# Appendix

**Herke van Hoof**[‡] **Jan Peters**[‡*] **Gerhard Neumann**[‡]
[‡]TU Darmstadt, Computer Science Department. [*]MPI for Intelligent Systems.

## 1 Derivation of REPS Solution

We start out with the constrained optimization problem

$$\max_{\pi,\mu_\pi} J(\pi) = \max_{\pi,\mu_\pi} \iint_{A\times S} \pi(\mathbf{a}|\mathbf{s})\mu_\pi(\mathbf{s})\mathcal{R}_{\mathbf{s}}^{\mathbf{a}} \mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s} \tag{1}$$

$$s.t. \quad \iint_{A\times S} \pi(\mathbf{a}|\mathbf{s})\mu_\pi(\mathbf{s})\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s} = 1 \tag{2}$$

$$\forall s' \quad \iint_{A\times S} \pi(\mathbf{a}|\mathbf{s})\mu_\pi(\mathbf{s})\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s} = \mu_\pi(\mathbf{s}') \tag{3}$$

$$\iint_{A\times S} \pi(\mathbf{a}|\mathbf{s})\mu_\pi(\mathbf{s}) \log \frac{\pi(\mathbf{a}|\mathbf{s})\mu_\pi(\mathbf{s})}{q(\mathbf{s},\mathbf{a})}\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s} \leq \epsilon. \tag{4}$$

For every constraint, we introduce a Lagrangian multiplier. Because (3) represents a continuum of constraints, we integrate over the value of this constraint multiplied by a state-dependent Lagrangian multiplier $V(\mathbf{s})$. We will write $p(\mathbf{s},\mathbf{a}) = \pi(\mathbf{a}|\mathbf{s})\mu_\pi(\mathbf{s})$ to keep the exposition brief. Therefore, the Lagrangian

$$\begin{aligned}L(p,\eta,V,\lambda) = &\iint_{A\times S} p(\mathbf{s},\mathbf{a})\mathcal{R}_{\mathbf{s}}^{\mathbf{a}}\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s}\\
&+ \lambda\left(1 - \iint_{A\times S} p(\mathbf{s},\mathbf{a})\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s}\right)\\
&+ \int_S V(\mathbf{s}')\left(\iint_{A\times S} p(\mathbf{s},\mathbf{a})\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s} - \mu_\pi(\mathbf{s}')\right)\mathrm{d}\mathbf{s}'\\
&+ \eta\left(\epsilon - \iint_{A\times S} p(\mathbf{s},\mathbf{a}) \log\frac{p(\mathbf{s},\mathbf{a})}{q(\mathbf{s},\mathbf{a})}\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s}\right).\end{aligned}$$

The Lagrangian can be re-shaped, using $\mu_\pi(\mathbf{s}) = \int_A p(\mathbf{s},\mathbf{a})\mathrm{d}\mathbf{a}$, in the more convenient form

$$\begin{aligned}L(p,\eta,V,\lambda) &= \lambda - \mathbb{E}_{p(\mathbf{s},\mathbf{a})}\left[V(\mathbf{s})\right] + \eta\epsilon\\
&+\mathbb{E}_{p(\mathbf{s},\mathbf{a})}\left[\mathcal{R}_{\mathbf{s}}^{\mathbf{a}} - \lambda + \int_S V(\mathbf{s}')\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{s}' - \eta\log\frac{p(\mathbf{s},\mathbf{a})}{q(\mathbf{s},\mathbf{a})}\right].\end{aligned}$$

To find the optimal $p$, we take the derivative of $L$ w.r.t. $p$ and set it to zero

$$\begin{aligned}0 =&\frac{\partial L}{\partial p(\mathbf{s},\mathbf{a})}\\
=&\mathcal{R}_{\mathbf{s}}^{\mathbf{a}} - \lambda + \int_S V(\mathbf{s}')\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{s}' - \eta\log\frac{p(\mathbf{s},\mathbf{a})}{q(\mathbf{s},\mathbf{a})} - \eta - V(\mathbf{s})\end{aligned}$$

therefore,

$$\eta\log\frac{p(\mathbf{s},\mathbf{a})}{q(\mathbf{s},\mathbf{a})} = \mathcal{R}_{\mathbf{s}}^{\mathbf{a}} - \lambda + \int_S V(\mathbf{s}')\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{s}' - \eta - V(\mathbf{s})$$

$$\begin{aligned}p(\mathbf{s},\mathbf{a}) =&q(\mathbf{s},\mathbf{a})\exp\left(\frac{\mathcal{R}_{\mathbf{s}}^{\mathbf{a}} - \int_S V(\mathbf{s}')\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{s}' - V(\mathbf{s})}{\eta}\right)\\
&\cdot\exp\left(\frac{-\lambda-\eta}{\eta}\right)\\
\propto&q(\mathbf{s},\mathbf{a})\exp\left(\frac{\mathcal{R}_{\mathbf{s}}^{\mathbf{a}} - \int_S V(\mathbf{s}')\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{s}' - V(\mathbf{s})}{\eta}\right).\end{aligned}$$

The function $V(\mathbf{s})$ resembles a value function, so that $\delta(\mathbf{s},\mathbf{a},V) = \mathcal{R}_{\mathbf{s}}^{\mathbf{a}} - \int_S V(\mathbf{s}')\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}\mathrm{d}\mathbf{s}' - V(\mathbf{s})$ can be identified as a Bellman error. Since $p(\mathbf{s},\mathbf{a})$ is a probability distribution we can identify $\exp(-\lambda/\eta - 1)$ to be a normalization factor

$$\begin{aligned}Z^{-1} &= \left(\iint_{A\times S} q(\mathbf{s},\mathbf{a})\exp\left(\delta(\mathbf{s},\mathbf{a},V)/\eta\right)\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{s}\right)^{-1}\\
&= \left(\mathbb{E}_{q(\mathbf{s},\mathbf{a})}\exp\left(\delta(\mathbf{s},\mathbf{a},V)/\eta\right)\right)^{-1}.\end{aligned}$$

## 2 The Dual and its Derivatives

We can re-insert the state-action probabilities in the Lagrangian to obtain the dual

$$\begin{aligned} g(\eta, V, \lambda) =& \lambda + \eta\epsilon \\ &+ \mathbb{E}_{p(\mathbf{s},\mathbf{a})}\left[\delta(\mathbf{s},\mathbf{a},V) - \lambda - \eta \log \frac{p(\mathbf{s},\mathbf{a})}{q(\mathbf{s},\mathbf{a})}\right] \\ =& \lambda + \eta\epsilon + \mathbb{E}_{p(\mathbf{s},\mathbf{a})}\left[-\lambda + \lambda\right] \\ &+ \mathbb{E}_{p(\mathbf{s},\mathbf{a})}\left[\delta(\mathbf{s},\mathbf{a},V) - \delta(\mathbf{s},\mathbf{a},V) + \eta\right] \\ =& \lambda + \eta\epsilon + \mathbb{E}_{p(\mathbf{s},\mathbf{a})}\eta \; d\mathbf{a}d\mathbf{s} \\ =& \lambda + \eta\epsilon + \eta = \eta\epsilon + \eta\log(Z) \\ =& \eta\epsilon + \eta \log\left(\mathbb{E}_{q(\mathbf{s},\mathbf{a})} \exp\left(\delta(\mathbf{s},\mathbf{a},V)/\eta\right)\right), \end{aligned}$$

where we used the identity

$$\begin{aligned} \exp\left(-\lambda/\eta - 1\right) &= Z^{-1} \\ \lambda + \eta &= \eta \log(Z). \end{aligned}$$

The expected value over $q$ can straightforwardly be approximated by taking the average of samples $1, \ldots, n$ taken from $q$. Note that $\lambda$ and $q$ do not appear in the final expression.

$$g(\eta, V) = \eta\epsilon + \eta \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(\delta(\mathbf{s}_i,\mathbf{a}_i,V)/\eta\right)\right).$$

When employing the kernel embedding, the Bellman error is written as

$$\delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha}) = \mathcal{R}_{\mathbf{s}_i}^{\mathbf{a}_i} + \boldsymbol{\alpha}^T(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)).$$

We define

$$w_i = \frac{\exp\left(\delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha})/\eta\right)}{\sum_{i=j}^{n} \exp\left(\delta(\mathbf{s}_j,\mathbf{a}_j,\boldsymbol{\alpha})/\eta\right)}$$

to keep equations brief and readable. The partial derivatives can be written as:

$$\begin{aligned} \frac{\partial g(\eta,\boldsymbol{\alpha})}{\partial \eta} =& -\frac{1}{\eta}\sum_{i=1}^{n} w_i \delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha}) + \epsilon \\ &+ \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(\delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha})/\eta\right)\right), \end{aligned}$$

$$\frac{\partial g(\eta,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} w_i \left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right),$$

and furthermore, for the Hessian we obtain

$$\begin{aligned} \frac{\partial^2 g(\eta,\boldsymbol{\alpha})}{\partial\eta\partial\eta} =& \frac{1}{\eta}\sum_{i=1}^{n} w_i \left(\delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha})\right)^2 \\ &- \frac{1}{\eta}\left(\sum_{i=1}^{n} w_i \delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha})\right)^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 g(\eta,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} =& -\frac{1}{\eta}\sum_{i=1}^{n} w_i \left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right) \\ &\cdot \sum_{i=1}^{n} w_i \left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right)^T + \\ &\sum_{i=1}^{n}\frac{w_i}{\eta}\left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right)\left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right)^T, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 g(\eta,\boldsymbol{\alpha})}{\partial\eta\partial\boldsymbol{\alpha}} =& -\frac{1}{\eta}\sum_{i=1}^{n} w_i \left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right) \\ &+ \sum_{i=1}^{n}\frac{w_i}{\eta}\delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha})\sum_{i=1}^{n} w_i \left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right) \\ &+ \frac{1}{\eta}\sum_{i=1}^{n} w_i \left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right) \\ &- \frac{1}{\eta}\sum_{i=1}^{n} w_i \delta(\mathbf{s}_i,\mathbf{a}_i,\boldsymbol{\alpha})\left(\mathbf{K}\beta(\mathbf{s}_i,\mathbf{a}_i) - \mathbf{k}_\mathrm{s}(\mathbf{s}_i)\right) \end{aligned}$$

.

## 3 Fitting a Generalizing Policy to State-Action Samples

To fit a generalizing policy $\tilde{\pi}(\mathbf{a}|\mathbf{s};\boldsymbol{\theta})$ to the samples-based policy $p(\mathbf{s}_i,\mathbf{a}_i) = \pi(\mathbf{a}_i|\mathbf{s}_i)\mu_\pi(\mathbf{s}_i)$ (defined only on samples $i \in \{1,\ldots,n\}$), we minize the expected Kullback-Leibler divergence

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mu_\pi(\mathbf{s})}\,\mathrm{KL}(\pi(\mathbf{a}|\mathbf{s})||\tilde{\pi}(\mathbf{a}|\mathbf{s}))\\ &= \int_S \mu_\pi(\mathbf{s})\int_A \pi(\mathbf{a}|\mathbf{s})\log\frac{\pi(\mathbf{a}|\mathbf{s})}{\tilde{\pi}(\mathbf{a}|\mathbf{s};\boldsymbol{\theta})}\mathbf{d}\mathbf{a}\mathbf{d}\mathbf{s}.\end{aligned}$$

This is a standard objective for matching two distributions. Note that the alternative Kullback-Leibler divergence $\mathrm{KL}(\tilde{\pi}(\mathbf{a}|\mathbf{s})||\pi(\mathbf{a}|\mathbf{s}))$ is undefined since $\pi(\mathbf{a}|\mathbf{s})$ is 0 at most places. Since the contribution to the integral is 0 for any $(\mathbf{s},\mathbf{a}) \notin \{(\mathbf{s}_1,\mathbf{a}_1),\ldots,(\mathbf{s}_n,\mathbf{a}_n)\}$, we can equivalently write:

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n \mu_\pi(\mathbf{s}_i)\pi(\mathbf{a}_i|\mathbf{s}_i)\log\frac{\pi(\mathbf{a_i}|\mathbf{s}_i)}{\tilde{\pi}(\mathbf{a}_i|\mathbf{s}_i;\boldsymbol{\theta})}\\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n \mu_\pi(\mathbf{s}_i)\pi(\mathbf{a}_i|\mathbf{s}_i)\log\frac{1}{\tilde{\pi}(\mathbf{a}_i|\mathbf{s}_i;\boldsymbol{\theta})}\\ &\quad + \sum_{i=1}^n \mu(\mathbf{s}_i)\pi(\mathbf{a}_i|\mathbf{s}_i)\log(\pi(\mathbf{a}_i|\mathbf{s}_i))\\ &= \arg\max_{\theta} \sum_{i=1}^n \mu_\pi(\mathbf{s}_i)\pi(\mathbf{a}_i|\mathbf{s}_i)\log\tilde{\pi}(\mathbf{a}_i|\mathbf{s}_i;\boldsymbol{\theta})\end{aligned}$$

where we used the fact that we can subtract terms constant in $\boldsymbol{\theta}$ and apply monotonously increasing functions to the terms to be minimized without changing the location of the minimum. Note that the final result is simply a weighted maximum-likelyhood estimate of $\boldsymbol{\theta}$. This result can be used to fit a parametric policy, or, as we demonstrate in the main material, a non-parametric policy to the weighted samples.

## 4 Optimization with Respect to $V$

In order to show that we can minimize the dual function $g$, we need to show that the optimal solution of the value function has the following form

$$V^* = \sum_{\tilde{\mathbf{s}}\in\tilde{\mathcal{S}}} \alpha_{\tilde{\mathbf{s}}} k_{\mathrm{s}}(\tilde{\mathbf{s}},\cdot) \tag{5}$$

We follow some steps in the proof of Schölkopf et al. [2001]. They consider arbitrary functions $c$ mapping to $\mathbb{R}\cup\{\infty\}$ of the form

$$c((\mathbf{s}_1,y_1,V(\mathbf{s}_1)),\ldots,(\mathbf{s}_m,y_m,V(\mathbf{s}_m))), \tag{6}$$

which typically defines an error function of function $V(\mathbf{s})$ on the samples $\mathbf{s}_i$ with desired output $y_i$. In our case, we do not have desired output values $y_i$ for our objective function. This is inconsequential as $c$ can be arbitrary, and so can be independent of all $y$ values.

Any function $V$ can be written as $V = \sum_{\tilde{\mathbf{s}}\in\tilde{\mathcal{S}}} \alpha_{\tilde{\mathbf{s}}} k_{\mathrm{s}}(\tilde{\mathbf{s}},\cdot) + v(\mathbf{s})$, where $v(\mathbf{s})$ is an additional bias term. If $V$ is constrained to be in the Hilbert space defined by $k$, Schölkopf et al. [2001] show that $c$ is independent of the bias term $v(\mathbf{s})$. This means that for any optimal $V'$ that is not of the proposed form, there is a $V^*$ of the proposed form that has the same objective value which is obtained by subtracting $v(\mathbf{s})$ from $V'$.

As the dual function $g$ satisfies the conditions to cost function $c$, for us this means that there is at least one $V^*$ optimizing $g$ of the proposed form. Note that it is inconsequential that the dual $g$ also depends on Langrangian parameter $\eta$. For any optimum $(\eta^*, V'^*)$, if $V'*$ is not of the proposed form, the projection $V^*$ of $V'^*$ on the proposed basis satisfies $g(\eta^*, V'^*) = g(\eta^*, V^*)$, so $(\eta^*, V^*)$ must be an optimum as well.