1.) $\sum_t \left[H(\Theta|\varepsilon_t, a_t) - H(\Theta|S_{t+1}, \varepsilon_t, a_t) \right]$

$I(X;Y) = H(X) - H(X|Y)$

2.) $\therefore I(\Theta; S_{t+1}|\varepsilon_t, a_t)$

$$= \mathbb{E}_{S_{t+1} \sim P(\cdot|\varepsilon_t, a_t)} \left[D_{KL}\left[P(\Theta|\varepsilon_t, a_t, S_{t+1}) \,||\, P(\Theta|\varepsilon_t, a_t) \right] \right]$$

$P(\Theta|\varepsilon_t, a_t) = p(\theta|\varepsilon_t)$

$I(X;Y) = E_Y \left[D_{KL}\left[P_{X|Y} \,||\, P_X \right] \right]$

info gain

3.) $r'(s_t, a_t, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}\left[p(\theta|\varepsilon_t, a_t, s_{t+1}) \,||\, p(\theta|\varepsilon_t) \right]$

4.) $p(\theta|\varepsilon_t, a_t, s_{t+1}) = \dfrac{p(\theta, \varepsilon_t, a_t, s_{t+1})}{p(\varepsilon_t, a_t, s_{t+1})}$ chain rule

$$= \frac{\cancel{p(a_t)}\, \cancel{p(\varepsilon_t|a_t)}\, p(\theta|\varepsilon_t, a_t)\, p(s_{t+1}|\varepsilon_t, a_t, \theta)}{\cancel{p(a_t)}\, \cancel{p(\varepsilon_t|a_t)}\, p(s_{t+1}|\varepsilon_t, a_t)}$$

$$= \frac{p(\theta|\varepsilon_t)\, p(s_{t+1}|\varepsilon_t, a_t, \theta)}{p(s_{t+1}|\varepsilon_t, a_t)}$$

$p(\theta|\varepsilon_t, a_t) = p(\theta|\varepsilon_t)$

5.) $p(s_{t+1}|\varepsilon_t, a_t) = \int_\theta \underbrace{p(s_{t+1}|\varepsilon_t, a_t, \theta)\, p(\theta|\varepsilon_t) d\theta}_{\text{hard to compute}}$

Calculating $p(\theta|D)$ is hard for given dataset D

$\therefore$ Approximate using $q(\theta;\phi)$ by minimizing

$D_{KL}\left[q(\theta;\phi) \,||\, p(\theta|D) \right]$

6.) Variational Lower bound derivation

$$KL\left[q(z) \,||\, p(z|x) \right] = \int_z q(z) \log \frac{q(z)}{p(z|x)}$$

$$= -\int_z q(z) \log \frac{p(z|x)}{q(z)}$$

$$= -\left[\int_z q(z) \log \frac{p(x,z)}{q(z)} - \int_z q(z) \log p(x)\right]$$

$$= -\int q(z) \log \left(\frac{p(x,z)}{q(z)}\right) + \log p(x) \int_z q(z)$$

$$= -L + \log p(x).$$

$$\therefore L = \log p(x) - KL\left[q(z) \| p(z|x)\right]$$

$$L \leq \log p(x) \qquad \text{as } KL \geq 0$$

Lower bound L hits the log probability iff the approx distro is perfectly close to the posterior distro. Hence maximise L

$$L[q(\theta;\phi), D] = \mathop{E}_{\theta \sim q(\cdot;\phi)}\left[\log p(D|\theta)\right] - D_{KL}\left[q(\theta;\phi) \| p(\theta)\right]$$

7.) Thus instead of calculating info gain in eq 3) we compute approx to it

$$\therefore r'(s_t, a_t, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}\left[q(\theta; \phi_{t+1}) \| q(\theta; \phi_t)\right]$$

ϕ_{t+1} and ϕ_t are new and old params representing the agent's belief

8.) ϕ_t is parametrized using BNNs

$$p(y|x) = \int_\theta p(y|x;\theta)\, q(\theta;\phi)\, d\theta$$

9) $$\log P(D) = \underbrace{\int_\theta q(\theta;\phi) \log \frac{p(\theta, D)}{q(\theta;\phi)}\, d\theta}_{L[q(\theta;\phi), D]} + D_{KL}\left[q(\theta;\phi) \| p(\theta;D)\right]$$

Implementation

11) $q(\theta;\phi) = \prod_{i=1}^{|\theta|} \mathcal{N}(\theta_i \mid \mu_i ; \sigma_i^2)$

$\phi = \{\mu, \sigma\}$ $\sigma \rightarrow$ covariance diagonal matrix

$\sigma = \log(1+e^{\rho})$ $\rho \in \mathbb{R}$ as $\sigma > 0$

$\boxed{\text{Eqn 6}}$

$$L[q(\theta;\phi), D] = \mathbb{E}_{\theta \sim q(\cdot,\phi)}[\log p(D|\theta)] - \underbrace{D_{KL}[q(\theta;\phi) \| p(\theta)]}$$

(VLB) variational lower bound

Approximated through sampling

$\mathbb{E}_{\theta\sim q(\cdot,\phi)}[\log p(D|\phi)] \approx \frac{1}{N} \sum_{i=1}^{N} \log p(D|\theta_i)$

with N samples drawn according to $\theta \sim q(\cdot;\phi)$

Refer stochastic gradient variational Bayes (SGVB)

The optimation of VLB is done at regular intervals by sampling D. This is done to break up intratrajectory sample correlation

12) ϕ for eqn 7) is calculated as

$$\phi' = \operatorname{argmin} \underbrace{\left[D_{KL}[q(\theta;\phi) \| q(\theta;\phi_{t-1})] - \mathbb{E}_{\theta\sim q(\cdot;\phi)}[\log(s_t | \xi_t, a_t;\theta)] \right]}_{L(q(\theta;\phi), s_t)}$$

12

Here, expectation over θ is replaced with samples $\theta \sim q(\cdot;\phi)$

13) To optimize 12) efficiently,

Compute $D_{KL}[q(\theta;\phi + \lambda\Delta\phi) \| q(\theta;\phi)]$

$\Delta\phi = H^{-1}(L) \nabla_{\phi} L(q(\theta;\phi), s_t)$

$H(L)$ is Hessian of $L(q(\theta;\phi), s_t)$

Since q is fully factorized Gaussian.

14) $D_{KL}[q(\theta;\phi) \| q(\theta;\phi')]$

$$= \frac{1}{2} \sum_{i=1}^{|\theta|} \left[\left(\frac{\sigma_i}{\sigma_i'}\right)^2 + 2\log\sigma_i' - 2\log\sigma_i + \frac{(\mu_i' - \mu_i)^2}{\sigma_i'^2} \right] - \frac{|\theta|}{2}$$

KL divergence is quadratic (approx) and the log-likelihood term can be seen as locally linear compared to the curved KL term, we approximate H by only calculating it for the KL term $l_{KL}(q(\theta;\phi))$

15.) $\frac{\partial^2 l_{KL}}{\partial \mu_i^2} = \frac{1}{\log^2(1+e^{\rho_i})}$ and $\frac{\partial^2 l_{KL}}{\partial \rho_i^2} = \frac{2e^{2\rho_i}}{(1+e^{\rho_i})^2} \frac{1}{\log^2(1+e^{\rho_i})}$

16) $D_{KL}[q(\theta;\phi+\lambda\Delta\phi) \| q(\theta;\phi)] \approx \frac{1}{2}\lambda^2 \nabla_\phi l^T H^{-1}(l_{KL}) \nabla_\phi l$

Here $H^{-1}(l_{KL})$ is diagonal

Algo

For each epoch n do

for each timestep t in trajectory do

Generate action $a_t \sim \pi_\alpha(s_t)$

and sample $s_{t+1} \sim P(\cdot \mid s_t, a_t)$, get $r_t(s_t, a_t)$

Add (s_t, a_t, s_{t+1}) in FIFO buffer R

Compute $D_{KL}[q(\theta;\phi'_{n+1}) \| q(\theta;\phi_{n+1})]$ by approximation $\nabla^T H^{-1} \nabla$ from (16) for diagonal BNNs

Divide $D_{KL}[q(\theta;\phi'_{n+1}) \| q(\theta;\phi_{n+1})]$ by median of previous KL divergences

Construct $r'(s_t, a_t, s_{t+1}) \leftarrow r(s_t, a_t) + \eta D_{KL}[q(\theta;\phi'_{n+1}) \| q(\theta;\phi_{n+1})]$

Minimize $D_{KL}[q(\theta;\phi_n) \| p(\theta)] - \mathbb{E}_{\theta \sim q(\cdot;\phi_n)}[\log p(D \mid \theta)]$ (6) with D sampled randomly from R leading to updated posterior $q(\theta;\phi_{n+1})$

Use rewards $\{r'(s_t, a_t, s_{t+1})\}$ to update π_α using any standard RL algo.