VII. Appendix

1) Continuous time policy gradient: Assume parameterizing the policy by θ, According to the bellman equation we have gradient of Q w.r.t to θ as:

$$\begin{aligned}
\nabla_\theta Q(\mathbf{s},\boldsymbol{\omega},d) &= \nabla_\theta \mathop{\mathbb{E}}_{\mathbf{s}',\boldsymbol{\omega}',d'}\Big[R(\mathbf{s},\boldsymbol{\omega},d)-\beta_h+\gamma(d)(Q(\mathbf{s}',\boldsymbol{\omega}',d')+\beta_E\mathcal{H}(\pi_\theta(.,.|\mathbf{s}')))\Big]\\
&= \nabla_\theta \mathop{\mathbb{E}}_{\mathbf{s}',\boldsymbol{\omega}',d'}\Big[R(\mathbf{s},\boldsymbol{\omega},d)-\beta_h+\gamma(d)(Q(\mathbf{s}',\boldsymbol{\omega}',d')-\beta_E\log(\pi_\theta(\boldsymbol{\omega}',d'|\mathbf{s}')))\Big]\\
&\quad \text{Using reparameterization trick we have}\\
&= \mathop{\mathbb{E}}_{\mathbf{s}',\epsilon'}\Big[\gamma(d)\left(\nabla_{\boldsymbol{\omega}}Q(\mathbf{s}',\boldsymbol{\omega}',d')\nabla_\theta f_\theta^{\boldsymbol{\omega}}(\mathbf{s}',\epsilon')+\nabla_d Q(\mathbf{s}',\boldsymbol{\omega}',d')\nabla_\theta f_\theta^{d}(\mathbf{s}',\epsilon')\right)\\
&\quad +\gamma(d)\nabla_\theta Q(s',\boldsymbol{\omega}',d')-\gamma(d)\nabla_\theta\beta_E\log(\pi_\theta(\boldsymbol{\omega}',d'|\mathbf{s}'))|_{\boldsymbol{\omega}'=f_\theta^{\boldsymbol{\omega}}(\mathbf{s}',\epsilon'),d'=f_\theta^{d}(\mathbf{s}',\epsilon')}\Big]\\
&\quad \text{We can recursively replace } \nabla_\theta Q(\mathbf{s}',\boldsymbol{\omega}',d') \text{ and obtain}\\
&= \mathop{\mathbb{E}}_{\mu_\pi}\Big[\sum_{i=0}^{\infty}(\prod_{j=0}^{i}\gamma(d_j))\Big(\nabla_{\boldsymbol{\omega}}Q(\mathbf{s}_{i+1},\boldsymbol{\omega}_{i+1},d_{i+1})\nabla_\theta f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_{i+1},\epsilon_{i+1})+\nabla_d Q(\mathbf{s}_{i+1},\boldsymbol{\omega}_{i+1},d_{i+1})\nabla_\theta f_\theta^{d}(\mathbf{s}_{i+1},\epsilon_{i+1})\\
&\quad -\beta_E\nabla_\theta\log(\pi_\theta(\boldsymbol{\omega}_{i+1},d_{i+1}|\mathbf{s}_{i+1}))\Big)|_{\boldsymbol{\omega}_{i+1}=f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_{i+1},\epsilon_{i+1}),d_{i+1}=f_\theta^{d}(\mathbf{s}_{i+1},\epsilon_{i+1})}\Big]\quad \mathbf{s}_0=\mathbf{s},\boldsymbol{\omega}_0=\boldsymbol{\omega},d_0=d
\end{aligned} \tag{22}$$

Hence,

Theorem 1 (Continuous-Time Continuous-Option Policy Gradient): Consider a CT-MDP and a sampling process for $\mathbf{s}_i,\boldsymbol{\omega}_i,d_i$ as described in Section III-C. The gradient of the objctive w.r.t. to the policy parameter is

$$\begin{aligned}
\nabla_\theta J_\pi = \mathop{\mathbb{E}}_{\mu_\pi}\Big[\sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1}\gamma(d_j)\right)&\Big(\nabla_{\boldsymbol{\omega}}Q(\mathbf{s}_i,\boldsymbol{\omega}_i,d_i)\nabla_\theta f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_i,\epsilon_i)+\nabla_d Q(\mathbf{s}_i,\boldsymbol{\omega}_i,d_i)\nabla_\theta f_\theta^{d}(\mathbf{s}_i,\epsilon_i))\\
&-\beta_E\nabla_\theta\log(\pi_\theta(\boldsymbol{\omega}_i,d_i|\mathbf{s}_i))\Big)|_{\boldsymbol{\omega}_i=f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_i,\epsilon_i),d_i=f_\theta^{d}(\mathbf{s}_i,\epsilon_i)}\Big]
\end{aligned}$$

where $\gamma(d_i)=e^{-\rho d_i}$ (note that $\prod_{j=0}^{-1}\gamma(d_j)=1$).
Since the sampling variables $\boldsymbol{\omega}_i,d_i,\ldots$ are Markov, we can assume that there is a discounted stationary distribution ζ^ρ from which we can sample them i.i.d. and obtain the same result.

Proof:

$$\begin{aligned}
J_\pi &= \mathop{\mathbb{E}}_{\mu_\pi}\Big[\sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1}\gamma(d_j)\right)(R(\mathbf{s}_i,\boldsymbol{\omega}_i,d_i)-\beta_h+\mathcal{H}(\pi_\theta(.,.|\mathbf{s}_i))\Big]\\
&= \mathop{\mathbb{E}}_{\mathbf{s}_0,\boldsymbol{\omega}_0,d_0}\Big[Q(\mathbf{s}_0,\boldsymbol{\omega}_0,d_0)-\log\pi_\theta(\boldsymbol{\omega}_0,d_0|\mathbf{s}_0))\Big]\\
\nabla_\theta J_\pi &= \nabla_\theta\mathop{\mathbb{E}}_{\mathbf{s}_0,\boldsymbol{\omega}_0,d_0}\Big[Q(\mathbf{s}_0,\boldsymbol{\omega}_0,d_0)-\log\pi_\theta(\boldsymbol{\omega}_0,d_0|\mathbf{s}_0))\Big]\\
&\quad \text{By reparameterizaiton trick}\\
&= \mathop{\mathbb{E}}_{\mathbf{s}_0,\epsilon_0}\Big[\nabla_\theta Q(\mathbf{s}_0,\boldsymbol{\omega}_0,d_0)\\
&\quad +\nabla_{\boldsymbol{\omega}}Q(\mathbf{s}_0,\boldsymbol{\omega}_0,d_0)\nabla_\theta f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_0,\epsilon_0)+\nabla_d Q(\mathbf{s}_0,\boldsymbol{\omega}_0,d_0)\nabla_\theta f_\theta^{d}(\mathbf{s}_0,\epsilon_0))\\
&\quad -\beta_E\nabla_\theta\log(\pi_\theta(\boldsymbol{\omega}_0,d_0|\mathbf{s}_0))|_{\boldsymbol{\omega}_0=f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_0,\epsilon_0),d_0=f_\theta^{d}(\mathbf{s}_0,\epsilon_0)}\Big]\\
&\quad \text{Replacing } \nabla_\theta Q(\mathbf{s}_0,\boldsymbol{\omega}_0,d_0) \text{ using equation 22 we have}\\
\nabla_\theta J_\pi &= \mathbb{E}_{\mu_\pi}\Big[\sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1}\gamma(d_j)\right)\Big(\nabla_{\boldsymbol{\omega}}Q(\mathbf{s}_i,\boldsymbol{\omega}_i,d_i)\nabla_\theta f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_i,\epsilon_i)+\nabla_d Q(\mathbf{s}_i,\boldsymbol{\omega}_i,d_i)\nabla_\theta f_\theta^{d}(\mathbf{s}_i,\epsilon_i))\\
&\quad -\beta_E\nabla_\theta\log(\pi_\theta(\boldsymbol{\omega}_i,d_i|\mathbf{s}_i))\Big)|_{\boldsymbol{\omega}_i=f_\theta^{\boldsymbol{\omega}}(\mathbf{s}_i,\epsilon_i),d_i=f_\theta^{d}(\mathbf{s}_i,\epsilon_i)}\Big]
\end{aligned}$$