# Variational Inference for Inverse Reinforcement Learning with Gaussian Processes: Supplementary Material

Paulius Dilkas (2146879)

27th December 2018

## 1 Preliminaries

For any matrix $\mathbf{A}$, we will use either $A_{i,j}$ or $[\mathbf{A}]_{i,j}$ to denote the element of $\mathbf{A}$ in row $i$ and column $j$.

For any vector $\mathbf{x}$, we write $\mathbb{R}_d[\mathbf{x}]$ to denote a vector space of polynomials with degree at most $d$, where variables are elements of $\mathbf{x}$, and coefficients are in $\mathbb{R}$.

In this paper, all references to measurability are with respect to the Lebesgue measure. Similarly, whenever we consider the existence of an integral, we use the Lebesgue definition of integration.

**Lemma 1.1** (Derivatives of probability distributions)**.**

1. $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}} = q(\mathbf{u})\frac{1}{2}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\intercal})(\mathbf{u} - \boldsymbol{\mu})$.
2. $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\boldsymbol{\Sigma}^{-\intercal} q(\mathbf{u}) + \frac{1}{2} q(\mathbf{u}) \boldsymbol{\Sigma}^{-\intercal} (\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-\intercal}$.
3. *For* $i = 0, \ldots, d$,
$$\frac{\partial q(\mathbf{r})}{\partial \lambda_i} = q(\mathbf{r}) \frac{1}{2} \operatorname{tr}\left( (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}\mathbf{u}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}) \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} \right),$$
*where*
$$\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} = \frac{1}{\lambda_i} \mathbf{K}_{\mathbf{u},\mathbf{u}}$$
*if* $i = 0$, *and*
$$\left[ \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} \right]_{j,k} = k_{\boldsymbol{\lambda}}(\mathbf{x}_{\mathbf{u},j}, \mathbf{x}_{\mathbf{u},k}) \left( -\frac{1}{2}(x_{\mathbf{u},j,i} - x_{\mathbf{u},k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \right)$$
*otherwise.*

*Proof.*

1.
$$\begin{aligned}
\frac{\partial q(\mathbf{u})}{\partial m} &= q(\mathbf{u}) \frac{\partial}{\partial \boldsymbol{\mu}} \left[ -\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{u} - \boldsymbol{\mu}) \right] \\
&= q(\mathbf{u}) \left( -\frac{1}{2} \right) (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\intercal})(\mathbf{u} - \boldsymbol{\mu}) \frac{\partial}{\partial \boldsymbol{\mu}} [\mathbf{u} - \boldsymbol{\mu}] \\
&= q(\mathbf{u}) \frac{1}{2} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\intercal})(\mathbf{u} - \boldsymbol{\mu}).
\end{aligned}$$

2.

$$\begin{aligned}
\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[ \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left( -\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}) \right) \right] \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[ \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \right] \exp\left( -\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}) \right) \\
&+ \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[ \exp\left( -\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}) \right) \right] \\
&= \frac{1}{(2\pi)^{m/2}} \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[ \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right] \exp\left( -\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}) \right) \\
&- \frac{1}{2} q(\mathbf{u}) \frac{\partial}{\partial \boldsymbol{\Sigma}} [(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})].
\end{aligned}$$

The two remaining derivatives can be taken with the help of *The Matrix Cookbook* [5]:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} \left[ \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right] = -\frac{1}{2}|\boldsymbol{\Sigma}|^{-3/2} \frac{\partial |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}|\boldsymbol{\Sigma}|^{-3/2}|\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^{-\intercal} = -\frac{1}{2|\boldsymbol{\Sigma}|^{1/2}}\boldsymbol{\Sigma}^{-\intercal}, \\
\frac{\partial}{\partial \boldsymbol{\Sigma}} [(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})] = -\boldsymbol{\Sigma}^{-\intercal}(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-\intercal}.
\end{aligned}$$

Substituting them back in gives

$$\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\boldsymbol{\Sigma}^{-\intercal} q(\mathbf{u}) + \frac{1}{2} q(\mathbf{u}) \boldsymbol{\Sigma}^{-\intercal}(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-\intercal}.$$

3. Using a result by Rasmussen and Williams [6],

$$\frac{\partial q(\mathbf{r})}{\partial \lambda_i} = q(\mathbf{r}) \frac{\partial}{\partial \lambda_i} \left[ -\frac{1}{2}\mathbf{u}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u} - \frac{1}{2}\log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| \right] = q(\mathbf{r})\frac{1}{2}\operatorname{tr}\left( (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}) \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} \right).$$

The remaining derivative is

$$\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} = \begin{cases} \frac{1}{\lambda_i}\mathbf{K}_{\mathbf{u},\mathbf{u}} & \text{if } i = 0, \\ \mathbf{L}_{\mathbf{u},\mathbf{u}} & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned}
[\mathbf{L}_{\mathbf{u},\mathbf{u}}]_{j,k} &= \frac{\partial}{\partial \lambda_i} k_{\boldsymbol{\lambda}}(\mathbf{x}_{\mathbf{u},j}, \mathbf{x}_{\mathbf{u},k}) \\
&= k_{\boldsymbol{\lambda}}(\mathbf{x}_{\mathbf{u},j}, \mathbf{x}_{\mathbf{u},k}) \frac{\partial}{\partial \lambda_i} \left[ -\frac{1}{2}(\mathbf{x}_{\mathbf{u},j}-\mathbf{x}_{\mathbf{u},k})^\intercal \boldsymbol{\Lambda}(\mathbf{x}_{\mathbf{u},j}-\mathbf{x}_{\mathbf{u},k}) - \mathbb{1}[j \neq k]\sigma^2 \operatorname{tr}(\boldsymbol{\Lambda}) \right] \\
&= k_{\boldsymbol{\lambda}}(\mathbf{x}_{\mathbf{u},j}, \mathbf{x}_{\mathbf{u},k}) \frac{\partial}{\partial \lambda_i} \left[ -\frac{1}{2}\sum_{l=1}^{d} \lambda_l (x_{\mathbf{u},j,l} - x_{\mathbf{u},k,l})^2 - \mathbb{1}[j \neq k]\sigma^2 \sum_{l=1}^{d} \lambda_l \right] \\
&= k_{\boldsymbol{\lambda}}(\mathbf{x}_{\mathbf{u},j}, \mathbf{x}_{\mathbf{u},k}) \left( -\frac{1}{2}(x_{\mathbf{u},j,i} - x_{\mathbf{u},k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \right).
\end{aligned}$$

□

## 1.1 Linear Algebra and Numerical Analysis

**Definition 1.2** (Norms)**.** For any finite-dimensional vector $\mathbf{x} = (x_1, \ldots, x_n)^\intercal$, its *maximum norm* is

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

whereas its *taxicab* (or *Manhattan*) *norm* is

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|.$$

Let $\mathbf{A}$ be a matrix. For any vector norm $\|\cdot\|_p$, we can also define its *induced norm* for matrices as

$$\|\mathbf{A}\|_p = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}.$$

In particular, for $p = \infty$, we have

$$\|\mathbf{A}\|_\infty = \max_i \sum_j |A_{i,j}|.$$

**Lemma 1.3** (Perturbation Lemma [4])**.** *Let $\|\cdot\|$ be any matrix norm, and let $\mathbf{A}$ and $\mathbf{E}$ be matrices such that $\mathbf{A}$ is invertible and $\|\mathbf{A}^{-1}\|\|\mathbf{E}\| < 1$, then $\mathbf{A}+\mathbf{E}$ is invertible, and*

$$\|(\mathbf{A}+\mathbf{E})^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1-\|\mathbf{A}^{-1}\|\|\mathbf{E}\|}.$$

# 2 Proofs

We primarily think of rewards as a vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, but sometimes we use a function notation $r(s)$ to denote the reward of a particular state $s \in \mathcal{S}$. The functional notation is purely a notational convenience.

MDP values are characterised by both a state and a reward function/vector. In order to prove the next theorem, we think of the value function as $V : \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$, i.e., $V$ takes a state $s \in \mathcal{S}$ and returns a function $V(s) : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}$ that takes a reward vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ and returns a value of the state $s$, $V_{\mathbf{r}}(s) \in \mathbb{R}$. The function $V(s)$ computes the values of all states and returns the value of state $s$.

**Proposition 2.1.** *MDP value functions $V(s) : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ (for $s \in \mathcal{S}$) are Lebesgue measurable.*

*Proof sketch.* For any reward vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$, the collection of converged value functions $\{V_{\mathbf{r}}(s) \mid s \in \mathcal{S}\}$ satisfy

$$V_{\mathbf{r}}(s) = \log \sum_{a\in\mathcal{A}} \exp\left(r(s) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{T}(s,a,s') V_{\mathbf{r}}(s')\right)$$

for all $s \in \mathcal{S}$. Let $s_0 \in \mathcal{S}$ be an arbitrary state. In order to prove that $V(s_0)$ is measurable, it is enough to show that for any $\alpha \in \mathbb{R}$, the set

$$\left\{ \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \;\middle|\; \begin{array}{l} V_{\mathbf{r}}(s_0) \in (-\infty, \alpha); \\ V_{\mathbf{r}}(s) \in \mathbb{R} \text{ for all } s \in \mathcal{S} \setminus \{s_0\}; \\ V_{\mathbf{r}}(s) = \log \sum_{a\in\mathcal{A}} \exp\left(r(s) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{T}(s,a,s') V_{\mathbf{r}}(s')\right) \text{ for all } s \in \mathcal{S} \end{array} \right\}$$

is measurable. Since this set can be constructed in Zermelo-Fraenkel set theory *without* the axiom of choice, it is measurable [3], which proves that $V(s)$ is a measurable function for any $s \in \mathcal{S}$. □

**Proposition 2.2.** *If the initial values of the MDP value function satisfy the following bound, then the bound remains satisfied throughout value iteration:*

$$|V_{\mathbf{r}}(s)| \leq \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}. \tag{1}$$

*Proof.* We begin by considering (1) without taking the absolute value of $V_{\mathbf{r}}(s)$, i.e.,

$$V_{\mathbf{r}}(s) \leq \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}, \tag{2}$$

and assuming that the initial values of $\{V_{\mathbf{r}}(s) \mid s \in \mathcal{S}\}$ already satisfy (2). For each $s \in \mathcal{S}$, the value of $V_{\mathbf{r}}(s)$ is updated via this rule:

$$V_{\mathbf{r}}(s) \coloneqq \log \sum_{a \in \mathcal{A}} \exp\left(r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') V_{\mathbf{r}}(s')\right).$$

Note that both log and exp are increasing functions, $\gamma > 0$, and the $\mathcal{T}$ function gives a probability (a non-negative number). Thus

$$\begin{aligned}
V_{\mathbf{r}}(s) &\leq \log \sum_{a \in \mathcal{A}} \exp\left(r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}\right) \\
&= \log \sum_{a \in \mathcal{A}} \exp\left(r(s) + \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|)}{1-\gamma} \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s')\right) \\
&= \log \sum_{a \in \mathcal{A}} \exp\left(r(s) + \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|)}{1-\gamma}\right)
\end{aligned}$$

by the definition of $\mathcal{T}$. Then

$$\begin{aligned}
V_{\mathbf{r}}(s) &\leq \log\left(|\mathcal{A}| \exp\left(r(s) + \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|)}{1-\gamma}\right)\right) \\
&= \log\left(\exp\left(\log|\mathcal{A}| + r(s) + \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|)}{1-\gamma}\right)\right) \\
&= \log|\mathcal{A}| + r(s) + \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|)}{1-\gamma} \\
&= \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|) + (1-\gamma)(\log|\mathcal{A}| + r(s))}{1-\gamma} \\
&\leq \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|) + (1-\gamma)(\log|\mathcal{A}| + \|\mathbf{r}\|_\infty)}{1-\gamma} \\
&= \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}
\end{aligned}$$

by the definition of $\|\mathbf{r}\|_\infty$.

The proof for

$$V_{\mathbf{r}}(s) \geq \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{\gamma - 1} \tag{3}$$

follows the same argument until we get to

$$\begin{aligned}
V_{\mathbf{r}}(s) &\geq \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|) + (\gamma-1)(\log|\mathcal{A}| + r(s))}{\gamma-1} \\
&\geq \frac{\gamma(\|\mathbf{r}\|_\infty + \log|\mathcal{A}|) + (\gamma-1)(-\log|\mathcal{A}| - \|\mathbf{r}\|_\infty)}{\gamma-1} \\
&= \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{\gamma-1},
\end{aligned}$$

where we use the fact that $r(s) \geq -\|\mathbf{r}\|_\infty - 2\log|\mathcal{A}|$. Combining (2) and (3) gives (1). □

**Theorem 2.3** (The Lebesgue Dominated Convergence Theorem [7]). *Let $(X, \mathcal{M}, \mu)$ be a measure space and $\{f_n\}$ a sequence of measurable functions on $X$ for which $\{f_n\} \to f$ pointwise a.e. on $X$ and the function $f$ is measurable. Assume there is a non-negative function $g$ that is integrable over $X$ and dominates the sequence $\{f_n\}$ on $X$ in the sense that*

$$|f_n| \leq g \text{ a.e. on } X \text{ for all } n.$$

*Then $f$ is integrable over $X$ and*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu.$$

**Lemma 2.4.** *Let $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (a, b) \subset \mathbb{R}$ be an arbitrary bounded function. Then, for $i = 0, \ldots, d$,*

$$\left.\frac{\partial q(\mathbf{r})}{\partial \lambda_i}\right|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$$

*has upper and lower bounds of the form $q(\mathbf{r})d(\mathbf{u})$, where $d(\mathbf{u}) \in \mathbb{R}_2[\mathbf{u}]$.*

*Proof.* Remember that

$$\frac{\partial q(\mathbf{r})}{\partial \lambda_i} = q(\mathbf{r})\frac{1}{2}\operatorname{tr}\left((\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^\intercal\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1})\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}\right)$$

by Lemma 1.1. We begin by producing constant upper and lower bounds for the elements of

$$\left.\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}\right|_{\lambda_i = c(\mathbf{r},\mathbf{u})}.$$

If $i = 0$, then each element of $\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_0}$ is of the form

$$\exp\left(-\frac{1}{2}(\mathbf{x}_j - \mathbf{x}_k)^\intercal\mathbf{\Lambda}(\mathbf{x}_j - \mathbf{x}_k) - \mathbb{1}[j \neq k]\sigma^2 \operatorname{tr}(\mathbf{\Lambda})\right),$$

i.e., without $\lambda_0$, so

$$\left.\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_0}\right|_{\lambda_0 = c(\mathbf{r},\mathbf{u})} = \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_0}$$

is already independent of $\mathbf{r}$ and $\mathbf{u}$—there is no need for any bounds.

If $i > 0$, then each element of $\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}$ is a constant multiple of $k_{\boldsymbol{\lambda}}(\mathbf{x}_j, \mathbf{x}_k)$, for some $\mathbf{x}_j$ and $\mathbf{x}_k$. Since $k_{\boldsymbol{\lambda}}(\mathbf{x}_j, \mathbf{x}_k)$ is a decreasing function of $\lambda_i$, and $c(\mathbf{r}, \mathbf{u}) > a$,

$$\begin{aligned}
k_{\boldsymbol{\lambda}}(\mathbf{x}_j, \mathbf{x}_k)|_{\lambda_i = c(\mathbf{r},\mathbf{u})} &= \lambda_0 \exp\Bigg( -\frac{1}{2}c(\mathbf{r},\mathbf{u})(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 c(\mathbf{r},\mathbf{u}) \\
&\qquad - \sum_{n \in \{1,\ldots,d\}\setminus\{i\}} \frac{1}{2}\lambda_n (x_{j,n} - x_{k,n})^2 + \mathbb{1}[j \neq k]\sigma^2\lambda_n \Bigg) \\
&< \lambda_0 \exp\Bigg( -\frac{1}{2}a(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 a \\
&\qquad - \sum_{n \in \{1,\ldots,d\}\setminus\{i\}} \frac{1}{2}\lambda_n (x_{j,n} - x_{k,n})^2 + \mathbb{1}[j \neq k]\sigma^2\lambda_n \Bigg),
\end{aligned}$$

which gives an upper bound on each element of

$$\left.\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}\right|_{\lambda_i = c(\mathbf{r},\mathbf{u})}.$$

A similar line of reasoning establishes lower bounds as well.

Combining the bounds with the observation that every element of $\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^\intercal\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal}$ is in $\mathbb{R}_2[\mathbf{u}]$ gives the required result. □

*Remark.* In order to find $\frac{\partial q(\mathbf{u})}{\partial t}$, where $t$ is the $i$th element of the vector $\boldsymbol{\mu}$, we can find $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}$ and simply take the $i$th element. A similar line of reasoning applies to matrices as well. Thus, we only need to consider derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

**Lemma 2.5.** *Let $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (a, b) \subset \mathbb{R}$ be an arbitrary bounded function. Then, for $i = 1, \ldots, m$, every element of*

$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}\right|_{\mu_i = c(\mathbf{r},\mathbf{u})}$$

*has upper and lower bounds of the form $q(\mathbf{u})d(\mathbf{u})$, where $d(\mathbf{u}) \in \mathbb{R}_1[\mathbf{u}]$.*

*Proof.* Using Lemma 1.1,

$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}\right|_{\mu_i = c(\mathbf{r},\mathbf{u})} = q(\mathbf{u})\frac{1}{2}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\intercal})(\mathbf{u} - \mathbf{c}(\mathbf{r}, \mathbf{u})),$$

where $\mathbf{c}(\mathbf{r}, \mathbf{u}) = (\mu_1, \ldots, \mu_{i-1}, c(\mathbf{r}, \mathbf{u}), \mu_{i+1} \ldots, \mu_m)^\intercal$. Since $c(\mathbf{r}, \mathbf{u})$ is bounded and $\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\intercal}$ is a constant matrix, we can use the bounds on $c(\mathbf{r}, \mathbf{u})$ to manufacture both upper and lower bounds on

$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}\right|_{\mu_i = c(\mathbf{r},\mathbf{u})}$$

of the required form. □

**Lemma 2.6.** *Let $i, j = 1, \ldots, m$, and let $\epsilon > 0$ be arbitrary. Furthermore, let*

$$c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (\Sigma_{i,j} - \epsilon, \Sigma_{i,j} + \epsilon) \subset \mathbb{R}$$

*be a function with a codomain arbitrarily close to $\Sigma_{i,j}$. Then every element of*

$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}}\right|_{\Sigma_{i,j} = c(\mathbf{r},\mathbf{u})}$$

*has upper and lower bounds of the form $q(\mathbf{u})d(\mathbf{u})$, where $d(\mathbf{u}) \in \mathbb{R}_2[\mathbf{u}]$.*

*Proof.* Using Lemma 1.1,

$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}}\right|_{\Sigma_{i,j} = c(\mathbf{r},\mathbf{u})} = -\frac{1}{2}\mathbf{C}(\mathbf{r}, \mathbf{u})^{-\intercal} + \frac{1}{2}\mathbf{C}(\mathbf{r}, \mathbf{u})^{-\intercal}(\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})^\intercal\mathbf{C}(\mathbf{r}, \mathbf{u})^{-\intercal},$$

where

$$[\mathbf{C}(\mathbf{r}, \mathbf{u})]_{k,l} = \begin{cases} c(\mathbf{r}, \mathbf{u}) & \text{if } (k, l) = (i, j), \\ \Sigma_{k,l} & \text{otherwise.} \end{cases}$$

We can also express $\mathbf{C}(\mathbf{r}, \mathbf{u})$ as $\mathbf{C}(\mathbf{r}, \mathbf{u}) = \boldsymbol{\Sigma} + \mathbf{E}(\mathbf{r}, \mathbf{u})$, where

$$[\mathbf{E}(\mathbf{r}, \mathbf{u})]_{k,l} = \begin{cases} c(\mathbf{r}, \mathbf{u}) - \Sigma_{i,j} & \text{if } (k, l) = (i, j), \\ 0 & \text{otherwise.} \end{cases}$$

We begin by establishing upper and lower bounds on $\mathbf{C}(\mathbf{r}, \mathbf{u})^{-1}$. For this, we use the maximum norm $\|\cdot\|_\infty$ on both vectors and matrices. We can apply Lemma 1.3 to $\boldsymbol{\Sigma}$ and $\mathbf{E}(\mathbf{r}, \mathbf{u})$ since

$$\|\mathbf{E}(\mathbf{r}, \mathbf{u})\|_\infty = \max_k \sum_l |[\mathbf{E}(\mathbf{r}, \mathbf{u})]_{k,l}| = |c(\mathbf{r}, \mathbf{u}) - \Sigma_{i,j}| < \epsilon$$

can be made arbitrarily small so that $\|\mathbf{\Sigma}^{-1}\|_\infty \|\mathbf{E}(\mathbf{r},\mathbf{u})\|_\infty < 1$. Then $\mathbf{C}(\mathbf{r},\mathbf{u})$ is invertible, and

$$\|\mathbf{C}(\mathbf{r},\mathbf{u})^{-1}\|_\infty \leq \frac{\|\mathbf{\Sigma}^{-1}\|_\infty}{1 - \|\mathbf{\Sigma}^{-1}\|_\infty \|\mathbf{E}(\mathbf{r},\mathbf{u})\|_\infty} < \frac{\|\mathbf{\Sigma}^{-1}\|_\infty}{1 - \|\mathbf{\Sigma}^{-1}\|_\infty \epsilon},$$

which means that

$$\max_k \sum_l \left|[\mathbf{C}(\mathbf{r},\mathbf{u})^{-1}]_{k,l}\right| < \frac{\|\mathbf{\Sigma}^{-1}\|_\infty}{1 - \|\mathbf{\Sigma}^{-1}\|_\infty \epsilon},$$

i.e., for any row $k$ and column $l$,

$$\left|[\mathbf{C}(\mathbf{r},\mathbf{u})^{-1}]_{k,l}\right| < \frac{\|\mathbf{\Sigma}^{-1}\|_\infty}{1 - \|\mathbf{\Sigma}^{-1}\|_\infty \epsilon},$$

which bounds all elements of $\mathbf{C}(\mathbf{r},\mathbf{u})^{-1}$ as required. Since every element of $(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal$ is in $\mathbb{R}_2[\mathbf{u}]$, and the elements of $\mathbf{C}(\mathbf{r},\mathbf{u})^{-1}$ are bounded, the desired result follows. □

**Lemma 2.7.**

$$\int \|\mathbf{r}\|_\infty q(\mathbf{r})\, d\mathbf{r} \leq a + \|\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\|_1, \tag{4}$$

*where $a$ is a constant independent of $\mathbf{u}$.*

*Proof.* Since $\|\mathbf{r}\|_\infty \leq \|\mathbf{r}\|_1$,

$$\int \|\mathbf{r}\|_\infty q(\mathbf{r})\, d\mathbf{r} \leq \int \|\mathbf{r}\|_1 q(\mathbf{r})\, d\mathbf{r} = \sum_{i=1}^{|\mathcal{S}|} \mathbb{E}[|r_i|].$$

As each $\mathbb{E}[|r_i|]$ is a mean of a folded Gaussian distribution,

$$\mathbb{E}[|r_i|] = \sigma_i \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\xi_i^2}{2\sigma_i^2}\right) + \xi_i \left(1 - 2\Phi\left(-\frac{\xi_1}{\sigma_1}\right)\right),$$

where $\xi_i = \left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\right]_i$, $\sigma_i = \sqrt{[\mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}}]_{i,i}}$[1], and $\Phi$ is the cumulative distribution function of the standard normal distribution. Furthermore,

$$\mathbb{E}[|r_i|] \leq \sigma_i \sqrt{\frac{2}{\pi}} + |\xi_i|,$$

as $\sigma_i$ is non-negative, and $\Phi(x) \in [0,1]$ for all $x$. Since

$$\sum_{i=1}^{|\mathcal{S}|} |\xi_i| = \|\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\|_1,$$

we can set

$$a = \sum_{i=1}^{|\mathcal{S}|} \sigma_i \sqrt{\frac{2}{\pi}}$$

to get (4). □

Our main theorem is a specialised version of an integral differentiation result by Chen [2].

**Theorem 2.8.** *Whenever the derivative exists,*

$$\frac{\partial}{\partial t} \iint V_{\mathbf{r}}(s) q(\mathbf{r}) q(\mathbf{u})\, d\mathbf{r}\, d\mathbf{u} = \iint \frac{\partial}{\partial t} [V_{\mathbf{r}}(s) q(\mathbf{r}) q(\mathbf{u})]\, d\mathbf{r}\, d\mathbf{u},$$

*where $t$ is any scalar part of $\boldsymbol{\mu}$, $\mathbf{\Sigma}$, or $\boldsymbol{\lambda}$.*

[1] The expression under the square root sign is non-negative because $\mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}}$ is a covariance matrix of a Gaussian distribution, hence also positive semi-definite, which means that its diagonal entries are non-negative.

*Proof.* Let

$$f(\mathbf{r}, \mathbf{u}, t) = V_{\mathbf{r}}(s) q(\mathbf{r}) q(\mathbf{u}),$$
$$F(t) = \iint f(\mathbf{r}, \mathbf{u}, t)\, d\mathbf{r}\, d\mathbf{u},$$

and fix the value of $t$. Let $(t_n)_{n=1}^{\infty}$ be any sequence such that $\lim_{n\to\infty} t_n = t$, but $t_n \neq t$ for all $n$. We want to show that

$$F'(t) = \lim_{n\to\infty} \frac{F(t_n) - F(t)}{t_n - t} = \iint \left.\frac{\partial f}{\partial t}\right|_{(\mathbf{r},\mathbf{u},t)} d\mathbf{r}\, d\mathbf{u}. \tag{5}$$

We have

$$\frac{F(t_n) - F(t)}{t_n - t} = \iint \frac{f(\mathbf{r}, \mathbf{u}, t_n) - f(\mathbf{r}, \mathbf{u}, t)}{t_n - t}\, d\mathbf{r}\, d\mathbf{u} = \iint f_n(\mathbf{r}, \mathbf{u})\, d\mathbf{r}\, d\mathbf{u},$$

where

$$f_n(\mathbf{r}, \mathbf{u}) = \frac{f(\mathbf{r}, \mathbf{u}, t_n) - f(\mathbf{r}, \mathbf{u}, t)}{t_n - t}.$$

Since

$$\lim_{n\to\infty} f_n(\mathbf{r}, \mathbf{u}) = \left.\frac{\partial f}{\partial t}\right|_{(\mathbf{r},\mathbf{u},t)},$$

(5) follows from Theorem 2.3 as soon as we show that both $f$ and $f_n$ are measurable and find a non-negative integrable function $g$ such that for all $n$, $\mathbf{r}$, $\mathbf{u}$,

$$|f_n(\mathbf{r}, \mathbf{u})| \leq g(\mathbf{r}, \mathbf{u}).$$

The MDP value function is measurable by Proposition 2.1. The result of multiplying or adding measurable functions (e.g., probability density functions (PDFs)) to a measurable function is still measurable. Thus, both $f$ and $f_n$ are measurable.

It remains to find $g$. For notational simplicity and without loss of generality, we will temporarily assume that $t$ is a parameter of $q(\mathbf{r})$. Then

$$|f_n(\mathbf{r}, \mathbf{u})| = |V_{\mathbf{r}}(s)| \left| \frac{q(\mathbf{r})|_{t=t_n} - q(\mathbf{r})}{t_n - t} \right| q(\mathbf{u})$$

since PDFs are non-negative. An upper bound for $|V_{\mathbf{r}}(s)|$ is given by Proposition 2.2, while

$$\frac{q(\mathbf{r})|_{t=t_n} - q(\mathbf{r})}{t_n - t} = \left.\frac{\partial q(\mathbf{r})}{\partial t}\right|_{t=c(\mathbf{r},\mathbf{u})}$$

for some function $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (\min\{t, t_n\}, \max\{t, t_n\})$ due to the mean value theorem (since $q$ is a continuous and differentiable function of $t$, regardless of the specific choices of $q$ and $t$).

We then have that

$$|f_n(\mathbf{r}, \mathbf{u})| \leq \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma} \left| \left.\frac{\partial q(\mathbf{r})}{\partial t}\right|_{t=c(\mathbf{r},\mathbf{u})} \right| q(\mathbf{u}).$$

The bound is clearly non-negative and measurable. It remains to show that it is also integrable. Depending on what $t$ represents, we can use one of the Lemmas 2.4, 2.5, and 2.6, which gives us two polynomials $p_1(\mathbf{u}), p_2(\mathbf{u}) \in \mathbb{R}_2[\mathbf{u}]$ such that

$$p_1(\mathbf{u}) q(\mathbf{r}) < \left.\frac{\partial q(\mathbf{r})}{\partial t}\right|_{t=c(\mathbf{r},\mathbf{u})} < p_2(\mathbf{u}) q(\mathbf{r}).$$

Then

$$\left| \left.\frac{\partial q(\mathbf{r})}{\partial t}\right|_{t=c(\mathbf{r},\mathbf{u})} \right| < q(\mathbf{r}) \max\{|p_1(\mathbf{u})|, |p_2(\mathbf{u})|\}.$$

We can now apply Lemma 2.7, which allows us to integrate out $\mathbf{r}$, and we are left with showing the existence of

$$\int \left(a + \|\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\|_1\right) \max\{|p_1(\mathbf{u})|, |p_2(\mathbf{u})|\} q(\mathbf{u})\, d\mathbf{u}, \tag{6}$$

where $a$ is a constant. The integral

$$\int \max\{|p_1(\mathbf{u})|, |p_2(\mathbf{u})|\} q(\mathbf{u})\, d\mathbf{u} = \int \max\{|p_1(\mathbf{u})q(\mathbf{u})|, |p_2(\mathbf{u})q(\mathbf{u})|\}\, d\mathbf{u}$$

exists because $p_1(\mathbf{u})q(\mathbf{u})$ and $p_2(\mathbf{u})q(\mathbf{u})$ are both integrable, hence their absolute values are integrable, and the maximum of two integrable functions is also integrable. Since $\|\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\|_1 \in \mathbb{R}_1[\mathbf{u}]$, a similar argument can be applied to the rest of (6) as well. □

# 3 Evidence Lower Bound

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{(\mathbf{u},\mathbf{r})\sim q_{\boldsymbol{\nu}}(\mathbf{u},\mathbf{r})}\left[\log \frac{p(\mathcal{D}, \mathbf{X}_{\mathbf{u}}, \mathbf{u}, \mathbf{r})}{q_{\boldsymbol{\nu}}(\mathbf{u}, \mathbf{r})}\right] \\ &= \iint q_{\boldsymbol{\nu}}(\mathbf{u}, \mathbf{r}) \log \frac{p(\mathcal{D}, \mathbf{X}_{\mathbf{u}}, \mathbf{u}, \mathbf{r})}{q_{\boldsymbol{\nu}}(\mathbf{u}, \mathbf{r})}\, d\mathbf{r}\, d\mathbf{u}.\end{aligned} \tag{7}$$

$$p(\mathcal{D}, \mathbf{X}_{\mathbf{u}}, \mathbf{u}, \mathbf{r}) = p(\mathbf{X}_{\mathbf{u}}) \times p(\mathbf{u}|\mathbf{X}_{\mathbf{u}}) \times p(\mathbf{r}|\mathbf{X}_{\mathbf{u}}, \mathbf{u}) \times p(\mathcal{D}|\mathbf{r}). \tag{8}$$

$$q_{\boldsymbol{\nu}}(\mathbf{u}, \mathbf{r}) = q(\mathbf{u}) \times q(\mathbf{r}|\mathbf{u}). \tag{9}$$

In this section we derive and simplify the ELBO for this (now fully specified) model. In order to derive the ELBO, let us go back to (7) and write[2]

$$\mathcal{L} = \mathbb{E}[\log p(\mathcal{D}, \mathbf{X}_{\mathbf{u}}, \mathbf{u}, \mathbf{r})] - \mathbb{E}[\log q_{\boldsymbol{\nu}}(\mathbf{u}, \mathbf{r})].$$

By substituting in (8) and (9), we get

$$\mathcal{L} = \mathbb{E}[\log p(\mathbf{X}_{\mathbf{u}}) + \log p(\mathbf{u}|\mathbf{X}_{\mathbf{u}}) + \log p(\mathbf{r}|\mathbf{X}_{\mathbf{u}}, \mathbf{u}) + \log p(\mathcal{D}|\mathbf{r})] - \mathbb{E}[\log q(\mathbf{u}) + \log q(\mathbf{r}|\mathbf{u})].$$

Note that $\mathbb{E}[\log p(\mathbf{X}_{\mathbf{u}})]$ is just a constant, so we can simply drop it from the expression. Furthermore, since $q(\mathbf{r}|\mathbf{u}) = p(\mathbf{r}|\mathbf{X}_{\mathbf{u}}, \mathbf{u})$, they cancel each other out. Then, we can substitute various terms with their definitions to get

$$\mathcal{L} = \mathbb{E}[\log \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})] + \mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} Q_{\mathbf{r}}(s_{i,t}, a_{i,t}) - V_{\mathbf{r}}(s_{i,t})\right] - \mathbb{E}[\log \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})].$$

Using the expressions for $Q_{\mathbf{r}}$ and the entropy of a normal distribution [1],

$$\mathcal{L} = \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| + \mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} r(s_{i,t}) - V_{\mathbf{r}}(s_{i,t}) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{T}(s_{i,t}, a_{i,t}, s')V_{\mathbf{r}}(s')\right].$$

We can simplify $\sum_{i=1}^{N}\sum_{t=1}^{T} r(s_{i,t})$ by defining a new vector $\mathbf{t} = (t_1, \ldots, t_{|\mathcal{S}|})^{\intercal}$, where $t_i$ is the number of times the state associated with the reward $r_i$ has been visited across all demonstrations. Then

$$\mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} r(s_{i,t})\right] = \mathbb{E}[\mathbf{t}^{\intercal}\mathbf{r}] = \mathbf{t}^{\intercal}\mathbb{E}[\mathbf{r}] = \mathbf{t}^{\intercal}\mathbb{E}\left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\right] = \mathbf{t}^{\intercal}\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}.$$

[2] At this point, we will drop the subscript denoting which variables the expectation is taken over. Also note that throughout the derivation equality is taken to mean 'equality up to an additive constant'.

This allows us to simplify $\mathcal{L}$ to

$$\mathcal{L} = \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| + \mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} - \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[V_{\mathbf{r}}(s_{i,t})] - \gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s_{i,t},a_{i,t},s')\mathbb{E}[V_{\mathbf{r}}(s')].$$

We use Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\intercal$, where $\mathbf{L} \in \mathbb{R}^{m\times m}$ is a lower triangular matrix with positive diagonal entries. As this decomposition is bijective as a mapping between positive-definite real matrices and $\mathbf{L}$ as described previously, we can construct any viable covariance matrix $\boldsymbol{\Sigma}$ from $\mathbf{L}$ (except matrices that are positive semi-definite but not positive definite, which is a reasonable compromise).

### 3.1 $\partial/\partial\boldsymbol{\mu}$

We begin by removing terms independent of $\boldsymbol{\mu}$:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}} = \frac{\partial}{\partial\boldsymbol{\mu}}[\mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}] - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\mu}}\mathbb{E}[V_{\mathbf{r}}(s_{i,t})] - \gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s_{i,t},a_{i,t},s')\frac{\partial}{\partial\boldsymbol{\mu}}\mathbb{E}[V_{\mathbf{r}}(s')].$$

Here

$$\frac{\partial}{\partial\boldsymbol{\mu}}\mathbb{E}[V_{\mathbf{r}}(s)] = \frac{\partial}{\partial\boldsymbol{\mu}}\iint V_{\mathbf{r}}(s)q(\mathbf{r})q(\mathbf{u})\,d\mathbf{r}\,d\mathbf{u} = \iint V_{\mathbf{r}}(s)q(\mathbf{r})\frac{\partial q(\mathbf{u})}{\partial\boldsymbol{\mu}}\,d\mathbf{r}\,d\mathbf{u} = \frac{1}{2}\mathbb{E}[V_{\mathbf{r}}(s)(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\intercal})(\mathbf{u}-\boldsymbol{\mu})]$$

by Theorem 2.8 and Lemma 1.1. Hence

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}} = \mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[V_{\mathbf{r}}(s_{i,t})(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\intercal})(\mathbf{u}-\boldsymbol{\mu})] - \gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s_{i,t},a_{i,t},s')\mathbb{E}[V_{\mathbf{r}}(s')(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\intercal})(\mathbf{u}-\boldsymbol{\mu})].$$

### 3.2 $\partial/\partial\boldsymbol{\Sigma}$

Similarly to the previous section,

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Sigma}} = \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\log|\boldsymbol{\Sigma}| - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathbb{E}[V_{\mathbf{r}}(s_{i,t})] - \gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s_{i,t},a_{i,t},s')\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathbb{E}[V_{\mathbf{r}}(s')],$$

where $\frac{\partial}{\partial\boldsymbol{\Sigma}}\log|\boldsymbol{\Sigma}| = \boldsymbol{\Sigma}^{-\intercal}$ by Petersen and Pedersen [5], and

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathbb{E}[V_{\mathbf{r}}(s)] = \iint V_{\mathbf{r}}(s)q(\mathbf{r})\frac{\partial q(\mathbf{u})}{\partial\boldsymbol{\Sigma}}\,d\mathbf{r}\,d\mathbf{u} = \frac{1}{2}\mathbb{E}[V_{\mathbf{r}}(s)(\boldsymbol{\Sigma}^{-\intercal}(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-\intercal}-\boldsymbol{\Sigma}^{-\intercal})],$$

by Theorem 2.8 and Lemma 1.1. Therefore,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Sigma}} = \frac{1}{2}\boldsymbol{\Sigma}^{-\intercal} - \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[V_{\mathbf{r}}(s_{i,t})(\boldsymbol{\Sigma}^{-\intercal}(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-\intercal}-\boldsymbol{\Sigma}^{-\intercal})] \\ - \gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s_{i,t},a_{i,t},s')\mathbb{E}[V_{\mathbf{r}}(s')(\boldsymbol{\Sigma}^{-\intercal}(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-\intercal}-\boldsymbol{\Sigma}^{-\intercal})].\end{aligned}$$

### 3.3 $\partial/\partial\lambda_j$

For $j = 0,\ldots,d$,

$$\frac{\partial\mathcal{L}}{\partial\lambda_j} = -\frac{1}{2}\frac{\partial}{\partial\lambda_j}\log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| + \mathbf{t}^\intercal\frac{\partial}{\partial\lambda_j}\left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\right]\boldsymbol{\mu} - \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial}{\partial\lambda_j}\mathbb{E}[V_{\mathbf{r}}(s_{i,t})] - \gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s_{i,t},a_{i,t},s')\frac{\partial}{\partial\lambda_j}\mathbb{E}[V_{\mathbf{r}}(s')],$$

where

$$\frac{\partial}{\partial \lambda_j} \log |\mathbf{K}_{\mathbf{u},\mathbf{u}}| = \operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right),$$
$$\frac{\partial}{\partial \lambda_j}\left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\right] = \frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} + \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}}{\partial \lambda_j} = \frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}$$

by Petersen and Pedersen [5], and

$$\frac{\partial}{\partial \lambda_j}\mathbb{E}[V_{\mathbf{r}}(s)] = \iint V_{\mathbf{r}}(s)\frac{\partial q(\mathbf{r})}{\partial \lambda_j} q(\mathbf{u})\, d\mathbf{r}\, d\mathbf{u} = \frac{1}{2}\mathbb{E}\left[V_{\mathbf{r}}(s) \operatorname{tr}\left((\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1})\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}\right)\right]$$

by Theorem 2.8 and Lemma 1.1. Thus,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda_j} =& -\frac{1}{2}\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right) + \mathbf{t}^{\intercal}\left(\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}}{\partial \lambda_j} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right)\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} \\ & -\frac{1}{2}\mathbb{E}\left[\operatorname{tr}\left((\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1})\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}\right)\sum_{i=1}^{N}\sum_{t=1}^{T} V_{\mathbf{r}}(s_{i,t}) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s_{i,t}, a_{i,t}, s')V_{\mathbf{r}}(s')\right],\end{aligned}$$

where the remaining derivatives can be found in Lemma 1.1.

# References


[1] N. A. Ahmed and D. V. Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Information Theory*, 35(3):688–692, 1989.

[2] R. Chen. The dominated convergence theorem and applications. National Cheng Kung University, 2016.

[3] H. Herrlich. *Axiom of choice*. Springer, 2006.

[4] W. Layton and M. Sussman. *Numerical linear algebra*. Lulu.com, 2014.

[5] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[6] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.

[7] H. Royden and P. Fitzpatrick. *Real Analysis*. Prentice Hall, 2010.