[$\pi=\{'_\circ V$: π is the greedy wrt. V]

1 MDP. Bellman eq. & operator $T^\pi V(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} \mathbb{E}_{s'\sim p(\cdot|s,a)} (r(s,a,s') + \gamma V(s'))$

T^π is mono., contraction wrt. $\|\cdot\|_\infty$. $T^\pi V^\pi = V^\pi$. $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} Q^\pi(a,s)$.

(action-value version) $\to$ fix point V^π. $\circ$ m+ form $T^\pi V = r^\pi + \gamma P^\pi V$.

$T^\pi Q(s,a) = \mathbb{E}_{s'\sim p(\cdot|s,a)} (r(s,a,s') + \gamma \mathbb{E}_{a'\sim\pi(\cdot|s')} Q(s',a'))$ and $(I-\gamma P^\pi)^{-1} \geq I$.

$T^\pi Q^\pi = Q^\pi$. $= I + \gamma P^\pi + \gamma^2 P^{\pi 2} + \cdots$

$(= \mathbb{E}_{s'\sim p(\cdot|s,a)} (r(s,a,s') + \gamma V^\pi(s')))$

Optim $V^*(s) \triangleq \max_\pi V^\pi(s)$. □1 (Policy improvement) $\pi'(a|s) = \begin{cases} 1 & a = \arg\max_a Q^\pi(s,a) \\ 0 & \text{ow.} \end{cases}$, $V^{\pi'} \geq V^\pi$.

□2 (optimal existence) Pf: $V^\pi = T^\pi V^\pi \leq T^{\pi'} V^\pi$. $V^\pi \leq \lim_k (T^{\pi'})^k V^\pi = V^{\pi'}$. (by □2)

(not unique) $\exists \pi^* \triangleq \{'_\circ V^*$ s.t. $V^* = V^{\pi^*}$.

Pf: $V^\pi \leq T^{\pi^*} V^*$, $\forall \pi$. $\Rightarrow V^* \leq V^{\pi^*}$. Bellman optimality eq. & operator $TV^* = V^*$.

$TV(s) = \max_a \mathbb{E}_{s'\sim p(\cdot|s,a)} (r(s,a,s') + \gamma V(s'))$

T is mono., contraction wrt. $\|\cdot\|_\infty$.

(a-v version) $Q^*(s,a) \triangleq \max_\pi Q^\pi(s,a)$, $V^*(s) = \max_a Q^*(s,a)$. $TQ^* = Q^*$.

$= \mathbb{E}_{s'\sim p(\cdot|s,a)} (r(s,a,s') + \gamma V^*(s')) = Q^{\pi^*}(s,a)$. $TQ(s,a) = \mathbb{E}_{s'\sim p(\cdot|s,a)} (r(s,a,s') + \gamma \max_{a'} Q(s',a'))$

VI. $V^{k+1} = TV^k = T^{\pi_{k+1}} V^k$. $\|V^k - V^*\|_\infty \leq \gamma^k \|V^0 - V^*\|_\infty$. $O(|S|^2|A|)$ per iter.

async. (Gauss-Seidel) $V(s) \leftarrow TV(s)$, $s = 1, 2, \ldots$.

PI. $\pi_{k+1} = \{'_\circ Q^{\pi_k}$. ($T^{\pi_{k+1}} V^{\pi_k} = TV^{\pi_k} \geq V^{\pi_k}$. $\|\cdot\|_\infty \sim$.

(first eval V^{π_k}) by □1 $V^{\pi_{k+1}} \geq V^{\pi_k}$. which is □1 Pf)

full exploration. if use a-v version can be model-free.

$\circ$ PI as Newton's solve (π_{k+1} is π for V^{π_k}): $F(V) := \max_\pi \{r^\pi + P^\pi V\} - V = 0$,

(Pf: $V^{\pi_{k+1}} \geq TV^{\pi_k} \geq V^{\pi_k}$, $V^{\pi_k} - (\gamma P^{\pi_{k+1}} - I)^{-1} F(V^{\pi_k})$ ← Jacobian $= V^{\pi_{k+1}}$.

$V^* - V^{\pi_k} \leq T^k (V^* - V^0)$)

$\circ$ TPI. $T^{\pi_k, m_k} V^{k-1} = V^k$, $\pi_{k+1} = \{'_\circ V^k$. API. $\|V^{\pi_k} - V^k\|_\infty \leq \delta$ eval,

$m_k \to \infty$ is PI, $m_k = 1$ is VI. $\|r^{\pi_{k+1}} + \gamma P^{\pi_{k+1}} V^k - TV^k\|_\infty \leq \varepsilon$ impv.

□3 (error amplification) $\pi = \{'_\circ V$, $V^\pi \geq V^* - \frac{2\gamma}{1-\gamma} \|V - V^*\|_\infty$ 1 (ε-optim.) Pf: 9.

(a-v version $\pi = \{'_\circ Q \sim -\frac{2}{1-\gamma} \|Q - Q^*\|_\infty$ 1)

Strongly polynomial PI ✓, VI ✗. LP ✓.

□4 (Strict progress) $\{\pi_k\}$ from PI, $\exists s$ s.t. $\forall k \geq \frac{1}{1-\gamma} \log \frac{1}{1-\gamma}$, $\pi_k(s) \neq \pi_0(s)$. Pf: 9.

$\Rightarrow$ at most $O(\frac{|S||A| - |S|}{1-\gamma} \log \frac{1}{1-\gamma})$ iters to optim.

2

MC. variation ctrl $\mathbb{E}f(x) = \mathbb{E}(f(x) + c(g(x) - \mathbb{E}g(x)))$, $c^* = -\frac{\mathrm{Cov}(f,g)}{\mathrm{V}(g)}$. importance sampling $\mathbb{E}_x f(x) = \mathbb{E}_Y \frac{f(y)p(y)}{q(y)}$, $q^*(y) = \frac{|f(y)|p(y)}{\mathbb{E}_x|f(x)|}$ (not computable)

Robbins-Monro algo. $x_{k+1} = x_k + \alpha_k(h(x_k) - x_k + e_k)$ bias-var tradeoff e.g. OLS or zero.
for $h(x) = x$.
e.g. $h(x) = \mu$, SGD $= x - \mathbb{E}_\xi \nabla f(x;\xi)$

□5 (RM convergence) $|h'(x)| \le \gamma < 1, \forall x$. $\sum_k^\infty \alpha_k = \infty$, $\sum_k^\infty \alpha_k^2 < \infty$. $\mathbb{E}(e_k|\mathcal{F}_k) = 0$, $\mathbb{E}(e_k^2|\mathcal{F}_k) < \infty$. $x_k \xrightarrow{a.s.} x^*$. ($\mathcal{F}_k = \{x_k, x_{k-1}, \cdots\}$) Pf: Dvoretzky Thm. (a.s.)

□5' (conv. rate) if $\alpha_k = \frac{1}{k}$, $\mathbb{E}\|x_k - x^*\|_2^2 \lesssim k^{-2\alpha\wedge\gamma}$ Pf: see.

sto. process $w_{k+1} = (1-a_k)w_k + \beta_k e_k$, e.a.s. $\beta \sim \gtrsim 0$
$w_k \xrightarrow{a.s.} 0$. let $x_k - x^* := w_k$.

(Pf of Dvore. $v_k := w_k^2$, $u_k := \alpha_k^2 w_k^2 + \beta_k^2 \mathbb{E}(e_k^2|\mathcal{F}_k)$.
$\mathbb{E}(v_k|\mathcal{F}_k) \le (1-\alpha_k)v_k + u_k$ with $\sum \alpha_k = \infty$, $\sum u_k < \infty$.
$S_k := v_k + \sum_{j=k}^\infty u_j$ is supermg. Doob $\Rightarrow S_k \xrightarrow{a.s.} S_\infty < \infty$
$\mathbb{E}(S_{k+1}|\mathcal{F}_k) \le S_k - \alpha_k v_k$. $\Rightarrow \sum \mathbb{E}\alpha_k v_k \le \mathbb{E}S_1 < \infty$. i.e. $\sum \alpha_k v_k < \infty$, a.s.
first is unbiased $w_\infty^2 = v_\infty = 0$ a.s. $\Leftarrow$ $\sum \alpha_k = \infty$
every X. short mse first > every, long mse first < every.)

(eval by MC)
primitive MC. $\nabla \tau_i \sim \pi_k$ from (s,a)
(policy impv by a-v) learn $Q^k(s,a) = \frac{1}{n}\sum_i^n (\sum_t^{T-1} \gamma^t r_t^i)$
$\pi_{k+1}(s) = \arg\max_a Q^k(s,a)$

use sub-traj. first/every visit.

every visit MC. $\nabla \tau \sim \pi$ for $t = T-1 \to 0$
eval $G = 0$ $G = \gamma G + r_t$
(or in incremental way $N(s_t) \mathrel{+}= 1$
$V(s_t) \mathrel{+}= \frac{G - V(s_t)}{N(s_t)}$) $G(s_t) \mathrel{+}= G$
$V(s_t) = G(s_t)/N(s_t)$
end

MC learn with greedy $\nabla \tau \sim \pi_k$ for $t = T-1 \to 0$
(may fail since lack of $G = 0$
exploration) if (s_t, a_t) not in $(s_0, a_0, \cdots, a_{t-1})$
exploitation tradeoff $N(s_t,a_t) \mathrel{+}= 1$; $Q(s_t,a_t) \mathrel{+}= \frac{G - Q(s_t,a_t)}{N(s_t,a_t)}$
$\pi_{k+1}(a|s_t) = \begin{cases} 1 \\ \text{or } \varepsilon \end{cases}$ $Q(s_t,a)$
end
end

ε-greedy. $\pi'(a|s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A|} \\ \frac{\varepsilon}{|A|} \end{cases}$ $a = \arg\max_a Q^\pi(s,a)$ □1' $\pi' = \varepsilon^1 Q^\pi$, $V^{\pi'} \gtrsim V^\pi$ when ε small.
ow.
Pf: $\sum_a \pi'(a|s) Q^\pi(s,a) \gtrsim \sum_a \pi(a|s) Q^\pi(s,a)$,

off-policy MC resample ratio $\rho_t^\pi(\tau_t) \triangleq \prod_{j=t+1}^\infty \frac{\pi(a_j|s_j)}{b(a_j|s_j)}$. $\Rightarrow \tau \sim b_k$, $W \xleftarrow{\text{i.e.}} \gamma^{\pi'} V^\pi \gtrsim V^\pi$.
$(b_k$ is soft $b_k(a|s) > 0, \forall a,s)$ $\frac{\rho_t^\pi(\tau_t)}{\rho_t^b(\tau_t)}$ $Q(s_t,a_t) \mathrel{+}= \frac{G_t \rho_t^\pi - Q(s_t,a_t)}{N(s_t,a_t)}$.
$W *= \frac{\pi_k(a_t|s_t)}{b_k(a_t|s_t)}$. $(Q^\pi = \mathbb{E}_{\tau\sim p^\pi}(G) = \mathbb{E}_{\tau\sim p^b}(\frac{p^\pi(\tau)}{p^b(\tau)}G))$

(conv. conditions are mild and thus omitted)

TD. eval by solve Bellman eq. stochastic and online. TD(0) $V_{(s)}^{t+1} = V_{(s)}^t + \alpha_t(s)$.
$(s,a,s') \sim \pi$. $(r(s,a,s') + \gamma V_{(s')}^t - V_{(s)}^t)$
(note is biased est. to V^π, $\neq V^t$)

• MC has high var ×, TD low var ✓. learns from
complete episode and (TD(0) one step) incomplete episode by bootstrapping.
no MDP structure × (exploit MDP structure) ✓
but unbias est. $V^\pi(s_t)$ first visit.

n-step. $(T^\pi)^n V(s) = \mathbb{E}_\pi(\sum_{k=0}^{n-1} \gamma^k r_k + \gamma^n V(s_n)|s_0 = s)$,
$G_{(s)}^\lambda \triangleq (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1} G_{(s)}^n$ ✓
$G_{(s)}^{t,n} \triangleq \sum_{k=0}^{n-1} \gamma^k r_k + \gamma^n V_{(s_n)}^t$, $V_{(s)}^{t+1} = V_{(s)}^t + \alpha_t(s)(G_{(s)}^{t,n} - V_{(s)}^t)$
TD(λ). $= \sum_{k=1}^\infty (\lambda\gamma)^k (r_k + \gamma V_{(s_{k+1})}^t - V_{(s_k)}^t)$ $(s_0 = s, a_0, \cdots, s_n) \sim \pi$
$+ V_{(s)}^t$
$\lambda = 0 \to$ TD(0). $\lambda = 1 - 0 \to G_{(s)}^\lambda = \sum_{k=0}^\infty \gamma^k r_k$ MC eval i.e. TD(1). (is a bias-var tradeoff)

(tabular VFA)
[or TD(0) tabular is RM to solve Beq.]
(+VFA eval) (if onehot i.e. tabular, SGD)
[TD(0) is semi SGD to BE problem and SGD to BE' $[\cdots \pi_t]$ (or π_{t+1})] (moving target)

(a-v version) $Q^{t+1}(s,a) = Q^t(s,a) + \alpha_t(s,a)(r(s,a,s') + \gamma Q^t(s',a') - Q^t(s,a))$. (RM for $Q^{t+1} \leftarrow f_e Q^t$. ($f_e$ when model known $\mathbb{E}_{s'\cdots}$)

SARSA. TD learn with ε-greedy. (con policy) $\nabla (S_t, a_t, r_t, S_{t+1}, a_t) \sim \pi_t$ TD(0) for $Q(S_t, a_t)$ $\pi_{t+1}(a|s_t) = \begin{cases} 1-\varepsilon \\ \varepsilon \end{cases}$ $Q^{t+1}(s_t,a)$

Q learn (off policy and online) $Q^{t+1}(s,a) = Q^t(s,a) + \alpha_t(s,a)(r + \gamma \max_{a'} Q^t(s',a') - Q(s,a))$ $(s,a,r,s') \sim b$. ∇ do the sampling and updng. (upd policy $\gamma \mathbb{E}_{a' \sim \pi_{target}} Q^t(s',a')$ is just $\max_{a'} Q^t(s',a')$ so no resampling. i.e. $f_e^{\pi} Q = f_e Q$.) s' i.e. S_{t+1}

(SARSA(λ) using TD(λ))

Fact: Jessen ineq. $\mathbb{E}_x \max_a Q(s',a) \geq \max_a \mathbb{E}_x Q(s',a)$.

double Q learn → one choose a^* but upd another. $\nabla (S_t, a_t, r_t, S_{t+1}) \sim b_t$ (e.g. ε-greedy wrt. $\frac{Q^{t,A}+Q^{t,B}}{2}$)

prob. $\frac{1}{2}$: $a^* = \arg\max_a Q^{t,A}(S_{t+1}, a)$; $Q^{t+1,A}(s_t,a_t) = Q^{t,A}(s_t,a_t) + \alpha_t(s_t,a_t)(r_t + \gamma Q^{t,B}(S_{t+1}, a^*) - Q^{t,A}(s_t,a_t))$

prob. $\frac{1}{2}$: $b^* = \arg\max_a Q^{t,B}(S_{t+1}, a)$; $Q^{t+1,B}(s_t,a_t) = \sim$ (to enhance explore SARSA $\neq$) unbiased.

3 VFA

value func approx. $V^{\pi}(s) \leftarrow V(s;w)$, e.g. lin. $\phi(s)'w$ or $NN_w(s)$. (not tabular) $Q^{\pi}(s,a) \leftarrow Q(s,a;w)$.

$\min_w \mathbb{E}_{s\sim D} (V(s;w) - V^{\pi}(s))^2$

SGD $\Rightarrow w_{t+1} = w_t + \alpha_t \cdot (V^{\pi}(s) - V(s;w_t)) \nabla_w V(s;w_t)$

(for lin. $\nabla_w V = \phi(S_t)$)

- TD(0) eval VFA perspectives $w_{t+1} = \arg\min_w \mathbb{E}_{s\sim D} (V(s,w) - \mathcal{T}^{\pi} V(s,w_t))^2$ by MC or TD eval. $G(s)$ / $r + \gamma V(s';w_t)$ (biased)

or Bellman error $\triangleq \mathbb{E}_{s\sim D} (V(s;w) - \mathcal{T}^{\pi} V(s;w))^2$ and SGD gives it. ($\|V - V^{\pi}\|_\infty \leq \frac{\|V - \mathcal{T}^{\pi} V\|_\infty}{1-\gamma}$.)

min semigrad replaces $\mathcal{T}^{\pi} V(s;w)$ by $\mathcal{T}^{\pi} V(s;w_t)$ also $\mathcal{T}^{\pi}$. f_e^{π} version) (MC conv. to the above min by SGD, like IRLS) to vanish $\nabla_w \mathcal{T}^{\pi} V(s;w)$.

~ interpreting Q learn. TD conv. to min projected Be if lin.)

learn with VFA. (a-v version ↓) SARSA-VFA $\nabla (S_t, a_t, r_t, S_{t+1}, a_{t+1})$ $a_{t+1} \sim \pi_t$ (to ϕ')

(both follow TD target $f_e^{\pi} Q$, SARSA est. by samples $\sim \pi$ and Q-learn use greedy π $f_i^{\pi} Q = f_e Q$. ($\pi := \varepsilon'_0 Q$)

TD eval VFA ← $w_{t+1} = w_t + \alpha_t (r_t + \gamma \phi(S_{t+1}, a_{t+1})' w_t - \phi(S_t, a_t)' w_t) \phi(S_t, a_t)$ with lin.

ε-greedy upd ← $\pi_{t+1} = \begin{cases} 1-\varepsilon_t \\ \varepsilon_t \end{cases}$ $\phi(S_t, a)' w_{t+1}$.

Q-learn-VFA similar. $\min_w \mathbb{E}_{(s,a)\sim D} (Q(s,a;w) - f_e Q(s,a;w))^2$ ← SGD.

And both upd. policy right after value para upd, which may be stab. issue in complex VFA e.g. deep Q learn) FQI. see $w_t =$ below. (i.e. approx VI with steps)

DQN. ∇ replay buffer D to capcity N: (pin target a while (policy to upd.) and minibatch)

(FQI with deep NN for VFA and adopts incremental learning by buffer) break seq. dependence

Q network $Q(s,a;w)$, target Q network $Q(s,a;\tilde{w})$; $\tilde{w} = w$; SGD iter number C; $k=0$; s_0

for $t=0 \to$? stop

$k+=1$

$(S_t, a_t, r_t, S_{t+1}) \sim b_t$ from S_t; add into D.

batch B of D

$w += \alpha \sum_{(s,a,r,s')_i \in B} (r_i + \gamma \max_{a'} Q(s_i', a'; \tilde{w}) - Q(s_i, a_i; w)) \nabla_w Q(s_i, a_i; w)$

if $k == C$: $\tilde{w} = w$; $k = 0$

end

end

(tricks: double, dueling, prioritized replay, clip..)

[□6 (Tsitsiklis - VanRoy)]

○ PBE $Q = \Pi \mathcal{T}_e Q$ where proj. $\Pi_{\mathcal{D}} f = \arg\min_{g \in \mathcal{Q}_\Pi} \|f-g\|_{L^2(\mathcal{D})} \triangleq \arg\min_{g \in \mathcal{Q}_\Pi} \mathbb{E}_{(s,a)\sim\mathcal{D}} (f-g)^2$ ($\mathcal{D}$ stn. sample from π)

$V = \Pi \mathcal{T} V$ $\mathcal{Q}_\Pi \triangleq \{Q(\cdot;w) \mid w \in \mathbb{R}^n\}$ subsp. □6 TD(0) if lin. VFA and on policy, eval π $\sum \alpha_t = \infty, \sum \alpha_t^2 < \infty$,

above actually $\mathbb{E}_{(s,a)\sim\mathcal{D}} (Q - \mathcal{T}_e Q_t)^2$

thus $\leq \mathbb{E}_{(s,a)\sim\mathcal{D}, s'\sim P(\cdot|s,a)} (Q - (r + \gamma \max_{a'} Q(s',a',w_t)))^2$ (i.e. TD(0) min PBe not (MSBe)) conv. to PBE solution.

Q-learn-VFA is noisy upperbound. in tabular case (Pf by mt) (i.e. semi SGD's conv. condition)

above means $Q(\cdot;w_{t+1}) = \Pi \mathcal{T}_e Q(\cdot;w_t)$ $\Pi = I$, Q-learn(VFA) → trinity. solve (P)BE $Q_{t+1} = Q_t + \alpha_t (\hat{\mathcal{T}} Q_t - Q_t + e_t)$

but only if □6 condition ow. $\Pi \mathcal{T}_e$ may not be contraction ⇏ PBE. (Baird counterex) ↑ above: $e_s, e_{s,a}$ by RM $\leq \mathbb{E}\mathbb{E}\cdots$ all conv. (note differ $\forall s, (r + \gamma \max) > 0$)

4 model/policy-based RL. simplex para. $\pi_s \in \Delta, \forall s$. Softmax para. $\pi_\theta(a|s) = \frac{e^{f_\theta(s,a)}}{\sum_{a'} e^{f_\theta(s,a')}}$ but VFA introduces error of est. $\mathcal{T}_e Q$ interacting with ($Q \in \mathcal{Q}_\Pi$) Q-error may diver. (biased but $\xrightarrow{a.s.} Q^*$ of BE)

$V^{\pi_\theta}(\mu) \triangleq \mathbb{E}_{s_0\sim\mu} V^{\pi_\theta}(s_0) = \mathbb{E}_{\tau\sim p^{\pi_\theta}_\mu} (\sum_{t=0}^{\infty} \gamma^t r_t)$, RL $\arg\max_\theta V^{\pi_\theta}(\mu)$.

$V^\pi(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi_\mu} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathbb{E}_{s'\sim P(\cdot|s,a)} r(s,a,s')$, where $d^\pi_\mu(s) \triangleq \mathbb{E}_{s_0\sim\mu} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi)$ (discounted state visit measure), $d^\pi_\mu = (1-\gamma)(I - \gamma P^\pi)^{-1} \mu$

advantage func $A^\pi(s,a) \triangleq Q^\pi(s,a) - V^\pi(s)$.

□7 (Performance difference lem.) $V^{\pi_1}(\mu) - V^{\pi_2}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_1}_\mu} \mathbb{E}_{a\sim\pi_1(\cdot|s)} A^{\pi_2}(s,a)$.

Pf: $V^{\pi_1} - V^{\pi_2} = (I - \gamma P^{\pi_1})^{-1} (\mathcal{T}^{\pi_1} V^{\pi_2} - V^{\pi_2})$, $\mathcal{T}^{\pi_1} V^{\pi_2}(s) = \mathbb{E}_{a\sim\pi_1(\cdot|s)} Q^{\pi_2}(s,a)$.

□8 (Policy gradient) $\nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau\sim p^{\pi_\theta}} (\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t,a_t) \nabla_\theta \log \pi_\theta(a_t|s_t))$ (Q^{π_θ} here can be viewed as eval)

Pf: $\nabla_\theta Q^{\pi_\theta}(s_0,a_0) = \gamma \mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)} \nabla_\theta V^{\pi_\theta}(s_1)$. $= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_\theta}_\mu} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} (Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s))$.

PPG. $\nabla_{\pi_s} V^\pi(\mu) = \frac{d^\pi_\mu(s)}{1-\gamma} Q^\pi(s,\cdot)$ by □8. (or $\sim A^{\pi_\theta}(s,a) \sim$) ○ note traj. expr $\mathbb{E}_{\tau\sim\pi_\theta} \sum_t G_t \nabla_\theta \log \pi_{\theta_t}$ if $\sum_a \pi_\theta(a|s) = 1$. → also uses, but visitation meas. expr $\nabla_\theta (\mathbb{E}_{s,a\sim d^{\pi_\theta}} Q^{\pi_\theta})$ ($\frac{1}{1-\gamma}$ omit)

upd. $\pi^{k+1} = \arg\min_{\pi\in\Pi} \{ -\alpha_k \langle \nabla_\pi V^\pi(\mu)|_{\pi=\pi_k}, \pi - \pi^k \rangle + \frac{1}{2} \|\pi - \pi^k\|^2 \}$

$\Rightarrow \pi_s^{k+1} = \mathrm{Proj}_\Delta (\pi_s^k + \frac{\alpha_k d^k_\mu(s)}{1-\gamma} Q^k(s,\cdot))$, $\forall s$. $O(\frac{1}{k})$ conv.

Softmax PG. $\nabla_{\theta_s} V^{\pi_\theta}(\mu) = \frac{d^\pi_\mu(s)}{1-\gamma} \pi_\theta(\cdot|s) A^{\pi_\theta}(s,\cdot)$ by □8 and $\frac{\partial \log \pi_\theta(a'|s')}{\partial \theta_{s,a}} = \mathbb{1}(s'=s)(\mathbb{1}(a'=a) - \pi_\theta(a|s))$. $\stackrel{\neq}{=}$ traj. $\nabla_\theta (\mathbb{E}_\tau \sum_t \gamma^t r_t)$.

upd. $\theta_s^{k+1} = \theta_s^k + \alpha_k \sim$. $O(\frac{1}{k})$.

PMA. $\pi^{k+1} = \arg\min_{\pi\in\Pi} \{ \sum_s (-\alpha_k \langle \sim_s \rangle + \frac{d^k_\mu(s)}{1-\gamma} D_h(\pi_s, \pi_s^k)) \}$ e.g. $h(\pi_s) = \frac{1}{2}\|\pi_s\|_2^2 \Rightarrow \pi_s^{k+1} = \mathrm{Proj}_\Delta(\pi_s^k +$ Proj QAscent. $\alpha_k Q^k(s))$,

(regularizer of visi. reweighted Bregman divergence) $\overset{\text{simplex}}{\pi_s^{k+1}} = \arg\min_{\pi_s\in\Delta} \{ -\alpha_k \langle Q^k(s,\cdot), \pi_s - \pi_s^k \rangle + D_h(\pi_s, \pi_s^k) \}$, $\forall s$. $h(\pi_s) \triangleq h(\pi_s^k) - \langle \nabla h(\pi_s^k), \pi_s - \pi_s^k \rangle$. $h(\pi_s) := \sum_a \pi_s(a) \log \pi_s(a)$, $D_h = \mathrm{KL}$.

$\alpha_k \to \infty$, PPG. (PMA) → PI. Expo QAscent. $\pi^{k+1}_{s,a} \propto \pi^k_{s,a} e^{\alpha_k Q^k_{s,a}}$

PQA. EQA pf: ◇ (coincides NPG) ($\mathrm{KL}(\pi_s^{k+1} \| \pi_s^k) \geq 0$ $= \log Z_s^k$.)

↘ both $O(\frac{1}{k})$ conv.

PPG. PQA conv. infinite iters.

(REINFORCE) ∇ for $k=0\to?$ sample $\mathcal{D}_k=\{\tau^i\}\sim\pi_{\theta_k}$ $(G_t^i:=\sum_{t'=t}^{T-1}\gamma^{t'-t}r_{t'}^i)$

$g_k=\frac{1}{|\mathcal{D}_k|}\sum_i\sum_t\gamma^t G_t^i\nabla_\theta\log\pi_{\theta_k}(a_t^i|s_t^i)$

$\theta_{k+1}=\theta_k+\alpha_k g_k$

PG + MC eval. learn π_θ via pg. learn in $\nabla_\theta J(\pi_\theta)$ via $V/Q(;w)$ eval.

actor-critic a-v a-c ∇ for $t=0\to?$ $(s_t,a_t,r_t,s_{t+1},a_{t+1})\sim\pi_\theta$

(TD, on policy) $\delta_t=r_t+\gamma Q(s_{t+1},a_{t+1};w)-Q(s_t,a_t;w)$

(v func, adv ver) (or using all the episode to upd.) c upd. $w\mathrel{+}=\alpha_t\delta_t\nabla_w Q(s_t,a_t;w)$

A2C. a upd. $\theta\mathrel{+}=\beta_t Q(s_t,a_t;w)\nabla_\theta\log\pi_\theta(a_t|s_t)$ end

$A(s_t,a_t)\approx$ e.g. MC, n-TD for w. note for π_θ if not normalized (need to Proj$_\Delta$ or softmax)

$r_t+\gamma V(s_{t+1};w)-V(s_t;w)$. MC $G(s,t)$ can't as multiplier to upd.

∇ $(s_t,a_t,r_t,s_{t+1})\sim\pi_\theta$

$\delta_t=r_t+\gamma V(s_{t+1};w)-V(s_t;w)$

$w\mathrel{+}=\alpha_t\delta_t\nabla_w V(s_t;w)$

$\theta\mathrel{+}=\beta_t\delta_t\nabla_\theta\log\pi_\theta(a_t|s_t)$

NPG. $\max_\theta V_{(\mu)}^{\pi_\theta}=\mathbb{E}_{\tau\sim p_\mu^{\pi_\theta}}r(\tau)$ if it's argmax thats $\|d\|\le\alpha$ $\Delta\theta\propto\nabla J(\theta)$.

Pf: $KL(P_\theta\|P_{\theta+d})\le\alpha\Leftrightarrow\frac12 d'F(\theta)d\le\alpha$. Fisher mat $F(\theta):=\mathbb{E}_{P_\theta}(\nabla_\theta\log P_\theta(\nabla_\theta\log P_\theta)')$

(expand $P_\theta(x)$ to second order)

TRPO. cmp. to ◻7. write $V_{(\mu)}^{\pi_\theta}$, $\max_\theta V_k(\theta):=V_{(\mu)}^{\pi_k}+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_k},a\sim\pi_\theta(\cdot|s)}A^{\pi_k}(s,a)$.

Facts: $V_{(\mu)}^{\pi_\theta}$ and $V^k(\theta)$ match at θ_k up to first derivative;

$V_{(\mu)}^{\pi_\theta}\ge V_k(\theta)-\frac{2\gamma\varepsilon}{(1-\gamma)^2}\max_s KL(\pi_{\theta_k}(\cdot|s)\|\pi_\theta(\cdot|s))$

$\varepsilon:=\max_{s,a}|A^{\pi_{\theta_k}}(s,a)|$.

$\Rightarrow\max_\theta V_k(\theta)\leftarrow\nabla_\theta V_{(\mu)}^{\pi_{\theta_k}}(\theta-\theta_k)$ s.t. $\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_k}}}KL(\pi_{\theta_k}(\cdot|s)\|\pi_\theta(\cdot|s))\le\delta$.

(replace by $\mathbb{E}$ version) $\frac12(\theta-\theta_k)'F(\theta_k)(\theta-\theta_k)\le\delta$.

$\approx$ NPG + line search (adaptive) in implementation.

average case: $\lim_{T\to\infty}\frac1T\mathbb{E}_\tau\sum_{t=0}^{T-1}\cdots=\mathbb{E}_{s\sim d_\mu^{\pi_\theta},a}\cdots$ stn. dis.

discounted case: $(1-\gamma)\mathbb{E}_\tau\sum_{t=0}^{\infty}\gamma^t\cdots=\mathbb{E}_{s\sim d_\mu^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\cdots$

where $w^*=\arg\min_w\mathbb{E}_{s,a}(\nabla_\theta\log\pi_\theta(a|s)'w-A^{\pi_\theta}(s,a))^2$.

(Pf by $\nabla_\theta\log\pi_\theta(a|s)'w^*=\sum_{a'}\nabla_\theta\pi_\theta(a|s)(A^{\pi_\theta}(s,a')+c_s)=A^{\pi_\theta}(s,a)$.)

in RL, $F(\theta)=\mathbb{E}_\tau\big(\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\cdot(\sum_{t=0}^{\infty}\nabla\log\pi_\theta(a_t|s_t))'\big)$ (ass!)

$=\mathbb{E}_\tau\big(\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)\nabla_\theta\log\pi_\theta(a_t|s_t)'\big)$.

$F(\theta)^\dagger\nabla_\theta V_{(\mu)}^{\pi_\theta}=\frac{1}{1-\gamma}w^*$, where γ stepsize. same with EQA.

for Softmax, $w_{s,a}^*=A^{\pi_\theta}(s,a)+c_s$, $\pi_\theta^{k+1}(a|s)=\pi_\theta^k(a|s)\cdot e^{\frac{\gamma}{1-\gamma}A^{\pi_\theta}(s,a)}/\sum_{a'}\cdots$

PPO. $V_k(\theta)-V_{(\mu)}^{\pi_k}\propto\mathbb{E}_{s\sim d_\mu^{\pi_k},a\sim\pi_\theta(\cdot|s)}\Big(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}A^{\pi_{\theta_k}}(s,a)\Big)$.

(surrogate of true target in small region around θ_k, under sampling from π_{θ_k})

dip. $r(\theta)\triangleq\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}$, $1-\varepsilon\ \underbrace{r(\theta)}_{1=r(\theta_k)}\ 1+\varepsilon$

thus θ won't move far using PG type method. $L^{clip}(\theta,\theta_k,s,a)\triangleq\min\{r(\theta)A,\ \mathrm{clip}_\varepsilon r(\theta)A\}$ (lower bound conservative)

∇ for $k=0\to?$

$\tau_n=\{s_t,a_t,r_t,s_{t+1}\}_{t\ge0}\sim\pi_{\theta_k}$

$\delta_t(w_k)=r_t+\gamma V(s_{t+1};w_k)-V(s_t;w_k)$

$A^\lambda(s_t,a_t)=\sum_{l\ge0}(\gamma\lambda)^l\delta_{t+l}(w_k)$; $G^\lambda(s_t)=A^\lambda(s_t,a_t)+V(s_t;w_k)$ ↖ actually $G^\lambda(s_t,a_t)$, e.g. $A<0$, $r>1+\varepsilon$, less.

$\theta_{k+1}=\arg\max_\theta\sum_{\tau_n}\sum_t L^{dip}(\theta,\theta_k,s_t,a_t)$

$w_{k+1}=\arg\min_w\sum_{\tau_n}\sum_t(V(s_t;w)-G^\lambda(s_t))^2$ ↖ when fit w we take average, not affecting $\max_\theta\cdots$.

end

in implem. use GDer. e.g. Adam,

$\theta\leftarrow\theta+\beta\sum_{i\in\mathcal{B}_\tau}\nabla_\theta L^{clip}(s_i,a_i;\theta,\theta_k)$; $w\leftarrow w+\alpha\frac{1}{|\mathcal{B}_\tau|}\sum_{i\in\mathcal{B}_\tau}(G^\lambda(s_i)-V(s_i;w))\nabla_w V(s_i;w)$

entropy regularization $V_\tau^\pi(\mu) \triangleq \frac{\mathbb{E}_{s\sim d_\mu^\pi}}{1-\gamma} \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim p(\cdot|s,a)} (r(s,a,s') + \tau H(\pi(\cdot|s)))$ $= \mathbb{E}_{s_0\sim\mu,\pi} \left(\sum_{t=0}^{\infty} \gamma^t (r_t - \tau \log \pi(a_t|s_t)\right.$

(like revising reward to encourage explore) Bellman op. $\mathcal{T}_\tau^\pi V(s) \triangleq \mathbb{E}_{a,s'} (r - \tau\log\pi(a|s) + \gamma V(s'))$.

$\max_\pi \mathcal{T}_\tau^\pi V \Rightarrow \pi(a|s) \propto e^{\frac{Q(s,a)}{\tau}}$. Q_τ^π def the same, $V_\tau^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} (Q_\tau^\pi(s,a) - \tau\log\pi(a|s))$,

□7' $V_\tau^{\pi_1}(\mu) - V_\tau^{\pi_2}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_1}(s)} (\mathbb{E}_{a\sim\pi(\cdot|s)} A_\tau^{\pi_2}(s,a) - \tau KL(\pi_1(\cdot|s)\|\pi_2(\cdot|s)))$ $\mathbb{E}_{a\sim\pi(\cdot|s)} A_\tau^\pi(s,a) = 0$, $A_\tau^\pi \triangleq Q_\tau^\pi - V_\tau^\pi - \tau\log\pi$

□2' $\mathcal{T}_\tau = \max_\pi \mathcal{T}_\tau^\pi$. $\exists \pi^*$ s.t. $V_\tau^{\pi^*} = V_\tau^*$ $\mathcal{T}_\tau^{\pi_1} V_\mu^{\pi_2}(s) - V_\mu^{\pi_2}(s)$ since $\pi^* \propto e^{\frac{Q^*}{\tau}}$, $A_\tau^* = 0$, $\forall s,a$.

Soft PI $\pi_{k+1}(\cdot|s) = \frac{e^{\frac{Q_\tau^{\pi_k}(s,\cdot)}{\tau}}}{\|\sim\|_1}$. $:= \max_\pi V_\tau^\pi$, $\mathcal{T}_\tau V_\tau^* = V_\tau^*$.

γ-conv. $\tau\to 0$, $\pi_\tau^* \to \ell_0' Q^*$,

□8' $\nabla_\theta V_\tau^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi(\cdot|s)} (A_\tau^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta) V_\tau^* \to \max_a Q^* = V^*$.

(ass1) also for Softmax para. the same. e.g. ent. sm. PG $\pi_{k+1}(a|s) \propto \pi_k(a|s) e^{\gamma \frac{d_\mu^{\pi_k}(s)}{1-\gamma} \pi_k(a|s) A_\tau^{\pi_k}(s,a)}$;

DPG. $a = \pi_\theta(s)$. □8'' $\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} (\nabla_\theta \pi_\theta(s) \cdot \nabla_a Q^{\pi_\theta}(s, \pi_\theta(s)))$ $\sim_{NPG} \sim \frac{\gamma}{1-\gamma} A_\tau^{\pi_k}(s,a)$ i.e. $\propto \pi_k^{1-\frac{\gamma\tau}{1-\gamma}} e^{\frac{\gamma}{1-\gamma} Q_\tau^{\pi_k}(s,a)}$

Pf: $\nabla_\theta (Q^{\pi_\theta}(s_0, \pi_\theta(s_0))) = \nabla_\theta \pi_\theta(s_0) \nabla_a Q^{\pi_\theta}(s,a)\big|_{a=\pi_\theta(s_0)} + \gamma \int_S \nabla_\theta V^{\pi_\theta}(s_1) p(s_1|s_0, a=\pi_\theta(s_0)) ds_1$.

DDPG. (a-c $Q(s,a;w) \to Q^{\pi_\theta}(s,a)$, by FQI.
b need add noise to π_θ. replay buffer for dependence.)

5 online planning. ('from now on', no $\pi(\cdot;\theta)$)

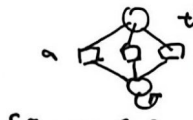
receding horizon planning (depth d). MCTS.

leaf estimated. building incrementally by e-e tradeoff.

Dyna-Q.

(combines model-based methods or estimated model i.e. simulation experience to improve efficiency of usage of data.)

MAB. ∀ SARSA/Q-learn



for $t=0\to$? $a_t \sim \varepsilon_t$-greedy Q^t (or b)

end $Q^{t+1}_{(a_t)} = Q^t(a_t) + \alpha_t(r_t - Q^t(a_t))$

regret $R_T \triangleq \sum_{t=1}^{T}(\mu_* - \mu_{a_t})$, $\mu_a \triangleq \mathbb{E}_{r\sim D_a}(r)$.

for evaluating algos. (note for offline (oracle) regret can be 0.)

□ (Hoeffding ineq.)

$\{a_1, \dots a_K\}$ $X_k \overset{iid}{\sim}$ sub-Gauss with v. $\mathbb{P}\left(\left|\frac{1}{n}\sum_i^n (X_i-\mu)\right| \geq t\right) \leq 2e^{-\frac{nt^2}{2v^2}}$.

$(\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{\lambda^2 v^2}{2}}, \forall \lambda \in \mathbb{R})$

algos: <nc, F> esp. Gauss. bd.

explore-first.

(N rounds each a) $\bar{\mu}_a = \frac{1}{N}\sum_{t=1}^{N} r_{a,t}$

then phase II $T-Nk$ rounds $\hat{a} = \operatorname{argmax}_a \bar{\mu}_a$.

$\mathbb{E}R_T \leq T^{\frac{2}{3}} O(k\log T)^{\frac{1}{3}}$ if $N = \left(\frac{T}{k}\right)^{\frac{2}{3}} O(\log T)^{\frac{1}{3}}$

ε-greedy. $\bar{\mu}_t(a) := \frac{\sum_i r_i \mathbb{1}_{\{a_i=a\}}}{\sum_i \mathbb{1}_{\{a_i=a\}}}$ (same as SARSA/Q-learn when a_t = sth.)

$\begin{cases} 1-\varepsilon+\frac{\varepsilon}{k} \\ \frac{\varepsilon}{k} \quad \dots \end{cases}$

~ : if $\varepsilon_t = \left(\frac{k}{t}\right)^{\frac{1}{3}}\log t^{\frac{1}{3}}$

UCB. $\operatorname{argmax}_a \bar{\mu}_t(a) + \text{radius}_t(a)$

by □/o,

$|\bar{\mu}_t(a) - \mu(a)| \leq v\cdot\sqrt{\frac{2}{n_t(a)}\log\frac{2}{\delta}}$ with Prob. $1-\delta$

as radius

typically, $\delta = n_t^{\beta}$, $U^{CB}_{t(a)} = \bar{\mu}_t(a) + C\sqrt{\frac{\log n_t}{2n_t(a)}}$.

$\mathbb{E}R_t \leq O(\sqrt{kt\log T})$ if $\delta = \frac{2}{T^4}$, $\forall t \leq T$.

pf: v^2: variance proxy

e.g. σ^2 Gauss. M^2 supp $\subset [-M, M]$

$\mathbb{P}(X \geq a) \leq \inf_{t>0} M_X(t)e^{-ta}$

Hoeffding ineq. can be derived by Chernoff bound.

- Bayes Bandits.

$\{\mu_a\}_{a=1}^k$ obey a priori dis.,

select a by $\mathbb{P}(\operatorname{argmax}_{a'} \mu_{a'} = a \mid \mathcal{H}_{t-1})$,

Thompson sampling. $\mathcal{H}_{t-1} = \{a_1, r_1, a_2, r_2, \dots, a_{t-1}, r_{t-1}\}$

(sample $(\mu_a)_{a=1}^k \sim P(\cdot \mid \mathcal{H}_{t-1})$ choose the largest.

even ind. setting i.e. $\bar{\mu}_t(a) \sim P(\mu_a \mid \mathcal{H}_{t-1})$, $a=1,\dots,k$.

≈ optim. $O(\sqrt{kT\log T})$

MCTS

(multi-step and conducts UCB search)

$n(s,a) \mathrel{+}= 1$, i.e. $Q(s,a) + C\sqrt{\frac{\log n(s)}{n(s,a)}}$

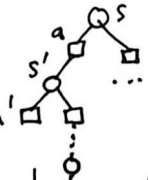
$Q(s,a) \mathrel{+}= \frac{1}{n(s,a)}(r(s,a,s') + \gamma V(s') - Q(s,a))$

if a is chosen,

a↑s ($V(s')$ is assumed)

note after this s' is not leaf.

(until s' meets termination conditions)



select – expand – propagate (traceback ↓ n-ahead

(use $V(s')$ or simulate) eg. $r_0 + \gamma r_1 + \dots + \gamma^n V(s_n)$

- AlphaGo Lee:

Selection $Q(s,a) + C\frac{P(s,a)}{1+n(s,a)}$ (P: supervised or RL get a learned net for policy)

Expand $\geq n_0$

Evaluation $V(s) = \lambda V_\theta(s) + (1-\lambda)$ & (fast rollout net as simulation)

BP. upd $Q(s,a)$, $n(s,a)$.

~ Zero: $Q + C\cdot\frac{P\sqrt{N(s)}}{1+N(s,a)}$ (PUCT)

same net has p, v two heads

learn from MCTS (search guided by p, use $n(s,a)$ as π)

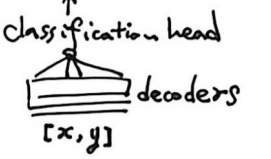
learn from outcomes (use π from MCTS till end)

$L = (\delta - v)^2 - \pi'\log p + c\|\theta\|^2$.

6 RL for LLMs. token-level MDP.

$S_t = [x, y_{<t}]$, $S_{t+1} = [S_t, a_t]$, $r(S_t, a_t) = \begin{cases} r(x,y) & t = T-1 \\ 0 \end{cases}$ (deterministic)

RM for RLHF:
- rule-based — e.g. safety or structural constraints, e.g. math-verify, coding
- discriminative — compare or rank based on human preference, e.g. $RM_\phi(x, y) \in \mathbb{R}$
- generative — e.g. LLM-as-a-judge, CoT



DPO. from $\max_\theta V^{\pi_\theta}(\mathcal{D})$

$\Rightarrow \pi^*_\beta(y|x) = \frac{e^{\frac{r_\phi(x,y)}{\beta}} \cdot \pi_{ref}(y|x)}{Z_\beta(x)}$

under BT model, $x \to y_{chosen}, y_{rejected}$

$\mathbb{P}(y_c > y_r | x) = \sigma(\beta \log \frac{\pi^*_\beta(y_c|x)}{\pi_{ref}(y_c|x)} - \beta \log \frac{\pi^*_\beta(y_r|x)}{\pi_{ref}(y_r|x)})$

thus, min DPO obj. $\min_\theta \mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}} -\log \sigma(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_{ref}(y_c|x)} - \beta \log \frac{\pi_\theta(y_r|x)}{\pi_{ref}(y_r|x)})$ (w/ LE way)

(off-line friendly, need high-qal. data)

Bradley-Terry model: $\log \frac{\mathbb{P}(y_c > y_r)}{1 - \mathbb{P}(y_c > y_r)} = r_\phi(x, y_c) - r_\phi(x, y_r)$ RM then in RLHF

i.e. $\min_\phi \mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}} -\log \sigma(r_\phi(x,y_c) - r_\phi(x,y_r))$.

PG. $\nabla_\theta V^{\pi_\theta}(\mathcal{D}) = \mathbb{E}_{x,y} \sum_{t=0}^{T-1} \nabla \log \pi_\theta(y_t|x, y_{<t}) \cdot (r_\phi(x,y) - \beta \sum_{t'=0}^{T-1} \log \frac{\pi_\theta(y_{t'}|x, y_{<t'})}{\pi_{ref}(y_{t'}|x, y_{<t'})})$

GRPO. sample $\{y_i\}_{i=1}^k$ for x, est. by $A^{\pi_{\theta_k}}_{t,i} = \frac{r_\phi(x,y_i) - \text{mean}_i(r_\phi(x,y_i))}{\text{std}_i(r_\phi(x,y_i))}$.

note $A^{\pi_{\theta_k}}_t = Q^{\pi_{\theta_k}}(x, y_{<t}), y_t) - V^{\pi_{\theta_k}}(x, y_{<t}))$

$\approx Q^{\pi_{\theta_k}}(x, y_{<t}), y_t) - V^{\pi_{\theta_k}}(x))$

$\to$ unbiased $r_\phi(x,y) - \text{mean}_i(r_\phi(x,y_i))$.

(note $V^{\pi_{\theta_k}}$ here needn't include βKL term)

PPO/TRPO

$\mathbb{E}_{x,y\sim\pi_{\theta_k}(\cdot|x)} \sum_{t=0}^{T-1} \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\theta_k}(y_t|x, y_{<t})} A^{\pi_{\theta_k}}(y_t, x, y_{<t})$ — ...

s.t. θ sufficiently close to θ_k. e.g. clipped obj.

$\mathbb{E}[\sum_t \min(r_t(\theta) A^{\pi_{\theta_k}}_t, \text{clip}(r_t(\theta), 1\pm\varepsilon) A^{\pi_{\theta_k}}_t) - \beta \text{KL}(\pi_\theta(\cdot|x) \| \pi_{ref}(\cdot|x))]$.

but vanilla PPO use GAE(λ) $\xrightarrow{\text{est.}} A^{\pi_{\theta_k}}_t$ and intermediate rewards is difficult for even RM trained.

(π_{ref} to control the shift to avoid knowledge forgetting)

- PG in LLMs: let $J(\theta) := V^{\pi_\theta}(\mathcal{D})$. When at θ_k to grad desc., in Pytorch etc. framework, θ can only be in static data, thus $\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}$ can not be optimized. Surrogate

$\mathcal{L}_k(\theta) \triangleq -\mathbb{E}_x \mathbb{E}_{y\sim\pi_{\theta_k}(\cdot|x)} (\sum_{t=0}^{T-1} \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} (r(x,y) - \beta \sum_{t'=0}^{T-1} \log \frac{\pi_{\theta_k}(a_{t'}|s_{t'})}{\pi_{ref}(a_{t'}|s_{t'})})$

has $\nabla_\theta \mathcal{L}_k(\theta_k) = -\nabla_\theta J(\theta_k)$.

at step k, sample $\{x_i\}_{i=1}^n \sim \mathcal{D}$, $\{y_{ij}\}_{j=1}^m \sim \pi_{\theta_k}(\cdot|x_i)$, batch

total policy loss $\mathcal{L}_k(\theta) := \frac{1}{nm} \sum_i \sum_j \mathcal{L}_k(\theta | x_i, y_{ij})$. $y_{ij} = [a^{(ij)}_0, \ldots, a^{(ij)}_{T_{ij}-1}]$.

REINFORCE. (by MC) $\mathcal{L}_k(\theta | x_i, y_{ij}) = -\sum_{t=0}^{T_{ij}-1} \frac{\pi_\theta}{\pi_{\theta_k}} \cdot r - \beta \Sigma$, unbiased but high variance.

reduce var: mod. $-\beta \sum_{t'=0}^{T-1} \cdots \to -\beta \sum_{t'=t}^{T-1} \ldots$ by noting that $\nabla_\theta J(\theta) = \mathbb{E}_{x,y} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta (r - \beta \sum_{t'=t}^{T-1} \log \frac{\pi_\theta}{\pi_{ref}})$.

(but $r(x_i, y_{ij})$ also has var.)

Pf: wrt. (s_t, a_t) $t' < t$ is fixed. but $t' > t$ and r term are varied.

A-C. (like in GRPO) $V^{\pi_\theta}(s) \triangleq \mathbb{E}_{y\sim\pi_\theta(\cdot|s)} r(s,y)$, $Q^{\pi_\theta}(s,a) = V^{\pi_\theta}([s,a])$.

then A-C trains a model to approx $Q^{\pi_{\theta_k}}$ (or $V^{\pi_{\theta_k}}$) in $\mathcal{L}_k$.

$(r(x_i, y_{ij}) | s^{(ij)}_t, a^{(ij)}_t) \to \hat{Q}(s^{(ij)}_t, a^{(ij)}_t)$ no var.

but has bias in Q's approx.)

i.e. $\mathbb{E} \nabla_\theta \mathcal{L}_k(\theta_k | x_i, y_{ij}) \neq -\nabla_\theta J(\theta_k)$.

$\hat{\mathcal{L}}_k(\theta) = \frac{1}{nm} \sum_i \sum_j (\underbrace{\sum_t \frac{\pi_\theta(a^{(ij)}_t | s^{(ij)}_t)}{\pi_{\theta_k}(a^{(ij)}_t | s^{(ij)}_t)} \hat{A}_k(s^{(ij)}_t, a^{(ij)}_t)}_{\hat{\mathcal{L}}_k(\theta | x_i, y_{ij})})$.

Lots of estimation of $\hat{A}_k(s_t, a_t)$.

(note if $\min_\theta \mathcal{L}_k(\theta)$ we may enlarge $\pi_\theta(a_t|s_t)$ for all sampled tokens a_t (ow. softmax or $\sum = 1$ would adjust) since $\hat{Q}^{\pi_{\theta_k}}(s_t, a_t) > 0$ and $\{a^{(ij)}_t\}_j | s^{(ij)}_t$ not covers all vocabulary)

□ (Token-level Policy Gradient)

□′ $\nabla_\theta J(\theta) = \mathbb{E}_{x,y} \sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot (\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} Q^{\pi_\theta}(s_t, a_t) - \beta \sum_{t'=t} \log \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_{ref}(a_{t'}|s_{t'})})$

(Pf by $\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} (r(x,y)|s_t, a_t) = Q^{\pi_\theta}(s_t, a_t)$.)

$= \mathbb{E}_{a_{T-1}\sim\pi_\theta(\cdot|s_{T-1})} \cdots \mathbb{E}_{a_0\sim\pi_\theta(\cdot|s_0)}$

□″ reduce a baseline $\hat{\mathcal{L}}_k(\theta | x_i, y_{ij}) = -\sum_{t=0}^{T_{ij}-1} (\frac{\pi_\theta(a^{(ij)}_t | s^{(ij)}_t)}{\pi_{\theta_k}(a^{(ij)}_t | s^{(ij)}_t)} \cdot (\hat{Q}(s^{(ij)}_t, a^{(ij)}_t) - b(s^{(ij)}_t)) - \beta \sum_{t'=t}^{T_{ij}-1} \log \frac{\pi_\theta \ldots}{\pi_{ref} \ldots})$.

and choose $b(s_t) = V^{\pi_\theta}(s_t)$

for $\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)} Q^{\pi_\theta}(s_t, a_t) = V^{\pi_\theta}(s_t)$.

7 Meta RL.

$\min_{\theta,\psi} \mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})} \mathcal{L}(\mathcal{D}^{test}_{\mathcal{T}}, \theta')$ s.t. $\theta' = f_\psi(\mathcal{D}^{tr}_{\mathcal{T}}, \theta)$.

(outer loop) (inner loop)

f_ψ { gradient-based / context-based (recurrence-) e.g.

Grad. MAML $f_\psi(\mathcal{D},\theta) = \theta - \alpha \nabla_\theta \mathcal{L}(\mathcal{D},\theta)$ (or multi-step) $\psi = \{\alpha_i\}_i$ or ϕ.

▽ sample $\mathcal{T}_i \sim p(\mathcal{T})$

for i sample $\{\tau^{\theta}_{ij}\}_{j=1}^{k}$ from $f(\mathcal{D}_{\mathcal{T}_i}, \theta)$ i.e. $\mathcal{T}_i$.

$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_i(f_\theta)$ $(\mathcal{L}_i(f_\theta) := \frac{1}{k}\sum_{j=1}^{k} \mathcal{L}_i(\tau^{\theta}_{ij}))$

end sample $\{\tau^{\theta_i'}_{ij}\}_{j=1}^{m}$ from $f(\mathcal{D}_{\mathcal{T}_i}, \theta_i')$ in $\mathcal{T}_i$.

upd. $\theta \mathrel{-}= \beta \nabla_\theta \sum_i \mathcal{L}_i(f_{\theta_i'})$. (FO: $\frac{\partial \theta_i'}{\partial \theta} = I - \alpha \frac{\partial^2 \mathcal{L}_i(f_\theta)}{\partial \theta^2}$ for simplicity, omit $\nabla_\theta \alpha \mathcal{L}_i(f_\theta)$ i.e. $\beta \nabla_{\theta_i'}(\cdots)$) $\theta_i' = \sim$ (optional)

MAESN. $\theta, \{(\mu_i, \sigma_i)\}_i$ · π_θ (a above, s below) $s \sim \mathcal{N}_i$

$\max_{\theta,\mu_i,\sigma_i} \sum_{\mathcal{T}_i} \mathbb{E}_{a_t\sim\pi(\cdot|s_t;\theta_i',\delta_i'),\ \delta_i'\sim\mathcal{N}(\mu_i',\sigma_i'^2)} (\sum_t r_i(s_t,a_t)) - \sum_{\mathcal{T}_i} KL(\mathcal{N}(\mu_i,\sigma_i) \| \mathcal{N}(0,I))$

s.t. $\mu_i' = \mu_i + \alpha_\mu \nabla_{\mu_i} \mathbb{E}_{a_t\sim\pi,\ \delta\sim\mathcal{N}} (\sum_t r_i(s_t,a_t))$ θ, μ_i, σ_i

$\sigma_i' = \sim$ · (optional)

Context. · POMDP belief $b(s)$.

$b_{t+1}(s') \propto \Omega(o|s',a) \sum_s P(s'|s,a) b_t(s)$.

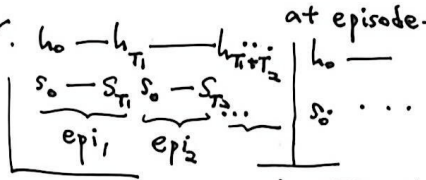
(action a and observe o) but still train at episode-level (though intros recur. structure in deep RL)

e.g. using TRPO.

(On-policy mostly, PEARL. but Pearl can be off-policy)

RL^2

ψ in RNN. e.g. GRU.



better when trial contains less episodes.

$\max_\psi \mathbb{E}(\sum_{\text{trial}} r)$. e.g. TRPO.

$\text{vec}(s_{t+1}, a_t, r_t, \mathbb{I}(\text{end\_epi}_t))$ $\xrightarrow{x\ (\text{input})}$ RNN $\longrightarrow$ FC & Softmax $\pi(a_{t+1})$

h_{t+1} ↻ (h_{t+2})

补-4

($\max_\theta \mathbb{E}_{x\sim\mathcal{D}} Q_w(x, \pi_\theta(x))$.)

· DDPG (DPG + DQN) use $\max_{a'}$ (greedy implicitly).

(soft upd i.e. $\tilde{\theta} \leftarrow$ (Polyak) $\tau\theta + (1-\tau)\tilde{\theta}$)

$\theta \mathrel{+}= \beta \cdot (\frac{1}{m}\sum_i \nabla_a Q(s,a;w)\big|_{a=\pi_\theta(s)} \nabla_\theta \pi_\theta(s))$ by □8''

$w \mathrel{-}= \alpha \nabla_w (\frac{1}{m}\sum_i (Q(s,a;w) - (r + \gamma Q(s', \pi_\theta(s'); \tilde{w})))^2)$

thus θ upd. just like Emax which leads to higher est. of Q_w.

(target i.e. previous w)

TD3 · (clipped double Q) $r + \gamma \min\{Q^A(s',a'), Q^B(s',a')\}$

· (delayed π upd.) actor and target net upd. / C step.

(target π smoothing) to avoid Q over-fitting. $a_t = \pi_\theta(s_t) + \text{clip}(\mathcal{N}_{(0,\sigma^2)}, -c, c)$

sample $\pi_\theta(s_t) + \mathcal{N}_t = a_t$ into $\mathcal{D}$. $\{sars'\}_t$

soft Q learn. $Q^{t+1}(s,a) = Q^t(s,a) + \alpha_t(s,a)(r + \gamma \cdot \underbrace{\tau \log \sum_{a'} e^{\frac{Q^t(s',a')}{\tau}}}_{V^t(s')} - Q^t(s,a))$.

temp. τ adjusting, target ent. $\mathcal{L}_\tau \triangleq \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi_\theta(\cdot|s)} -\tau(\log \pi_\theta(a|s) + \mathcal{H}_0)$.

$\{sars'\}$

SAC. Q target $\tilde{Q} = r + \gamma(\min_{A,B} Q_{\tilde{w}}(s', \cdot) - V^t_\pi(s'))$

(off-policy i.e. replay $\mathcal{D}$.) $(\mathbb{E}_{sars'\sim\mathcal{D}} (Q_w(s,a) - \tilde{Q}(r,s'))^2) \triangleq \mathcal{L}_w$ $a'\sim\pi_\theta(s')$ $-\tau \log \pi_\theta(a'|s')$.

ent. a-c $-\alpha \nabla_w \mathcal{L}_w$. $-\beta \nabla_\theta \mathcal{L}_\theta$. where repara. (e.g. Gauss)

by □8', $\mathcal{L}_\theta \triangleq \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi_\theta(\cdot|s)} (\tau \log \pi_\theta(a|s) - Q_w(s,a))$ $a = \mu_\theta(s) + \sigma_\theta(s) \cdot \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$.

VariBAD

(online)

$q_\varphi(z|\tau_{<t}) = \mathcal{N}(\mu_t, \sigma_t^2)$. e.g. GRU.

$p_\theta(\tau|z) = \prod_t p_\theta(s_{t+1}|s_t, a_t, z)$. $p_\theta(r_{t+1}|s_t, a_t, z)$.

VAE 学习 task repre.

$\pi_\psi(a_t|s_t, q_\varphi(z|\tau_{<t}))$ (i.e. use $\mu_{\varphi,t}, \sigma_{\varphi,t}$) $(\theta, \varphi|x)$

$\mathcal{L}_{(\theta,\varphi,\psi)} = \mathbb{E}_\tau\left(J(\psi,\varphi) + \lambda \sum_{t\in\tau} ELBO_t(\varphi,\theta)\right)\{x, z\}_i$

where RL $J = \mathbb{E}_{\tau\sim\pi_\psi, q_\varphi}\left(\sum_t \gamma^t r_t + \alpha H(\pi_\psi)\right)$

recon. $ELBO_t = \mathbb{E}_{z\sim q_\varphi(\cdot|\tau_{<t})}(\log p_\theta(\tau|z)) - KL(q_\varphi(\cdot|\tau_{<t}) \| q_\varphi(\cdot|\tau_{<t-1}))$.

(KL has closed form of Gaussian)

PEARL

(off-line)

$z \sim q_\varphi(z|\text{context}_i)$ (from π in T_i).

$\pi_\theta(a|s, z)$

θ, ω by SAC.

φ by $\nabla_\varphi \sum_i(\mathcal{L}_i^{critic} + \beta KL(q_\varphi \| p(z)))$.

• VAE:

$x \xrightarrow[q_\varphi(z|x)]{en.} z \xrightarrow[p_\theta(x|z)]{de.} \hat{x}$

(AE deterministic z, $\hat{x}$. f_θ recon. $\mathcal{L} = \|x - \hat{x}\|_2^2$)

(if not have this term it comes to AE)

$\mathcal{L} = -\mathbb{E}_{z\sim q_\varphi(\cdot|x)}(\log p_\theta(x|z)) + \beta \cdot KL(q_\varphi(z|x) \| p(z))$.

(recon.: neg. log. likelihood) e.g. Softmax → Cross Ent. Gaussian → MSE.

$\beta > 1$ info. bottleneck $I(x; z)$ (i.e. disentangle)

$\beta < 1$ relieve posterior collapse ($q_\varphi \approx p(z)$)

(grad thru z_i) repara. $p(z) = \mathcal{N}(0, I)$, $q_\varphi(z|x) = \mu_\varphi(x) + \sigma_\varphi(x)\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$.

i.e. $\nabla_{\theta,\varphi} \mathcal{L}_{recon.} = \nabla_{\theta,\varphi} \mathbb{E}_\varepsilon(\log p_\theta(x|\mu_\varphi(x) + \sigma_\varphi(x)\varepsilon))$.

Gen. model. (max) $\log p_\theta(x)$

$= \log \int \frac{p_\theta(x|z)p(z)}{q_\varphi(z|x)} q_\varphi(z|x)\,dz$

Jensen. $\geq \mathbb{E}_{q_\varphi(z|x)} \log p_\theta(x|z) - KL(q_\varphi \| p(z))$

$\triangleq$ ELBO.

note eq. here $\log p_\theta(x) = ELBO + KL(q_\varphi(z|x) \| p_\theta(z|x))$.

(i.e. $p_\theta(z|x)$ is difficult to calculate due to $\frac{1}{Z}$ then use q_φ to approx)

(in Bayes) $Z \approx p_\theta(x)$

(max ELBO i.e. min $KL(q_\varphi(z|x) \| p_\theta(x, z))$).

intuition: $\nabla_\theta \log p_\theta(x) = \frac{\int p(z)\nabla_\theta p_\theta(x|z)\,dz}{p_\theta(x)} \overset{but}{=} \int p_\theta(z|x)\nabla_\theta \log p_\theta(x|z)\,dz$

$\nabla_\theta \log \int p(z) p_\theta(x|z)\,dz \geq \mathbb{E}_{z\sim p(z)} \nabla_\theta \log p_\theta(x|z)$.

if no encoder we still can not get $p_\theta(z|x)$, note intro. $q_\varphi(z|x)$ we remain two KL (sum of $KL(q_\varphi \| p_\theta(z|x)) - KL(q_\varphi \| p(z))$.)

---

w {Randomness}. (repara. explicitly)

↓

'search in recurrence' △ z 每个 π_θ 时 fix.

评价 cmp to: MAESN $\pi_\theta(a|s, z)$, $z \sim \mathcal{N}(\mu_i, \sigma_i^2)$

▽ 没有 history coherent. (grad 训练) 是仅有的 hist. 作用.

RL²

▽ episode 内不传导 for cost. (upd. hist./hidden)

▽ (GRU) RNN 改 LLM + Muload.

▽ 无 randomness (不对 T_i 作分布 inference. i.e. π^* 与 task infer. disentangle) $(\cdot|z)$

(MetaQlearn also GRU $\pi_\theta(a_t|s_t, z_t)$. explicit: 似, 以 score 选 相似 task 'match' (based on TD3 (offline).) z use GRU but also explicit π_θ upd. 'critic net')

task infer. $q_\varphi(z|\cdot)$

- PEARL △ encoder $q_\varphi(z|\text{Context})$ for infer.
  - ▽ 排列对称. $c_i = \{s_i, a_i, r_i, s_i'\}$. $\prod_i \Phi_\psi(z|c_i)$.
- VariBAD VAE construct τ.
  - ▽ $\pi_\psi(a_t|s_t, \mu_{\varphi,t}, \sigma_{\varphi,t})$ 本质又退回了 recurrence. (只是 hidden 的提取 use recons. 进行)