« Alireza Gargoori Motlagh _ 98102176 »

« Homework 3 _ Reinforcement Learning »

1

(a) One of the main requirements of LQR to converge is that environment must be fully-observable and we have x_t in each time step. Also, the quadratic approximation for the cost function and linear approximation for dynamics must be reasonable. Another thing is that in short time horizons LQR may lead to suboptimal solutions. Also, the system neends to be controllable in full coordinates.

(b) When the environment is partially observable, we have POMDP and therefore we can estimate states through the observations using HMMs : $P(s_t \mid O_{1:t}) = P(O_t \mid s_t) \sum_{s_{t-1}} P(s_t \mid s_{t-1}) P(s_{t-1} \mid O_{1:t-1})$

or Variational inference : treating hidden states as latent variables.

After estimating the states, we can replace them and run LQR ;

it can be shown that optimal policy is : $u_t = K_t E[x_t \mid O_{1:t}] + k_t$

(C) Model-free methods suffer from sample-efficiency while model-based methods suffer from significant bias, since complex unknown dynamics cannot always be modeled accurately enough to produce effective; therefore combining these methods could result in a better unbiased sample-efficient model.

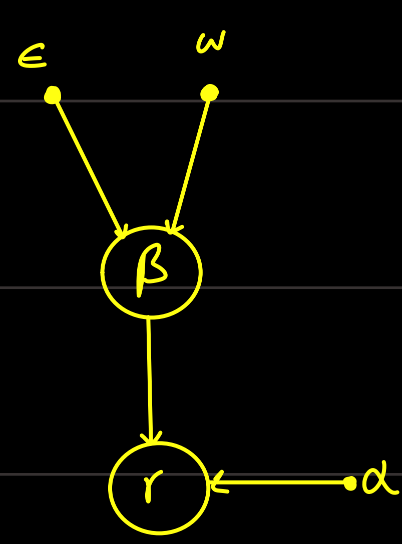
To combine LQR with a deep model-free model, we can
1 use the model-free method (such as TPO, TRPO, SAC, DQN, etc.) to learn an initial control policy through interacting with environment 2; then this learned policy could be used to collect trajectories for fitting the system's dynamics. 3 LQR assumes a linear dynamic for the system; so using the collected data, we fit a linear model to state-action to model the dynamics. 4 After designing a reasonable cost function and a quadratic approximation for that, 5 we can now run LQR for trajectory optimization and finding the optimal control policy. 6 Then we use the LQR solution by using it as a starting policy to fine-tune model-free policy and iterate the above steps (2-6) to improve the policy.

(d) In the stochastic system, we have $x_{t+1} \sim \mathbb{P}(x_{t+1} \mid x_t, u_t)$; e.g. a normal distribution; therefore we can model the uncertainty by using stochastic dynamics. However, in iLQR we can have a nonlinear estimation for the mean & std of this distribution and then approximate it as a Linear-quadratic for dynamics and cost using taylor expansion around the current state-action pair. Other nonlinear parameterization of other distributions are possible, yet very hard to solve.

To ensure exploration, we can add a regularization term to the cost function which accounts for exploration, such as entropy or information gain. These terms could be expanded through quadratic approximation as well and the stochastic iLQR would both capture the optimal control policy and enough exploration.

2

(a) $r \sim \text{Gamma}(\alpha, \beta) \longrightarrow P(r \mid \alpha, \beta) = \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}$



$\beta \sim \text{Gamma}(\epsilon, \omega) \longrightarrow P(\beta \mid \epsilon, \omega) = \dfrac{\omega^{\epsilon}}{\Gamma(\epsilon)} \beta^{\epsilon-1} e^{-\omega\beta}$

$$P(\beta \mid r_1, \alpha, \epsilon, \omega) \propto P(r_1 \mid \alpha, \beta)\, P(\beta \mid \epsilon, \omega)$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r} \frac{\omega^{\epsilon}}{\Gamma(\epsilon)} \beta^{\epsilon-1} e^{-\omega\beta} = \frac{f(r_1, \omega, \alpha, \epsilon)}{\Gamma(\alpha+\epsilon)} \beta^{\alpha+\epsilon-1} e^{-(r_1+\omega)\beta}$$

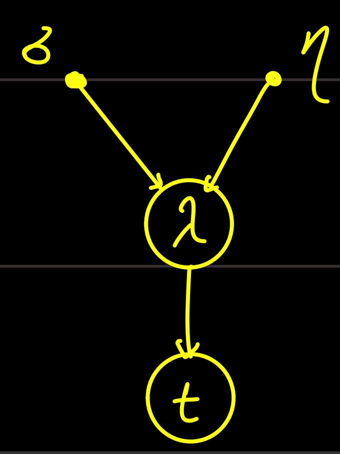
$\longrightarrow$ We can confirm that the $P(\beta \mid \epsilon, \omega)$ is a conjugate prior for the likelihood and therefore:

$$P(\beta \mid r_1, \alpha, \epsilon, \omega) = \frac{(\omega')^{\epsilon'}}{\Gamma(\alpha+\epsilon)} \beta^{\overbrace{\alpha+\epsilon}^{\epsilon'}-1} e^{-\overbrace{(r_1+\omega)}^{\omega'}\beta} = \frac{(r_1+\omega)^{\alpha+\epsilon}}{\Gamma(\alpha+\epsilon)} \beta^{\alpha+\epsilon-1} e^{-(r_1+\omega)\beta}$$

$$\longrightarrow \boxed{\beta \mid r_1, \alpha, \epsilon, \omega \sim \text{Gamma}(\epsilon' = \alpha+\epsilon,\ \omega' = r_1+\omega)}$$

- $t \sim \text{Exp}(\lambda) \implies p(t|\lambda) = \lambda e^{-\lambda t}$

$\lambda \sim \text{Gamma}(\delta, \eta) \implies p(\lambda|\delta,\eta) = \dfrac{\eta^{\delta}}{\Gamma(\delta)} \lambda^{\delta-1} e^{-\eta\lambda}$



$$p(\lambda|t_1,\delta,\eta) \propto p(t_1|\lambda)\, p(\lambda|\delta,\eta)$$

$$= \lambda e^{-\lambda t_1} \frac{\eta^{\delta}}{\Gamma(\delta)} \lambda^{\delta-1} e^{-\eta\lambda} = f(\eta,\delta,t_1)\, \lambda^{\delta} e^{-(t_1+\eta)\lambda}$$

→ We can confirm that the $p(\lambda|t_1,\delta,\eta)$ is a conjugate prior for the likelihood and therefore:

$$p(\lambda|t_1,\delta,\eta) = \frac{(\eta')^{\delta'}}{\Gamma(\delta')} \lambda^{\delta'-1} e^{-\eta'\lambda} \quad \begin{cases} \delta' = \delta+1 \\ \eta' = t_1+\eta \end{cases}$$

$$\boxed{\begin{gathered} \lambda|t_1,\delta,\eta \sim \text{Gamma}(\delta'=\delta+1,\ \eta'=t_1+\eta) \\ p(\lambda|t_1,\delta,\eta) = \frac{(t_1+\eta)^{\delta+1}}{\Gamma(\delta+1)} \lambda^{\delta} e^{-(t_1+\eta)\lambda} \end{gathered}}$$

(b) $p(t_2|t_1) = \int p(t_2, \lambda | t_1)\, d\lambda = \int p(t_2|\lambda, t_1)\, p(\lambda|t_1)\, d\lambda =$

$$\int p(t_2|\lambda)\, p(\lambda|t_1)\, d\lambda = \int \lambda e^{-\lambda t_2} \frac{\eta'^{\delta'}}{\Gamma(\delta')} \lambda^{\delta'-1} e^{-\eta'\lambda}\, d\lambda =$$

$$\frac{\eta'^{\delta'}}{\Gamma(\delta')} \int_0^\infty \lambda^{\delta'} e^{-(\eta'+t_2)\lambda}\, d\lambda = \qquad \left(\Gamma(\delta) = (\delta-1)!\right) \; \delta' \in \mathbb{N}$$

$$\frac{\eta'^{\delta'}}{(\delta-1)!} \frac{1}{(\eta'+t_2)} \underbrace{\int_0^\infty \lambda^{\delta'} (\eta'+t_2)\, e^{-(\eta'+t_2)\lambda}\, d\lambda}_{\mathrm{I}} =$$

I : Integration by parts $\begin{cases} u(\lambda) := \lambda^{\delta'} \\ v(\lambda) = e^{-(\eta'+t_2)\lambda} \end{cases}$

$$\mathrm{I} = \int_0^\infty u(\lambda) v'(\lambda)\, d\lambda = u(\lambda) v(\lambda) \Big|_0^\infty - \int_0^\infty u'(\lambda) v(\lambda)\, d\lambda$$

$$= \underbrace{\lambda^{\delta'} e^{-(\eta'+t_2)\lambda} \Big|_0^\infty}_{=0 \; \left(\text{since } \delta' \in N,\; \eta', t_2 > 0\right)} - \frac{\delta'}{\eta'+t_2} \underbrace{\int_0^\infty \lambda^{\delta'-1} (\eta'+t_2)\, e^{-(\eta'+t_2)\lambda}\, d\lambda}_{\mathrm{II}}$$

$$\mathrm{II} = \underbrace{\lambda^{\delta'-1} e^{-(\eta'+t_2)\lambda} \Big|_0^{\infty}}_{=0} - \frac{\delta'-1}{\eta'+t_2} \underbrace{\int_0^{\infty} \lambda^{\delta'-2} (\eta'+t_2) e^{-(\eta'+t_2)\lambda} \, d\lambda}_{\mathrm{III}}$$

$$\mathrm{III} = \underbrace{\lambda^{\delta'-2} e^{-(\eta'+t_2)\lambda} \Big|_0^{\infty}}_{=0} - \frac{\delta'-2}{\eta'+t_2} \underbrace{\int_0^{\infty} \lambda^{\delta'-2} (\eta'+t_2) e^{-(\eta'+t_2)\lambda} \, d\lambda}_{\mathrm{IV}}$$

$\vdots$

$$\delta' \text{ times} = \underbrace{\lambda^{0} e^{-(\eta'+t_2)} \Big|_0^{\infty}}_{=0} - \underline{\frac{1}{\eta'+t_2} \int_0^{\infty} (\eta'+t_2) e^{-(\eta'+t_2)\lambda} \, d\lambda}$$

$$= \frac{-1}{\eta'+t_2} e^{-(\eta'+t_2)\lambda} \Big|_0^{\infty} = \frac{1}{\eta'+t_2}$$

$$\Rightarrow \boxed{\mathrm{I} = \frac{\delta'}{\eta'+t_2} \left(\frac{\delta'-1}{\eta'+t_2} \left(\frac{\delta'-2}{\eta'+t_2} \left(\cdots \frac{1}{\eta'+t_2} \right)\right)\right) = \frac{\delta'!}{(\eta'+t_2)^{\delta'}}}$$

$$\Rightarrow p(t_2 \mid t_1) = \frac{\eta'^{\delta'}}{(\delta-1)!} \frac{1}{(\eta'+t_2)} \frac{\delta!}{(\eta'+t_2)^{\delta'}} = \frac{\eta'^{\delta'} \delta'}{(\eta'+t_2)^{\delta'+1}}$$

$$\longrightarrow p(t_2 \mid t_1) = \frac{\eta'^{\,\delta'}\,\delta'}{(\eta' + t_2)^{\delta'+1}} \;\; \frac{\times\, \eta'^{-(\delta'+1)}}{\times\, \eta'^{-(\delta'+1)}} \;\; = \frac{\eta'^{-1}\,\delta'}{\left(1 + \frac{t_2}{\eta'}\right)^{\delta'+1}}$$

$$= \frac{\delta'}{\eta'}\left(1 + \frac{t_2}{\eta'}\right)^{-(\delta'+1)}$$

$$\Longrightarrow \boxed{\begin{aligned} & p(t_2 \mid \delta', \eta') = \frac{\delta'}{\eta'}\left(1 + \frac{t_2}{\eta'}\right)^{-(\delta'+1)} \\ & t_2 \sim \text{Lomax}(\delta', \eta') \end{aligned}}$$

(c) Risky: Since the rewards could get to infinity, in this scenario the agent would always choose to play. Risk-free: In contrast to the previous, since the time it takes could be anything from zero to infinity, the agent would always chose to not play.

Risk-neutral: we need to compare $\mathbb{E}\left[\frac{r}{t}\right]$ with K;

$$\mathbb{E}[r/t] = \mathbb{E}[r]\,\mathbb{E}\left[\frac{1}{t}\right]$$

$$\mathbb{E}[r] = \int r \int p(r \mid \alpha, \beta)\, p(\beta \mid \epsilon, \omega)\, d\beta\, dr = R$$

$$\mathbb{E}\left[\frac{1}{t}\right] = \int \frac{1}{t} \int p(t \mid \lambda)\, p(\lambda \mid \sigma, \eta)\, d\lambda\, dt = T$$

if $\frac{R}{T} \geq K \Longrightarrow$ would play

else $\Longrightarrow$ wouldn't play.

3

(a) $$\log p(O_{1:T}) \geq \sum_t \mathbb{E}_{(s_t,a_t)\sim q}\left[r(s_t,a_t) + H(q(a_t|s_t))\right]$$

$t=T$: $$\mathbb{E}_{(s_T,a_T)\sim q}\left[r(s_T,a_T) - \log q(a_T|s_T)\right] =$$

$$\mathbb{E}_{(s_T,a_T)\sim q}\left[\log(\exp(r(s_T,a_T)) - \log q(a_T|s_T)\right] =$$

$$\mathbb{E}_{s_T\sim q(s_T)}\left[\mathbb{E}_{a_T\sim q(a_T|s_T)}\left[\log(\exp(r(s_T,a_T)) - \log q(a_T|s_T)\right]\right]$$

$$= \mathbb{E}_{s_T\sim q(s_T)}\left[-D_{KL}\left(q(a_T|s_T) \,\|\, \frac{1}{Z}\exp(r(s_T,a_T))\right)\right]$$

We wanted to maximize lower bound; since $D_{KL} \geq 0$, we need to have:

$$q(a_T|s_T) = \arg\max_q -D_{KL}\left(q(a_T|s_T) \,\|\, \frac{1}{Z}\exp(r(s_T,a_T))\right)$$

$$= \arg\min_q D_{KL}\left(q(a_T|s_T) \,\|\, \frac{1}{Z}\exp(r(s_T,a_T))\right) \Rightarrow$$

$$q(a_t \mid s_t) = \frac{1}{Z} \exp\left(r(s_T, a_T)\right)$$

$$Z = \int \exp\left(r(s_T, a_T)\right) da_T$$

$$\Longrightarrow \quad q(a_t \mid s_t) = \frac{\exp\left(r(s_T, a_T)\right)}{\int \exp\left(r(s_T, a_T)\right) da_T}$$

☆ for $t = T$ we have: $Q(s_T, a_T) = r(s_T, a_T)$ and therefore:

$$V(s_T) = \log \int \exp\left(Q(s_T, a_T)\right) da_T = \log \int \exp\left(r(s_T, a_T)\right) da_T$$

$$\longrightarrow \quad \boxed{q(a_T \mid s_T) = \frac{\exp\left(Q(s_T, a_T)\right)}{\exp\left(V(s_T)\right)} = \exp\left(Q(s_T, a_T) - V(s_T)\right)}$$

☆ for $t < T$:

$$q(a_t \mid s_t) = \arg\max_q \mathbb{E}_{s_t \sim q(s_t)} \Big[\mathbb{E}_{a_t \sim q(a_t \mid s_t)} \Big[\overbrace{r(s_t, a_t) + \mathbb{E}_{p(s_{t+1} \mid a_t, s_t)} \left[V(s_{t+1}) \right] }^{Q(s_t, a_t)} \Big] - \log q(a_T \mid s_T) \Big]$$

$= \dots$ (next page)

$$\longrightarrow q(a_t|s_t) = \arg\max_q \mathbb{E}_{s_t \sim q(s_t)} \left[\mathbb{E}_{a_t \sim q(a_t|s_t)} \left[Q(s_t, a_t) - \log q(a_t|s_t) \right] \right]$$

$$= \arg\min_q \mathbb{E}_{s_t \sim q(s_t)} \left[D_{KL}\left(q(a_t|s_t) \,\|\, \frac{1}{Z} \exp\left(Q(s_t,a_t)\right)\right) \right]$$

$D_{KL} \geq 0$ and $D_{KL}(p\|q) = 0$ when $p=q$ $\Rightarrow$

$$q(a_t|s_t) = \frac{1}{Z} \exp\left(Q(s_t,a_t)\right), \quad Z = \int \exp\left(Q(s_t,a_t)\right) da_t$$

$$\Rightarrow q(a_t|s_t) = \frac{\exp\left(Q(s_t,a_t)\right)}{\int \exp\left(Q(s_t,a_t)\right) da_t}$$

$$V(s_t) = \log \int \exp\left(Q(s_t,a_t)\right) da_t \Rightarrow$$

$$q(a_t|s_t) = \frac{\exp\left(Q(s_t,a_t)\right)}{\exp\left(V(s_t)\right)} = \exp\left(Q(s_t,a_t) - V(s_t)\right)$$

$$\Rightarrow \boxed{q(a_t|s_t) = \exp\left(Q(s_t,a_t) - V(s_t)\right) = \exp\left(A(s_t,a_t)\right)}$$

(b) $\pi_\theta(s_t, a_t) = \pi_\theta(a_t | s_t)\, \pi_\theta(s_t)$

$$J(\theta) = \sum_t \mathbb{E}_{(s_t,a_t)\sim\pi_\theta(s_t,a_t)}\left[r(s_t,a_t) + H(\pi_\theta(a_t|s_t))\right]$$

$$\Longrightarrow \nabla_\theta J(\theta) = \nabla_\theta \sum_t \mathbb{E}_{(s_t,a_t)\sim\pi_\theta(s_t,a_t)}\left[r(s_t,a_t) + H(\pi_\theta(a_t|s_t))\right]$$
$$= \sum_t \nabla_\theta \mathbb{E}_{(s_t,a_t)\sim\pi_\theta(s_t,a_t)}\left[r(s_t,a_t) - \log(\pi_\theta(a_t|s_t))\right]$$

$$\Longrightarrow \nabla_\theta J(\theta) = \nabla_\theta \sum_\tau \pi_\theta(\tau) \sum_t \left(r(s_t,a_t) - \log \pi_\theta(a_t|s_t)\right)$$
$$= \underbrace{\nabla_\theta \sum_\tau \pi_\theta(\tau) \sum_t r(s_t,a_t)}_{\mathrm{I}} - \underbrace{\nabla_\theta \sum_\tau \pi_\theta(\tau) \sum_t \log \pi_\theta(a_t|s_t)}_{\mathrm{II}}$$

I : $\nabla_\theta \sum_\tau \pi_\theta(\tau)\, r(\tau) = \sum_\tau \nabla_\theta \pi_\theta(\tau)\, r(\tau) = \sum_\tau \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau)\, r(\tau)$

$$= \mathbb{E}_{\tau\sim\pi_\theta(\tau)}\Big[\underbrace{\nabla_\theta \log \pi_\theta(\tau)}_{*} \sum_t r(s_t,a_t)\Big]$$

☆ $\pi_\theta(\tau) = p(s_1) \prod_t p(s_{t+1} | s_t, a_t)\, \pi_\theta(a_t|s_t) \Longrightarrow \nabla_\theta \log \pi_\theta(\tau) = \sum_t \nabla_\theta \log \pi(a_t|s_t)$

$$\Longrightarrow \boxed{\mathrm{I} = \mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_t r(s_t,a_t)\right]}$$

$\mathrm{II}:$

$$\nabla_\theta \sum_\tau \pi_\theta(\tau) \sum_t \log \pi_\theta(a_t|s_t) =$$

$$\sum_\tau \nabla_\theta \pi_\theta(\tau) \sum_t \log \pi_\theta(a_t|s_t) + \sum_\tau \pi_\theta(\tau) \underbrace{\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)}_{= \nabla_\theta \log \pi_\theta(\tau)} =$$

$$\sum_\tau \nabla_\theta \pi_\theta(\tau) \sum_t \log \pi_\theta(a_t|s_t) + \sum_\tau \underbrace{\pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau)}_{\nabla_\theta \pi_\theta(\tau)} =$$

$$\sum_\tau \nabla_\theta \pi_\theta(\tau) \left(\sum_t \log \pi_\theta(a_t|s_t) + 1 \right) =$$

$$\sum_\tau \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) \left(\sum_t \log \pi_\theta(a_t|s_t) + 1 \right) =$$

$$\mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(\tau) \left(\sum_t \log \pi_\theta(a_t|s_t) + 1 \right) \right]$$

$$\overset{*}{\Longrightarrow} \boxed{\mathrm{II} = \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[\sum \nabla_\theta \log \pi_\theta(a_t|s_t) \left(\sum_t \log \pi_\theta(a_t|s_t) + 1 \right) \right]}$$

$$\nabla_\theta J(\theta) = \mathrm{I} - \mathrm{II} \Rightarrow$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_t r(s_t, a_t)\right] -$$

$$\mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_t \log \pi_\theta(a_t|s_t) + 1\right] =$$

$$\mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t=1}^{T}\left[r(s_t, a_t) - \log \pi_\theta(a_t|s_t)\right] - 1\right)\right]$$

Due to causality, that is the policy at $\underline{t}$ cannot affect the previous rewards and policies, we can rewrite the equation as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T}\left[r(s_{t'}, a_{t'}) - \log \pi_\theta(a_{t'}|s_{t'})\right] - 1\right)\right]$$

$$= \mathbb{E}_{(s_t, a_t) \sim \pi_\theta(s_t, a_t)}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T}\left[r(s_{t'}, a_{t'}) - \log \pi_\theta(a_{t'}|s_{t'})\right] - 1\right)\right]$$

$\Longrightarrow$ We can sample N trajectories and approximate $\nabla_\theta J(\theta)$ as:

$$\boxed{\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_i \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\left(\sum_{t'=t}^{T}\left[r(s_{i,t'}, a_{i,t'}) - \log \pi_\theta(a_{i,t'}|s_{i,t'})\right] - 1\right)}$$

(c) $\pi_\theta(a_t|s_t) = \exp\left(Q(s_t,a_t) - V(s_t)\right)$ *

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_i \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T}\left[r(s_{t'},a_{t'}) - \log \pi_\theta(a_t|s_t)\right] - 1\right) \quad **$$

* in ** $\Rightarrow$

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_i \sum_{t=1}^{T} \nabla_\theta \left(Q(s_t,a_t) - V(s_t)\right)\left(\sum_{t'=t}^{T}\left[r(s_{t'},a_{t'}) - Q(s_{t'},a_{t'}) + V(s_{t'})\right] - 1\right)$$

$$= \frac{1}{N}\sum_i \sum_{t=1}^{T} \left(\nabla_\theta Q(s_t,a_t) - \nabla_\theta V(s_t)\right)\left(\sum_{t'=t}^{T}\left[r(s_{t'},a_{t'}) - Q(s_{t'},a_{t'}) + V(s_{t'})\right] - 1\right)$$

$$\Rightarrow \boxed{\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_i \sum_{t=1}^{T} \left(\nabla_\theta Q(s_t,a_t) - \nabla_\theta V(s_t)\right)\left(\sum_{t'=t}^{T}\left[\underbrace{r(s_{t'},a_{t'}) - Q(s_{t'},a_{t'})}_{-E[V(s_{t'+1})]} + V(s_{t'})\right] - 1\right)}$$

(d) ☆ MaxEnt RL grad. :

$$\boxed{\begin{aligned}\nabla_\theta J(\theta) &= E_{\tau\sim\pi_\theta(\tau)}\left[\sum_{t=1}^{T}\left(\nabla_\theta Q(s_t,a_t) - \nabla_\theta V(s_t)\right)\hat{A}(s_t,a_t)\right] \\ &\simeq \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\nabla_\theta Q(s_t^{(i)},a_t^{(i)}) - \nabla_\theta V(s_t^{(i)})\right)\hat{A}(s_t^{(i)},a_t^{(i)})\end{aligned}}$$

☆ soft Q-learning grad. :

$$\boxed{\begin{aligned}\nabla_\theta J(\theta) &= E_{\tau\sim\pi_\theta(\tau)}\left[\sum_{t=1}^{T}\nabla_\theta Q(s_t,a_t)\,\hat{A}(s_t,a_t)\right] \\ &\simeq \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta Q(s_t^{(i)},a_t^{(i)})\,\hat{A}(s_t^{(i)},a_t^{(i)})\end{aligned}}$$

(both taken from RL & Control as Probabilistic Inference: Tutorials and Review)

There is an extra term, $\nabla_\theta V(s_t)$ in MaxEnt RL gradient which resemble the insufficiency of policy gradient to resolve the addition or substraction of an action-independent constant.
This term could be eliminated for a particular choice of $\hat{A}(s_t, a_t)$, which in the paper "equivalence between policy gradients and soft Q-learning" is explained in detail.
In general, Q-learning methods are more sample-efficient but their Q-function estimate may be inaccurate; but in the sense of entropy-regularized Q-learning has a better estimate of this function.