Supplementary Material for Probability Functional Descent

A. Proofs and Computations

Lemma 1. *Let $J : \mathcal{P}(X) \to \mathbb{R}$. Then $\Psi : X \to \mathbb{R}$ is an influence function of J at μ if and only if*

$$\frac{d}{d\epsilon} J(\mu + \epsilon\chi)\Big|_{\epsilon=0^+} = \int_X \Psi(x)\,\chi(dx).$$

Proof. The left-hand side equals (1), which equals (2). □

Theorem 1 (Chain rule). *Let $J : \mathcal{P}(X) \to \mathbb{R}$ be continuously differentiable, in the sense that the influence function Ψ_μ exists and $(\mu, \nu) \mapsto \mathbb{E}_{x\sim\nu}[\Psi_\mu(x)]$ is continuous. Let the parameterization $\theta \mapsto \mu_\theta$ be differentiable, in the sense that $\frac{1}{||h||}(\mu_{\theta+h} - \mu_\theta)$ converges to a weak limit as $h \to 0$. Then*

$$\nabla_\theta J(\mu_\theta) = \nabla_\theta \mathbb{E}_{x\sim\mu_\theta}[\hat{\Psi}(x)],$$

where $\hat{\Psi} = \Psi_{\mu_\theta}$ is treated as a function $X \to \mathbb{R}$ that is not dependent on θ.

Proof. Without loss of generality, assume $\theta \in \mathbb{R}$, as the gradient is simply a vector of one-dimensional derivatives. Let $\chi_\epsilon = \frac{1}{\epsilon}(\mu_{\theta+\epsilon} - \mu_\theta)$, and let $\chi = \lim_{\epsilon\to 0}\chi_\epsilon$ (weakly). Then

$$\begin{aligned}\frac{d}{d\theta} J(\mu_\theta) &= \frac{d}{d\epsilon} J(\mu_{\theta+\epsilon})\Big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon} J(\mu_\theta + \epsilon\chi_\epsilon)\Big|_{\epsilon=0}.\end{aligned}$$

Assuming for now that

$$\frac{d}{d\epsilon} J(\mu_\theta + \epsilon\chi_\epsilon)\Big|_{\epsilon=0} = \frac{d}{d\epsilon} J(\mu_\theta + \epsilon\chi)\Big|_{\epsilon=0},$$

we have by Lemma 1 that

$$\begin{aligned}\frac{d}{d\theta} J(\mu_\theta) &= \int_X \hat{\Psi}\,d\chi \\ &= \int_X \hat{\Psi}\,d\Big(\lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mu_{\theta+\epsilon} - \mu_\theta)\Big) \\ &= \lim_{\epsilon\to 0}\int_X \hat{\Psi}\,d\Big(\frac{1}{\epsilon}(\mu_{\theta+\epsilon} - \mu_\theta)\Big) \\ &= \frac{d}{d\theta}\int_X \hat{\Psi}\,d\mu_\theta,\end{aligned}$$

where the interchange of limits is by the definition of weak convergence (recall we assumed that X is compact, so $\hat{\Psi}$ is continuous and bounded by virtue of being continuous).

The equality we assumed is the definition of a stronger notion of differentiability called Hadamard differentiability of J. Our conditions imply Hadamard differentiability via Proposition 2.33 of Penot (2012), noting that the map $(\mu, \chi) \mapsto \int_X \Psi_\mu\,d\chi$ is continuous by assumption. □

Theorem 2 (Fenchel–Moreau representation). *Let $J : \mathcal{M}(X) \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous. Then the maximizer of $\varphi \mapsto \mathbb{E}_{x\sim\mu}[\varphi(x)] - J^\star(\varphi)$, if it exists, is an influence function for J at μ. With some abuse of notation, we have that*

$$\Psi_\mu = \operatorname*{arg\,max}_{\varphi\in\mathcal{C}(X)} \Big[\mathbb{E}_{x\sim\mu}[\varphi(x)] - J^\star(\varphi)\Big].$$

Proof. We will exploit the Fenchel–Moreau theorem, which applies in the setting of locally convex, Hausdorff topological vector spaces (see e.g. Zalinescu (2002)). The space we consider is $\mathcal{M}(X)$, the space of signed, finite measures equipped with the topology of weak convergence, of which $\mathcal{P}(X)$ is a convex subset. $\mathcal{M}(X)$ is indeed locally convex and Hausdorff, and its dual space is $\mathcal{C}(X)$ (see e.g. Aliprantis & Border (2006), section 5.14).

We now show that a maximizer φ^* is an influence function. By the Fenchel–Moreau theorem,

$$J(\mu) = J^{\star\star}(\mu) = \sup_{\varphi\in\mathcal{C}(X)} \Big[\int_X \varphi\,d\mu - J^\star(\varphi)\Big],$$

and

$$J(\mu + \epsilon\chi) = \sup_{\varphi\in\mathcal{C}(X)} \Big[\int_X \varphi\,d\mu + \epsilon\int_X \varphi\,d\chi - J^\star(\varphi)\Big].$$

Because J is differentiable, $\epsilon \mapsto J(\mu + \epsilon\chi)$ is differentiable, so by the envelope theorem (Milgrom & Segal, 2002),

$$\frac{d}{d\epsilon} J(\mu + \epsilon\chi)\Big|_{\epsilon=0} = \int_X \varphi^*\,d\chi,$$

so that φ^* is an influence function by Lemma 1.

The abuse of notation stems from the fact that not all influence functions are maximizers. This is true, though, if

$J(\mu) = \infty$ if $\mu \notin \mathcal{P}(X)$:

$$\begin{aligned}
&\int_X \Psi_\mu \, d\mu - J^\star(\Psi_\mu) \\
&\quad = \int_X \Psi_\mu \, d\mu - \sup_{\nu \in \mathcal{P}(X)} \Big[\int_X \Psi_\mu \, d\nu - J(\nu) \Big] \\
&\quad = \inf_{\nu \in \mathcal{P}(X)} \Big[- \int_X \Psi_\mu \, d(\nu - \mu) + J(\nu) \Big] \\
&\quad = \inf_{\nu \in \mathcal{P}(X)} \Big[- \frac{d}{d\epsilon} J(\mu + \epsilon(\nu - \mu)) \Big|_{\epsilon=0} + J(\nu) \Big] \\
&\quad \geq J(\mu),
\end{aligned}$$

since the convex function $f(\epsilon) = J(\mu + \epsilon(\nu - \mu))$ lies above its tangent line:

$$f(1) \geq f(0) + 1 \cdot f'(0).$$

Since $J(\mu) = J^{\star\star}(\mu)$, we have that

$$\int_X \Psi_\mu \, d\mu - J^\star(\Psi_\mu) \geq \sup_{\varphi \in \mathcal{C}(X)} \Big[\int_X \varphi \, d\mu - J^\star(\varphi) \Big].$$

□

The following lemma will come in handy in our computations.

Lemma 2. *Suppose $J : \mathcal{M}(X) \to \overline{\mathbb{R}}$ has a representation*

$$J(\mu) = \sup_{\varphi \in \mathcal{C}(X)} \Big[\int_X \varphi \, d\mu - K(\varphi) \Big],$$

where $K : \mathcal{C}(X) \to \overline{\mathbb{R}}$ is proper, convex, and lower semicontinuous. Then $J^\star = K$.

Proof. By definition of the convex conjugate, $J = K^\star$. Then $J^\star = K^{\star\star} = K$, by the Fenchel–Moreau theorem. □

We note that when applying this lemma, we will often implicitly define the appropriate extension of J to $\mathcal{M}(X)$ to be $J(\mu) = \sup_{\varphi \in \mathcal{C}(X)} [\int \varphi \, d\mu - K(\varphi)]$. The exact choice of extension can certainly affect the exact form of the convex conjugate; see Ruderman et al. (2012) for one example of this phenomenon.

Proposition 2. *Suppose μ has density $p(x)$ and ν has density $q(x)$. Then the influence function for J_{JS} is*

$$\Psi_{\mathrm{JS}}(x) = \frac{1}{2} \log \frac{p(x)}{p(x) + q(x)}.$$

Proof. The result follows from Lemma 1:

$$\begin{aligned}
&\frac{d}{d\epsilon} J_{\mathrm{JS}}(\mu + \epsilon\chi) \Big|_{\epsilon=0} \\
&\quad = \frac{1}{2} \int_X \frac{d}{d\epsilon} \Big[(p + \epsilon\chi) \log \frac{p + \epsilon\chi}{\frac{1}{2}(p + \epsilon\chi) + \frac{1}{2}q} \\
&\qquad\qquad + q \log \frac{q}{\frac{1}{2}(p + \epsilon\chi) + \frac{1}{2}q} \Big]_{\epsilon=0} dx \\
&\quad = \frac{1}{2} \int_X \Big[\log \frac{p}{\frac{1}{2}p + \frac{1}{2}q} + 1 - \frac{p}{p+q} - \frac{q}{p+q} \Big] \chi \, dx \\
&\quad = \frac{1}{2} \int_X \Big[\log \frac{p}{p+q} + \log 2 \Big] \chi \, dx.
\end{aligned}$$

□

Proposition 3. *The convex conjugate of J_{JS} is*

$$J^\star_{\mathrm{JS}}(\varphi) = -\tfrac{1}{2} \mathbb{E}_{x \sim \nu} [\log(1 - e^{2\varphi(x) + \log 2})] - \tfrac{1}{2} \log 2.$$

Proof.

$$\begin{aligned}
J^\star_{\mathrm{JS}}(\varphi) &= \sup_{\mu \in \mathcal{M}(X)} \Big[\int_X \varphi \, d\mu - J_{\mathrm{JS}}(\mu) \Big] \\
&= \sup_p \int_X \Big[\varphi p - \frac{1}{2} p \log \frac{p}{\frac{1}{2}p + \frac{1}{2}q} - \frac{1}{2} q \log \frac{q}{\frac{1}{2}p + \frac{1}{2}q} \Big] \, dx.
\end{aligned}$$

Setting the integrand's derivative w.r.t. p to 0, we find that pointwise, the optimal p satisfies

$$\varphi = \frac{1}{2} \log \frac{p}{\frac{1}{2}p + \frac{1}{2}q}.$$

We eliminate p in the integrand. Notice that the first two terms in the integrand cancel after plugging in p. Since

$$\frac{q}{\frac{1}{2}p + \frac{1}{2}q} = 2\Big(1 - \frac{p}{p+q}\Big) = 2(1 - 2e^{2\varphi}),$$

we obtain that

$$J^\star_{\mathrm{JS}}(\varphi) = -\frac{1}{2} \int_X q \log(1 - 2e^{2\varphi}) \, dx - \frac{1}{2} \log 2.$$

□

Proposition 5. *Suppose μ has density $p(x)$ and ν has density $q(x)$. The influence function for J_{NS} is*

$$\Psi_{\mathrm{NS}}(x) = \log \frac{p(x)}{q(x)}.$$

Proof. The result follows from Lemma 1:

$$\begin{aligned}
&\frac{d}{d\epsilon} J_{\mathrm{NS}}(\mu + \epsilon\chi)\Big|_{\epsilon=0} \\
&\quad = \frac{d}{d\epsilon} \int_X (p + \epsilon\chi) \log \frac{p + \epsilon\chi}{q}\, dx\Big|_{\epsilon=0} \\
&\quad = \int_X \left[\chi \log \frac{p}{q} + \chi\right] dx \\
&\quad = \int_X \left[\log \frac{p}{q} + 1\right] d\chi \\
&\quad = \int_X \left[\log \frac{p}{q}\right] d\chi.
\end{aligned}$$

□

Proposition 7. *The influence function for J_{W} is the Kantorovich potential corresponding to the optimal transport from μ to ν.*

Proof. See Santambrogio (2015), Proposition 7.17. □

Proposition 8. *The convex conjugate of J_{W} is*

$$J_{\mathrm{W}}^{\star}(\varphi) = \mathbb{E}_{x\sim\nu}[\varphi(x)] + \{||\varphi||_L \leq 1\}.$$

Proof. Using Kantorovich–Rubinstein duality, we have that

$$\begin{aligned}
J_{\mathrm{W}}(\mu) &= \sup_{||\varphi||_L \leq 1} \left[\int_X \varphi\, d\mu - \int_X \varphi\, d\nu\right] \\
&= \sup_{\varphi} \left[\int_X \varphi\, d\mu - \int_X \varphi\, d\nu - \{||\varphi||_L \leq 1\}\right],
\end{aligned}$$

where we use the notation

$$\{A\} = \begin{cases} 0 & A \text{ is true,} \\ \infty & A \text{ is false.} \end{cases}$$

By Lemma 2,

$$J_{\mathrm{W}}^{\star}(\varphi) = \int_X \varphi\, d\nu + \{||\varphi||_L \leq 1\}.$$

□

Proposition 10. *The influence function for J_{VI} is*

$$\Psi_{\mathrm{VI}}(z) = \log \frac{q(z)}{p(x|z)p(z)}.$$

Proof. The result follows from Lemma 1:

$$\begin{aligned}
&\frac{d}{d\epsilon} J_{\mathrm{VI}}(q + \epsilon\chi)\Big|_{\epsilon=0} \\
&\quad = \frac{d}{d\epsilon} \int (q(z) + \epsilon\chi(z)) \log \frac{q(z) + \epsilon\chi(z)}{p(z|x)}\, dz\Big|_{\epsilon=0} \\
&\quad = \int \left[\chi(z) \log \frac{q(z) + \epsilon\chi(z)}{p(z|x)} + \chi(z)\right] dz\Big|_{\epsilon=0} \\
&\quad = \int \left[\log \frac{q(z)}{p(z|x)} + 1\right] \chi(z)\, dz \\
&\quad = \int \left[\log \frac{q(z)}{p(x|z)p(z)} + \log p(x) + 1\right] \chi(z)\, dz \\
&\quad = \int \log \frac{q(z)}{p(x|z)p(z)}\, \chi(z)\, dz.
\end{aligned}$$

□

Proofs continue on the following page.

Proposition 13. *The influence function for J_{RL} is*

$$\Psi_{\mathrm{RL}}(s,a) = -\frac{\sum_{t=0}^{\infty}\gamma^t p_t^\pi(s)}{\pi(s)}(Q^\pi(s,a) - V^\pi(s)),$$

where Q^π is the state-action value function, V^π is the state value function, and p_t^π is the marginal distribution of states after t steps, all under the policy π.

Proof. First, we note that

$$\begin{aligned} &\frac{d}{d\epsilon}(\pi+\epsilon\chi)(a|s)\Big|_{\epsilon=0} \\ &\quad = \frac{d}{d\epsilon}\frac{\pi(a,s)+\epsilon\chi(s,a)}{\pi(s)+\epsilon\chi(s)}\Big|_{\epsilon=0} \\ &\quad = \frac{\chi(s,a)-\chi(s)\pi(a|s)}{\pi(s)}, \end{aligned}$$

where we abuse notation to denote $\chi(s) = \int \chi(s,a')\,da'$.

We have

$$-J_{\mathrm{RL}} = \mathbb{E}\Big[\sum_{t=1}^{\infty}\gamma^{t-1}r_t\Big],$$

or, plugging in the measure,

$$-J_{\mathrm{RL}} = \int \sum_{t=1}^{\infty}\gamma^{t-1}r_t\, p_0(s_0)\prod_{j=1}^{\infty}p(s_j,r_j|s_{j-1},a_j)\prod_{k=1}^{\infty}\pi(a_k|s_{k-1}).$$

The integral is over all free variables; we omit them here and in the following derivation for conciseness.

In computing $\frac{d}{d\epsilon}J_{\mathrm{RL}}(\pi+\epsilon\chi)|_{\epsilon=0}$, the product rule dictates that a term appear for every k, in which $\pi(a_k|s_{k-1})$ is replaced with $\frac{d}{d\epsilon}(\pi+\epsilon\chi)(a_k|s_{k-1})|_{\epsilon=0}$. Hence:

$$\begin{aligned} &-\frac{d}{d\epsilon}J_{\mathrm{RL}}(\pi+\epsilon\chi)\Big|_{\epsilon=0} \\ &\quad = \int \sum_{t=1}^{\infty}\gamma^{t-1}r_t\, p_0(s_0)\prod_{j=1}^{\infty}p(s_j,r_j|s_{j-1},a_j) \\ &\qquad \times \sum_{k=1}^{\infty}\frac{\chi(s_{k-1},a_k)-\chi(s_{k-1})\pi(a_k|s_{k-1})}{\pi(s_{k-1})}\prod_{\substack{\ell=1\\ \ell\neq k}}^{\infty}\pi(a_\ell|s_{\ell-1}) \\ &\quad = \sum_{k=1}^{\infty}\int \sum_{t=1}^{\infty}\gamma^{t-1}r_t\, p_0(s_0)\prod_{j=1}^{\infty}p(s_j,r_j|s_{j-1},a_j) \\ &\qquad \times \frac{\chi(s_{k-1},a_k)-\chi(s_{k-1})\pi(a_k|s_{k-1})}{\pi(s_{k-1})}\prod_{\substack{\ell=1\\ \ell\neq k}}^{\infty}\pi(a_\ell|s_{\ell-1}), \end{aligned}$$

reordering the summations. Note that for $t < k$, the summand vanishes:

$$\begin{aligned}
&\int \prod_{j=k}^{\infty} p(s_j, r_j | s_{j-1}, a_j) \\
&\qquad \times \left(\chi(s_{k-1}, a_k) - \chi(s_{k-1})\pi(a_k|s_{k-1})\right) \prod_{\ell=k+1}^{\infty} \pi(a_\ell|s_{\ell-1}) \\
&= \int \left(\chi(s_{k-1}, a_k) - \chi(s_{k-1})\pi(a_k|s_{k-1})\right) \\
&= \int \left(\chi(s_{k-1}) - \chi(s_{k-1})\right) \\
&= 0,
\end{aligned}$$

since all the variables $a_k, r_k, s_k, a_{k+1}, r_{k+1}, s_{k+1}, \ldots$ integrate away to 1. This yields:

$$\begin{aligned}
&-\frac{d}{d\epsilon} J_{\mathrm{RL}}(\pi + \epsilon\chi)\Big|_{\epsilon=0} \\
&\quad = \sum_{k=1}^{\infty} \int \sum_{t=k}^{\infty} \gamma^{t-1} r_t \, p_0(s_0) \prod_{j=1}^{\infty} p(s_j, r_j | s_{j-1}, a_j) \\
&\qquad \times \frac{\chi(s_{k-1}, a_k) - \chi(s_{k-1})\pi(a_k|s_{k-1})}{\pi(s_{k-1})} \prod_{\substack{\ell=1 \\ \ell \neq k}}^{\infty} \pi(a_\ell|s_{\ell-1}).
\end{aligned}$$

Then, substituting the marginal distribution (note s_{k-1} is not integrated)

$$p_{k-1}^{\pi}(s_{k-1}) = \int \prod_{j=1}^{k-1} p(s_j, r_j | s_{j-1}, a_j) \prod_{\ell=1}^{k-1} \pi(a_\ell|s_{\ell-1}),$$

we obtain

$$\begin{aligned}
&-\frac{d}{d\epsilon} J_{\mathrm{RL}}(\pi + \epsilon\chi)\Big|_{\epsilon=0} \\
&\quad = \sum_{k=1}^{\infty} \int \sum_{t=k}^{\infty} \gamma^{t-1} r_t \, p_{k-1}^{\pi}(s_{k-1}) \prod_{j=k}^{\infty} p(s_j, r_j | s_{j-1}, a_j) \\
&\qquad \times \frac{\chi(s_{k-1}, a_k) - \chi(s_{k-1})\pi(a_k|s_{k-1})}{\pi(s_{k-1})} \prod_{\ell=k+1}^{\infty} \pi(a_\ell|s_{\ell-1}).
\end{aligned}$$

Let us rename the integration variables by decreasing their indices by $k-1$:

$$\begin{aligned}
&-\frac{d}{d\epsilon} J_{\mathrm{RL}}(\pi + \epsilon\chi)\Big|_{\epsilon=0} \\
&\quad = \sum_{k=1}^{\infty} \int \sum_{t=1}^{\infty} \gamma^{t+k-2} r_t \, p_{k-1}^{\pi}(s_0) \prod_{j=1}^{\infty} p(s_j, r_j | s_{j-1}, a_j) \\
&\qquad \times \frac{\chi(s_0, a_1) - \chi(s_0)\pi(a_1|s_0)}{\pi(s_0)} \prod_{\ell=2}^{\infty} \pi(a_\ell|s_{\ell-1}).
\end{aligned}$$

Substituting in

$$V^{\pi}(s_0) = \int \sum_{t=1}^{\infty} \gamma^{t-1} r_t \prod_{j=1}^{\infty} p(s_j, r_j | s_{j-1}, a_j) \prod_{\ell=1}^{\infty} \pi(a_\ell|s_{\ell-1}),$$

$$Q^{\pi}(s_0, a_1) = \int \sum_{t=1}^{\infty} \gamma^{t-1} r_t \prod_{j=1}^{\infty} p(s_j, r_j | s_{j-1}, a_j) \prod_{\ell=2}^{\infty} \pi(a_\ell|s_{\ell-1}),$$

we obtain

$$-\frac{d}{d\epsilon}J_{\mathrm{RL}}(\pi+\epsilon\chi)\Big|_{\epsilon=0} = \sum_{k=1}^{\infty}\int \gamma^{k-1}\,p_{k-1}^{\pi}(s_0)\frac{Q^{\pi}(s_0,a_1)\chi(s_0,a_1)-V^{\pi}(s_0)\chi(s_0)}{\pi(s_0)}.$$

Finally, by Lemma 1, we obtain that

$$\Psi_{\mathrm{RL}}(s,a) = -\frac{\sum_{k=0}^{\infty}\gamma^k p_k^{\pi}(s)}{\pi(s)}(Q^{\pi}(s,a)-V^{\pi}(s)).$$

□

Proposition 16. *The convex conjugate of J_{RL} is*

$$J_{\mathrm{RL}}^{\star}(\varphi) = (1-\gamma)\mathbb{E}_{p_0(s)}V_{\varphi}(s) + \{V_{\varphi}\text{ exists}\},$$

where V_{φ} is the unique solution to $\varphi = -\mathcal{A}V_{\varphi}$, if it exists.

Proof. As mentioned in the text, we set the arbitrary distribution $\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t^{\pi}(s)$. In doing so, $\pi(s,a)$ becomes a state-action *occupancy measure* that describes the frequency of encounters of the state-action pair (s,a) over trajectories governed by the policy $\pi(a|s)$. It is known that there is a bijection between occupancy measures $\pi(s,a)$ and policies $\pi(a|s)$ (Syed et al., 2008; Ho & Ermon, 2016).

We can enforce this setting by redefining

$$J_{\mathrm{RL}}(\pi) = -\mathbb{E}\sum_{t=1}^{\infty}\gamma^{t-1}r_t + \Big\{\forall s : \pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t^{\pi}(s)\Big\},$$

where again $\{\cdot\}$ is the convex indicator function. This equation can be rewritten as

$$J_{\mathrm{RL}}(\pi) = -\mathbb{E}_{\pi(s,a)}R(s,a) + \Big\{\forall s' :\ \pi(s') = (1-\gamma)p_0(s') + \gamma\mathbb{E}_{\pi(s,a)}p(s'|s,a)\Big\},$$

where $R(s,a) = \mathbb{E}_{p(s',r|s,a)}[r]$. The constraint is known as the *Bellman flow equation*. This formulation is convex, as it is the sum of an affine function and an indicator of a convex set (indeed, an affine subspace).

We recall $-\varphi = \mathcal{A}V_{\varphi}$, where $\mathcal{A}V(s,a) = \mathbb{E}_{p(s',r|s,a)}[r+\gamma V(s')] - V(s)$. Now, V_{φ} is uniquely defined by φ if a solution to the equation exists. To see this, note that V_{φ} is the fixed point of the Bellman operator $\mathcal{T}^a$ defined by

$$(\mathcal{T}^a V)(s) = (R+\varphi)(s,a) + \gamma\mathbb{E}_{p(s'|s,a)}V(s'),$$

which is contractive and therefore has a unique fixed point. A representation of V_{φ} may be obtained via fixed point iteration using $\mathcal{T}^a$ for an arbitrary action a:

$$V_{\varphi}(s) = \lim_{k\to\infty}(\mathcal{T}^a)^k 0 = \mathbb{E}^a\sum_{t=1}^{\infty}\gamma^{t-1}(R+\varphi)(s_t,a),$$

where the expectation is taken under the deterministic policy a.

We rewrite J_{RL} using a Lagrange multiplier $V(s)$

$$\begin{aligned}
J_{\mathrm{RL}}(\pi) &= -\mathbb{E}_{\pi(s,a)}R(s,a) + \sup_V \int V(s')\Big[\pi(s') - (1-\gamma)p_0(s') - \gamma\mathbb{E}_{\pi(s,a)}p(s'|s,a)\Big]\,ds' \\
&= \sup_V -\mathbb{E}_{\pi(s,a)}R(s,a) + \mathbb{E}_{\pi(s)}V(s) - (1-\gamma)\mathbb{E}_{p_0(s)}V(s) - \gamma\mathbb{E}_{\pi(s,a)}\mathbb{E}_{p(s'|s,a)}V(s') \\
&= \sup_{\varphi}\mathbb{E}_{\pi(s,a)}\varphi(s,a) - (1-\gamma)\mathbb{E}_{p_0(s)}V_{\varphi}(s) - \{V_{\varphi}\text{ exists}\}.
\end{aligned}$$

Note that $(1-\gamma)\mathbb{E}_{p_0(s)}V_\varphi(s) + \{V_\varphi \text{ exists}\}$ is convex in φ; this stems from the fact that

$$V_{\alpha\varphi+(1-\alpha)\varphi'} = \alpha V_\varphi + (1-\alpha)V_{\varphi'}.$$

The result follows from Lemma 2.

□