_\theta(a_t|s_t)))]$

Rem: $\nabla_\theta \log(\pi_\theta(a|s)) = \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} =: S$ = "score fct", e.g.
for log-linear $\pi_\theta(a|s)$: $S = \phi(s,a) - \mathbb{E}_{a \sim \pi_\theta(.|s)}[\phi(s,a)]$

Rem: $\mathbb{E}_{a \sim \pi_\theta(.|s)}[S] = 0$

REINFORCE Estim: Generate $\tau = (s_0, a_0, \dots) \sim \pi_\theta$
↳ $\hat\nabla_\theta J(\pi_\theta) = (\sum_{t=0}^\infty \gamma^t r_t) \cdot (\sum_{t=0}^\infty \nabla_\theta \log(\pi_\theta(a_t|s_t)))$

Rem: REIN. has high variance due to corr btw R(τ) & $\{\pi_\theta(a_t|s_t)\}_{t=1}^\infty$

Rem: $\pi(a_{t_2}|s_{t_2})$ does not affect $\sum_{t=0}^{t_1} r(s_t,a_t)$ if $t_1 < t_2$!

PG Thm: $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim p_\theta}[\sum_{t=0}^\infty \gamma^t Q^{\pi_\theta}(s_t,a_t)\nabla_\theta \log(\pi_\theta(a_t|s_t))]$

↳ Difference: Action-Value Fct: $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim p_\theta}[\sum_{t=0}^\infty (\sum_{t'=t}^\infty \gamma^{t'} r(s_{t'},a_{t'}))\nabla_\theta \log(\dots)]$
REINFORCE: $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim p_\theta}[\sum_{t=0}^\infty (\sum_{t'=0}^\infty \gamma^{t'} r(s_{t'},a_{t'}))\nabla_\theta \log(\dots)]$

Proof: $\nabla_\theta V^{\pi_\theta}(s_0) = \nabla_\theta \sum_{a_0} \pi(a_0|s_0) Q^{\pi_\theta}(s_0,a_0) = \sum_{a_0}(\nabla_\theta \pi_\theta) Q^{\pi_\theta} + \pi_\theta(\nabla_\theta Q^{\pi_\theta})$
$(Q^{\pi_\theta} = r(s_0,a_0) + \gamma\sum_{s_1} P(s_1|s_0,a_0)V^{\pi_\theta}(s_1))$ ⟹ $= \sum_{a_0} \pi_\theta \nabla\log(\pi_\theta) Q^{\pi_\theta} + \gamma \sum_{a_0,s_1} P(s_1|s_0,a_0)\nabla V^{\pi_\theta}(s_1)$
Induction ⟹ $\nabla_\theta J(\pi_\theta) = \sum_{s_0} \mu(s_0)\nabla V^{\pi_\theta}(s_0) = \mathbb{E}_{\tau \sim p_\theta}[Q^{\pi_\theta}(s_0,a_0)\nabla\log(\pi_\theta(a_0|s_0)) + \gamma\mathbb{E}_{\tau\sim p_\theta}[\nabla V^{\pi_\theta}(s_1)]]$
Induction → $= \mathbb{E}_{\tau \sim p_\theta}[\sum_{t=0}^\infty \gamma^t Q^{\pi_\theta}(s_t,a_t)\nabla\log(\pi_\theta(a_t|s_t))]$ □

Reward to go:

REIN-Estim: generate $\tau = (s_0, a_0, \dots) \sim \pi_\theta$. Construct:
$\hat\nabla_\theta J(\pi_\theta) = \sum_{t=0}^\infty \gamma^t G_t \cdot \nabla_\theta \log(\pi_\theta(a_t|s_t))$, $G_t = \sum_{i=t}^\infty \gamma^{i-t} r_i$
↳ Unbiased estim of PG. reduces variance

Baselines: $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim p_\theta}[\sum_{t=0}^\infty \gamma^t [Q^{\pi_\theta}(s_t,a_t) - b(s_t)]\nabla_\theta \log(\pi_\theta(a_t|s_t))]$

Rem: $\forall b(s)$: $\mathbb{E}_{a \sim \pi_\theta}[b(s)\nabla_\theta \log(\pi_\theta(a|s))] = 0$, e.g. $b(s) = V^\pi(s)$

Def: Advantage Fct: $A^{\pi_\theta}(s,a) =: Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$

Def: Discounted state visitation distr. under π:
$d_\mu^\pi(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t P(s_t = s | s_0 \sim \mu, \pi)$
⟹ $\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^{\pi_\theta}}[\mathbb{E}_{a \sim \pi_\theta(.|s)}[Q^{\pi_\theta}(s,a)\nabla_\theta \log(\pi_\theta(a|s))]]$
$= \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^{\pi_\theta}}[\mathbb{E}_{a \sim \pi_\theta(.|s)}[A^{\pi_\theta}(s,a)\nabla_\theta \log(\pi_\theta(a|s))]]$

4.3 Policy Gradient Methods: Exact: $\theta_{t+1} \leftarrow \theta_t + \alpha_t \nabla_\theta J(\pi_{\theta_t})$

Stochastic: $\theta_{t+1} \leftarrow \theta_t + \alpha_t \hat\nabla_\theta J(\pi_\theta)$ $\tau_T = (s_0, a_0, r_0, \dots, s_T, a_T, r_T)$

Monte Carlo PG Method: Algo: Init policy param $\theta \in \mathbb{R}^d$ & step size $\alpha > 0$, baseline b(.). ∀episodes do: {Generate $\tau_T \sim \pi_\theta$, ∀steps $t \in 1:T$ do: {$G_t = \sum_{i=t}^T \gamma^{i-t} r_i$, $A_t = G_t - b(s_t)$, $\theta \leftarrow \theta + \alpha\gamma^t A_t \cdot \nabla_\theta \log(\pi_\theta(a_t|s_t))$}}

Rem: Challenges: Nonconvex (concave, bec. maximization) landscape, sensitive to stepsize choice, High variance/bias of policy grad esti.

Convergence:
Of exact PG: If $J(\pi_\theta)$ is L-smooth & $\alpha_t = \frac{1}{L}$: $\min_{t=0,\dots,T-1}\|\nabla_\theta J(\pi_{\theta_t})\|_2^2 \le \frac{2L(J(\pi_{\theta^*}) - J(\pi_{\theta_0}))}{T}$
of stochastic PG: if $J(\pi_\theta)$ is L-smooth & $\hat\nabla_\theta J(\pi_\theta)$ is unbiased & has bounded variance σ^2: with proper stepsize: $\min_{t=0,\dots,T-1}\mathbb{E}[\|\nabla_\theta J(\pi_{\theta_t})\|_2^2] = \mathcal{O}\left(\sqrt{\frac{L(J(\pi_{\theta^*}) - J(\pi_{\theta_0}))\sigma^2}{T}}\right)$

Other Challenges:
· Stepsize: {too large: terrible policy → poor data; too small: similar policy → inefficient explor.}
· Sample efficiency: cannot reuse data from prev. policies
· high variance: SPG estim suffers from high var

4.4 Natural Policy Gradient Method (NPG)

NPG: $\theta_{t+1} \leftarrow \theta_t + \eta(F_{\theta_t})^\dagger \nabla_\theta J(\pi_{\theta_t})$ where:

F_θ: Fisher Information Matrix:
$F_\theta = \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi(.|s)}[\nabla_\theta \log(\pi_\theta(a|s)) \cdot \nabla_\theta \log(\pi_\theta(a|s))^T]$

PG: $\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi(.|s)}[\nabla_\theta \log(\pi_\theta(a|s)) A^{\pi_\theta}(s,a)]$

Moore-Penrose inverse: $A^\dagger$ s.t. $AA^\dagger A = A$, $A^\dagger A A^\dagger = A^\dagger$
If cols of A lin indep: $A^\dagger = (A^T A)^{-1} A^T$
if rows — " —: $A^\dagger = A^T (AA^T)^{-1}$

Interpretation: $\theta_{t+1} \approx \arg\max_\theta \{J(\pi_\theta), \text{ s.t. } KL(p_{\theta_t}(\tau) \| p_\theta(\tau)) \le \delta\}$
↳ Approx: - objective w/ first-order Taylor: $\nabla_\theta J(\pi_{\theta_t})^T(\theta - \theta_t)$
- constraint w/ second-order Taylor: $\frac{1}{2}(\theta - \theta_t)^T F_{\theta_t}(\theta - \theta_t) \le \delta$

Equivalent form of NPG: It holds that $(1-\gamma)(F_\theta)^\dagger \nabla_\theta J(\pi_\theta) = w^*(\theta) \in \arg\min_w \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(.|s)}[(w^T \nabla_\theta \log(\pi_\theta(a|s)) - A^{\pi_\theta}(s,a))^2]$

⟹ $\theta_{t+1} \leftarrow \theta_t + \frac{\eta}{1-\gamma} w^*(\theta_t)$

Compatible Fct Approx: Def $A_{w^*}(s,a) = w^{*T}\nabla_\theta \log(\pi_\theta(a|s))$

Then: $\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma} F_\theta \cdot w^* = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi(.|s)}[\nabla_\theta \log(\pi_\theta(a|s)) \cdot A_{w^*}(s,a)]$
unbiased

Generally $A^{\pi_\theta}(s,a) \approx w^T \cdot \phi(s,a)$ not unbiased

Sample-based NPG: Algo: Init $\theta_0 \in \mathbb{R}^d$, $\eta > 0$, $\alpha > 0$
For $t = 1:T$ do: { Init w_0, $\pi_t = \pi_{\theta_t}$, for $n = 0:(N-1)$ do {
Obtain $s \sim d_\mu^{\pi_t}$, $a \sim \pi_t(.|s)$, obtain unbiased estim $\hat{A}$ for A^{π_t},
Update w: $w = w - \alpha(w^T \nabla_\theta \log(\pi_t(a|s)) - \hat{A}(s,a)) \cdot \nabla_\theta \log(\pi_t(a|s))$
} Set $w_t = w$ (or avg), Update $\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma} w_t$ }

4.5 Beyond PG/NPG:

Trust Region Policy Optim (TRPO):
$\max_\theta \mathbb{E}_{s \sim d^{\pi_{\theta_t}}, a \sim \pi_{\theta_t}(.|s)}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a)\right]$ s.t. $\mathbb{E}_{s \sim d^{\pi_{\theta_t}}}[KL(\pi_\theta(.|s) \| \pi_{\theta_t}(.|s))] \le \delta$ "max J(π) by PDL"

⟹ Surrogate objective can be viewed as lin approx: $J(\pi) = J(\pi_t) + \frac{1}{1-\gamma}\mathbb{E}_{s \sim d^{\pi_t}, a \sim \pi(.|s)}[A^{\pi_t}(s,a)]$

Thm: TR Update Performance
Let π_{t+1}, π_t be the two consecutive updates of TRPO. It holds that: $J(\pi_{t+1}) - J(\pi_t) \ge -\frac{\sqrt{2\delta}\gamma}{(1-\gamma)^2} A_t$
where $A_t = \max_s |\mathbb{E}_{a \sim \pi(.|s)}[A^{\pi_t}(s,a)]|$

Proof Sketch: (i) $\mathbb{E}_{s \sim d^{\pi_t}, a \sim \pi(.|s)}[A^{\pi_t}(s,a)] \ge 0$ by def
(ii) prove $\|d^{\pi'}(.) - d^\pi(.)\|_1 \le \frac{2\gamma}{1-\gamma}\mathbb{E}_{s \sim d^\pi}[D_{TV}(\pi'(.|s) \| \pi(.|s))]$
where $D_{TV}(p\|q) = \frac{1}{2}\sum_a |p(a) - q(a)| \le \sqrt{D_{KL}(p\|q)/2}$
(iii) Rest: Performance difference Lemma & Jensen

Proximal Policy Optim: (PPO)
$\max_\theta \mathbb{E}_{s \sim d^{\pi_{\theta_t}}, a \sim \pi(.|s)}\left[\min\left\{\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a), \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}; 1-\varepsilon, 1+\varepsilon\right) \cdot A^{\pi_{\theta_t}}(s,a)\right\}\right]$

where: $\text{clip}(x; 1-\varepsilon, 1+\varepsilon) = \begin{cases}1-\varepsilon & \text{if } x < 1-\varepsilon \\ 1+\varepsilon & \text{if } x > 1+\varepsilon \\ x & \text{otherwise}\end{cases}$

↳ PPO penalizes large deviations from current π directly inside the obj. fct.

4.6 Recap/Results so far:

Discrete: $\pi_\theta(a|s) = \theta_{s,a} \Rightarrow \frac{\partial J(\pi_\theta)}{\partial\theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) Q^{\pi_\theta}(s,a)$

Softmax: $\pi_\theta(a|s) \propto \exp(\theta_{s,a}) \Rightarrow \frac{\partial J(\pi_\theta)}{\partial\theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s) Q^{\pi_\theta}(s,a)$

Proof Sketch: $\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\sum_s d(s)\sum_a Q^\pi(s,a)\nabla\pi_\theta(a|s)$
↳ Discrete: $\frac{\partial\pi_\theta(a|s)}{\partial\theta_{s',a'}} = \delta_{ss'}\delta_{aa'}$, Softmax: $\frac{\partial\pi(a|s)}{\partial\theta_{s',a'}} = \pi_\theta(a|s)\delta_{ss'}\delta_{aa'} - \pi_\theta(a|s)\pi_\theta(a'|s)\delta_{ss'}$

4.6 Performance Difference Lemma

Thm: $\forall \pi, \pi'$: $J(\pi) - J(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^\pi, a \sim \pi(.|s)}[A^{\pi'}(s,a)]$

$V^\pi(s) - V^{\pi'}(s) = \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)|s_0 = s] - V^{\pi'}(s)$

Proof
$= \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^\infty \gamma^t(r(s_t,a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t))|s_0 = s] - V^{\pi'}(s)$
$= \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^\infty \gamma^t(r(s_t,a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t))|s_0 = s]$
$= \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^\infty \gamma^t(r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[V^{\pi'}(s_{t+1})] - V^{\pi'}(s_t))|s_0 = s]$
$= \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^\infty \gamma^t(Q^{\pi'}(s_t,a_t) - V^{\pi'}(s_t))|s_0 = s]$
$= \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^\infty \gamma^t A^{\pi'}(s_t,a_t)|s_0 = s]$

$J(\pi^*) - J(\pi) = \mathcal{O}(\|\nabla J(\pi)\|^p)$

4.7 Global Conv. of PG Methods:

Based on prev. results: $J(\pi^*) - J(\pi) = \mathcal{O}(\|\nabla J(\pi)\|)$

Policy Optim. under direct/discrete Param:
Def: $\Pi = \Delta(\mathcal{A})^{|S|} = \{\pi: \pi(a|s) \ge 0, \sum_{a\in\mathcal{A}}\pi(a|s) = 1\ \forall s\}$
$\max_{\pi\in\Pi} J(\pi) := \mathbb{E}_{s\sim\mu}[V^\pi(s)]$
↳ If $\pi \in \Pi$ optimal: first order optimality cond: $\forall\bar\pi\in\Pi$: $\langle\bar\pi - \pi, \nabla J(\pi)\rangle \le 0 \Leftrightarrow \max_{\bar\pi\in\Pi}\langle\bar\pi - \pi, \nabla J(\pi)\rangle = 0$

Gradient Mapping Domination:
$J(\pi^*) - J(\pi) \le \|\frac{d_\mu^{\pi^*}}{d_\mu^\pi}\|_\infty \cdot \max_{\bar\pi\in\Pi}\langle\bar\pi - \pi, \nabla J(\pi)\rangle$

Implications: · Any first order stationary point is thus globally optimal.
· The $\|\dots\|_\infty$ term is called the distr. mismatch coeff.
↳ captures the hardness of explor. problem.
↳ $\max_\pi \|\dots\|_\infty \le \frac{1}{1-\gamma}\|\frac{d_\mu^{\pi^*}}{\mu}\|_\infty$ since $d_\mu^\pi(s) \ge (1-\gamma)\mu(s)$

$J(\pi^\star) - J(\pi) = \frac{1}{1-\gamma}\sum_s d_\mu^{\pi^\star}(s)\sum_a \pi^\star(a|s)A^\pi(s,a)$

Proof:
$= \frac{1}{1-\gamma}\sum_s \frac{d_\mu^{\pi^\star}(s)}{d_\mu^\pi(s)} d_\mu^\pi(s)\sum_a \pi^\star(a|s)A^\pi(s,a)$
$\le \frac{1}{1-\gamma}\|\frac{d_\mu^{\pi^\star}}{d_\mu^\pi}\|_\infty \cdot \max_{\bar\pi\in\Pi}\sum_{s,a} d_\mu^\pi(s)\bar\pi(a|s)A^\pi(s,a)$
$= \frac{1}{1-\gamma}\|\frac{d_\mu^{\pi^\star}}{d_\mu^\pi}\|_\infty \cdot \max_{\bar\pi\in\Pi}\sum_{s,a} d_\mu^\pi(s)(\bar\pi(a|s) - \pi(a|s))A^\pi(s,a)$
$= \frac{1}{1-\gamma}\|\frac{d_\mu^{\pi^\star}}{d_\mu^\pi}\|_\infty \cdot \max_{\bar\pi\in\Pi}\sum_{s,a} d_\mu^\pi(s)(\bar\pi(a|s) - \pi(a|s))Q^\pi(s,a)$
$= \|\frac{d_\mu^{\pi^\star}}{d_\mu^\pi}\|_\infty \cdot \max_{\bar\pi\in\Pi}\langle\bar\pi - \pi, \nabla J(\pi)\rangle$

Projected Policy Gradient Method:
$\pi_{t+1} = \text{Proj}_\Pi(\pi_t + \eta\nabla J(\pi_t))$ where $\text{Proj}_\Pi(\pi) = \arg\min_{\pi'\in\Pi}\|\pi - \pi'\|_2^2$ } Gradient Ascent & proj. onto simplex

Def: Generalized Gradient Mapping: $G(\pi_t) = \frac{1}{\eta}(\pi_{t+1} - \pi_t)$
↳ If $J(\pi)$ is L-smooth: $\min_{t\le T}\|G(\pi_t)\|_2^2 \le \frac{2L}{T}(J(\pi^*) - J(\pi_0))$

Global Conv. of Projected Policy Grad. Method:

Assume access to exact gradient & let $\eta = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$. Then:

$J(\pi^*) - \max_{t<T} J(\pi_t) \le \frac{8\sqrt{\gamma|S||\mathcal{A}|}}{(1-\gamma)^3\sqrt{T}} \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty$ large, slow rate

Proof sketch:

- Show that the objective $J(\pi)$ is L-smooth with $L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$ and $J(\pi) \le \frac{1}{1-\gamma}$.
- Invoke convergence on gradient mapping: $\min_{t\le T}\|G(\pi_t)\|_2^2 \le \frac{2L(J(\pi^*)-J(\pi_0))}{T}$.
- Invoke the relationship between gradient mapping and approximation of stationary point [4]:

$$\max_{\bar\pi\in\Pi}\langle\bar\pi - \pi_{t+1}, \nabla J(\pi_{t+1})\rangle \le (1+L\eta)\cdot\|G(\pi_t)\|_2\cdot\|\pi_{t+1}-\pi_t\|_2.$$

- Use the gradient dominance for global convergence.

Rem: PG update can be viewed as:

$\pi_{t+1} := \mathrm{Proj}_\Pi(\pi_t + \eta\nabla J(\pi_t)) = \arg\max_{\pi\in\Pi}\{\langle\nabla J(\pi_t), \pi\rangle - \frac{1}{2\eta}\|\pi-\pi_t\|_2^2\}$

$= \arg\max_{\pi\in\Pi}\{\langle Q^{\pi_t}, \pi\rangle_{d_\mu^{\pi_t}} - \frac{1}{2\eta}\|\pi-\pi_t\|_2^2\}$

where: $\frac{\partial J(\pi)}{\partial\pi(a|s)} = \frac{1}{1-\gamma}d_\mu^\pi(s)Q^\pi(s,a)$ & $\langle .,.\rangle_{d_\mu^\pi}$ is reweighted inner prod.

Policy Mirror Descent:

$\pi_{t+1} = \arg\max_{\pi\in\Pi}\{\langle Q^{\pi_t}, \pi\rangle_{d_\mu^{\pi_t}} - \frac{1}{\eta_t}\sum_s d_\mu^{\pi_t}(s) D(\pi(.|s), \pi_t(.|s))\}$

Projected Q-Ascent: set $D(p,q) = \frac{1}{2}\|p-q\|_2^2$:

$\pi_{t+1}(.|s) = \mathrm{Proj}_{\Delta(\mathcal{A})}(\pi_t(.|s) + \eta_t Q^{\pi_t}(s,.))\ \forall s\in S$

Exponentiated Q-Ascent: set $D(p,q) = KL(p\|q) = \sum_a p_a \log(\frac{p_a}{q_a})$

$\pi_{t+1}(.|s) = \pi_t(a|s)\frac{\exp(\eta_t Q^t(s,a))}{\sum_{a'}\pi_t(a'|s)\exp(\eta_t Q^t(s,a'))}\ \forall a\in\mathcal{A}, s\in S$

↳ Akin to NPG. As $\eta_t\to\infty$: reduces to policy iteration update.

4.8 Global Conv. of NPG Methods:

Let $\pi_\theta(a|s) = \exp(\theta_{s,a})/\sum_{a'}\exp(\theta_{s,a'})$ & $\pi_t \equiv \pi_{\theta_t}$

↳ $\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma}A^{\pi_t}$ & $\pi_{t+1}(a|s) = \pi_t(a|s)\frac{\exp(\eta A^{\pi_t}(s,a)/(1-\gamma))}{Z_t(s)}$

Convergence of Tabular NPG: $\forall\eta\ge(1-\gamma)^2\log(|\mathcal{A}|)$ & $T>0$:

$J(\pi^*) - J(\pi_T) \le \frac{2}{(1-\gamma)^2 T}$ Rem: Dim-free conv.! No dep on $|\mathcal{A}|, |S|$

No dep. on distr mismatch coeff

Lemma: $J(\pi) - J(\pi_t) = \frac{1}{\eta}\mathbb{E}_{s\sim d_\mu^\pi}[KL(\pi(.|s)\|\pi_t(.|s)) - KL(\pi(.|s)\|\pi_{t+1}(.|s)) + \log(Z_t(s))]$

↳ Proof: Performance difference Lemma + Update rule $\Rightarrow A^{\pi_t}(s,a) = \frac{1-\gamma}{\eta}\log(\frac{\pi_{t+1}(a|s)Z_t(s)}{\pi_t(a|s)})$

Proof of Tabular NPG Conv:

Setting $\pi = \pi^\star$ in the previous lemma and telescoping from $t = 0, \dots, T-1$

$$\frac{1}{T}\sum_{t=0}^{T-1} J(\pi^\star) - J(\pi_t) \le \frac{1}{\eta T}\mathbb{E}_{s\sim d_\mu^{\pi^\star}}[KL(\pi^\star(\cdot|s)\|\pi_0(\cdot|s))] + \frac{1}{\eta T}\sum_{t=0}^{T}\mathbb{E}_{s\sim d_\mu^{\pi^\star}}[\log Z_t(s)]$$

Setting $\pi = \pi_{t+1}$ in the previous lemma, we have

$$J(\pi_{t+1}) - J(\pi_t) \ge \frac{1}{\eta}\mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[\log Z_t(s)] \ge \frac{1-\gamma}{\eta}\mathbb{E}_{s\sim\mu}[\log Z_t(s)] \ge 0, \forall\mu. \text{ (why?)}$$

Combining these two equations and the fact that $J(\pi) \ge \frac{1}{1-\gamma}$ implies that

$$\frac{1}{T}\sum_{t=0}^{T-1} J(\pi^\star) - J(\pi_t) \le \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.$$

Let $\pi_\theta(a|s) = \frac{\exp(\theta^T\phi(s,a))}{\sum_{a'}\exp(\theta^T\phi(s,a'))}$, $\pi_t = \pi_{\theta_t}$

↳ $\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma}w_t$ with $w_t\in\arg\min_w \mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(.|s)}[(w^T\bar\phi(s,a) - A^{\pi_\theta}(s,a))^2]$, $\bar\phi = \phi(s,a) - \mathbb{E}[\phi(s,a)]$

↳ $\pi_{t+1}(a|s) = \pi_t(a|s)\frac{\exp(\eta w_t^T\phi(s,a)/(1-\gamma))}{Z_t(s)}$

Conv of sample based NPG: $\mathbb{E}[\min_{t\le T} J(\pi_{\theta^*}) - J(\pi_{\theta_t})] \le \mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{\frac{2\log(|\mathcal{A}|)}{T}} + \sqrt{\varepsilon_{stat}} + \sqrt{\varepsilon_{bias}}\right)$ where

ε_{stat}: how close w_t to a $w^*(\theta_t)$ statistical error

ε_{bias}: how good the best policy in the class

5. Multi-Agent RL & Markov Games:

5.1 From Single Agent to Multiple Agents:

Issues: exp growing search space, Non stationarity of env, Non-Markovian

Markov Games: Multiple competing Agents & full state info

5.2 Normal form Games & Repeated Games:

Normal Form Games: players/agents $\mathcal{I}$, $|\mathcal{I}|=n$, joint act. $\underline{a} = (a_1,\dots,a_{|\mathcal{I}|})$, Reward $r_i(\underline{a}) \Rightarrow$ NFG (/one shot game)

If repeated ∞ many times: repeated game

Strategies: $\pi_i\in\Delta(\mathcal{A}_i)\subseteq\mathbb{R}^{|\mathcal{A}_i|}$: $\pi_i(a_i)$ prob. agent i selects act. a_i

pure strategy (deterministic policy): only play one action

mixed strategy (stochastic): a distr over set of actions

Strategy Profile: concat. of players' strategies $\underline{\pi}$

Expected Payoff of player i: $r_i(\underline{\pi}) = \mathbb{E}_{\underline{a}\sim\underline{\pi}}[r_i(\underline{a})]$

$= \sum_{\underline{a}} r_i(\underline{a})\prod_{j\in\mathcal{I}}\pi_j(a_j) = \sum_{\underline{a}} r_i(\underline{a})\mathbb{P}(\underline{a}|\underline{\pi})$ (indep. policies)

Zero-Sum Games: In 2-player-game: $r_1(a_1,a_2) = -r_2(a_1,a_2)$

↳ Payoff can be represented with matrix A: $A(i,j) = r_1(a_i,a_j)$

↳ Expected Payoff of P1: $r_1(\pi_1,\pi_2) = (\pi_1)^T A\pi_2$

Response Models: Fix $\underline{\pi}_{-i} \triangleq (\pi_1,\dots,\pi_{i-1},\pi_{i+1},\dots,\pi_n)$. Best response: π_i s.t. $r_i(\pi_i,\underline{\pi}_{-i}) \ge r_i(\tilde\pi_i,\underline{\pi}_{-i})\ \forall\tilde\pi_i \Rightarrow \pi_i$ = "optimal strategy"

Softmax response: $\pi_i(a_i) \propto \exp(\lambda r_i(a_i,\underline{\pi}_{-i}))$

Dominant Strategy Equilibrium: π_i optimal $\forall i$

↯ may not exist!

Nash Equilibrium: $\underline{\pi}^*$: no player can improve exp. payoff by changing policy if the others stick to their policy.

$\Rightarrow r_i(\underline{\pi}^*) \ge r_i(\pi_i,\underline{\pi}_{-i})\ \forall i$ — in mixed strategies. ⊕ NE in ZSG eff. comp. ⊖: Computing: intractable in general

Thm: In a normal form game with finite players & actions, ∃NE

NE in 2P ZSG: $\max_{\pi_1\in\Delta^{d_1}}(\pi_1)^T A\pi_2$ & $\min_{\pi_2\in\Delta^{d_2}}(\pi_1)^T A\pi_2$

NE $= (\pi_1^*,\pi_2^*)$ s.t. $(\pi_1)^T A\pi_2^* \le (\pi_1^*)^T A\pi_2^* \le (\pi_1^*)^T A\pi_2\ \forall\pi_1,\pi_2$

$\Leftrightarrow$ saddle point of fct $f(\pi_1,\pi_2) = (\pi_1)^T A\pi_2$

Minimax Thm: $X\subseteq\mathbb{R}^{d_1}$, $Y\subseteq\mathbb{R}^{d_2}$, $f: X\times Y\to\mathbb{R}$ cont. s.t. $f(.,y)$ is convex $\forall y$ & $f(x,.)$ is concave $\forall x$, then: $\min_x\max_y f(x,y) = \max_y\min_x f$

Iterated Best Response: Algo: $\forall t=1\dots$ do { each player i updates π_i^{t+1} s.t.: $r_i(\pi_i^{t+1},\underline{\pi}_{-i}^t) \ge r_i(\pi_i,\underline{\pi}_{-i}^t)\ \forall\pi_i$ }

Rem: In general: does not converge.

Potential Games: The potential fct $\Phi: \mathcal{A}\to\mathbb{R}$ s.t. $\forall i\in\mathcal{I}$: $\forall a_{-i}\in\mathcal{A}_{-i}$, $\forall a_i,\hat a_i\in\mathcal{A}_i$

$r_i(a_i,a_{-i}) - r_i(\hat a_i,a_{-i}) = \Phi(a_i,a_{-i}) - \Phi(\hat a_i,a_{-i})$

Prop: If a potential game is finite, ∃ pure NE. If players use IBR seq (or parallel), the dynamic will terminate @ NE after finite steps

Fictious Play: IBR only considers last round ∴

↳ FP: each agent i counts opponents actions $N_t(j,a_j)$

Agent i assumes: $\tilde\pi_j^t(a_j) = N_t(j,a_j)/\sum_{\bar a_j\in\mathcal{A}_j}N_t(j,\bar a_j)$

$\Rightarrow a_i^{t+1} = \arg\max_{a_i} r_i(a_i,\underline{\tilde\pi}_{-i}^t)$

Problem: not guaranteed to converge

Good: converges for zero-sum games.

Thm: For 2P ZSG the emp. distr. of FP conv. to a NE:

$(\tilde\pi_1^t,\tilde\pi_2^t)\to(\pi_1^*,\pi_2^*)$ where (π_1^*,π_2^*) is a NE. Conv. rate: $\mathcal{O}(\frac{1}{\sqrt{T}})$

Grad. Ascent: $\pi_i^{t+1}(a_i) = \pi_i^t(a_i) + \alpha_i^t\frac{\partial r_i(\underline{\pi})}{\partial\pi_i(a_i)}\Big|_{\underline{\pi}=\underline{\pi}^t}$ & project to valid prob distr

Rem: $\frac{\partial r_i(\underline{\pi})}{\partial\pi_i(a_i)}\Big|_{\underline{\pi}=\underline{\pi}^t} = \frac{\partial}{\partial\pi_i(a_i)}\left(\sum_{\underline{a}} r_i(\underline{a})\prod_j\pi_j(a_j)\right)\Big|_{\underline{\pi}=\underline{\pi}_t}$

$= \sum_{\underline{a}_{-i}} r(a_i,\underline{a}_{-i})\prod_{j\ne i}\pi_j^t(a_j)$

Grad Asc. in 2P ZSG:

Goal: $\min_{\pi_2}\max_{\pi_1}(\pi_1)^T A\pi_2$ GDA = GD Ascent

↳ $\pi_2^{t+1} = \mathcal{P}_{\Delta^{d_2}}(\pi_2^{t+1} - \alpha_2^t A^T\pi_1^t)$, $\pi_1^{t+1} = \mathcal{P}_{\Delta^{d_1}}(\pi_1^t + \alpha_1^t A\pi_2^t)$

Setting (solution concept)	Best response	Fictitious play	GDA
Normal form games (NE)	No	No	No
Potential games (NE)	Yes	Yes	Yes
Zero-sum two-player games (NE)	No	Yes[1]	Yes/No[2]

5.3 Markov Games:

Finite Markov Games: MDP w/ multiple agents, each w/ their own reward. FMG $= (S,\underline{\mathcal{A}},P,\underline{r},\gamma)$. A MG is an extension of NF game w/ multiple stages & shared state $s\in S$.

Joint action: $\underline{a} = (a_i)_i$, $a_i\in\mathcal{A}_i$, Transition fct: $P(s'|s,\underline{a})$

Reward fct: $r_i(s,\underline{a})$, stationary policy: $\pi_i(a_i|s)$, disc. fact. γ

	action	state	transition	reward	policy	multi-stage
normal form game	$a_i\in\mathcal{A}_i$	no	no	$r_i(\boldsymbol{a})$	$\pi_i(a)$	no
Markov game	$a_i\in\mathcal{A}_i$	$s\in\mathcal{S}$	$P(s'\mid s,a)$	$r_i(s,\boldsymbol{a})$	$\pi_i(a_i\mid s)$	yes

Value Fct: $V_i^{\underline{\pi}}(s) = \mathbb{E}[\sum_{t=0}^\infty\gamma^t r_i(s^t,\underline{a}^t)|s^0=s, \underline{a}^t\sim\underline{\pi}(.|s^t), s^{t+1}\sim P(.|s^t,\underline{a}^t)]$

$Q_i^{\underline{\pi}}(s,\underline{a}) = \mathbb{E}[\sum_{t=0}^\infty\gamma^t r_i(s^t,\underline{a}^t)|s^0=s, \underline{a}^0=\underline{a}, \underline{a}^t\sim\underline{\pi}(.|s^t), s^{t+1}\sim P(.|s^t,\underline{a}^t)]$

$= r_i(s,\underline{a}) + \gamma\sum_{s'}P(s'|s,\underline{a})V_i^{\underline{\pi}}(s')$

Response Model:

$r_i(s,\underline{\pi}(.|s)) = \sum_{\underline{a}} r_i(s,\underline{a})\prod_j\pi_j(a_j|s)$,

$P(s'|s,\underline{\pi}(.|s)) = \sum_{\underline{a}}P(s'|s,\underline{a})\prod\pi_j(a_j|s)$

Best Response Policy: chose π_i that maximizes utility given $\underline{\pi}_{-i}$. Best Response can be computed by solving MDP w/ $P'(s'|s,a) = P(s'|s,a_i,\underline{\pi}_{-i}(.|s))$ & $r'(s,a_i) = r_i(s,a_i,\underline{\pi}_{-i}(s))$

Nash Equilibrium: $\underline{\pi}^*$ optimal, no player can improve

$\Rightarrow V_i^{\underline{\pi}^*}(s) \ge V_i^{\pi_i,\underline{\pi}_{-i}^*}(s)\ \forall\pi_i$

Def: ε-NE: $\underline{\pi}$ a ε-NE if: $V_i^{\underline{\pi}}(s) + \varepsilon \ge \max_{\pi_i}V_i^{\underline{\pi}}(s)\ \forall i$

Thm: All finite Markov games with a discounted ∞-horizon have a NE. Proof (idea): Construct a new normal-form game w/ each player & state pair in the original Markov Game, i.e. (i,s) as an agent in the new game

Nonlinear Optim.: minimize $_{\underline{\pi},V}\sum_{i\in\mathcal{I}}\sum_s(V_i(s) - Q_i(s,\underline{\pi}(.|s)))$ subject to:

(i) $V_i(s)\ge Q_i(s,a_i,\underline{\pi}_{-i}(.|s))\ \forall i,s,a_i$, (ii) $\sum_{a_i}\pi_i(a_i|s) = 1\ \forall i,s$,

(iii) $\pi_i(a_i|s)\ge 0\ \forall i,s,a$, $Q_i(s,\underline{\pi}(.|s)) = r_i(s,\underline{\pi}(.|s)) + \gamma\sum_{s'}P(s'|s,\underline{\pi}(.|s))\cdot V_i(s')$

$\underline{\pi}(.|s)$ has product structure ⇒ non-linear!

Thm: A joint policy $\underline{\pi}^*$ is a NE with val fct V^* iff $(\underline{\pi}^*,V^*)$ is a global min to the above non-linear programming.

Lemma: In an MDP V^* optimal value w/ optim policy π^* iff: $V^*(s) = r(s,\pi^*(.|s)) + \gamma\sum_{s'}P(s'|s,\pi^*(.|s))V^*(s')\ \forall s\in S$ & $V^*(s)\ge r(s,a) + \gamma\sum_{s'}P(s'|s,a)V^*(s')\ \forall s\in S, a\in\mathcal{A}$

Proof Thm: NE ⇒ NO: C2&C3 trivial, C1 ⇒ OPT ≥ 0, Lemma ⇒ C1 is feasible & OPT = 0. NO ⇒ NE: OPT ≥ 0 from C_1, OPT = 0 if take NE as a sol, Lemma, OPT = 0, C_1,C_2,C_3 ⇒ $\underline{\pi}^*,V^*$ is NE

Fictious Play in Markov Games:

Count opponent's actions at state s: $N_t(j,a_j,s)\ \forall j\ne i, \forall s\in S$

$\Rightarrow\tilde\pi_j^t(a_j|s) = N_t(j,a_j,s)/\sum_{\bar a_j\in\mathcal{A}_j}N_t(j,\bar a_j,s)$

$\Rightarrow\arg\max_{a_i}[Q_i^t(s,a_i,\underline{\tilde\pi}_{-i}^t(.|s)) = \sum_{\underline{a}_{-i}\in\mathcal{A}_{-i}}\prod_{j\ne i}\tilde\pi_j^t(a_j|s)Q_i^t(s,a_i,\underline{a}_{-i})]$

Policy Gradient Methods / Gradient Ascent:

$\pi_i^{t+1}(a_i|s) = \pi_i^t(a_i|s) + \alpha_i^t\frac{\partial V_i^{\underline{\pi}}(s)}{\partial\pi_i(a_i|s)}\Big|_{\underline{\pi}=\underline{\pi}_t}$ ⇒ Project to valid probd.

PG Algos in LQ (Linear Quadratic) Games: Lin. dyn. for S-trans.

$s\in\mathbb{R}^m$, $a_i\in\mathbb{R}^{d_i}$, $A\in\mathbb{R}^{m\times m}$, $B_i\in\mathbb{R}^{d_i\times m}$: $s^{t+1} = As^t + \sum_{i=1}^n B_i^T a_i^t$

↳ linear feedback: $a_i = \pi_i(s) = -K_i s$ with $K\in\mathbb{R}^{m\times d_i}$

↳ Loss-fct: $Q_i\in\mathbb{R}^{m\times m}$, $R_i\in\mathbb{R}^{d_i\times d_i}$: $\ell(K_1,\dots,K_n) = \mathbb{E}[(s^t)^T Q_i s^t + (a_i^t)^T R_i a_i^t]$, $s^0\sim\mathcal{D}_0$ ← init state distr.

Non-Convergence: $(K_1^*,\dots,K_n^*)$ NE if $\forall i$: $f(K_1^*,\dots,K_i^*,\dots,K_n^*)\le f(K_1^*,\dots,K_i,\dots,K_n^*)\ \forall K_i$

PG Algo: $K_i^{t+1} = K_i^t - \alpha_i\frac{\partial f}{\partial K_i}(K_1^t,\dots,K_n^t)$

Thm: ∃ LQ Game that the set of init cond in a neighborhood of the Nash eq. from which gradient converges to the Nash eq is of measure zero

Nash-Q-Learning: ∀stage t do: { Agent i takes action a_i^t & observes every player's reward, every player's action & the next state s_{t+1}. Computes the NE: $(\tilde\pi_1,\dots,\tilde\pi_n)$ for a normal form game which player j has the reward $r_j(\underline{a}) = Q_j^t(s^{t+1},\underline{a})$.

Compute the value of NE ∀ players i: Nash $Q_i^t = \sum_{\underline{a}}\prod_j\tilde\pi_j(a_j)Q_i^t(s^{t+1},\underline{a})$

Update: $Q_i^{t+1}(s^t,\underline{a}^t) = (1-\alpha_t)Q_i^t + \alpha_t[r_i(s^t,\underline{a}^t) + \gamma\,\mathrm{Nash}Q_i^t]$ }

Convergence of Nash-Q-Learning Thm: Assume: (i) every $s\in S$ & a_i $\forall i$ are visited ∞-often, (ii) $0\le\alpha_t<1$, $\sum_{t=0}^\infty\alpha_t = \infty$, $\sum_{t=0}^\infty\alpha_t^2<\infty$, (iii) Every NFG in Nash-Q-Learning with reward $r_j(\underline{a}) = Q_j^t(s^{t+1},\underline{a})$ has a global optimal point (may not exist, very strong assumption) (every agent receives its highest payoff) $\Rightarrow(Q_1^t,\dots,Q_n^t)\xrightarrow{conv.}(Q_1^*,\dots,Q_n^*)$

Rem: Diff to Q-Learn. in single agent MDP: not $\max_a Q$ but NashQ

5.4 Two-Player Zero-Sum Markov Games:

$r_1(s,a_1,a_2) = -r_2(s,a_1,a_2) = r(s,a_1,a_2)\ \forall s\in S$

$V^{\pi_1,\pi_2}(s) = \mathbb{E}[\sum_{t=0}^\infty\gamma^t r(s_t,a_1^t,a_2^t)|s_0=s, a_1^t\sim\pi_1(.|s_t), a_2^t\sim\pi_2(.|s_t), s_{t+1}\sim P(.|s_t,a_1^t,a_2^t)]$

Thm: $\max_{\pi_1}\min_{\pi_2}V^{\pi_1,\pi_2}(s) = \min_{\pi_2}\max_{\pi_1}V^{\pi_1,\pi_2}(s) =: V^*(s)\Rightarrow$ NE

V^* satisfies Bellman's type eq.:

Bellman Operators in Two-Player ZSMG: Def:

$[\mathcal{T}_{\pi_1,\pi_2}V](s) = r(s,\pi_1(s),\pi_2(s)) + \gamma\sum_{s'}P(s'|s,\pi_1(s),\pi_2(s))V(s')$

$\mathcal{T}_{\pi_1}V = \min_{\pi_2}\mathcal{T}_{\pi_1,\pi_2}V$, $\hat{\mathcal{T}}_{\pi_2}V = \max_{\pi_1}\mathcal{T}_{\pi_1,\pi_2}V$,

$\mathcal{T}V = \max_{\pi_1}\mathcal{T}_{\pi_1}V$, $\hat{\mathcal{T}}V = \min_{\pi_2}\hat{\mathcal{T}}_{\pi_2}V \Rightarrow\mathcal{T}\equiv\hat{\mathcal{T}}\Rightarrow$ FP of $\mathcal{T}$ & $\hat{\mathcal{T}}$: V^*

VI for 2P ZSMG:

∀t do: { Apply Bellman operator $\mathcal{T}$ @ each iter: $V^{t+1} = \mathcal{T}V^t$ }

Thm: $\|V^t - V^*\|_\infty\le\gamma^t\|V^0 - V^*\|_\infty$

PI for 2P ZSMG: π_1 is "greedy": $\pi_1\in G(V)$ iff $\forall s\in S$:

$\pi_1(.|s)\in\arg\max_{\pi_1(.|s)}\min_{\pi_2(.|s)}[r(s,\pi_1(s),\pi_2(s)) + \gamma\sum_{s'}P(s'|s,\pi_1(s),\pi_2(s))\cdot V(s')]$

Algo: ∀t do: { find $\pi_1^t\in G(V^{t-1})$ (sol of |S| lin. progr.), compute $V^t = \min_{\pi_2}V^{\pi_1^t,\pi_2}$ } (sol. MDP: $P = \mathbb{E}_{a_1\sim\pi_1^t}[P(.|s,a_1,a_2)]$, $r = -\mathbb{E}_{a_1\sim\pi_1^t}[r(s,a_1,a_2)]$)

⊕: Comp NE, Simple to implement

⊖: Comp. expensive, Model-Based

PG in 2P ZSMG (Model-free) ∀i do:

Sample $\tau = \{(s^t,a_1^t,a_2^t)\}_{t=0}^{H-1}\sim\pi_1^i,\pi_2^i$. Updates:

$\pi_1^{i+1}\leftarrow\Pi[\pi_1^i + \alpha(\sum_{t=0}^{H-1}r(s^t,a_1^t,a_2^t))\cdot\sum_{t=0}^{H-1}\nabla\log(\pi_1^i(a_1^t|s^t))]$

$\pi_2^{i+1}\leftarrow\Pi[\pi_2^i - \beta(\sum_{t=0}^{H-1}r(s^t,a_1^t,a_2^t))\cdot\sum_{t=0}^{H-1}\nabla\log(\pi_2^i(a_2^t|s^t))]$

where Π is the euclidean proj. to the set of policies

Thm: PG in 2P ZSMGs requires $\mathcal{O}(1/\varepsilon^{12.5})$ trajectories to converge to an ε-NE. ⊕ each player needs to learn only her individual experienced payoffs, efficient & simple ⊖: huge sample complexity

Optimistic GDA with actor critic:

At each stage i: $\tau = \{(s^t,a_1^t,a_2^t)\}_{t=0}^{H-1}\sim\pi_1^i,\pi_2^i$.

↳ $\hat Q^i(s,a_1)\leftarrow\frac{\sum_{t=0}^{H-1}\mathbb{1}\{s^t=s,a_1^t=a_1\}\cdot(r(a_1^t,a_2^t,s^t) + \gamma V^{i-1}(s^{t+1}))}{\sum_{t=0}^{H-1}\mathbb{1}\{s^t=s,a_1^t=a_1\}}$ & $\hat Q_2(s,a_2)\leftarrow\dots$

↳ At each state s, optimistic GA (GD for A2) uses $\hat Q^i(s,a)$ to update $\pi^i(.|s)$

Convergence: Optimistic GD/GA with actor-critic in 2P ZSG requires $\mathcal{O}(1/\varepsilon^4)$ traj. to conv. to an ε-NE

Improvement: NPG w/ actor-critic in 2P ZSG req. $\mathcal{O}(1/\varepsilon^2)$ samples to converge to an ε-NE.

6. Imitation Learning:

6.1 Introduction: Learning from Demonstration:

Given an expert's demonstrations $\{(s_i,\pi_E(s_i))\}$, reward signal is unobserved. Transition Model may be (un)known.

Goals & Approaches: (i) recover π_E directly: Imitation Learning (ii) recover reward fct: $r^{\pi_E}(s,a)$: inverse RL (IRL)

Categories: Expert demonstration: only access to offline expert's demonstration. Interactive Expert: Access to the environment & a queryable online expert. Interactive Simulator: Access to the environment & offline expert's demonstrations.

Goal: learn a policy $\pi\in\Pi$ that minimizes the Imitation Gap: $J(\pi_E) - J(\pi)$

6.2 Offline Immitation Learning: Behavioral Cloning

Assume $\exists\pi_E$. Input: offline data from expert's demonstration $\mathcal{D} = \{(s_i,a_i)\}_{i=1}^n$ where $a_i\sim\pi_E(s_i)$. Idea: directly learn π_E via Supervised learning. MLE: $\max_{\pi\in\Pi}\sum_{(s,a)\in\mathcal{D}}\log(\pi(a|s))$

Risk Minimization: $\min_\theta\mathbb{E}_{s\sim d^{\pi_E}, a\sim\pi_E(.|s)}[\ell(\pi_\theta(.|s),\pi_E(.|s))]$

Rem: MLE ⟺ min KL

Theoretical Guarantees for Behavior Cloning:

Thm: assume $r(s,a)\in[0,1]\ \forall s,a$. Suppose behavioral cloning returns an immitated policy $\hat\pi$ s.t. $\mathbb{E}_{s\sim d^{\pi_E}}[D_{KL}(\pi_E(.|s),\hat\pi(.|s))]\le\varepsilon$

Then: $J(\pi_E) - J(\hat\pi)\le\frac{2\sqrt{2}}{(1-\gamma)^2}\sqrt{\varepsilon}$

Proof:

Recall the advantage defined as $A^{\hat\pi}(s,a) = Q^{\hat\pi}(s,a) - V^{\hat\pi}(s)$ and $\mathbb{E}_{a\sim\hat\pi(\cdot|s)}A^{\hat\pi}(s,a) = 0, \forall s$.

Note also that $|A^{\hat\pi}(s,a)|\le\frac{1}{1-\gamma}$ since $0\le r(s,a)\le 1$.

Based on performance difference lemma, we have

$$J(\pi_E) - J(\hat\pi) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_E},\,a\sim\pi_E(\cdot|s)}A^{\hat\pi}(s,a)$$
$$= \frac{1}{1-\gamma}\left[\mathbb{E}_{s\sim d^{\pi_E},\,a\sim\pi_E(\cdot|s)}A^{\hat\pi}(s,a) - \mathbb{E}_{s\sim d^{\pi_E},\,a\sim\hat\pi(\cdot|s)}A^{\hat\pi}(s,a)\right]$$
$$\le\frac{1}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi_E}}\|\hat\pi(\cdot|s) - \pi_E(\cdot|s)\|_1$$
$$= \frac{1}{(1-\gamma)^2}\cdot 2\mathbb{E}_{s\sim d^{\pi_E}}[D_{TV}(\hat\pi(\cdot|s),\pi_E(\cdot|s))]\quad\text{(by definition)}$$
$$\le\frac{1}{(1-\gamma)^2}\cdot 2\mathbb{E}_{s\sim d^{\pi_E}}\left[\sqrt{2D_{KL}(\hat\pi(\cdot|s),\pi_E(\cdot|s))}\right]\quad\text{(Pinsker's inequality)}$$
$$\le\frac{2\sqrt{2}}{(1-\gamma)^2}\sqrt{\mathbb{E}_{s\sim d^{\pi_E}}[D_{KL}(\hat\pi(\cdot|s),\pi_E(\cdot|s))]}\quad\text{(Jensen's inequality)}$$

$D_{TV}(p,q) = \frac{1}{2}\sum_a|p(a) - q(a)|$

Corollary: Let Π be a discrete & realizable policy class, i.e. $\pi_E\in\Pi$. With prob ≥ 1-δ, behavioral cloning based on MLE returns a policy that obeys: $J(\pi_E) - J(\hat\pi_{MLE})\le\mathcal{O}\left(\frac{1}{(1-\gamma)^2}\sqrt{\frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}}\right)$

MLE guarantees: $\mathbb{E}_{s\sim d^{\pi_E}}[\|\hat\pi_{MLE} - \pi_E\|_{TV}^2]\le\log(|\Pi|/\delta)/|\mathcal{D}|$

Rem on BC: ⊕: Simple & effective, ⊖: No long-term planning, Cascading errors, Possible mismatch b/w train & test distr.

Difference to Supervised Learning: would like: $\min_\theta\mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_E(.|s)}[\ell(\pi_\theta(.|s),\pi_E(.|s))]$

6.3 Online Immitation Learning:

Dataset Aggregation: iteratively build up a policy via supervised Learning on aggregated $\mathcal{D}$ from π_E

Policy Aggregation: iteratively build up a policy by mixing newly trained policies.

Algo: Init π_0, $\forall t=1{:}T$ do: { Generate $\tau\sim\pi_{t-1}$, collect new data $\mathcal{D}_t = \{(s,\pi_E(s))|s\in\tau\}$ based on expert's feedback,

DA: run Behavioral Cloning with $\mathcal{D} = \mathcal{D}_1\cup\dots\cup\mathcal{D}_t\Rightarrow$ obtain π_t

PA: run BC with $\mathcal{D}_t\Rightarrow$ obtain $\hat\pi_t\Rightarrow\pi_t = \beta\hat\pi_t + (1-\beta)\pi_{t-1}$ }

Value Aggregation: reason about long term cost instead of agreement with a demonstrator. Assume access to $A^{\pi_E}(s,a)$

Algo: Init π_0, $\forall t=1{:}T$ do { sample $s_t\sim d^{\pi_t}$, collect $A^{\pi_E}(s_t,a)$ $\forall a$, Policy update: $\pi_{t+1} = \arg\min_{\pi\in\Pi}\sum_{j=0}^t\mathbb{E}_{a\sim\pi(.|s_j)}[A^{\pi_E}(s_j,a)]$ }

Online No-regret Learning:

Goal: minimize Regret: $R_T := \sum_{t=1}^T\ell_t(\underline{x}_t) - \min_{\underline{x}\in X}\sum_{t=1}^T\ell_t(\underline{x})$ (decision, loss)

Follow-the-(Regularized) Leader Algo: (FTRL)

$\underline{x}_t = \arg\min_{\underline{x}\in X}\sum_{i=1}^t\ell_i(\underline{x}) + \lambda R(\underline{x})$

6.4 Inverse RL:

RL: recovers a nearly optimal policy from reward fct.

IRL: recovers a reward fct & nearly optim pol. from demonstr. by E.

↳ avoids distr shift issue in Sup. Learn.

↳ Uses MDP structure for the learning from E-demonstr., while BC does not use MDP structure

IRL Objective: Assume E is optim. some $r(s,a)$ in mind. True reward fct is unknown; π_E is optimal policy of $\mathcal{M} = (S,\mathcal{A},P,r,\gamma)$

↳ Find reward fct $r(.,.)$: $S\times\mathcal{A}\to[-1,1]$ that explains experts behaviour:

$\pi_E\in\arg\max_{\pi\in\Pi}\mathbb{E}[\sum_{t=0}^\infty\gamma^t r(s_t,a_t)|s_0\sim\mu,\pi]$

↳ $\mathbb{E}[\sum_{t=0}^\infty\gamma^t r(s_t,a_t)|s_0\sim\mu,\pi_E]\ge\mathbb{E}[\sum_{t=0}^\infty\gamma^t r(s_t,a_t)|s_0\sim\mu,\pi]\ \forall\pi$

Challenges: • IRL does not necessarily have a unique rew. sol.

Thm: (Reward shaping) An expert policy π_E optimal in the MDP $\mathcal{M}$ with reward r is optimal also in the MDP $\mathcal{M}$ with reward fct $\hat r(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}[\Phi(s')] - \Phi(s)$ where $\Phi: S\to\mathbb{R}$ is called potential fct.

• Computationally expensive, In reality, don't observe π_E but trajectories $\tau\sim\pi_E$. Infeasible if π_E is not optimal.

Feature based reward: assume $r(s,a) = \underline{w}^T\cdot\underline{\phi}(s,a)$, $\|\underline{w}\|_2\le 1$

↳ $\mathbb{E}[\sum_{t=0}^\infty\gamma^t r(s_t,a_t)|\pi] = \mathbb{E}[\sum_{t=0}^\infty\gamma^t\underline{w}^T\phi(s_t,a_t)|\pi] = \underline{w}^T\mathbb{E}[\dots] = \underline{w}^T\underline{\mu}(\pi)$

↳ $\underline{\mu}(\pi)$ can be estimated from sampled trajectories.

↳ Goal: find $\underline{w}$ s.t.: $\underline{w}^T\underline{\mu}(\pi_E)\ge\underline{w}^T\underline{\mu}(\pi)\ \forall\pi\in\Pi$

Max Margin IRL: Find separating hyperplane given a set of points, where $\underline{\mu}(\pi_E)$ is given a label 1 & $\{\mu(\pi^j): j=0,\dots,J-1\}$ a label of -1: $\max_{\beta,\underline{w}}\beta$ s.t. $\|\underline{w}\|_2\le 1$, $\underline{w}^T\underline{\mu}(\pi_E)\ge\underline{w}^T\mu^j + \beta$, $j=1,\dots,J-1$

Feature Expectation Matching:

Directly find π s.t. $\underline{\mu}(\pi) = \underline{\mu}(\pi_E)$. If $\|\underline{\mu}(\pi_E) - \underline{\mu}(\pi)\|_2\le\varepsilon$, then: $\underline{w}^T\underline{\mu}(\pi_E) - \underline{w}^T\underline{\mu}(\pi)\le\|\underline{w}\|_2\|\underline{\mu}(\pi_E) - \underline{\mu}(\pi)\|_2\le\varepsilon$

Maximum Entropy Inverse RL:

$\max_{\pi\in\Pi}\mathbb{E}_{s\sim d^\pi, a\sim\pi(.|s)}[-\log(\pi(a|s))]$ s.t. $\mathbb{E}_{s\sim d^\pi, a\sim\pi(.|s)}[\phi(s,a)] = \mathbb{E}_{s\sim d^{\pi_E}, a\sim\pi_E(.|s)}[\phi(s,a)]$

↳ Lagrangian dual: $\min_w\max_{\pi\in\Pi}w^T(\mathbb{E}_{s,a\sim\lambda^\pi}[\phi(s,a)] - \mathbb{E}_{s,a\sim\lambda^{\pi_E}}[\phi(s,a)]) + \mathbb{E}_{s,a\sim\lambda^\pi}[-\log(\pi(a|s))]$

where λ^π = state-action occupancy measure

Algo: Alternatively update: (i) Inner max can be solved via soft policy iteration or entropy-reg. NPG. (ii) Grad is of the form $\mathbb{E}_{s,a\sim\lambda^{\pi^*}}[\phi(s,a)] - \mathbb{E}_{s,a\sim\lambda^{\pi_E}}[\phi(s,a)]$, where π^* is the sol to the inner max.

Feasibility problem: find π s.t. $\lambda^\pi(s,a) = \lambda^{\pi_E}(s,a)\ \forall s,a\Rightarrow$ not unique!

↳ $\max_\pi\sum_s d^\pi(s)H^\pi(s)$ s.t. $\lambda^\pi(s,a) = \lambda^{\pi_E}(s,a)\ \forall s,a\in S\times\mathcal{A}$

where: $H^\pi(s) = -\sum_a\pi(a|s)\log(\pi(a|s))$

D_θ: real → 1, fake → 0; G_ϕ: goal: $D(G(z))\to 1$

6.5 General Adversarial Networks:

GAN: $\min_{G_\phi}\max_{D_\theta}\mathbb{E}_{x\sim p_D}[\log(D_\theta(x))] + \mathbb{E}_z[\log(1 - D_\theta(G_\phi(z)))]$

Wasserstein GAN: $\min_{G_\phi}\max_{f_\theta: 1\text{-Lipschitz}}\mathbb{E}_{x\sim p_D}[f_\theta(x)] - \mathbb{E}_z[f_\theta(G(z))]$

$D_\theta: \mathcal{X}\to(0,1)$, $G_\phi: \mathcal{Z}\to\mathcal{X}$, $\mathcal{X}$: Images, $\mathcal{Z}$: $z\sim p(z)$, e.g. $\mathcal{N}(0,1)$

GAIL: Generative Adversarial Learning:

GAIL Objective:

$\min_\theta\max_\phi\mathbb{E}_{s,a\sim\lambda^{\pi_\theta}}[\log(C_\phi(s,a))] + \mathbb{E}_{s,a\sim\lambda^{\pi_E}}[\log(1 - C_\phi(s,a))] - c\cdot H(\pi_\theta)$

C_ϕ is the discriminator network: how likely $(s,a)\sim\pi_\theta$ & not π_E

↳ $C_\phi = \{(s,a)\sim\pi_\theta\to 1;\ (s,a)\sim\pi_E\to 0\}$

Primal-Dual Perspective:

$\max_\pi\mathbb{E}_{s,a\sim\lambda^\pi}[-\log(\pi(a|s))]$ s.t. $\lambda^\pi(s,a) = \lambda^{\pi_E}(s,a)\ \forall s,a$ (Primal)

Dual: $\min_r\max_\pi h(r) + \mathbb{E}_{s,a\sim\lambda^{\pi_E}}[r(s,a)] - \mathbb{E}_{s,a\sim\lambda^\pi}[r(s,a)] + \mathbb{E}_{s,a\sim\lambda^\pi}[-\log(\pi(a|s))]$ (h: convex regularizer)

↳ max-min formulation reduces to: $\max_{\pi\in\Pi} - h^*(\lambda^{\pi_E} - \lambda^\pi) + \mathbb{E}_{s,a\sim\lambda^\pi}[-\log(\pi(a|s))]$

↳ Let $g(x) = -x - \log(1 - e^{-x})$ & $h(r) = \{\mathbb{E}_{s,a\sim\lambda^{\pi_E}}[g(r(s,a))]$ if $r(s,a)\ge 0$; ∞ otherwise

↳ Fenchel conjugate of h:

$h^*(\lambda^{\pi_E} - \lambda^\pi) = \max_{D\in[0,1]}\mathbb{E}_{s,a\sim\lambda^\pi}[\log(D(s,a))] + \mathbb{E}_{s,a\sim\lambda^{\pi_E}}[\log(1 - D(s,a))]$

Fenchel Conj.: best lin approx: $f^*(y) = \max_{x\in\mathbb{R}^n}\{\langle y,x\rangle - f(x)\}$

↳ $h^*(x) = \max_r\{\langle x,r\rangle - h(r)\}\Rightarrow -h^*(-x) = \min_r h(r) + \langle x,r\rangle$

↳ can learn policy from demonstrations solving the following saddle point problem: $\min_{\pi\in\Pi}\max_{D\in[0,1]}\mathbb{E}_{s,a\sim\lambda^\pi}[\log(D(s,a))] + \mathbb{E}_{s,a\sim\lambda^{\pi_E}}[\log(1 - D(s,a))] - \mathbb{E}_{s,a\sim\lambda^\pi}[-\log(\pi(a|s))]$

GAIL with f-divergence:

$\min_{\pi\in\Pi}D_f(\lambda^{\pi_E},\lambda^\pi)$, $\lambda^{\pi_E},\lambda^\pi$ are state-action visitation distr.

Let $f(.)$ be convex & $f(1) = 0$. The f-div. is def:

$D_f(\mu,\nu) = \int\mu(x)f\left(\frac{\mu(x)}{\nu(x)}\right)dx$

Theoretical Guarantees: Assume $|r(s,a)|\le 1\ \forall s,a$.

Suppose GAIL returns an imitated policy π s.t. $D_f(\lambda^{\pi_E},\lambda^\pi)\le\varepsilon$. Then: $J(\pi_E) - J(\pi)\le\mathcal{O}(\frac{1}{1-\gamma}\sqrt{\varepsilon})$

Proof: $J(\pi_E) - J(\hat\pi) = \frac{1}{1-\gamma}\mathbb{E}_{\lambda^{\pi_E}}[r(s,a)] - \frac{1}{1-\gamma}\mathbb{E}_{\lambda^\pi}[r(s,a)]$

$\le\frac{1}{1-\gamma}\|\lambda^{\pi_E} - \lambda^\pi\|_1 = \frac{2}{1-\gamma}D_{TV}(\lambda^{\pi_E},\lambda^\pi)$

Rem: Offline IL: simple, scaleable, suffers from cascading error due to distr. shifts. Interactive IL: Alleviates cascading error, requires expensive expert queries. Inverse RL: explains expert's behavior, but poor sample efficiency & may not scale

Method	Reward learning	Access to environment	Interactive demonstrations	Pre-collected demonstrations
Offline IL	NO	NO	NO	YES
Interactive IL	NO	YES	YES	MAYBE
Inverse RL	YES	YES	NO	YES

7. Deep RL - Practice

Value-Based RL: V^*, Q^*, MC, SARSA, Q-learning ⊕ low varian. ⊖ doesn't scale

Policy-Based RL: π^*, PG, NPG, TRPO, PPO, ⊕ scales to large cont. ⊖ high var & smpl. ineff

7.1 Actor Critic Methods:

Actor uses PG to update learning policy. Critic uses TDL to estim val. fct.

↳ $\nabla_\theta J(\pi_\theta)\approx\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}[\mathbb{E}_{a\sim\pi_\theta(.|s)}[Q_w(s,a)\nabla_\theta\log(\pi_\theta(a|s))]]$ ($A_w(s,a)$)

↳ Actor: adj policy param $\underline{\theta}$ using PG based on low var value fct estim. updated by critic

Critic: update param $\underline{w}$ s.t. $Q_w(s,a)\approx Q^{\pi_\theta}(s,a)$ & $A_w\approx A_\theta$

↳ $\nabla_\theta J(\pi_\theta)$ maybe biased bec. of $Q_w\to$ careful choice

Compatible fct Approx: Suppose (i) $\nabla_w Q_w(s,a)|_{w=w^*} = \nabla_\theta\log(\pi_\theta(a|s))$

(ii) $\underline{w}^*\in\arg\min_{\underline{w}}\mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(.|s)}[(Q_w(s,a) - Q^{\pi_\theta}(s,a))^2]$,

⇒ exact: $\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(.|s)}[\nabla_\theta\log(\pi_\theta(a|s))Q_{w^*}(s,a)]$

Online Action Value AC:

Algo: Init $\underline{\theta}_0$, $\underline{w}_0$ state $s_0\sim\mu_0$, $a_0\sim\pi_{\theta_0}(.|s_0)$. $\forall t=1{:}T$: { Obtain $(r_t,s_{t+1},a_{t+1})\sim\pi_{\theta_t}$, Compute PG estim:

$\hat\nabla_\theta J(\pi_{\theta_t}) = Q_{w_t}(s_t,a_t)\nabla_\theta\log(\pi_{\theta_t}(a_t|s_t))$.

Actor update: $\underline{\theta}_{t+1} = \underline{\theta}_t + \alpha_t\hat\nabla_\theta J(\pi_{\theta_t})$

Compute TD: $\delta_t = r_t + \gamma Q_{w_t}(s_{t+1},a_{t+1}) - Q_{w_t}(s_t,a_t)$

Critic update: $\underline{w}_{t+1} = \underline{w}_t - \beta_t\delta_t\nabla_w Q_{w_t}(s_t,a_t)$ }

Advantage Actor-Critic: (A2C)

Algo: Init $\underline{\theta}_0$, $\underline{w}_0$, $s_0\sim\mu_0$. $\forall t=1{:}T$ do: { $a_t\sim\pi_{\theta_t}(.|s_t)\to$ obtain $(r_t,s_{t+1})\to$ Estim A: $\delta_t = r_t + \gamma V_{w_t}(s_{t+1}) - V_{w_t}(s_t)\to$ comp. PG: $\hat\nabla_\theta J(\pi_{\theta_t}) = \delta_t\nabla_\theta\log(\pi_{\theta_t}(a_t|s_t))\to$ Actor update: $\underline{\theta}_{t+1} = \underline{\theta}_t + \alpha_t\hat\nabla_\theta J(\pi_{\theta_t})$

→ Critic Update: $\underline{w}_{t+1} = \underline{w}_t - \beta_t\delta_t\nabla_w V_{w_t}(s_t)$ }

Extensions: Generalized Adv. Estim:

$\hat A_t^k(s_t,a_t) = r(s_t,a_t) + \gamma r(s_{t+1},a_{t+1}) + \dots + \gamma^{k-1}r(s_{t+k-1},a_{t+k-1}) + \gamma^k V_w(s_{t+k}) - V_w(s_t)$

$\hat A_t^{GAE}(s_t,a_t) = (1-\lambda)\sum_{k=1}^\infty\lambda^{k-1}\hat A_t^k(s_t,a_t)$

Soft AC: $\max_\pi\mathbb{E}[\sum_{t=0}^\infty\gamma^t r(s_t,a_t) + \lambda\cdot H(\pi(.|s_t))]$, $H = \mathbb{E}_{a\sim\pi(.|s)}[-\log(\pi(a|s))]$

7.2 Value Based Deep RL:

Recap Q-Learning: $\underline{w}_{t+1} = \underline{w}_t + \alpha_t[r_t + \gamma\max_a Q_{w_t}(s_{t+1},a) - Q_{w_t}(s_t,a_t)]\nabla_w Q_{w_t}(s_t,a_t)$

↳ Q-Learning is not a stochastic gradient descent method

↳ Naive deep QL could diverge due to sample correlation & moving targets.

Deep Q-Networks (DQN): experience replay & target network

Main Idea: $\min_{\underline{w}}\ell(\underline{w}) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}[(r + \gamma\max_{a'}Q(s',a';w^-) - Q(s,a;w))^2]$ (w^- held fixed & updated periodically)

Double DQN:

$\min_{\underline{w}}\ell(\underline{w}) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}[(r + \gamma Q(s',\arg\max_{a'}Q(s',a';\underline{w});w^-) - Q(s,a;w))^2]$

$\mathbb{E}[\max_i X_i]\ge\max_i\mathbb{E}[X_i]$, Jensen: $\varphi(\mathbb{E}[X])\le\mathbb{E}[\varphi(X)]$

8. DRL - Theory:

8.1 From DL Theory to DRL Theory:

Neural Tangent Kernel: SL: $\min_{w\in\mathbb{R}^m}\ell(w)$ where:

$\ell(w) = \frac{1}{2}\sum_{i=1}^n(f(x_i;w) - y_i)^2$.

↳ NTK: $K_{ij}(w) = \langle\nabla_w f(x_i;w),\nabla_w f(x_j;w)\rangle$

$K_{ij}(w_0) = \mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbb{1}_m)}[\langle\nabla_w f(x_i;w_0),\nabla_w f(x_j;w_0)\rangle]\succ 0$ (if m suff. large)

↳ Key Insights: Gradient flow: $\frac{dw(t)}{dt} = -\nabla\ell(w(t))$

def $\underline{u}(t) = f(w(t);x) - y\Rightarrow\frac{du}{dt} = -K(w(t))u(t)$

PL Cond: $\|\nabla\ell(w)\|^2 = (f(w;\underline{x}) - \underline{y})^T K(w)(f(w;\underline{x}) - \underline{y})\ge 2\cdot\lambda_{min}(K(w))\cdot\ell(w)$

Others: $K(w_0)\succ\mu_0$ for random w_0 & large $m = poly(n)$

· $\lambda_{min}(K(w)) - \lambda_{min}(K(w_0)) = \mathcal{O}(\frac{1}{\sqrt{m}})$ for $w\in\mathcal{B}$

· $K(w(t))\succ\mu$

· GD/SGD conv to global optim.

8.2 Convergence Analysis of Natural TD Learning

TD Learning with NNs: Def $x = \phi(s)\in\mathbb{R}^d$,

$\hat V(x;W,\alpha) = \frac{1}{\sqrt{m}}\sum_{i=1}^m\alpha_i\,\mathrm{ReLU}(W_i^T x)$

Symmetric Init: $\alpha_i = -\alpha_{i+m/2}\sim\mathcal{U}(\{\pm 1\})$

$W_i(0) = W_{i+m/2}(0)\sim\mathcal{N}(0,\mathbb{1}_d)$

Neural TDL: $W(t+1) = W(t) + \eta_t[r(x_t) + \gamma\hat V_t(x_{t+1}) - \hat V_t(x_t)]\nabla_W\hat V_t(x_t)$.

Goal: minimize mean-square Bellman Error

$\min_W\mathbb{E}_{x\sim\mu}[(\hat V(x;W,\alpha) - (r(x) + \gamma\mathbb{E}_{x'|x}[\hat V(x';W,\alpha)]))^2]$

↳ Stochastic semi-gradient method.

· With NN approx, MSBE becomes non-convex

Neural Tangent Kernel: Approx $\hat V(x;W,\alpha)$:

$\hat V(x;W,\alpha)\approx\hat V(x;W(0),\alpha) + \frac{1}{\sqrt{m}}\sum_{i=1}^m\alpha_i\mathbb{1}\{W_i^T(0)x\ge 0\}x^T\cdot[W_i - W_i(0)]$

$\approx\frac{1}{\sqrt{m}}\sum_{i=1}^m\alpha_i\mathbb{1}\{W_i^T(0)x\ge 0\}x^T W_i$

↳ $K(x,y) = \mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbb{1}_d)}[\mathbb{1}\{w_0^T x\ge 0\}\mathbb{1}\{w_0^T y\ge 0\}x^T y]$

Assumption: $V(x) = \mathbb{E}[v^T(w_0)x\cdot\mathbb{1}\{w_0^T x\ge 0\}]$ where $\sup_w\|v(w)\|_2\le\bar v<\infty$

NTDL with Regularization:

Projection Free NTD: $W(t+1) = W(t) + \eta\cdot g_t$ early stopping

Max-Norm NTD: $W_i(t+1) = \mathrm{Proj}_{\mathcal{B}(W_i(0),R)}[W_i(t) + \eta\cdot g_i]$ Regul: Max-Norm $\|W_i(t) - W_i(0)\|_2\le\frac{R}{\sqrt{m}}$

Convergence of NTD:

Assume $V(.)\in\mathcal{F}_{NTK}$ (dense in cont. fct. over compact state sp.)

↳ $\mathbb{E}[|\hat V_T - V|_\mu\mathbb{1}_\mathcal{E}]\le\varepsilon$ where $P(\mathcal{E})>1-\delta$

↳ Projection-free: NN width: $m = poly(\bar v)/\varepsilon^6$, $T = \frac{poly(R)}{\varepsilon^4}$

Max-Norm: $m = poly(R)/\varepsilon^2$, Sample compl: $T = \frac{poly(R)}{\varepsilon^6}$

↳ Projection Radius R

MinNormSol: $\bar W = [W_i(0) + \alpha_i\frac{v(W_i(0))}{\sqrt{m}}]_{i\in[m]}$

8.3 Conv. Analysis of Neural AC:

PG Methods: PG: $\theta_{t+1} = \theta_t + \eta_t\nabla V^{\pi_{\theta_t}}(\mu)$, $\cdot = \mathbb{E}_{s,a}[\nabla_\theta\log(\pi_\theta)Q^{\pi_\theta}]$

NPG: $\theta_{t+1} = \theta_t + \eta_t[G^{\pi_{\theta_t}}(\mu)]^{-1}\nabla_\theta V^{\pi_{\theta_t}}(\mu)$

↳ $G^{\pi_\theta}(\mu) := \mathbb{E}_{\pi_\theta}[\nabla_\theta\log(\pi_\theta)\cdot\nabla_\theta\log(\pi_\theta)^T]$

↳ $\arg\min_w\mathbb{E}_{s,a}[(w^T\nabla\log(\pi_\theta) - Q^{\pi_\theta})^2] = [G^{\pi_\theta}(\mu)]^{-1}\nabla_\theta V^{\pi_\theta}(\mu)$

Entropy Regularized PG Methods:

$\max_\pi V_\lambda^\pi(\mu) := V^\pi(\mu) + \lambda H^\pi(\mu)$, $H^\pi(\mu) = \mathbb{E}_{s_0\sim\mu}[\sum_{t=0}^\infty - \gamma^t\log(\pi(a_t|s_t))]$

ERPG: $\theta_{t+1} = \theta_t + \eta_t\nabla_\theta V_\lambda^{\pi_{\theta_t}}(\mu)$

ERNPG: $\theta_{t+1} = \theta_t + \eta_t[G^{\pi_{\theta_t}}(\mu)]^{-1}\nabla_\theta V_\lambda^{\pi_{\theta_t}}(\mu)$

↳ with averaging: $w_t\in\arg\min_{w: \|w\|_2\le R}\ell(w,\theta_t)$

$\theta_{t+1} - \theta_0 = (1 - \eta_t\lambda)(\theta_t - \theta_0) + \eta_t w_t$

$\ell(w,\theta) = \mathbb{E}_{s,a}[(\nabla_\theta^T\log(\pi_\theta(a|s))w - Q_\lambda^{\pi_\theta}(s,a))^2]$

Rem: Using entropy regularization, you get for free: $\inf_{t\ge 0}\inf_{s,a}\pi_t(a|s)\ge\pi_{min}>0\Rightarrow$ You visit all states ∞ often.

9. Advanced Topics:

Algorithm 1 Upper Confidence Bound Value Iteration (UCB-VI)

Initialize: dataset $\mathcal{D} := \emptyset$; $Q_h(s,a) := H$ for $h\in[H]$; $V_{H+1}(s) := 0$.
for episode $k = 1$ **to** K **do**
  (Part 1: Value iteration with bonus)
  for $h = H$ **to** 1 **do**
    for $(s,a)\in\mathcal{S}\times\mathcal{A}$ **do**
      if $(s,a)\in\mathcal{D}$ **then**
        Compute $\hat{\mathbb{P}}_h(s'|s,a) := \frac{N_h(s,a,s')}{N_h(s,a)}$ for all $s'\in\mathcal{S}$, where $N_h(s,a,s')$
        $Q_h(s,a) := \min\{H, r_h(s,a) + (\hat{\mathbb{P}}_h V_{h+1})(s,a) + b(N_h(s,a))\}$;
    for $s\in\mathcal{S}$ **do**
      $V_h(s) := \max_{a\in\mathcal{A}}Q_h(s,a)$;
  (Part 2: Executing the greedy policy)
  for $h = 1$ **to** H **do**
    Take action $a_h := \arg\max_{a\in\mathcal{A}}Q_h(s_h,a)$;
    Observe next state s_{h+1}, reward $r_h(s_h,a_h)$;
    $\mathcal{D} := \mathcal{D}\cup\{(h,s_h,a_h,s_{h+1})\}$;

$b_h(s,a) = H\sqrt{\frac{\log(SAT/\delta)}{N_h(s,a)}}$ & $N_h(s,a)$ are counting nbrs

Thm: With prob ≥ 1-δ, the regret of UCB-VI satisfies: Regret$(K)\le C_1 H^{3/2}\sqrt{SALT} + C_2 H^2 S^2 A L^2$

↳ $L = \log(HSAT/\delta)$, $T = KH$

With probability $1-\delta$, $Q_h^k(s,a)\ge Q_h^*(s,a)$ and $V_h^k(s)\ge V_h^*(s)$ for all (k,h,s,a).

- Follows by induction.
- Recall the Hoeffding inequality: if $Z_1,\dots,Z_n$ be independent random variables with $Z_i\in[a,b]$ for all i, then $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\ge t\right)\le 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$.
- By Hoeffding inequality, we have with probability $1-\delta$, for all s,a,h,k,

$$\left|(\hat P_h^k(\cdot|s,a) - P_h(\cdot|s,a))\cdot V_{h+1}^*\right|\le H\sqrt{\frac{\log(SAHK/\delta)}{N_h^k(s,a)}}.$$

- Suppose $V_{h+1}^k(s)\ge V_{h+1}^*(s)$ for all s, then either $Q_h^k(s,a) = H\ge Q_h^*(s,a)$ or

$$Q_h^k(s,a) - Q_h^*(s,a) = b_h^k(s,a) + \hat P_h^k(\cdot|s,a)\cdot V_{h+1}^k - P_h(\cdot|s,a)\cdot V_{h+1}^*$$
$$\ge b_h^k(s,a) + (\hat P_h^k(\cdot|s,a) - P_h(\cdot|s,a))\cdot V_{h+1}^*$$
$$\ge 0.$$

$\mathrm{Regret}(K) := \sum_{k=1}^K[V_1^*(s_1) - V_1^{\pi^k}(s_1)]\le\sum_{k=1}^K[V_1^k(s_1) - V_1^{\pi^k}(s_1)]$
$\le\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{s,a\sim d_h^{\pi^k}}[b_h^k(s,a) + (\hat P_h^k(\cdot|s,a) - P_h(\cdot|s,a))\cdot\hat V_{h+1}^k]$
$\le O(1)\cdot\sum_k\sum_h b_h^k + O(\delta KH)$
$\le\tilde O(1)H\sum_k\sum_h\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}} + \dots$
$\le\tilde O(1)H\sum_h\sum_{(s,a)}\sum_{i=1}^{N_h(s,a)}\sqrt{\frac{1}{i}} + \dots$
$\le\tilde O(1)H^{3/2}\sqrt{SAT} + \dots$ (Cauchy-Schwartz)