

# 1 REINFORCE versus Reparameterization

The REINFORCE trick is just the application of the following useful identity:

$$\nabla_{\phi} \log q_{\phi}(x) = \frac{1}{q_{\phi}(x)} \nabla_{\phi} q_{\phi}(x) \quad \rightarrow \quad q_{\phi}(x) \nabla_{\phi} \log q_{\phi}(x) = \nabla_{\phi} q_{\phi}(x).$$

We can then use this identity to evaluate expectations using Monte-Carlo:

$$\nabla_{\phi} \mathbb{E}_{q_{\phi}(x)} [f(x)] = \int f(x) \nabla_{\phi} q_{\phi}(x) dx = \mathbb{E}_{q_{\phi}(x)} [f(x) \nabla_{\phi} \log q_{\phi}(x)]$$

Conversely, a unique reparameterization trick must be formulated for every distribution that you apply it to, and cannot be applied to discrete distributions. For the univariate normal distribution, find a literal reparameterization such that we can sample from white noise (i.e. a Gaussian with mean zero, and variance one), and then evaluate the expectation over this distribution. This looks something like,

$$\nabla_{\phi} \mathbb{E}_{x \sim q_{\phi}(x)} [f(x)] = \nabla_{\phi} \mathbb{E}_{\epsilon \sim N(0,1)} [f(\mu_{\phi} + \epsilon \sigma_{\phi})] = \mathbb{E}_{\epsilon \sim N(0,1)} [\nabla_{\phi} f(\mu_{\phi} + \epsilon \sigma_{\phi})] \quad (1)$$

In general, the reparameterization is known to be much lower variance than the REINFORCE estimator, but this difference can vary drastically when you introduce trust region methods and control variates depending on how accurate the control variate is, and how small the trust region.

## 2 Finite Time Horizon Problems: Reverse KL

This section will discuss how one might produce an approximately optimal policy ( $\pi_{\phi}(a_t|s_t) \approx \pi(a_t|s_t, o_t, \dots, o_{T-1})$ ) by minimizing the reverse KL divergence between the optimal policy and our parameterized approximation. This section contains proofs and derivations relevant to the reverse KL applied to RLAI. More specifically, this section gives information on the un-marginalized objective, the marginalized objective, the REINFORCE trick, and the reparameterization trick. While largely technical, it does highlight some of the advantages, and disadvantages with this objective which will be useful later when we discuss how to create a practical algorithm. This section will rely on many of the usual tools found in variational inference, and makes heavy use of the reinforce trick, and zero expectation score function trick.

### 2.1 Derivation of the Un-marginalized Reverse KL Objective

For the following we assume that  $\tau_t := (a_t, s_t, s_{t+1}, r_t)$ , and  $(s_{0:T}, a_{0:T-1}, r_{0:T-1}) := \tau$ . We also make a distinction between  $q_{\phi}(\tau)$  and  $p(\tau, O)$  which are defined in the following way:

$$q_{\phi}(\tau) = p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \pi_{\phi}(a_t|s_t),$$

$$p(\tau, O) = p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \left(1 - \frac{1}{Z} \exp r_t(a_t, s_t)\right)^{1-O_t} \left(\frac{1}{Z} \exp r_t(a_t, s_t)\right)^{O_t} \pi_0(a_t|s_t).$$

We wish to prove the following:

$$\min_{\phi} D_{KL}(q_{\phi}(\tau) || p(\tau|O=1)) = \min_{\phi} - \mathbb{E}_{\tau \sim q_{\phi}} \left[ \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_{\phi}(a_t|s_t)}{\pi_0(a_t|s_t)} \right]$$

This is relatively easy to show if we consider the functional form of the KL divergence above,

$$\begin{aligned}
\min_{\phi} D_{KL}(q_{\phi}(\tau) \parallel p(\tau|O)) &= \min_{\phi} \left[ \int_{\tau} q_{\phi}(\tau) \log \left( \frac{q_{\phi}(\tau)}{p(\tau|O)} \right) d\tau \right] \\
&= \min_{\phi} \left[ \int_{\tau} q_{\phi}(\tau) \log \left( \frac{q_{\phi}(\tau)}{p(\tau, O)} \right) d\tau - \int_{\tau} q_{\phi}(\tau) \log(p(O)) d\tau \right] \\
&= \min_{\phi} \left[ \int_{\tau} q_{\phi}(\tau) \log \left( \frac{q_{\phi}(\tau)}{p(\tau, O)} \right) d\tau - \log(p(O)) \right] \\
&= \min_{\phi} \left[ \int_{\tau} q_{\phi}(\tau) \log \left( \frac{q_{\phi}(\tau)}{p(\tau, O)} \right) d\tau \right]
\end{aligned}$$

Next we can look at the log term and note, that it can be broken up into each time step, and additionally that the dynamics between the the joint and approximate distribution cancel out. Below we assume that anywhere  $O$  is written, we have  $O_{0:T-1} = 1$  (i.e. the posterior is over optimal policies).

$$\begin{aligned}
\log \left( \frac{q_{\phi}(\tau)}{p(\tau, O)} \right) &= \log(q_{\phi}(\tau)) - \log(p(\tau, O)) = \log(q_{\phi}(\tau)) - \log(p(O|\tau)) - \log(p(\tau)) \\
&= \sum_{t=0}^{T-1} (\log(\pi_{\phi}(a_t|s_t)) - r(a_t, s_t) - \log(\pi_0(a_t|s_t))) = - \sum_{t=0}^{T-1} \left( r(a_t, s_t) - \frac{\log(\pi_0(a_t|s_t))}{\log(\pi_{\phi}(a_t|s_t))} \right)
\end{aligned}$$

With this expression in mind, simply replace the log within the previous integral to get the desired result:

$$\begin{aligned}
\min_{\phi} \left[ \int_{\tau} q_{\phi}(\tau) \log \left( \frac{q_{\phi}(\tau)}{p(\tau, O)} \right) d\tau \right] &= \min_{\phi} - \left[ \int_{\tau} q_{\phi}(\tau) \sum_{t=0}^{T-1} \left( r(a_t, s_t) - \frac{\log(\pi_0(a_t|s_t))}{\log(\pi_{\phi}(a_t|s_t))} \right) d\tau \right] \\
&= \min_{\phi} - \mathbb{E}_{\tau \sim q_{\phi}} \left[ \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_{\phi}(a_t|s_t)}{\pi_0(a_t|s_t)} \right]
\end{aligned}$$

## 2.2 The Reverse KL as a stochastic lower bound

This objective can also be derived as a stochastic lower bound on the marginal log probability of optimality. To see this, rewrite, and then lower-bound the marginal likelihood of these samples under the true data generating process  $p$ :

$$\begin{aligned}
\log p(O) &= \log \int p(O, \tau_{0:T-1}) d\tau = \log \int p(O, \tau_{0:T-1}) \frac{q_{\phi}(\tau_{0:T-1})}{q_{\phi}(\tau_{0:T-1})} d\tau_{0:T-1} \\
&= \log \left( \mathbb{E}_{\tau_{0:T-1} \sim q_{\phi}} \left[ \frac{p(O, \tau_{0:T-1})}{q_{\phi}(\tau_{0:T-1})} \right] \right) \geq \mathbb{E}_{\tau_{0:T-1} \sim q_{\phi}} \left[ \log \left( \frac{p(O, \tau_{0:T-1})}{q_{\phi}(\tau_{0:T-1})} \right) \right].
\end{aligned}$$

## 2.3 Derivation of the Marginalized Reverse KL gradient

For the following we assume for conciseness that  $\tau_t := (a_t, s_t)$ , and that  $(s_{0:T}, a_{0:T-1}) := \tau$ . We again make a distinction between  $q_{\phi}(\tau)$  and  $\pi(\tau)$  which are defined in the following way:

$$q_{\phi}(\tau) = p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \pi_{\phi}(a_t|s_t), \quad \pi_{\phi}(\tau) = \prod_{t=0}^{T-1} \pi_{\phi}(a_t|s_t).$$

In this setting we will actually derive a gradient estimator for the original objective defined above which is much lower variance. From this gradient estimator we can infer an objective whose minimum is the same as the one above up to a constant factor. This derivation is closely related to that of the policy gradient theorem for those interested. We wish to prove the following:

$$\nabla_{\phi} \mathbb{E}_{q(\tau)} \left[ \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_{\phi}(a_t|s_t)}{\pi_0(a_t|s_t)} \right] = \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi}(\tau_{0:T-1})} \left[ \nabla_{\phi} \log \pi_{\phi}(s_t, a_t) \left( \sum_{t'=t}^T r(s_{t'}, a_{t'}) - \log \frac{\pi_{\phi}(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} \right) \right]$$

We start by passing the gradient into the expectation under the assumption that the gradient can be passed into the integral, which should hold so long as the expectation is bounded.

$$\begin{aligned}
\nabla_\phi J_{q,p}(\phi) &= \int \nabla_\phi \left[ \left( \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_\phi(a_t|s_t)}{\pi_0(a_t|s_t)} \right) q_\phi(\tau) \right] d\tau \\
&= \int \nabla_\phi q_\phi(\tau) \left( \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_\phi(a_t|s_t)}{\pi_0(a_t|s_t)} \right) d\tau + \int q_\phi(\tau) \nabla_\phi \left( \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_\phi(a_t|s_t)}{\pi_0(a_t|s_t)} \right) d\tau \\
&= \int \nabla_\phi \log q_\phi(\tau) \left( \sum_{t=0}^{T-1} r(s_t, a_t) - \log \frac{\pi_\phi(a_t|s_t)}{\pi_0(a_t|s_t)} \right) q_\phi(\tau) d\tau - \int \left( \sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \right) q_\phi(\tau) d\tau \\
&= \int \left( \sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \right) \left( \sum_{t'=0}^{T-1} r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} \right) q_\phi(\tau) d\tau \\
&\quad - \int \left( \sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \right) q_\phi(\tau) d\tau \\
&= \int \left( \sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \right) \left( \sum_{t'=0}^{T-1} r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} - 1 \right) q_\phi(\tau) d\tau
\end{aligned}$$

Note that  $s_t$  and  $s'_t$  correspond to the *same* random variable (and same for  $a_t$  and  $a'_t$  for all  $t$ . The 's are introduced only to index the double sum below correctly.

Here we use the score function trick in line two to transform the integral of a gradient into the expectation over a gradient. We then combine terms and simplify, noting that the gradient of the dynamics is zero, and can thus be ignored. We then convolve these sums and further simplify using d separation of the probabilistic graphical model.

$$\begin{aligned}
J_{q,p}(\phi) &= \int \sum_{t=0}^{T-1} \sum_{t'=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \left( r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} - 1 \right) q_\phi(\tau) d\tau, \\
&= \sum_{t=0}^{T-1} \int \sum_{t'=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \left( r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} - 1 \right) q_\phi(\tau) d\tau, \\
&= \sum_{t=0}^{T-1} \sum_{t'=0}^{T-1} \int \nabla_\phi \log \pi_\phi(a_t|s_t) \left( r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} \right) q_\phi(\tau) d\tau, \\
&= \sum_{t=0}^{T-1} \sum_{t'=t}^{T-1} \int \nabla_\phi \log \pi_\phi(a_t|s_t) \left( r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} \right) q_\phi(\tau) d\tau \\
&\quad + \sum_{t=0}^{T-1} \sum_{t'=0}^{t-1} \int \nabla_\phi \log \pi_\phi(a_t|s_t) \left( r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} \right) q_\phi(\tau) d\tau, \\
&= \sum_{t=0}^{T-1} \sum_{t'=t}^{T-1} \int \nabla_\phi \log \pi_\phi(a_t|s_t) \left( r(s_{t'}, a_{t'}) - \log \frac{\pi_\phi(a_{t'}|s_{t'})}{\pi_0(a_{t'}|s_{t'})} \right) q_\phi(\tau) d\tau.
\end{aligned}$$

Where the last line follows from the Markov property of the graphical model which gives  $t' < t \rightarrow r(s_{t'}, a_{t'}) - \log \pi_\phi(a_{t'}|s_{t'}) + \log \pi_0(a_{t'}|s_{t'})$  is independent of  $\nabla_\phi \log \pi_\phi(a_t|s_t)$ . This means that we can decompose the

expectation such that for all  $t' < t$ ,

$$\begin{aligned}
\int \nabla_{\phi} \log \pi_{\phi}(a_t | s_t) f(a_{t'}, s_{t'}) q_{\phi}(\tau) d\tau &= \int f(a_{t'}, s_{t'}) \left( \int \nabla_{\phi} \log \pi_{\phi}(a_t | s_t) \pi_{\phi}(a_t | s_t) da_t \right) \frac{q_{\phi}(\tau)}{\pi_{\phi}(a_t | s_t)} d\tau \setminus a_t \\
&= \int f(a_{t'}, s_{t'}) \left( \int \nabla_{\phi} \pi_{\phi}(a_t | s_t) da_t \right) \frac{q_{\phi}(\tau)}{\pi_{\phi}(a_t | s_t)} d\tau \setminus a_t \\
&= \int f(a_{t'}, s_{t'}) \nabla_{\phi} \left( \int \pi_{\phi}(a_t | s_t) da_t \right) \frac{q_{\phi}(\tau)}{\pi_{\phi}(a_t | s_t)} d\tau \setminus a_t \\
&= \int f(a_{t'}, s_{t'}) \nabla_{\phi} (1) \frac{q_{\phi}(\tau)}{\pi_{\phi}(a_t | s_t)} d\tau \setminus a_t \\
&= \int f(a_{t'}, s_{t'}) (0) \frac{q_{\phi}(\tau)}{\pi_{\phi}(a_t | s_t)} d\tau \setminus a_t = 0
\end{aligned}$$

Therefore the inner expectation is a constant for all  $t'$  less than  $t$ , and the score function estimator is zero in expectation at these terms in the series. This sum of (entropy regularized) rewards ahead is often referred to as the Q function or the advantage (if we include a baseline). If we define this term according to the following expectation,

$$Q(s_t, a_t) = \mathbb{E}_{q_{\phi}(\tau_{t+1:T}) | \tau_t = \{a_t, s_t\}} \left[ \sum_{t'=t}^T r(s_{t'}, a_{t'}) - \log \frac{\pi_{\phi}(a_{t'} | s_{t'})}{\pi_0(a_{t'} | s_{t'})} \right],$$

Then we can concisely write the desired result:

$$\begin{aligned}
\nabla_{\phi} J_{q,p}(\phi) &= \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi}(\tau_{t:T-1})} \left[ \nabla_{\phi} \log \pi_{\phi}(a_t | s_t) \left( \sum_{t'=t}^T r(s_{t'}, a_{t'}) - \log \frac{\pi_{\phi}(a_{t'} | s_{t'})}{\pi_0(a_{t'} | s_{t'})} \right) \right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi}(a_t, s_t)} [\nabla_{\phi} \log \pi_{\phi}(a_t | s_t) Q(s_t, a_t)].
\end{aligned}$$

## 2.4 Control Variates and the Value Function

As in the case of VAEs, one can introduce a parameterized, zero mean mean function to reduce the variance of the gradient estimator. This function in the context of RL is referred to as the baseline, and is generally a function of the state. This gradient tends to look something like:

$$\nabla_{\phi} J_{q,p}(\phi) = \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi}(a_t, s_t)} [\nabla_{\phi} \log \pi_{\phi}(a_t | s_t) Q(s_t, a_t) - b(s_t)].$$

Further, it can actually be shown that the optimal baseline is in fact the integral of our Q function, defined above as the value function:

$$V(s_t) = \int_a Q(a, s_t) \pi_{\phi}(a | s_t) da \quad (2)$$

This yields what is referred to as the advantage, and both in theory and practice avoids many of the pitfalls of learning a policy only via the q function.

$$\nabla_{\phi} J_{q,p}(\phi) = \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi}(a_t, s_t)} [\nabla_{\phi} \log \pi_{\phi}(a_t | s_t) Q(s_t, a_t) - V(s_t)] = \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi}(a_t, s_t)} [\nabla_{\phi} \log \pi_{\phi}(a_t | s_t) A(s_t, a_t)].$$

The value function will be discussed further in a later section where we will consider its relationship to the backwards message in a particle filter. We can however make a couple notes on how critics are estimated or learned in practice however.

### 3 Finite Time Horizon Problems: Forward KL

This section details derivations for how to produce an approximate optimal policy under the forward KL, and like the previous section includes the derivation of its relationship to the marginal log likelihood. The un-marginalized gradient of the objective is also derived along side its estimator as before. In this case we take advantage of tools from self normalized importance weighting to produce our estimator, and briefly discuss some of the issues present.

#### 3.1 Un-marginalized Objective Derivation for the Forward KL

This objective, represents the expected KL divergence between a target distribution, and an approximate distribution under samples from the target distribution:

$$KL(p(\tau|O)||q_\phi(\tau)). \quad (3)$$

Again, our latent variables are given as the associated state-action pairs, while the observed variables represents the optimality distribution. In order to evaluate the expression above, we apply a standard importance weighting scheme, beginning with the removal constant terms within the expectation.

$$\begin{aligned} KL(p(\tau|O)||q_\phi(\tau)) &= \int p(\tau|O) \log \frac{p(\tau|O)}{q_\phi(\tau)} d\tau = - \int p(\tau|O) \log q_\phi(\tau) d\tau + \int p(\tau|O) \log p(\tau|O) d\tau \\ &= - \int p(\tau|O) \log q_\phi(\tau) d\tau + c = \int \log q_\phi(\tau) p(\tau|O) \frac{q_\phi(\tau)}{q_\phi(\tau)} d\tau + c \\ &= \int \frac{1}{p(O)} \log q_\phi(\tau) \frac{p(\tau, O)}{q_\phi(\tau)} q_\phi(\tau) d\tau + c = \mathbb{E}_{\tau \sim q_\phi} \left[ \frac{1}{Z} \log q_\phi(\tau) \frac{p(\tau, O)}{q_\phi(\tau)} \right] + c \end{aligned}$$

#### 3.2 The forward KL as a stochastic upper bound

Similar to the previous RKL, we can also use this divergence to get a bound on the marginal log likelihood of the observed random variables and then attempt to iteratively tighten this bound. Unlike the previous example however, we arrive at a stochastic upper bound that we must minimize.

$$\begin{aligned} KL(p(\tau|O)||q_\phi(\tau)) \geq 0 &\Rightarrow \mathbb{E}_{\tau \sim p} \left[ \log \frac{p(\tau|O)}{q_\phi(\tau)} \right] \geq 0 \\ &\Rightarrow \mathbb{E}_{\tau \sim p} \left[ \log \frac{p(\tau, O)}{p(O)q_\phi(\tau)} \right] \geq 0 \Rightarrow \mathbb{E}_{\tau \sim p} \left[ \log \frac{p(\tau, O)}{q_\phi(\tau)} \right] \geq \mathbb{E}_{\tau \sim p} [\log p(O)] \\ &\Rightarrow \mathbb{E}_{\tau \sim p} \left[ \log \frac{p(\tau, O)}{q_\phi(\tau)} \right] \geq \log p(O) \Rightarrow \mathbb{E}_{\tau \sim q_\phi} \left[ \frac{p(\tau|O)}{q_\phi(\tau)} \log \frac{p(\tau, O)}{q_\phi(\tau)} \right] \geq \log p(O) \end{aligned}$$

This upper bound is re-written to exclude terms that do not include the proposal distribution parameters  $\phi$ , which will become zero when differentiated. We can then apply self normalized importance weighting in order for us to sample from a known distribution in order to approximate the expectation of another.

#### 3.3 Un-marginalized gradient of the Forward KL Objective

Crucially, because the original problem is over a distribution not dependent on  $\phi$ , we can directly pass the gradient into the expectation without the REINFORCE trick, or the reparameterization trick. This gives the following expression for the gradient:

$$\nabla_\phi KL(p(\tau|O=1)||q_\phi(\tau)) = \mathbb{E}_{\tau \sim q_\phi} \left[ \frac{p(\tau, O)}{q_\phi(\tau)} \frac{\nabla_\phi \log q_\phi(\tau)}{Z} \right]$$

This actually gives us a simple algorithm that can evaluate the gradient, and thereby minimize the expected KL between our two distributions following samples generated from our policy interacting with the simulator.

Based upon this approach, we can apply self normalized importance weighting to avoid explicitly computing the constant  $Z$ . This is done in the following way:

$$\nabla_{\phi} KL(p(\tau|O)||q_{\phi}(\tau)) \approx \sum_{i=1}^m [w_i \nabla_{\phi} \log q_{\phi}(\tau_i)] = \frac{1}{Z} \sum_{i=1}^m \left[ \frac{p(\tau_i, O_i)}{q_{\phi}(\tau_i)} \nabla_{\phi} \log q_{\phi}(\tau_i) \right].$$

Where for optimal trajectories ( $O = O_{t=0:T-1}$ ), the un-normalized weights are defined as,

$$\begin{aligned} w_i &= \frac{1}{Z} \frac{p(\tau^i, O)}{q_{\phi}(\tau^i)} = \frac{1}{Z} \frac{p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \left(1 - \frac{1}{Z_0} \exp r_t(a_t, s_t)\right)^{1-O_t} \left(\frac{1}{Z_0} \exp r_t(a_t, s_t)\right)^{O_t} \pi_0(a_t|s_t)}{p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \pi_{\phi}(a_t|s_t)} \\ &= \frac{1}{Z} \prod_{t=0}^{T-1} \frac{\left(1 - \frac{1}{Z_0} \exp r_t(a_t, s_t)\right)^{1-O_t} \left(\frac{1}{Z_0} \exp r_t(a_t, s_t)\right)^{O_t} \pi_0(a_t|s_t)}{\pi_{\phi}(a_t|s_t)} \\ &= \frac{1}{Z} \prod_{t=0}^{T-1} \frac{\frac{1}{Z_0} \exp r_t(a_t, s_t) \pi_0(a_t|s_t)}{\pi_{\phi}(a_t|s_t)} = \frac{1}{\hat{Z}} \exp \left[ \sum_{t=0}^{T-1} r(a_t^i, s_t^i) - \log \left( \frac{\pi_{\phi}(a_t^i|s_t^i)}{\pi_0(a_t^i|s_t^i)} \right) \right] \end{aligned}$$

Where the first line follows from cancellation of the dynamics, and the second follows from the fact that we are only considering  $O_t = 1$ . In order to avoid computation of  $Z$  explicitly, we take advantage of the consistent but biased self normalized importance weighting estimator. This means that the weights  $w_i$  are replaced with the following:

$$\hat{w}_i = \frac{w_i}{\sum_{j=1}^m w_j} = \frac{\exp \left[ \sum_{t=0}^{T-1} r(a_t^i, s_t^i) - \log \left( \frac{\pi_{\phi}(a_t^i|s_t^i)}{\pi_0(a_t^i|s_t^i)} \right) \right]}{\sum_{j=1}^m \exp \left[ \sum_{t=0}^{T-1} r(a_t^j, s_t^j) - \log \left( \frac{\pi_{\phi}(a_t^j|s_t^j)}{\pi_0(a_t^j|s_t^j)} \right) \right]}. \quad (4)$$

In this case we now have a more biased estimator of the gradient, but also one that is low variance when compared to its un-marginalized RKL counterpart,

$$\nabla_{\phi} KL(p(\tau|O)||q_{\phi}(\tau)) \approx \sum_{i=1}^n \frac{\exp \left[ \sum_{t=0}^{T-1} r(a_t^i, s_t^i) - \log \left( \frac{\pi_{\phi}(a_t^i|s_t^i)}{\pi_0(a_t^i|s_t^i)} \right) \right]}{\sum_{j=1}^n \exp \left[ \sum_{t=0}^{T-1} r(a_t^j, s_t^j) - \log \left( \frac{\pi_{\phi}(a_t^j|s_t^j)}{\pi_0(a_t^j|s_t^j)} \right) \right]} \nabla_{\phi} \log q_{\phi}(\tau_i) \quad (5)$$