Supplement

Here we provide the derivations that were omitted in the main text. The following two identities will be used repeatedly. Since μ is stationary, we have

$$\textstyle\sum_x \mu\left(x, \mathbf{w}\right) \pi\left(x'|x, \mathbf{w}\right) = \mu\left(x', \mathbf{w}\right) \tag{42}$$

Since $\sum_{x'} \pi\left(x'|x, \mathbf{w}\right) = 0$ for all $\mathbf{w}$, we can differentiate and obtain

$$\textstyle\sum_{x'} \nabla_{\mathbf{w}} \pi\left(x'|x, \mathbf{w}\right) = 0 \tag{43}$$

Again, we will suppress the dependence on $\mathbf{w}$.

Proof of Theorem 1:

Differentiating the Bellman equation (2) yields

$$\begin{aligned} \nabla_{\mathbf{w}} c + \nabla_{\mathbf{w}} v\left(x\right) &= \textstyle\sum_{x'} \nabla_{\mathbf{w}} \pi\left(x'|x\right) \left(\log \dfrac{\pi\left(x'|x\right)}{p\left(x'|x\right)} + v\left(x'\right)\right) \\ &\quad + \pi\left(x'|x\right) \left(\dfrac{\nabla_{\mathbf{w}} \pi\left(x'|x\right)}{\pi\left(x'|x\right)} + \nabla_{\mathbf{w}} v\left(x'\right)\right) \\ &= \textstyle\sum_{x'} \nabla_{\mathbf{w}} \pi\left(x'|x\right) \left(\log \dfrac{\pi\left(x'|x\right)}{p\left(x'|x\right)} + v\left(x'\right)\right) + \pi\left(x'|x, \mathbf{w}\right) \nabla_{\mathbf{w}} v\left(x'\right) \end{aligned} \tag{44}$$

To obtain the last equation we used (43). Now we move $\nabla_{\mathbf{w}} v\left(x\right)$ on the right side of (44), multiply by $\mu\left(x\right)$ and sum over x. Noting that

$$\textstyle\sum_{x,x'} \mu\left(x\right) \pi\left(x'|x\right) \nabla_{\mathbf{w}} v\left(x'\right) = \sum_x \mu\left(x\right) \nabla_{\mathbf{w}} v\left(x\right) \tag{45}$$

which follows from (42), the LMDP policy gradient is as given in (5).

Proof of Theorem 2:

Using the identity $\nabla_{\mathbf{w}} \pi = \pi \nabla_{\mathbf{w}} \log \pi$, equation (5) can also be written as

$$\nabla_{\mathbf{w}} c = \textstyle\sum_{x,x'} \mu\left(x, x'\right) \nabla_{\mathbf{w}} \log \pi\left(x'|x\right) \left(\log \dfrac{\pi\left(x'|x\right)}{p\left(x'|x\right)} + v\left(x'\right)\right) \tag{46}$$

With the policy parameterization (7), we have

$$\begin{aligned} \nabla_{\mathbf{w}} \log \pi\left(x'|x\right) &= \nabla_{\mathbf{w}} \left(\log p\left(x'|x\right) - \mathbf{w}^{\mathsf{T}} \mathbf{f}\left(x'\right) - \log \textstyle\sum_y p\left(y|x\right) \exp\left(-\mathbf{w}^{\mathsf{T}} \mathbf{f}\left(y\right)\right)\right) \\ &= -\mathbf{f}\left(x'\right) + \textstyle\sum_y \dfrac{p\left(y|x\right) \exp\left(-\mathbf{w}^{\mathsf{T}} \mathbf{f}\left(y\right)\right)}{\sum_s p\left(s|x\right) \exp\left(-\mathbf{w}^{\mathsf{T}} \mathbf{f}\left(s\right)\right)} \mathbf{f}\left(y\right) \\ &= \Pi\left[\mathbf{f}\right]\left(x\right) - \mathbf{f}\left(x'\right) \end{aligned} \tag{47}$$

Substituting (47) in (46) and using the fact that

$$\textstyle\sum_x \mu\left(x\right) \log\left(\sum_y p\left(y|x\right) \exp\left(-\mathbf{w}^{\mathsf{T}} \mathbf{f}\left(y\right)\right)\right) \sum_{x'} \nabla_{\mathbf{w}} \pi\left(x'|x\right) = 0 \tag{48}$$

which follows from (43), the gradient is as given in (9).

Proof of Theorem 4:

Using (47), equation (17) can be written as

$$\begin{aligned} \nabla_{\mathbf{w}} c &= \textstyle\sum_{x,x'} \mu\left(x, x'\right) \nabla_{\mathbf{w}} \log \pi\left(x'|x\right) \left(\nabla_{\mathbf{w}} \log \pi\left(x'|x\right) - \Pi\left[\mathbf{f}\right]\left(x\right)\right)^{\mathsf{T}} \left(\mathbf{w} - \mathbf{r}\right) \\ &= G\left(\mathbf{w}\right) \left(\mathbf{w} - \mathbf{r}\right) - \textstyle\sum_{x,x'} \mu^{\pi}\left(x\right) \nabla_{\mathbf{w}} \pi\left(x'|x\right) \Pi\left[\mathbf{f}\right]\left(x\right)^{\mathsf{T}} \left(\mathbf{w} - \mathbf{r}\right) \end{aligned} \tag{49}$$

The second term is zero because of (43), thus we have equation (19).