A Supplementary material for T. Osogami, "Robustness and risk-sensitivity in Markov decision processes," in *Advances in Neural Information Processing Systems*, 25, 2012

A.1 Proof of Theorem 5

By the translation invariance and the recursiveness of ERM, we have

$$\begin{aligned}
\mathsf{ERM}_\gamma\left[\tilde{L}(\pi)\right] &= \mathsf{ERM}_\gamma\left[\sum_{i=0}^{N-1}\mathsf{ERM}_\gamma[C_i]\right] && (40)\\
&= \mathsf{ERM}_\gamma[C_0] + \mathsf{ERM}_\gamma\left[\sum_{i=1}^{N-1}\mathsf{ERM}_\gamma[C_i]\right] && (41)\\
&= \mathsf{ERM}_\gamma[C_0] + \mathsf{ERM}_\gamma\left[\mathsf{ERM}_\gamma\left[\sum_{i=1}^{N-1}\mathsf{ERM}_\gamma[C_i] \mid S_1\right]\right]. && (42)
\end{aligned}$$

In the right-hand side of (40), the inner ERM_γ is with respect to $f^\pi(\cdot|S_i)$; the outer is with respect to p^π. In the right-hand side of (42), the first ERM_γ is with respect to $f^\pi(\cdot|S_0)$; the second is with respect to $p^\pi(\cdot|S_0)$; the third is with respect to $p^\pi(\cdot|S_i)$ for $i>0$; the last is with respect to $f^\pi(\cdot|S_i)$. By (34), the first term of (42) can be represented as follows:

$$\mathsf{ERM}_\gamma[C_0] = \max_{f^\pi(\cdot|S_0)\in\mathcal{P}(f_0^\pi(\cdot|S_0))}\left\{\mathsf{E}_{f^\pi(\cdot|S_0)}[C_0] - \gamma\, H\left(f^\pi(\cdot|S_0)||f_0^\pi(\cdot|S_0)\right)\right\}. \quad (43)$$

Analogously, the second term of (42) can be represented as follows:

$$\begin{aligned}
&\mathsf{ERM}_\gamma\left[\mathsf{ERM}_\gamma\left[\sum_{i=1}^{N-1}\mathsf{ERM}_\gamma[C_i] \mid S_1\right]\right]\\
&= \max_{p^\pi(\cdot|S_0)\in\mathcal{P}(p_0^\pi(\cdot|S_0))}\left\{\mathsf{E}_{p^\pi(\cdot|S_0)}\left[\mathsf{ERM}_\gamma\left[\sum_{i=1}^{N-1}\mathsf{ERM}_\gamma[C_i] \mid S_1\right]\right] - \gamma\, H\left(p^\pi(\cdot|S_0)||p_0^\pi(\cdot|S_0)\right)\right\} \quad (44)
\end{aligned}$$

Applying the above argument to the term inside $\mathsf{E}_{p^\pi(\cdot|S_0)}$, we obtain, after simplification, that

$$\begin{aligned}
&\mathsf{ERM}_\gamma\left[\tilde{L}(\pi)\right]\\
&= \max_{\substack{p^\pi\in\mathcal{P}_1(p_0^\pi)\\ f^\pi\in\mathcal{P}_1(f_0^\pi)}}\left\{\mathsf{E}_{p^\pi,f^\pi}\left[\sum_{i=0}^{1}C_i\right]\right.\\
&\qquad \left.+\mathsf{E}_{p^\pi}\left[\mathsf{ERM}_\gamma\left[\sum_{i=1}^{N-1}\mathsf{ERM}_\gamma[C_i \mid S_1]\right] - \gamma\sum_{i=0}^{1}\left(H\left(f^\pi(\cdot|S_i)||f_0^\pi(\cdot|S_i)\right) + H\left(p^\pi(\cdot|S_i)||p_0^\pi(\cdot|S_i)\right)\right)\right]\right\}, \quad (45)
\end{aligned}$$

where $p^\pi \in \mathcal{P}_\ell(p_0^\pi)$ denotes that $p^\pi(\cdot|s_i) \in \mathcal{P}(p_0^\pi(\cdot|s_i)), \forall s_i \in \mathcal{S}_i, i=0,\ldots,\ell$. In establishing (45), we exchange the expectation and the max operator, because expectation is monotonic (i.e., $\mathsf{E}[Y] \geq \mathsf{E}[Z]$ if Y stochastically dominates Z).

Then we recursively apply the above process to the expression

$$\mathsf{ERM}_\gamma\left[\sum_{i=\ell}^{N-1}\mathsf{ERM}_\gamma[C_i \mid S_\ell]\right] \quad (46)$$

for $\ell = 1,\ldots,N-1$ to complete the proof of the theorem. In (39), notice that the summation of $H(p^\pi(\cdot|S_i)||p_0^\pi(\cdot|S_i))$ is from $i=0$ to $i=N-2$, while $H(f^\pi(\cdot|S_i)||f_0^\pi(\cdot|S_i))$ is summed up to $i=N-1$. This is because, in the last step of the recursion, we have

$$\mathsf{ERM}_\gamma\left[\mathsf{ERM}_\gamma[C_{N-1} \mid S_{N-1}]\right] \quad (47)$$

$$= \quad \mathsf{ERM}_\gamma[C_{N-1} \mid S_{N-1}] \tag{48}$$

$$= \quad \max_{f^\pi(\cdot|S_{N-1}) \in \mathcal{P}(f_0^\pi(\cdot|S_{N-1}))} \left\{ \mathsf{E}_{f^\pi(\cdot|S_{N-1})}[C_{N-1} \mid S_{N-1}] - \gamma \, H\left(f^\pi(\cdot|S_{N-1}) || f_0^\pi(\cdot|S_{N-1})\right) \right\}. \tag{49}$$

In the right-hand side of (47), the outer ERM_γ is with respect to $p^\pi(\cdot|S_{N-1})$. Because $\mathsf{ERM}_\gamma[C_{N-1} \mid S_{N-1}]$ is independent of S_N, we obtain (48), where ERM_γ is with respect to $f^\pi(\cdot|S_{N-1})$.

A.2 Remarks

A.2.1 Continuous cost

Throughout the paper, we assumed that the cost is discrete for simplicity, but all of our results carry over to the case where the cost is continuous and its support is infinite by simply making both of the following changes:

- replace "probability mass function" with "probability density function," and
- replace "$\sum$" with "$\int dx$."

In particular, (34) holds for the continuous case.

If the changes look nontrivial around (25)-(29), notice that, by letting

$$F(x|s,a) \quad = \quad \int_{-\infty}^{x} f(y|s,a)\, dy \tag{50}$$

and following the steps shown for the discrete case, we have

$$\max_{\mathbf{f} \in \mathcal{U}_f} \int_{x \in \mathcal{X}(s,a)} x \, f(x|s,a)\, dx \quad = \quad -\int_{-\infty}^{x} x \, dg(1 - F(x|s,a)) \tag{51}$$

$$= \quad \int x \int_{1-F(x|s,a)}^{1} \frac{1}{u}\, dH(u)\, dF(x|s,a), \tag{52}$$

where the first equality follows from (24), and the second one follows from (20). Hence, we have

$$\max_{\mathbf{f} \in \mathcal{U}_f} \int_{x \in \mathcal{X}(s,a)} x \, f(x|s,a)\, dx \quad = \quad \int_0^1 \frac{1}{u} \int_{F^{-1}(1-s)}^{\infty} x \, dF(x|s,a)\, dH(u), \tag{53}$$

which gives (29).

A.2.2 Implications with respect to efficient algorithms

Existing algorithms such as those in [11] for robust MDPs iteratively solve "inner" optimization problems to consider the worst case. The knowledge of a relation to a risk-sensitive MDP allows us to essentially replace the steps of solving "inner" optimization problems with calculation of the value of risk measures. Notice that the right-hand side of (34), i.e. the value of the optimal objective value, would be very hard to calculate without the knowledge of the equivalence to its left-hand side. The knowledge of the relations between risk-sensitive MDPs and robust MDPs allows us to identify a class of robust MDPs whose optimal policy can be found without explicitly solving the "inner" optimization problems.