	Standard MDP	Maximum Entropy MDP
State:	S	S
Actions:	A	A
Reward:	$r(s,a)$	$r(s,a) + \alpha \mathcal{H}(\pi(\cdot\|s))$
Transition:	$P(s'\|s,a)$	$P(s'\|s,a)$
Discount:	γ	γ
optimal policy:	$\pi^* = \arg\max_\pi \sum_t \gamma^t E_{s_t,a_t\sim\pi}[r(s_t,a_t)]$	$\pi^*_{\text{soft}} = \arg\max_\pi \sum_t \gamma^t E_{s_t,a_t}[r(s_t,a_t) + \alpha \mathcal{H}(\pi(\cdot\|s_t))]$

Definition:

optimal policy: $\pi^*_{\text{soft}} = \arg\max_\pi \sum_t \gamma^t E_{s_t,a_t\sim\pi}[r(s_t,a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))]$

value function: $V^\pi_{\text{soft}}(s_t) = \sum_{t'=t}^{\infty} \gamma^{t'} E_{s_{t'},a_{t'}\sim\pi}[r(s_{t'},a_{t'}) + \alpha \mathcal{H}(\pi(\cdot|s_{t'}))]$

Action-value function: $Q^\pi_{\text{soft}}(s_0,a_0) = r(s_0,a_0) + \sum_{t=1}^{\infty} \gamma^t E_{s_t,a_t}[r(s_t,a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))]$

Theorem

$$Q^\pi_{\text{soft}}(s,a) = r(s,a) + \gamma E_{s'}[V^\pi_{\text{soft}}(s')]$$

$$V^\pi_{\text{soft}}(s) = \alpha \mathcal{H}(\pi(\cdot|s)) + \sum_a \pi(a|s) Q^\pi_{\text{soft}}(s,a)$$

Easy to see

Theorem: Energy Form

Given a policy π with value function $Q^\pi_{\text{soft}}, V^\pi_{\text{soft}}$

the greedy policy wrt π is $\tilde{\pi}(s) = \arg\max_{\pi'(\cdot)} \sum_a \pi'(a|s) Q^\pi_{\text{soft}}(s,a) + \alpha \mathcal{H}(\pi'(\cdot|s))$

$$Q^\pi_{\text{soft}}(s,\tilde{\pi}(s)) = \alpha \log \sum_a \exp\left(Q^\pi_{\text{soft}}(s,a)/\alpha\right)$$

$$\tilde{\pi}(a|s) = \frac{\exp(Q^\pi_{\text{soft}}(s,a)/\alpha)}{\sum_a \exp(Q^\pi_{\text{soft}}(s,a)/\alpha)} = \exp\left\{\frac{1}{\alpha}\left[Q^\pi_{\text{soft}}(s,a) - Q^\pi_{\text{soft}}(s,\tilde{\pi}(s))\right]\right\}$$

$$\tilde{\pi}(s) = \arg\max_{\pi'(s)} \sum_a \pi'(a|s) Q^\pi_{\text{soft}}(s,a) + \mathcal{H}(\pi'(\cdot|s))$$

$$= \arg\max_{\pi'(s)} \sum_a \pi'(a|s)\left[Q^\pi_{\text{soft}}(s,a) - \alpha \log \pi'(a|s)\right]$$

$$\left.\begin{array}{ll} \min\limits_{x} & -\sum\limits_i x_i (a_i - \alpha \log x_i) \\ \text{s.t.} & \sum\limits_i x_i = 1 \end{array}\right\} \Rightarrow \text{convex}$$

(optimal value is negated when flip max to min)

$$\mathcal{L}(x,v) = -\sum_i x_i(a_i - \alpha \log x_i) + v\left(\sum_i x_i - 1\right)$$

$$\frac{\partial \mathcal{L}}{\partial x_i} = -a_i + \alpha(1+\log x_i) + v = 0$$

$$x_i^*(v) = \exp\left(\frac{a_i - v}{\alpha} - 1\right)$$

$$g(v) = -\sum_i \exp\left(\frac{a_i - v}{\alpha} - 1\right)(v+\alpha) + v\left(\sum_i \exp\left(\frac{a_i - v}{\alpha} - 1\right) - 1\right)$$

$$= -\alpha \sum_i \exp\left(\frac{a_i - v}{\alpha} - 1\right) - v$$

$$dg/dv = \sum_i \exp\left(\frac{a_i - v}{\alpha} - 1\right) - 1 = \exp\left(-\frac{v}{\alpha}\right)\sum_i \exp\left(\frac{a_i}{\alpha} - 1\right) - 1 = 0$$

$$v^* = \alpha \log \sum_i \exp\left(\frac{a_i}{\alpha} - 1\right)$$

$$x_i^* = \frac{\exp\left(\frac{a_i}{\alpha} - 1\right)}{\exp\left(\frac{v^*}{\alpha}\right)} = \frac{\exp\left(\frac{a_i}{\alpha} - 1\right)}{\sum_j \exp\left(\frac{a_j}{\alpha} - 1\right)} = \frac{\exp(a_i/\alpha)}{\sum_j \exp(a_j/\alpha)}$$

$$p^* = -\sum_i \frac{\exp(a_i/\alpha)}{\sum_j \exp(a_j/\alpha)}\left(a_i - \alpha \log \frac{\exp(a_i/\alpha)}{\sum_j \exp(a_j/\alpha)}\right)$$

$$= -\frac{1}{Z}\sum_i \exp(a_i/\alpha)(\alpha \log Z)$$

$$= -\alpha \log Z$$

$$= -\alpha \log \sum_i \exp(a_i/\alpha)$$

$$p^* = -\alpha \log \sum_i \exp(a_i/\alpha) \longrightarrow x_i^* = \exp\left(\frac{a_i}{\alpha} + \frac{p^*}{\alpha}\right)$$

Theorem: Monotonic Improvement

given a policy π with value function Q^{π}_{soft}

let one-step greedy policy $\tilde{\pi}(a|s) = \frac{\exp(Q^{\pi}_{soft}(s,a)/\alpha)}{\sum_{\tilde{a}} \exp(Q^{\pi}_{soft}(s,\tilde{a})/\alpha)}$

then $Q^{\tilde{\pi}}_{soft}(s,a) \geq Q^{\pi}_{soft}(s,a) \;\forall (s,a)$

Thus the optimal policy must have form $\pi^*(a|s) = \frac{\exp(Q^*_{soft}(s,a))}{\sum_{\tilde{a}} \exp(Q^*_{soft}(s,a))}$

$$Q^{\pi}_{soft}(s_0,a_0) = E_{s_1}\left[r_0 + \gamma\left(\alpha \mathcal{H}(\pi(s_1)) + E_{a_1 \sim \pi}\left[Q^{\pi}_{soft}(s_1,a_1)\right]\right)\right]$$

$$\leq E_{s_1}\left[r_0 + \gamma\left(\alpha \mathcal{H}(\tilde{\pi}(s_1)) + E_{a_1 \sim \tilde{\pi}}\left[Q^{\pi}_{soft}(s_1,a_1)\right]\right)\right]$$

$$\leq \dots$$

$$\leq Q^{\tilde{\pi}}_{soft}(s_0,a_0)$$

Theorem: Bellman Optimality

By the monotonic improvement theorem $\pi^*(a|s) = \frac{\exp(Q^{\pi^*}_{soft}(s,a))}{\sum_{\tilde{a}} \exp(Q^{\pi^*}_{soft}(s,\tilde{a}))}$

then $\pi^* = \arg\max_{\pi} \sum_a \pi(a|s) Q^{\pi^*}_{soft}(s,a) + \alpha \mathcal{H}(\pi(s))$

$$V^{\pi^*}_{soft}(s) = \sum_a \pi^*(a|s) Q^{\pi^*}_{soft}(s,a) + \alpha \mathcal{H}(\pi^*(s))$$
$$= \max_{\pi} \sum_a \pi(a|s) Q^{\pi^*}_{soft}(s,a) + \alpha \mathcal{H}(\pi(s))$$
$$= \alpha \log \sum_a \exp(Q^{\pi^*}_{soft}(s,a)/\alpha)$$

$$Q^{\pi^*}_{soft}(s,a) = r(s,a) + \gamma E_{s'}[V^{\pi^*}_{soft}(s')]$$
$$= r(s,a) + \gamma E_{s'}[\alpha \log \sum_{a'} \exp(Q^{\pi^*}_{soft}(s,a)/\alpha)]$$

At each state, the optimal policy π^* is greedy w.r.t. $Q^{\pi^*}_{soft}$

Theorem: Bellman Backup and uniqueness of π^*

define the soft value iteration as

$$(TQ)(s,a) = r(s,a) + \gamma E_{s'}[\alpha \log \int_{a'} \exp(Q^{\pi^*}_{soft}(s',a')/\alpha) da']$$

Then T is a γ-contraction

Thus there's a unique Q^*_{soft} satisfying

$$Q^*(s,a) = r(s,a) + \gamma E_{s'}[\alpha \log \int_{a'} \exp(Q^*_{soft}(s',a')/\alpha) da']$$

and the optimal policy is $\pi^*(a|s) = \frac{\exp(Q^*_{soft}(s,a)/\alpha)}{\int_{\tilde{a}} \exp(Q^*_{soft}(s,\tilde{a})/\alpha) d\tilde{a}}$

let $\|Q_1 - Q_2\|_\infty = \max_{s,a} |Q_1(s,a) - Q_2(s,a)| = \epsilon$

$$\alpha \log \int_A \exp \frac{Q_1(s,a)}{\alpha} da \le \alpha \log \int_A \exp \frac{Q_2(s,a)+\epsilon}{\alpha} da$$
$$= \alpha \log \int_A \exp \frac{\epsilon}{\alpha} \exp \frac{Q_2(s,a)}{\alpha} da$$
$$= \epsilon + \alpha \log \int_A \exp \frac{Q_2(s,a)}{\alpha} da \quad \forall s$$

$$\alpha \log \int_A \exp \frac{Q_1(s,a)}{\alpha} da \ge -\epsilon + \alpha \log \int_A \exp \frac{Q_2(s,a)}{\alpha} da \quad \forall s$$

$$-\epsilon \le \alpha \log \int_A \exp \frac{Q_1(s,a)}{\alpha} da - \alpha \log \int_A \exp \frac{Q_2(s,a)}{\alpha} da \le \epsilon \quad \forall s$$

$$\|TQ_1 - TQ_2\|_\infty = \max_{s,a} \gamma \left| E_{s'}\left[\alpha \log \int_A \exp \frac{Q_1(s,a)}{\alpha} da\right] - E_{s'}\left[\alpha \log \int_A \exp \frac{Q_2(s,a)}{\alpha} da\right] \right|$$

$$= \max_{s,a} \gamma \left| E_{s'}\left[\alpha \log \int_A \exp \frac{Q_1(s,a)}{\alpha} da - \alpha \log \int_A \exp \frac{Q_1(s,a)}{\alpha} da\right] \right|$$

$$\le \gamma \epsilon$$

$$= \gamma \|Q_1 - Q_2\|_\infty$$

- Soft Q-Iteration

$$Q_{soft}(s,a) \leftarrow r(s,a) + \gamma \cdot \alpha \log \int_A \exp \frac{Q(s',a')}{\alpha} da'$$

$\alpha \log \sum_{a'} \exp \frac{Q(s',a')}{\alpha} da'$ for discrete action space

$$= r(s,a) + \gamma \alpha \log E_{q(a')}\left[\frac{\exp Q(s',a')/\alpha}{q(a')}\right]$$

Initialize Q^θ and target $Q^{\bar\theta} = Q^\theta$

Initialize replay buffer D

while true

Sample $(s,a,r,s') \sim D$

Sample $a_i' \sim q(a')$ — q can be any distribution, for example uniform

Compute $t = r + \gamma \log \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\exp(Q^{\bar\theta}(s',a_i')/\alpha)}{q(a_i')}\right]$ — but scales poorly to high dimension

one choice is $q(a'|s) = \frac{\exp(Q(s,a)/\alpha)}{\int_A \exp Q(s,a)/\alpha \, da}$

$\min_\theta \|Q^\theta(s,a) - t\|_2^2$ — but this will require SVGD for continuous action space

for each k iteration: $\bar\theta \leftarrow \theta$

- Soft Actor Critic (Standard)

Compute Q_θ π_ϕ V_ψ

policy $\pi_\phi(a|s) = \frac{\exp(Q_\theta(s,a)/\alpha)}{\int_a \exp(Q_\theta(s,a)/\alpha)\, da}$

$\downarrow$

$\min_\phi D_{KL}\left(\pi_\phi(\cdot|s) \,\|\, \frac{\exp(Q_\theta(s,a)/\alpha)}{\int_a \exp(Q_\theta(s,a)/\alpha)\, da}\right)$

$\downarrow$

$\min_\phi E_{a\sim\pi_\phi(s)}\left[\log \pi_\phi(a|s) - Q_\theta(s,a)/\alpha\right]$

$= \min_\phi \int_a \pi_\phi(a|s) \cdot \log \frac{\pi_\phi(a|s)}{\exp(Q_\theta(s,a)/\alpha)/Z_\theta(s)}\, da$

$= \min_\phi E_{a\sim\pi_\phi(s)}\left[\log \pi_\phi(a|s) - Q_\theta(s,a)/\alpha + \log Z_\theta(s)\right]$

$= \min_\phi E_{a\sim\pi_\phi(s)}\left[\log \pi_\phi(a|s) - Q_\theta(s,a)/\alpha\right]$

let $\pi_\phi(s) = \mu_\phi(s) + \sigma_\phi(s)\cdot\epsilon \quad \epsilon\sim N(0,I)$

$\min_\phi E_{\epsilon\sim N}\left[\alpha \log \pi_\phi(\mu_\phi(s)+\sigma_\phi(s)\cdot\epsilon) - Q_\theta(s, \mu_\phi(s)+\sigma_\phi(s)\cdot\epsilon)\right]$

Initialize V_ψ, $V_{\bar\psi}=V_\psi$, Q_{θ_1}, Q_{θ_2}, π_ϕ

while true {

sample $(s,a,r,s') \sim D$

learn Q to approximate $Q^{\pi^*}_{soft}(s,a) = r(s,a) + \gamma E_{s'}\left[V^{\pi^*}_{soft}(s')\right]$

compute target $y(s) = r(s,a) + \gamma V_{\bar\psi}(s')$

$\min_{\theta_1} \|Q_{\theta_1}(s,a) - y(s)\|_2^2$ $\min_{\theta_2} \|Q_{\theta_2}(s,a) - y(s)\|_2^2$

sample $\epsilon\sim N(0,I)$ $\tilde a = \mu_\phi(s) + \sigma_\phi(s)\cdot\epsilon$

learn V to approximate $V^{\pi^*}_{soft}(s) = \alpha H(\pi^*(s)) + \int_a \pi(a|s)\cdot Q^{\pi^*}_{soft}(s,a)\, da$

$= E_{a\sim\pi(s)}\left[Q^{\pi^*}_{soft}(s,a) - \alpha \log \pi^*(a|s)\right]$

compute target $y(s) = \min\left(Q_{\theta_1}(s,\tilde a), Q_{\theta_2}(s,\tilde a)\right) - \alpha \log \pi^*(\tilde a|s)$

$\min_\psi \|V_\psi(s) - y(s)\|_2^2$

learn π to approximate $\pi^*(a|s) = \frac{\exp(Q^*(s,a)/\alpha)}{\int_a \exp(Q^*(s,\tilde a)/\alpha)\, d\tilde a}$

$\min_\phi \alpha \log \pi_\phi(\tilde a|s) - \min\left(Q_{\theta_1}(s,\tilde a), Q_{\theta_2}(s,\tilde a)\right)$

$\bar\psi = \rho\bar\psi + (1-\rho)\psi$

}

· Soft Actor-Critic (Modern)

Initialize $Q_{\bar{\theta}_1} = Q_{\theta_1}$ $Q_{\bar{\theta}_2} = Q_{\theta_2}$ π_ϕ

while true {

sample $(S, a, r, S') \sim D$

learn Q to approximate $Q^{\pi}_{soft}(s,a) = r(s,a) + \gamma E_{s'}\left[V^{\pi}_{soft}(s')\right]$

$$= r(S,a) + \gamma E_{s'}\left[\int_{a'} \pi^*(a'|s') Q^{\pi^*}_{soft}(s,a') + \alpha \mathcal{H}(\pi^*(s'))\right]$$

$$= r(S,a) + \gamma E_{s'}\; E_{a' \sim \pi(s)}\left[Q^{\pi^*}_{soft}(s',a') - \alpha \log \pi^*(a'|s')\right]$$

Sample $a' \sim \pi_\phi(a'|s) \sim \mu_\phi(s') + b_\phi(s') \cdot \epsilon$

Compute target $y(s) = r(s,a) + \gamma\left[\min\left(Q_{\bar{\theta}_1}(s',a'), Q_{\bar{\theta}_2}(s',a')\right) - \alpha \log \pi^*(a'|s')\right]$

$\min_{\theta_1} \| Q_{\theta_1}(s,a) - y(s) \|_2^2$ $\min_{\theta_2} \| Q_{\theta_2}(s,a) - y(s) \|_2^2$

learn π_ϕ to approximate $\pi^*(a|s) = \dfrac{\exp(Q^*(s,a)/\alpha)}{\int_{\bar{a}} \exp(Q^*(s,\bar{a})/\alpha)\, d\bar{a}}$

Sample $a \sim \pi_\phi(a|s)$

$\min_\phi \log \pi_\phi(a|s) - \min\left(Q_{\theta_1}(s,\bar{a}), Q_{\theta_2}(s,\bar{a})\right)$

$\bar{\theta}_1 = \rho \bar{\theta}_1 + (1-\rho)\theta_1$

$\bar{\theta}_2 = \rho \bar{\theta}_2 + (1-\rho)\theta_2$

}

- Automating Entropy Adjustment

An Intuition (Based on dual gradient method explanation in EE364B
the dual gradient is basically adaptive penalty for violation of constraint)

Constraint: $E_{a\sim\pi}[\log \pi(a|s)] \leq -\bar{\mathcal{H}}$: α

update $\alpha \mathrel{+}= \underbrace{E_{a\sim\pi}[\log \pi(a|s)] + \bar{\mathcal{H}}}$ → $\bar{\mathcal{H}} = -\dim A$

if this is positive, the constraint is violated
then penalty is increased