$V_\pi^\tau(s) = E_\pi\left[\sum \gamma^t r_{\tau,t} \mid s_0 = s\right] \quad q_\pi^\tau(s,a) = E_\pi\left[r_0 + \sum_{t=1} \gamma^t r_{\tau,t} \mid s_0 = s, a_0 = a\right]$

$V_\pi^\tau(s) = E_{a\sim\pi}\left[q_\pi^\tau(s,a) - \tau \log \pi(a|s)\right]$

$= E_{a\sim\pi}\left[q_\pi^\tau(s,a)\right] - H(\pi(\cdot|s))$

<每遇到一个随机变量a就惩罚一次.>

► Soft action value $q_\pi^\tau(s,a)$: [a_0 is chosen, thus entropy equal to 0] $\leftarrow \pi(a|s_0) = 1(a=a_0)$

$$q_\pi^\tau(s,a) = \mathbb{E}_\pi\left[r(s_0,a_0,s_1) + \sum_{t=1}^{\infty} \gamma^t r_\tau(s_t,a_t,s_{t+1}) \mid s_0 = s, a_0 = a\right]$$

For any $v \in \mathbb{R}^{|\mathcal{S}|}$, the soft Bellman optimality operator $\mathcal{T}^\tau$ is defined by

$$[\mathcal{T}^\tau v](s) = \max_{\pi(\cdot|s)} \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}[r_\tau(s,a,s') + \gamma v(s')]$$
$$= \max_{\pi(\cdot|s)} \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}[r(s,a,s') + \gamma v(s')] - \tau \log \pi(a|s)\right]$$
$$= \tau \log\left(\|\exp(q(s,\cdot)/\tau)\|_1\right),$$

where the maximum value is attained iff $\pi(\cdot|s) \propto \exp(q(s,\cdot)/\tau)$

► $\mathcal{T}^\tau$ is γ-contraction with respect to ℓ_∞-norm.

另解[这里] 最后证明(1)

1°: $[\mathcal{T}^\tau v](s) = \tau \log(\|\exp(q(s,\cdot)/\tau)\|_1)$

$\arg\max_\pi(\cdots) = \frac{1}{C}\cdot \exp(q(s,\cdot)/\tau)$

① $\sum_a \pi(a|s)[q(s,a) - \tau \log \pi(a|s)]$ 对 $\pi(a|s)$ 求导:

$\Rightarrow q(s,a) - \tau \log \pi(a|s) - \pi(a|s)\cdot\tau\cdot\frac{1}{\pi(a|s)} = 0.$

$\Rightarrow \frac{q(s,a)}{\tau} = \log \pi(a|s) + 1$? ✓ $\pi(a|s) = \exp[\frac{q(s,a)}{\tau}] \div e$ 大概率通过 $\sum_a \pi(a|s) = 1$ 约掉.. 拉格朗日化为无约束 应该能得到课件结论

to get $\pi(\cdot|s) = \frac{1}{C}\cdot\exp[\frac{q(s,\cdot)}{\tau}]$ & $\sum_a \pi(a|s) = 1$ 同求

$[\mathcal{T}^\tau v](s) = \sum_{a_i} \frac{1}{C}\exp[\frac{q(s,a_i)}{\tau}][q(s,a_i) - q(s,a_i) + \tau\cdot\log C]$

$= \frac{1}{C}\cdot\sum_{a_i}\exp[\frac{q(s,a_i)}{\tau}]\cdot\tau\log C$

$Z(s) \leftarrow C = \sum_{a_i}\exp[\frac{q(s,a_i)}{\tau}]$ 正规化

或者①: $E_{a\sim\pi}[q(s,a) - \tau\log\pi(a|s)] = \sum_a \pi(a|s)[\tau\log\pi(a|s) - \log\exp[q(s,a)]]$

$= -\tau\sum_a \pi(a|s)[\log\pi(a|s) - \log\exp[\frac{q(s,a)}{\tau}]]$

$\max[\ ] = \max[\ ] \Leftrightarrow \min KL(\pi(\cdot|s)\,\|\,\frac{e^{q(s,a)/\tau}}{Z(s)})$

$\log\frac{\exp[\cdot]}{Z(s)} + \log Z(s)$

2°: $\mathcal{T}^\tau$ 是压缩算子.

先 $\mathcal{T}_\pi^\tau$ 是压缩算子吧. $r(s,a,s') - \tau\log\pi(a|s)$ 由于 $H(z) = E_{s\sim\pi}[-\log\pi(a|s)]$

$[\mathcal{T}_\pi^\tau v](s) = E_{a\sim\pi(\cdot|s),s'}[\overbrace{r_\tau(s,a,s')}^{} + \gamma V(s')]$

$[\mathcal{T}_\pi^\tau v_1 - \mathcal{T}_\pi^\tau v_2](s) = E_{a,s'}[\gamma V_1(s') - \gamma V_2(s')]$

$\begin{cases} \|\mathcal{T}_\pi^\tau v_1 - \mathcal{T}_\pi^\tau v_2\|_\infty = \gamma\max_s E_{a,s'}[V_1(s') - V_2(s')] \le \gamma\max_s E_{a,s'}|V_1(s') - V_2(s')| & ① \\ \gamma\|v_1 - v_2\|_\infty = \gamma\max_s |V_1(s) - V_2(s)| & ② \end{cases}$

假设 $s_* = \arg\max_s |V_1(s) - V_2(s)|$

②: $\gamma\cdot|V_1(s_*) - V_2(s_*)|$

①: $E_{s'}|V_1(s') - V_2(s')| < |V_1(s_*) - V_2(s_*)|$

$E_a[\cdots] < E_a|V_1(s_*) - V_2(s_*)| =$

$\max_s[\ ] < \max_s|\ \ | =$ Done.

再看 $\mathcal{T}^\tau$: $[\mathcal{T}^\tau v](s) = \max_\pi E_{a,s'}[r_\tau + \gamma V(s')]$

$[\mathcal{T}^\tau v_1 - \mathcal{T}^\tau v_2](s) \le \max_\pi E_{a,s'}[\gamma V_1(s') - \gamma V_2(s')]$ $\max(A_1) - \max(A_2) \le \max(A_1 - A_2)$

同上可证. ✓.

- Soft Policy Evaluation ①
- Soft Policy Improvement ②
- Soft Policy Iteration ③

① $q_{\pi_{new}}^\tau = \mathcal{T}_\pi^\tau q_{\pi_{old}}^\tau$ (证明有解 $\mathcal{T}_\pi^\tau$ 收缩)

(网上更好) $q_{\pi_{old}}^\tau(s,a) = E_{s'}[r(s,a,s') + \gamma V_{\pi_{old}}^\tau(s')]$

$= E_{s'}[r(s,a,s') + \gamma E_{a'\sim\pi}[q_{\pi_{old}}^\tau(s',a') - \tau\log\pi(a'|s')]$

$= E_{s'}[r(s,a,s') + \gamma E_{a'\sim\pi}[q_{\pi_{old}}^\tau(s',a')] + \gamma H(\pi)]$

Given π 下常数

∴ 右侧(Bellman Backup)仍然收敛 $\Rightarrow$ Q 收敛

经典RL中如何收敛? 其实就 r 换 r_τ 不影响压缩算子证明

② $\pi' = \exp[\frac{q_\pi^\tau(s,\cdot)}{\tau}]/Z(s)$ s.t. $q_\pi^\tau(s,a) \le q_\pi^\tau(s,a)$ 证明有提升

$V_\pi^\tau(s) = E_a[q_\pi^\tau(s,a) - \log\pi(a|s)]$

$= -KL(\pi(\cdot|s)\,\|\,\exp[\frac{q_\pi^\tau(s,\cdot)}{\tau}]/Z) + \log(Z)$

$\le E_{a\sim\pi'}[q_\pi^\tau(s,a) - \log\pi'(a|s)] = \log(Z)$

$= H(\pi') + E_{a\sim\pi'}[q_\pi^\tau(s,a)]$

$q_\pi^\tau(s,a) = E_{s_1}[r_0 + \gamma V_\pi^\tau(s_1)]$

$\le E_{s_1}[r_0 + \gamma H(\pi') + \gamma E_{a_1\sim\pi'}[q_\pi^\tau(s_1,a_1)]$

$= E_{s_1,a_1}[r_0 + \gamma H(\pi') + \gamma\cdot r_1 + E_{s_2}[V_\pi^\tau(s_2)]$

$\cdots$

$\le E_{s_1,\cdots}[r_0 + \sum\gamma^t(r_t + H(\pi_t))] = q_{\pi'}^\tau(s,a)$

③ Policy Evaluation / Policy Improvement 收敛到最优 Policy Iteration

$V_k^\tau(s) = \max_\pi V_\pi^\tau(s) = \tau\log(\|\exp\frac{q_k^\tau(s,\cdot)}{\tau}\|_1)$

$[\mathcal{T}^\tau v](s) = \max_\pi[\mathcal{T}_\pi^\tau v](s)$

- $\mathcal{T}_\pi^\tau v = v$, Policy Evaluation
- $\mathcal{T}^\tau v = v$. Optimal

[经典RL]

非参数
- 策略/值迭代 (压强多项式收敛)
- 采样法
  - RM法: $x_{k+1} = x_k + \alpha_k(h(x_k) - x_k + e_k)$ → for $x = h(x)$ 的解; RL中 $z_k = q_k + \text{noise}$, $x = h(x)$; 多为TD算法
  - 蒙特卡洛 MC: 按bk采样 $b_k = \text{eg greedy } \pi_k$; $q_{k+1} = q_k + \frac{1}{\alpha_k}(\frac{\sum_{i=0}^{n} r_i}{n} - q_k)$ 样本回报

$q_{k+1} = q_k + \frac{1}{\alpha_k}[\text{TD target} - q_k]$

- SARSA: $r + \gamma q_k(s',a')$
- Q-learning (SARS): $r + \gamma\cdot\max_{a'} q_k(s',a')$

TD(n): $\text{Target} = \sum r_i\gamma^t + \gamma^n q_k(s',a')$

$n\to\infty$, TD $\Leftrightarrow$ MC.

参数化
采样估计 V or Q 后让一个函数学习它 e.g. $\theta_{k+1} = \theta_k + \frac{1}{\alpha_k}[\text{估计一函数}]\cdot$ 函数梯度
当然也有无梯度法.

$V(s)$ to $V(s;\theta)$ (V于A) eg $V(s;\theta) = \phi(s)^T\theta$

$\pi(a|s)$ to $\pi_\theta(a|s)$
- 自然梯度: $\Delta\theta = \alpha\cdot F(\theta)^{-1}\nabla J(\theta)$ → 信息阵
- TPRO
- ppo
- DDPG

[正则RL]

大局