University of Glasgow | School of Computing Science

Variational Inference for Inverse Reinforcement Learning with Gaussian Processes

Paulius Dilkas (2146879)

22nd March 2019

ABSTRACT

The inverse reinforcement learning (IRL) problem asks us to find a reward function of a Markov decision process that explains observed behaviour. Many approaches are only able to construct reward functions as linear combinations of state features. Out of those that can handle nonlinearity, none can provide a full posterior distribution of rewards. Providing variance estimates for rewards would allow one to judge how well the model has learned its policy and discover any weak spots the model may have. We show how to perform variational inference (VI) on a Gaussian process-based IRL model in order to approximate the posterior distribution of rewards. We prove the correctness of the approach and demonstrate the model's behaviour in practice. Being able to provide full posterior probability distributions in IRL unlocks many new research frontiers ranging from integrating recent developments in VI to make the models more efficient and flexible, to developing complex reinforcement learning agents that can explicitly search for opportunities to fix their weaknesses.

1. INTRODUCTION

Imagine using a machine learning (ML) algorithm to teach a robot how to move around people so that it learns to predict where people are going and adjust its path accordingly. The ML algorithm would use data about various possible situations. But do we have enough data to ensure reasonably optimal behaviour? Perhaps the robot behaves well in most situations, but fails in less common scenarios. Can the ML model itself describe its weaknesses so that we could ensure it is exposed to sufficiently many uncommon or difficult situations?

This learning problem [11, 12] as well as many others have benefited from an approach called *inverse reinforcement learning* (IRL) (also known as inverse optimal control). IRL proposes a way to learn behaviour from *demonstrations* that typically come from human actions. More formally, the IRL problem asks us to find a reward function for a Markov decision process (MDP), where demonstrations are encoded as sets of state-action pairs.

IRL is an important problem because adjusting the reward function by hand is often unwieldy, since human behaviour often depends on many factors in complicated ways [2]. Moreover, learning the reward function rather than the policy itself makes the model more transferable to new environments—a minor change in the environment can reorganise the whole policy but only have a local effect in the reward structure [10, 15]. IRL has been used to teach helicopters how to perform tricks [1], predict taxi destinations [33], and make driving safer and more efficient by predicting pedestrian movement [34] and the driver's intentions [28].

However, most IRL models in the literature make a convenient yet unjustified assumption that the reward function can be expressed as a linear combination of features [2, 18, 32]. This assumption makes the models unable to represent many reward structures. Out of the non-linear models proposed to date, none can answer the questions posed in the first paragraph. Quite often, the models assume that rewards have no variance [15, 10]. In this paper, we show how that assumption can be lifted by switching from maximum likelihood estimation to *variational inference* (VI), i.e., we approximate the posterior distribution of the model by optimising the parameters of a simpler distribution to make it similar to the posterior. This approach can prove useful in four major ways:

1. By working with full distributions instead of point estimates, we can expect more precise reward predictions.
2. Variance estimates can be used to guide what data should be collected next, i.e., if the rewards of some states have abnormally high variance, we might want to expose the model to more data visiting those and surrounding states.
3. Variances estimates can also be used to judge whether we can trust the predictions of the model or, perhaps, the model could benefit from some adjustments or more data.
4. By adopting a more Bayesian approach, we automatically incorporate Occam's razor into the model that guards against overfitting [10].

Our main contribution is a lengthy proof in Section 5 that shows how VI can be applied to the maximum-entropy IRL model with Gaussian processes (GPs) proposed by Levine et al. [15]. We describe how we adapt the model and set up the VI problem in Section 4. Finally, in Section 6 we examine the convergence properties of our model and its ability to deduce optimal policies in practice.

2. THE PROBLEM

In this section, we introduce definitions and mathematical details relevant to the problem. We begin by formally defining what the problem is.

Definition 2.1 (MDP). *A* Markov decision process *is a set* $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathbf{r}\}$*, where* $\mathcal{S}$ *and* $\mathcal{A}$ *are sets of states and actions, respectively;* $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ *is a function* *defined so that $\mathcal{T}(s,a,s')$ is the probability of moving to state s' after taking action a in state s; $\gamma \in [0,1)$ is the discount factor; and $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ is the reward vector*[1].

Definition 2.2 (IRL). *Given an MDP without rewards $\mathcal{M} \setminus \{\mathbf{r}\}$, an $|\mathcal{S}| \times d$ feature matrix $\mathbf{X}$ (where d is the number of features), and a set of expert demonstrations $\mathcal{D} = \{\zeta_i\}_{i=1}^N$, where each demonstration $\zeta_i = \{(s_{i,t}, a_{i,t})\}_{t=1}^T$ is a multiset of state-action pairs representing optimal actions executed by an expert, find the reward function that maximises the probability of observing the demonstrations, i.e.,*

$$\arg\max_{\mathbf{r}} p(\mathcal{D} \mid \mathbf{r}).$$

The optimal (deterministic) *policy* $\pi : \mathcal{S} \to \mathcal{A}$ (i.e., a choice of actions for each state that maximises reward over time) is usually constructed by defining a *value (utility) function* $V_{\mathbf{r}} : \mathcal{S} \to \mathbb{R}$ that measures how good a state is based on the reward $\mathbf{r}$ as well as the structure of the MDP. One can then find $V_{\mathbf{r}}$ by applying the Bellman backup operator until convergence to every $s \in \mathcal{S}$ (the technique is known as *value iteration*) [26]:

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

However, we follow previous work on GP IRL [15, 10], and use a *linearly solvable* (or *maximum causal entropy*) MDP with a stochastic policy that defines probability distributions over actions (instead of suggesting a single action for each state) [32]. This type of MDP can be solved by applying the 'soft' version of the operator [15, 16]:

$$V_{\mathbf{r}}(s) \coloneqq \log \sum_{a \in \mathcal{A}} \exp\left(r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') V_{\mathbf{r}}(s') \right). \quad (1)$$

With this model, we can express the likelihood as [10, 15]

$$\begin{aligned} p(\mathcal{D} \mid \mathbf{r}) &= \prod_{i=1}^N \prod_{t=1}^T p(a_{i,t} \mid s_{i,t}) \\ &= \exp\left(\sum_{i=1}^N \sum_{t=1}^T Q_{\mathbf{r}}(s_{i,t}, a_{i,t}) - V_{\mathbf{r}}(s_{i,t}) \right), \end{aligned}$$

where

$$Q_{\mathbf{r}}(s,a) = r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') V_{\mathbf{r}}(s').$$

As we want learned rewards to generalise to previously unseen states and for states to have a notion of similarity, the IRL definition also includes d features associated with each state. In this paper, we will focus on modelling rewards as a function from feature space to $\mathbb{R}$ using a *Gaussian process*. A GP is defined as a collection of random variables, any finite number of which has a joint Gaussian distribution [22]. We write $r \sim \mathcal{GP}(0,k)$ to say that r is a GP with mean 0 and covariance function k. *Covariance functions* (also known as *kernels*) take two state feature vectors as input and quantify how similar the two states are, in a sense that we would expect high covariance scores to be associated with similar rewards.

[1]Depending on the situation, we will sometimes represent rewards as a function $r : \mathcal{S} \to \mathbb{R}$.

A common way to scale GPs to larger data sets is by selecting m points in the feature space–called *inducing points*—and focus most of the training effort on them [17]. Let $\mathbf{X_u}$ be the $m \times d$ matrix of features at inducing points, and let $\mathbf{u}$ be the rewards at those states. Then the full joint probability distribution can be factorised as

$$p(\mathcal{D}, \mathbf{u}, \mathbf{r}) = p(\mathbf{u}) \times p(\mathbf{r} \mid \mathbf{u}) \times p(\mathcal{D} \mid \mathbf{r}), \quad (2)$$

where

$$\begin{aligned} p(\mathbf{u}) &= \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{K_{u,u}}) \\ &= \frac{1}{(2\pi)^{m/2} |\mathbf{K_{u,u}}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{u}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u} \right) \\ &= \exp\left(-\frac{1}{2} \mathbf{u}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u} - \frac{1}{2} \log |\mathbf{K_{u,u}}| - \frac{m}{2} \log 2\pi \right) \end{aligned}$$

is the GP prior [22]. The GP posterior is then a multivariate Gaussian [15] defined as

$$p(\mathbf{r} \mid \mathbf{u}) = \mathcal{N}(\mathbf{r}; \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}, \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{r},\mathbf{u}}). \quad (3)$$

The matrices such as $\mathbf{K}_{\mathbf{r},\mathbf{u}}$ are called *covariance matrices* (also *kernel matrices* and *Gram matrices*) and are defined as $[\mathbf{K}_{\mathbf{r},\mathbf{u}}]_{i,j} = k(\mathbf{x}_{\mathbf{r},i}, \mathbf{x}_{\mathbf{u},j})$, where $\mathbf{x}_{\mathbf{r},i}$ and $\mathbf{x}_{\mathbf{u},j}$ denote feature vectors for the ith state in $\mathcal{S}$ and the jth inducing point, respectively [10].

Our goal is then to use VI in order to approximate $p(\mathbf{u}, \mathbf{r} \mid \mathcal{D})$. Let $q(\mathbf{u}, \mathbf{r})$ denote our approximation. VI aims to optimise this approximation by minimising the *Kullback-Leibler* (KL) divergence between the original probability distribution and our approximation. KL divergence (asymmetrically) measures the difference between two probability distributions and can be defined as [4]

$$\begin{aligned} D_{\mathrm{KL}}(q(\mathbf{u}, \mathbf{r}) \parallel p(\mathbf{u}, \mathbf{r} \mid \mathcal{D})) &= \mathbb{E}_{q(\mathbf{u},\mathbf{r})}[\log q(\mathbf{u}, \mathbf{r}) - \log p(\mathbf{u}, \mathbf{r} \mid \mathcal{D})] \\ &= \mathbb{E}_{q(\mathbf{u},\mathbf{r})}[\log q(\mathbf{u}, \mathbf{r}) - \log p(\mathcal{D}, \mathbf{u}, \mathbf{r})] \\ &\quad + \mathbb{E}_{q(\mathbf{u},\mathbf{r})}[\log p(\mathcal{D})]. \end{aligned}$$

The last term is both hard to compute and constant with respect to $q(\mathbf{u}, \mathbf{r})$ [4], so we can remove it from our optimisation objective. The negation of what remains is known as the *evidence lower bound* (ELBO) and is defined as [3, 4]

$$\begin{aligned} \mathcal{L} &= \mathbb{E}\left[\log \frac{p(\mathcal{D}, \mathbf{u}, \mathbf{r})}{q(\mathbf{u}, \mathbf{r})} \right] \\ &= \iiint \log \frac{p(\mathcal{D}, \mathbf{u}, \mathbf{r})}{q(\mathbf{u}, \mathbf{r})} q(\mathbf{u}, \mathbf{r}) \, d\mathbf{r} \, d\mathbf{u}. \end{aligned} \quad (4)$$

Thus, instead of minimising KL divergence, we focus on maximising $\mathcal{L}$ by optimising the values of parameters to be defined in Section 4. Note that in this paper all integrals should be interpreted as definite integrals over the entire sample space. Also, all expected values are with respect to $q(\mathbf{u}, \mathbf{r})$.

3. BACKGROUND

Here we introduce a few definitions and results from linear algebra, numerical analysis, and measure theory that will be used later in the paper. Namely, we will use several different vector and matrix norms, consider how an inverse of a matrix changes with a small perturbation, and use Lebesgue's dominated convergence theorem in order to justify the validity of our approach.

Definition 3.1 (Norms). *For any finite-dimensional vector* $\mathbf{x} = (x_1, \dots, x_n)^\intercal$, *its* maximum norm *(*$\ell_\infty$-norm*) is*

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

whereas its ℓ_1-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 3.2 (Perturbation Lemma [14]). *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*

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

Theorem 3.3 (Dominated Convergence Theorem [24]). *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*

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

4. THE MODEL

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. Moreover, we use $\operatorname{tr}(\mathbf{A})$ to denote its *trace* and $\operatorname{adj}(\mathbf{A})$ for its *adjugate* (or *classical adjoint*). 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.

We keep the covariance function the same as in the work by Levine et al. [15], which is a version of the automatic relevance detection kernel [15]:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \lambda_0 \exp\bigg(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{\Lambda}(\mathbf{x}_i - \mathbf{x}_j) - \mathbb{1}[i \neq j]\sigma^2 \operatorname{tr}(\mathbf{\Lambda}) \bigg).$$

Here, λ_0 is the overall 'scale' factor for how similar or distant the states are, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ is a diagonal matrix that determines the relevance of each feature (where d denotes the number of features), $\mathbb{1}$ is defined as

$$\mathbb{1}[b] = \begin{cases} 1 & \text{if } b \text{ is true} \\ 0 & \text{otherwise,} \end{cases}$$

and σ^2 is set to $10^{-2}/2$ (as the original paper noted that the value makes little difference to the performance of the algorithm [15]). We will write $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_d)^\intercal$ to refer to both λ_0 and $\mathbf{\Lambda}$ at the same time.

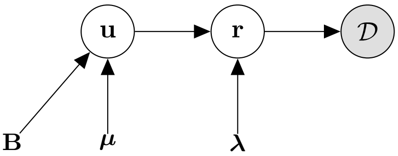

Figure 1: Our VI problem expressed as a (simplified) Bayesian network. The only observed variable (representing the demonstrations) is in a gray circle, modelled latent variables are in white circles, and the variational parameters are at the bottom.

Ideally, we would like to model $\boldsymbol{\lambda}$ with an approximating distribution. However, due to how $p(\mathbf{u})$ has $\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}$ in its expression, and the ELBO is defined as an expectation, we are unable to show that the ELBO is well-defined. More generally, we pose the following problem, which is open to the best of our knowledge:

Open Problem 4.1. *Let* $\mathbf{A}$ *be a* $n \times n$ *matrix of coefficients,* X *be a random variable, and* $\mathbf{M}$ *be an* $n \times n$ *matrix such that* $M_{i,j} = f(X, A_{i,j})$, *where* f *is an arbitrary function. Under what circumstances does* $\mathbb{E}[\mathbf{M}^{-1}]$ *exist?*

While there are some obvious examples of when the required expected value exists (e.g., $f(X, A_{i,j}) = A_{i,j}X$ for an invertible $\mathbf{A}$ and many distributions of X), it would be particularly interesting to know whether the answer is 'always'. A proof of such a result would allow us to model $\boldsymbol{\lambda}$ instead of treating it as a variational parameter, and would thus guard against overfitting. For now, $\boldsymbol{\lambda}$ will have to be treated as a variational parameter.

It remains to decide on the model for $\mathbf{u}$ and $\mathbf{r}$. As is commonly done when applying VI to GPs [6], we set

$$q(\mathbf{u}, \mathbf{r}) = q(\mathbf{u})q(\mathbf{r} \mid \mathbf{u}), \tag{5}$$

where $q(\mathbf{r} \mid \mathbf{u}) = p(\mathbf{r} \mid \mathbf{u})$ and $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}, \mathbf{\Sigma})$

Ong et al. [19] have recently suggested that, in order to make variational approximation of a multivariate Gaussian more scalable, the covariance matrix should be decomposed as $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\intercal + \mathbf{D}^2$, where $\mathbf{B}$ is a lower triangular $m \times p$ matrix with positive diagonal entries, and $\mathbf{D}$ is a diagonal matrix. Typically, we would set p so that $p \ll m$ to get an efficient approximation. However, as our goal is precision rather than scalability, we will set $p = m$ and $\mathbf{D} = \mathbf{O}_m$ in order to retain full covariance structure.

The resulting model is summarised in Figure 1. We rely on $p(\mathcal{D} \mid \mathbf{r})$ as the only link between data and the model. Since the expression for $q(\mathbf{r} \mid \mathbf{u})$ has both $\mathbf{u}$ and covariance matrices in it, $\mathbf{r}$ depends on both $\mathbf{u}$ and the parameters of the kernel, $\boldsymbol{\lambda}$. The two remaining dependencies stem from the fact that the approximating distribution for $\mathbf{u}$ is $\mathcal{N}(\boldsymbol{\lambda}, \mathbf{B}\mathbf{B}^\intercal)$.

As we want to restrict some parameters (namely, $\boldsymbol{\lambda}$ and the diagonal of $\mathbf{B}$) to positive values, we express them as exponentials and later adjust their derivatives accordingly. Specifically, we can set $\lambda_i = e^{\lambda'_i}$ and optimise λ'_i using the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \lambda'_i} = e^{\lambda'_i} \frac{\partial \mathcal{L}}{\partial \lambda_i}.$$

This way, we restrict λ_i to positive values while allowing λ'_i to range over $\mathbb{R}$.

Finally, the parameters are initialised as follows:

$$\begin{aligned} \mu_i &\sim \mathcal{U}(0,1) \quad \text{for } i = 1, \dots, m, \\ \lambda_0 &\sim \chi^2_5, \\ \lambda_i &\sim \chi^2_1 \qquad\quad \text{for } i = 1, \dots, d, \\ \operatorname{diag}(\mathbf{B}) &\sim \chi^2_4, \\ \text{the rest of } \mathbf{B} &\sim \mathcal{N}(0,1). \end{aligned}$$

The initialisation of $\boldsymbol{\mu}$ mirrors the initialisation of $\mathbf{r}$ in previous work by Levine et al. [15]. While they have constant initial values for $\boldsymbol{\lambda}$, we sample from χ^2 distributions centred around those values (5 for λ_0 and 1 for any other λ_i). The distributions for initial values of $\mathbf{B}$ are simply set to provide a reasonable spread of positive values for the diagonal, and both positive and negative values for all other entries in the matrix.

4.1 Evidence Lower Bound

In this section, we derive and simplify the ELBO for this (now fully specified) model. Note that in order to keep the derivation simple, we drop all constant terms in the expression of $\mathcal{L}$, i.e., equality is taken to mean 'equality up to an additive constant'. Also note that all expected values are with respect to $(\mathbf{u}, \mathbf{r}) \sim q(\mathbf{u}, \mathbf{r})$.

In order to derive the ELBO, let us go back to (4) and write

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

By substituting in (2) and (5), we get

$$\begin{aligned} \mathcal{L} &= \mathbb{E}[\log p(\mathbf{u}) + \log p(\mathbf{r} \mid \mathbf{u}) + \log p(\mathcal{D} \mid \mathbf{r})] \\ &\quad - \mathbb{E}[\log q(\mathbf{u}) + \log q(\mathbf{r} \mid \mathbf{u})]. \end{aligned}$$

Since $q(\mathbf{r} \mid \mathbf{u}) = p(\mathbf{r} \mid \mathbf{u})$, they cancel each other out. Also notice that

$$\begin{aligned} \mathbb{E}[\log p(\mathbf{u}) - \log q(\mathbf{u})] &= -D_{\mathrm{KL}}(q(\mathbf{u}) \parallel p(\mathbf{u})) \\ &= -\frac{1}{2}(\operatorname{tr}(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} - m \\ &\quad + \log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| - \log|\boldsymbol{\Sigma}|), \end{aligned}$$

by the definition of KL divergence between two multivariate Gaussians [7]. Hence,

$$\begin{aligned} \mathcal{L} &= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} Q_{\mathbf{r}}(s_{i,t}, a_{i,t}) - V_{\mathbf{r}}(s_{i,t})\right] \\ &\quad - \frac{1}{2}\left(\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} + \log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| - \log|\boldsymbol{\Sigma}|\right). \end{aligned}$$

Using the expressions for $Q_{\mathbf{r}}$ we get

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

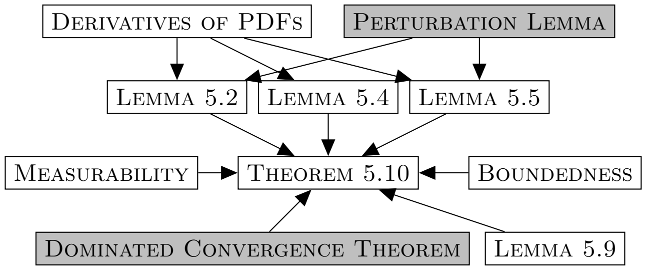

Figure 2: A graphical representation of dependencies between our theoretical results. An arrow from A to B means that A was used to prove B. Results from the literature are in gray.

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

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

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

$$\begin{aligned} \mathcal{L} &= \mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} - \mathbb{E}[v] \\ &\quad - \frac{1}{2}\left(\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} + \log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| - \log|\boldsymbol{\Sigma}|\right), \end{aligned}$$

where

$$v = \sum_{i=1}^{N}\sum_{t=1}^{T} V_{\mathbf{r}}(s_{i,t}) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s_{i,t}, a_{i,t}, s') V_{\mathbf{r}}(s').$$

5. THEORETICAL JUSTIFICATION

The typical way to optimise a quantity (the ELBO, in this case) involves computing its gradient. Unfortunately, the term $\mathbb{E}[v]$ in $\mathcal{L}$ complicates the situation. The goal of this section is to show how Theorem 3.3 can be applied to our model in order to derive the gradient anyway[2]. After showing that the theorem applies to our situation, we can estimate $\nabla \mathbb{E}[v]$ with

$$\begin{aligned} \nabla \mathbb{E}[v] &= \nabla \iint q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u}) \, d\mathbf{r} \, d\mathbf{u} \\ &= \iint \nabla[v q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u})] \, d\mathbf{r} \, d\mathbf{u} \\ &= \iint \frac{\nabla[v q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u})]}{q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u})} q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u}) \, d\mathbf{r} \, d\mathbf{u} \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \frac{\nabla[v q(\mathbf{r}_s \mid \mathbf{u}_s) q(\mathbf{u}_s)]}{q(\mathbf{r}_s \mid \mathbf{u}_s) q(\mathbf{u}_s)}, \end{aligned}$$

which can be computed by drawing S samples $\{(\mathbf{u}_s, \mathbf{r}_s)\}_{s=1}^{S}$ from $q(\mathbf{u}, \mathbf{r})$.

Our main goal is Theorem 5.10, which allows us to move differentiation inside the integral. In order to prove it, we use a number of intermediate results. We start by stating a

[2]This technique is inspired by black box VI [21], but takes a more detailed look at the problem and requires significantly more work to prove correctness.

few derivatives of probability density functions (PDFs) and covariance matrices, and bound their values with some easy-to-deal-with polynomials. We then provide a sketch proof of the measurability of MDP value functions, which is non-obvious due to their non-trivial definition. Afterwards, we establish bounds for the value functions, and, after another quick lemma, tackle the main proof of this paper. See Figure 2 for an overview of how these results fit together.

Before that, however, we define a few extra variables in order to simplify expressions of derivatives:

$$\begin{aligned}
&\mathbf{U} = (\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\intercal,\\
&\mathbf{S} = \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1},\\
&\boldsymbol{\Gamma} = \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{S}\mathbf{K}_{\mathbf{r},\mathbf{u}},\\
&\mathbf{R} = \mathbf{S}\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}}{\partial \lambda_i} - \frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{r}}}{\partial \lambda_i} + \left(\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal}{\partial \lambda_i} - \mathbf{S}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}\right)\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}},\\
&Q = (\mathbf{u}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}).
\end{aligned}$$

Also note that throughout this section the word 'constant' means 'constant with respect to $\mathbf{u}$ and $\mathbf{r}$'.

LEMMA 5.1 (DERIVATIVES OF PDFS).

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

2. (a) $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}} = \frac{1}{2} q(\mathbf{u})(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1})$.

(b) $\frac{\partial q(\mathbf{u})}{\partial \mathbf{B}} = q(\mathbf{u})(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{B}$.

3. For $i = 0, \dots, d$,

(a)

$$\begin{aligned}
\frac{\partial q(\mathbf{r}\mid\mathbf{u})}{\partial \lambda_i} = \frac{1}{2} q(\mathbf{r}\mid\mathbf{u})(&|\boldsymbol{\Gamma}|^{-1}\operatorname{tr}(\mathbf{R}\operatorname{adj}(\boldsymbol{\Gamma}))\\
&- (\mathbf{r}-\mathbf{S}\mathbf{u})^\intercal \boldsymbol{\Gamma}^{-1}\mathbf{R}\boldsymbol{\Gamma}^{-1}(\mathbf{r}-\mathbf{S}\mathbf{u})).
\end{aligned}$$

(b) For any covariance matrix $\mathbf{K}$,

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

where

$$L_{j,k} = k(\mathbf{x}_j, \mathbf{x}_k)\left(-\frac{1}{2}(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2\right).$$

LEMMA 5.2. *Let $i \in \{0, \dots, d\}$ and $\epsilon > 0$ be arbitrary. Furthermore, let $c : \mathbb{R}^{|S|} \times \mathbb{R}^m \to (\lambda_i - \epsilon, \lambda_i + \epsilon) \subset \mathbb{R}$ be a function with a codomain arbitrarily close to λ_i. Then*

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

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

PROOF. Remember that

$$\begin{aligned}
\frac{\partial q(\mathbf{r}\mid\mathbf{u})}{\partial \lambda_i} = \frac{1}{2} q(\mathbf{r}\mid\mathbf{u})(&|\boldsymbol{\Gamma}|^{-1}\operatorname{tr}(\mathbf{R}\operatorname{adj}(\boldsymbol{\Gamma}))\\
&- (\mathbf{r}-\mathbf{S}\mathbf{u})^\intercal \boldsymbol{\Gamma}^{-1}\mathbf{R}\boldsymbol{\Gamma}^{-1}(\mathbf{r}-\mathbf{S}\mathbf{u})).
\end{aligned}$$

by Lemma 5.1. Let $\mathbf{K}$ be any covariance matrix and

$$\mathbf{A} = \frac{1}{\lambda_0}\mathbf{K}.$$

First, we can easily deduce that

$$\mathbf{K}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} \to \mathbf{K} \tag{6}$$

and

$$\left.\frac{\partial \mathbf{K}}{\partial \lambda_i}\right|_{\lambda_i = c(\mathbf{r},\mathbf{u})} \to \frac{\partial \mathbf{K}}{\partial \lambda_i} \tag{7}$$

as $\epsilon \to 0$. For (6), the result is obvious if $i = 0$. Otherwise, it follows from the continuity of the exponential function. For (7), observe that if $i = 0$, then

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

as an expression has no λ_0 in it, so

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

Finally, if $i > 0$, then each element of

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

is a constant multiple of the corresponding element of

$$\mathbf{K}|_{\lambda_i = c(\mathbf{r},\mathbf{u})},$$

so the same reasoning as for (6) applies.

Now, we will show that $\mathbf{K}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ exists and

$$\lim_{\epsilon \to 0} \mathbf{K}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \mathbf{K}.$$

If $i = 0$, then $\mathbf{K}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = c(\mathbf{r},\mathbf{u})\mathbf{A}$. Therefore[3],

$$\begin{aligned}
\mathbf{K}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} &= \frac{1}{c(\mathbf{r},\mathbf{u})}\mathbf{A}^{-1}\\
&\to \frac{1}{\lambda_0}\mathbf{A}^{-1} = \frac{1}{\lambda_0}\left(\frac{1}{\lambda_0}\mathbf{K}\right)^{-1} = \mathbf{K}^{-1}
\end{aligned}$$

as $\epsilon \to 0$. For $i > 0$, let

$$S = \sum_{n \in \{1,\dots,d\}\setminus\{i\}} \frac{\lambda_n}{2}(x_{j,n} - x_{k,n})^2 + \mathbb{1}[j \neq k]\sigma^2\lambda_n$$

and $\delta = c(\mathbf{r},\mathbf{u}) - \lambda_i$ so that $c(\mathbf{r},\mathbf{u}) = \lambda_i + \delta$, and $\lim_{\epsilon \to 0}\delta = 0$.

[3]Note that since $\lambda_0 \neq 0$, $c(\mathbf{r},\mathbf{u}) \neq 0$ for small enough ϵ.

Then,

$$k(\mathbf{x}_j, \mathbf{x}_k)|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \lambda_0 \exp\Bigg(-\frac{1}{2} c(\mathbf{r},\mathbf{u})(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 c(\mathbf{r},\mathbf{u}) - S \Bigg)$$
$$= \lambda_0 \exp\Bigg(-\frac{1}{2}(\lambda_i + \delta)(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 (\lambda_i + \delta) - S \Bigg)$$
$$= \lambda_0 \exp\Bigg(-\frac{1}{2}(\mathbf{x}_j - \mathbf{x}_k)^\intercal \mathbf{\Lambda}(\mathbf{x}_j - \mathbf{x}_k) - \mathbb{1}[j \neq k]\sigma^2 \operatorname{tr}(\mathbf{\Lambda}) - \frac{\delta}{2}(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \delta \Bigg)$$
$$= k(\mathbf{x}_j, \mathbf{x}_k) \exp\Bigg(-\frac{\delta}{2}(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \delta \Bigg)$$
$$= k(\mathbf{x}_j, \mathbf{x}_k) + k(\mathbf{x}_j, \mathbf{x}_k) \Bigg(\exp\Bigg(-\frac{\delta}{2}(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \delta \Bigg) - 1 \Bigg).$$

Thus, we can express $\mathbf{K}|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ as $\mathbf{K}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \mathbf{K} + \mathbf{E}$, where $\mathbf{E}$ is defined by

$$E_{j,k} = k(\mathbf{x}_j, \mathbf{x}_k) \Bigg(\exp\Bigg(-\frac{\delta}{2}(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \delta \Bigg) - 1 \Bigg).$$

By this definition,

$$\lim_{\epsilon \to 0} E_{j,k} = 0.$$

Then, since $\mathbf{K}$ is invertible,

$$\mathbf{K}^{-1}_{\lambda_i = c(\mathbf{r},\mathbf{u})} \to \mathbf{K}^{-1}$$

by continuity of $\mathbf{A} \mapsto \mathbf{A}^{-1}$.

This is enough to prove constant upper and lower bounds on $\mathbf{S}$, $\mathbf{\Gamma}$, and $\mathbf{R}$ (all with λ_i replaced with $c(\mathbf{r},\mathbf{u})$), which means that $(\mathbf{r} - \mathbf{S}\mathbf{u})^\intercal \mathbf{\Gamma}^{-1} \mathbf{R} \mathbf{\Gamma}^{-1} (\mathbf{r} - \mathbf{S}\mathbf{u})|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ has upper and lower bounds in $\mathbb{R}_2[\mathbf{u}]$. Furthermore, having convergence results for arbitrary covariance matrices and their inverses means that

$$\lim_{\epsilon \to 0} \mathbf{\Gamma}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \mathbf{\Gamma}.$$

Hence,

$$\lim_{\epsilon \to 0} \det(\mathbf{\Gamma})|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \det(\mathbf{\Gamma}).$$

Assuming that $\mathbf{\Gamma}$ is invertible so that $q(\mathbf{r} \mid \mathbf{u})$ exists,

$$\det(\mathbf{\Gamma})|_{\lambda_i = c(\mathbf{r},\mathbf{u})} \neq 0$$

for small enough ϵ, and, thus, $\det(\mathbf{\Gamma})^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ exists and is bounded.

Recall that

$$\mathbf{\Gamma} = \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{S}\mathbf{K}_{\mathbf{r},\mathbf{u}} = \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{r},\mathbf{u}}.$$

We have already demonstrated that for $\mathbf{K} \in \{\mathbf{K}_{\mathbf{r},\mathbf{r}}, \mathbf{K}_{\mathbf{r},\mathbf{u}}, \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\}$,

$$\lim_{\epsilon \to 0} \mathbf{K}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \mathbf{K}.$$

We can use a general fact that if $\lim_{x \to 0} f(x) = f$, then $f(x) = f + g(x)$ for some function g such that $\lim_{x \to 0} g(x) = 0$ to define matrices $\mathbf{E}_{\mathbf{r},\mathbf{r}}$, $\mathbf{E}_{\mathbf{r},\mathbf{u}}$, and $\mathbf{E}_{\mathbf{u},\mathbf{u}}$ such that

$$\begin{aligned}
\mathbf{K}_{\mathbf{r},\mathbf{r}}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} &= \mathbf{K}_{\mathbf{r},\mathbf{r}} + \mathbf{E}_{\mathbf{r},\mathbf{r}}, & & \mathbf{E}_{\mathbf{r},\mathbf{r}} \to \mathbf{O}_{|\mathcal{S}|}, \\
\mathbf{K}_{\mathbf{r},\mathbf{u}}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} &= \mathbf{K}_{\mathbf{r},\mathbf{u}} + \mathbf{E}_{\mathbf{r},\mathbf{u}}, & \text{and} \quad & \mathbf{E}_{\mathbf{r},\mathbf{u}} \to \mathbf{O}_{|\mathcal{S}|,m}, \\
\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} &= \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} + \mathbf{E}_{\mathbf{u},\mathbf{u}}, & & \mathbf{E}_{\mathbf{u},\mathbf{u}} \to \mathbf{O}_m.
\end{aligned}$$

as $\epsilon \to 0$. Then, $\mathbf{\Gamma}|_{\lambda_i = c(\mathbf{r},\mathbf{u})} = \mathbf{\Gamma} + \mathbf{E}$, where

$$\begin{aligned}
\mathbf{E} &= \mathbf{E}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{E}_{\mathbf{r},\mathbf{u}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{E}_{\mathbf{u},\mathbf{u}} (\mathbf{K}_{\mathbf{r},\mathbf{u}} + \mathbf{E}_{\mathbf{r},\mathbf{u}}) \\
&\quad - \mathbf{E}_{\mathbf{r},\mathbf{u}}^\intercal (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} + \mathbf{E}_{\mathbf{u},\mathbf{u}})(\mathbf{K}_{\mathbf{r},\mathbf{u}} + \mathbf{E}_{\mathbf{r},\mathbf{u}}) \\
&\quad \to \mathbf{O}_{|\mathcal{S}|}
\end{aligned}$$

as $\epsilon \to 0$. Thus, Lemma 3.2 shows that $\mathbf{\Gamma}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ exists and provides constant bounds on its elements.

Since we already know that $\mathbf{\Gamma}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ and $\det(\mathbf{\Gamma})|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$ are bounded, so is $\operatorname{adj}(\mathbf{\Gamma}) = \det(\mathbf{\Gamma})\mathbf{\Gamma}^{-1}|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$. Thus, we have constant bounds on $|\mathbf{\Gamma}|^{-1} \operatorname{tr}(\mathbf{R} \operatorname{adj}(\mathbf{\Gamma}))|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$, which means that

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

has the required bounds. □

Remark 5.3. *In order to find a derivative such as* $\frac{\partial q(\mathbf{u})}{\partial \mu_i}$, *we can find* $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}$ *and simply take the ith element. A similar line of reasoning applies to matrices as well. Thus, we only need to consider derivatives with respect to* $\boldsymbol{\mu}$ *and* $\mathbf{\Sigma}$.

Lemma 5.4. *Let* $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (a, b) \subset \mathbb{R}$ *be an arbitrary bounded function. Then, for* $i = 1, \dots, m$, *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 5.1,

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

where $\mathbf{c}(\mathbf{r},\mathbf{u}) = (\mu_1, \dots, \mu_{i-1}, c(\mathbf{r},\mathbf{u}), \mu_{i+1} \dots, \mu_m)^\intercal$. Since $c(\mathbf{r},\mathbf{u})$ is bounded and $\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-\intercal}$ 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 5.5. *Let* $i, j = 1, \dots, m$, *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*

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

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 5.1,

$$\frac{\partial q(\mathbf{u})}{\partial \mathbf{\Sigma}}\bigg|_{\Sigma_{i,j}=c(\mathbf{r},\mathbf{u})} = \frac{1}{2}q(\mathbf{u})(\mathbf{C}(\mathbf{r},\mathbf{u})^{-\intercal}\mathbf{U}\mathbf{C}(\mathbf{r},\mathbf{u})^{-\intercal} - \mathbf{C}(\mathbf{r},\mathbf{u})^{-\intercal}),$$

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 $\mathbf{C}(\mathbf{r},\mathbf{u}) = \mathbf{\Sigma} + \mathbf{E}(\mathbf{r},\mathbf{u})$, 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 3.2 to $\mathbf{\Sigma}$ 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

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

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 $\mathbf{U} = (\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})^\intercal$ is in $\mathbb{R}_2[\mathbf{u}]$, and the elements of $\mathbf{C}(\mathbf{r},\mathbf{u})^{-1}$ are bounded, the desired result follows. □

REMARK 5.6. *MDP values are characterised by both a state and a reward function/vector. In this section, 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, $V_\mathbf{r}(s) \in \mathbb{R}$. *Given a reward vector, the function* $V(s)$ *computes the values of all states and returns the value of state* s.

PROPOSITION 5.7 (MEASURABILITY). *MDP value functions* $V(s) : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ *(for* $s \in \mathcal{S}$*) are Lebesgue measurable.*

PROOF. 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

$$V_\mathbf{r}(s) = \log \sum_{a \in \mathcal{A}} \exp\left(r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') V_\mathbf{r}(s')\right) \quad (8)$$

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

$$\left\{ \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \;\middle|\; \begin{array}{l} V_\mathbf{r}(s_0) \in (-\infty, \alpha); \\ V_\mathbf{r}(s) \in \mathbb{R} \text{ for all } s \in \mathcal{S} \setminus \{s_0\}; \\ (8) \text{ is satisfied by all } s \in \mathcal{S} \end{array} \right\}$$

is measurable. Since this set can be constructed in Zermelo-Fraenkel set theory *without* the axiom of choice, it is measurable [9], which proves that $V(s)$ is a measurable function for any $s \in \mathcal{S}$. □

PROPOSITION 5.8 (BOUNDEDNESS). *If the initial values of the MDP value function satisfy the following bound, then the bound remains satisfied throughout value iteration:*

$$|V_\mathbf{r}(s)| \le \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}. \quad (9)$$

PROOF. We begin by considering (9) without taking the absolute value of $V_\mathbf{r}(s)$, i.e.,

$$V_\mathbf{r}(s) \le \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}, \quad (10)$$

and assuming that the initial values of $\{V_\mathbf{r}(s) \mid s \in \mathcal{S}\}$ already satisfy (10). Recall that for each $s \in \mathcal{S}$, the value of $V_\mathbf{r}(s)$ is updated by applying (1). Note that both log and exp are increasing functions, $\gamma > 0$, and the $\mathcal{T}$ function gives a probability (a non-negative number). Thus

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

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

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

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

The proof for

$$V_\mathbf{r}(s) \ge \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{\gamma - 1} \quad (11)$$

follows the same argument until we get to

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

where we use the fact that $r(s) \ge -\|\mathbf{r}\|_\infty - 2\log|\mathcal{A}|$. Combining (10) and (11) gives (9). □

LEMMA 5.9.

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

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

PROOF. Since $\|\mathbf{r}\|_\infty \le \|\mathbf{r}\|_1$,

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

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 $\xi_i = \left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\right]_i$, $\sigma_i = \sqrt{[\mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}}]_{i,i}}$[4], and Φ is the cumulative distribution function of the standard Gaussian. Furthermore,

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

as σ_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 the desired result. □

Our main theorem is a specialised version of an integral differentiation result by Timoney [27].

THEOREM 5.10. *Whenever the derivative exists,*

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

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

PROOF. Let

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

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 $t_n \ne t$ for all n. We want to show that

$$F'(t) = \lim_{n\to\infty} \frac{F(t_n) - F(t)}{t_n - t} = \iint \left.\frac{\partial f}{\partial t}\right|_{(\mathbf{r},\mathbf{u},t)} d\mathbf{r}\, d\mathbf{u}. \quad (12)$$

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

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

(12) follows from Theorem 3.3 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, $\mathbf{r}$, $\mathbf{u}$,

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

The MDP value function is measurable by Proposition 5.7. The result of multiplying or adding measurable functions (e.g., probability density functions) 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 $q(\mathbf{r} \mid \mathbf{u})$. Then

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

since PDFs are non-negative. An upper bound for $|V_\mathbf{r}(s)|$ is given by Proposition 5.8, while

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

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

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

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 5.2, 5.4, and 5.5, which gives us two polynomials $p_1(\mathbf{u}), p_2(\mathbf{u}) \in \mathbb{R}_2[\mathbf{u}]$ such that

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

Then

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

We can now apply Lemma 5.9, which allows us to integrate out $\mathbf{r}$, and we are left with showing the existence of

$$\int \left(a + \|\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\|_1\right) \max\{|p_1(\mathbf{u})|, |p_2(\mathbf{u})|\} q(\mathbf{u})\, d\mathbf{u}, \quad (13)$$

where a is a constant. The integral

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

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 (13) as well. □

[4] 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.

6. EVALUATION

In order to fully understand the model's behaviour, we focus on a three-state MDP where the agent can deterministically move from any state to any other state. More formally, we set $\mathcal{S} = \{s_1, s_2, s_3\}$, $\mathcal{A} = \{a_1, a_2\}$,

$$\mathcal{T}(s_1, a_1, s_2) = 1, \quad \mathcal{T}(s_1, a_2, s_3) = 1,$$
$$\mathcal{T}(s_2, a_1, s_1) = 1, \quad \mathcal{T}(s_1, a_2, s_3) = 1,$$
$$\mathcal{T}(s_3, a_1, s_1) = 1, \quad \mathcal{T}(s_1, a_2, s_2) = 1,$$

all other values of $\mathcal{T}$ to zero, and $\gamma = 0.9$. We also set the inducing points to be equal to the three states in $\mathcal{S}$, add a single feature $f : \mathcal{S} \to \mathbb{R}$ such that

$$f(s_1) = 1, \quad f(s_2) = 2, \quad f(s_3) = 3,$$

and create two demonstrations $\zeta_1 = \{(s_1, a_1)\}$ and $\zeta_2 = \{(s_3, a_2)\}$ that correspond to moving from s_1 and s_3 to s_2. Therefore, we would expect the reward of s_2 to be higher than the other two rewards in order to reflect this.

Unfortunately, we are not able to use $\frac{\partial \mathcal{L}}{\partial \mathbf{B}}$ in order to optimise $\mathbf{B}$. We illustrate the problem in Figure 3. On the left side of the figure, we plot how $q(\mathbf{u})$ behaves as a function of a diagonal and a non-diagonal element of $\mathbf{B}$. Both functions have maximum values that can be attained by following the corresponding derivatives. However, when these derivatives are used to estimate $\frac{\partial \mathbb{E}[v]}{\partial \mathbf{B}}$, the resulting derivatives no longer match their corresponding functions, although the functions themselves still have optimal values: at or below 1 for $B_{1,1}$ and at 0 for $B_{2,1}$. This leads us to consider two restrictions of the model in order to investigate its convergence behaviour:

- In **Scenario 1**, we remove $\frac{\partial \mathbb{E}[v]}{\partial \mathbf{B}}$ from $\frac{\partial \mathcal{L}}{\partial \mathbf{B}}$. This essentially optimises $\mathbf{B}$ to match $\mathbf{K}_{\mathbf{u},\mathbf{u}}$, because then $\frac{\partial \mathcal{L}}{\partial \mathbf{B}}$ becomes $-\frac{\partial}{\partial \mathbf{B}} D_{\mathrm{KL}}(q(\mathbf{u}) \parallel p(\mathbf{u}))$, i.e., we are optimising $\mathbf{B}$ to minimise the difference between the prior and the posterior of $\mathbf{u}$.

- In **Scenario 2**, we set $\mathbf{B} = \mathbf{I}_m$, and do not optimise it at all.

We plot how $\mathcal{L}$ as well as policies $\pi(a_1 \mid s_1)$, $\pi(a_2 \mid s_3)$, and $\pi(a_1 \mid s_2)$ converge over a number of iterations in Figure 4. The first two policies correspond to actions taken in the set of demonstrations $\mathcal{D}$, so we would expect to see their probabilities converge to values above 0.5. Note that due to the stochastic nature of our MDP model, we do not expect to see any probability reach exactly 1. The third policy, however, has no relevant data in $\mathcal{D}$, so the maximum causal entropy framework would put the probability at around 0.5.

As we ran the algorithm with two stopping conditions—300 iterations and the ℓ_1-norm of the change in parameter values being below 0.01—note that the algorithm terminated early in Scenario 2 but not in Scenario 1, although there is clear convergence in both cases. An important difference is that $\pi(a_1 \mid s_1)$ and $\pi(a_2 \mid s_3)$ converge closer and closer to 1 with the slightly-restricted model of Scenario 1, but stabilise at 0.6 in Scenario 2 due to an important part of the model being completely fixed. In both cases, $\pi(a_1 \mid s_2)$ converges to 0.5 as expected.

Figure 5 shows how the parameters of the model converge in both scenarios. The algorithm seems to converge fine in Scenario 2, but most of the parameters fail to stabilise in Scenario 1. Most importantly, the diagonal values of the $\mathbf{B}$ matrix diverge to positive infinity, leading to higher variance in ELBO estimates seen in Figure 4. The non-diagonal values, however, converge just fine. Both λ_0 and λ_1 also diverge to positive infinity, while the elements of $\boldsymbol{\mu}$ diverge, but in 'the right' direction: as μ_2 increases while μ_1 and μ_3 decrease, the policies in Figure 4 converge to their optimal values.

This leaves us with two models: one that converges to reasonable-but-suboptimal values, and one that diverges to infinite variance but also provides correct policies.

7. RELATED WORK

The IRL problem itself was originally proposed by Russell in 1998 [25]. Most of the early approaches had the aforementioned reward linearity assumption. One of the first papers on the subject by Ng and Russell [18] introduced several linear programming algorithms and identified an important issue: there are typically many reward functions that can explain the data equally well. This problem was solved by Ziebart et al. [32] with the introduction of IRL based on the principles of maximum causal entropy in a linearly-solvable MDP.

Levine et al. [15] were the first to lift the linearity assumption without imposing additional restrictions on the problem. They do, however, model rewards as having no variance—our work removes this restriction without any compromises.

Recently, Jin et al. [10] have adapted the model proposed by Levine et al. [15] to use deep GPs, harnessing the power of deep learning to make the model less dependent on what features are provided. Although they use VI, their approximating distribution for rewards at inducing points is simply the Dirac δ function, which is essentially equivalent to the assumption of no variance.

An alternative to GPs for modelling nonlinear functions is, of course, neural networks. Wulfmeier et al. [31] have shown how they can be used in the IRL setting. While this approach benefits from constant-time inference and the ability to learn complex features from data, neural networks often need significantly more data for the weights across all layers to stabilise.

7.1 Variational Inference

Since the focus of our work was on proving feasibility rather than ensuring performance, we simply used a combination of Gaussians for our variational approximation. However, while VI was initially focused on approximating distributions using simplistic models where all variables are independent [4], the last few years have brought many advances in approximating more complex distributions and greatly reducing the computational complexity of the task. Adapting some of them to IRL is a nontrivial but highly valuable undertaking.

For approximating complex distributions, Rezende and Mohamed [23] suggest using *normalising flows*, i.e., a collection of invertible functions—parametrised by additional variational parameters—that are applied to latent variables. A major challenge in applying their work to IRL is related to variational parameters being used to compute the MDP value function, i.e., how can we take the derivative of the value function with respect to such a variational parameter? Alternatively, perhaps one can construct a different model that would make this question moot.

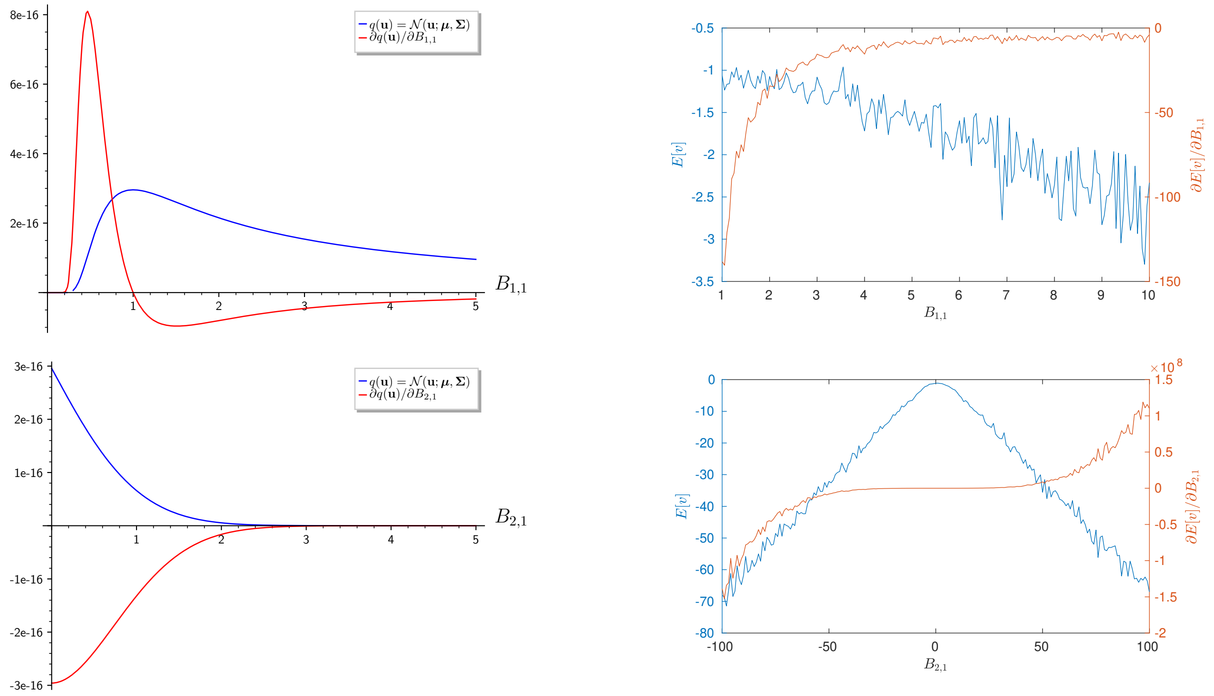

Figure 3: How a well-defined problem with a correct solutions becomes a well-defined problem with an incorrect solution. In each plot, the function we are trying to optimise is in blue, and its derivative is in red. The plots on the left are for $q(\mathbf{u})$ and its derivatives with respect to $B_{1,1}$ (at the top) and $B_{2,1}$ (at the bottom), whereas the plots on the right are for $\mathbb{E}[v]$ and its corresponding derivatives.

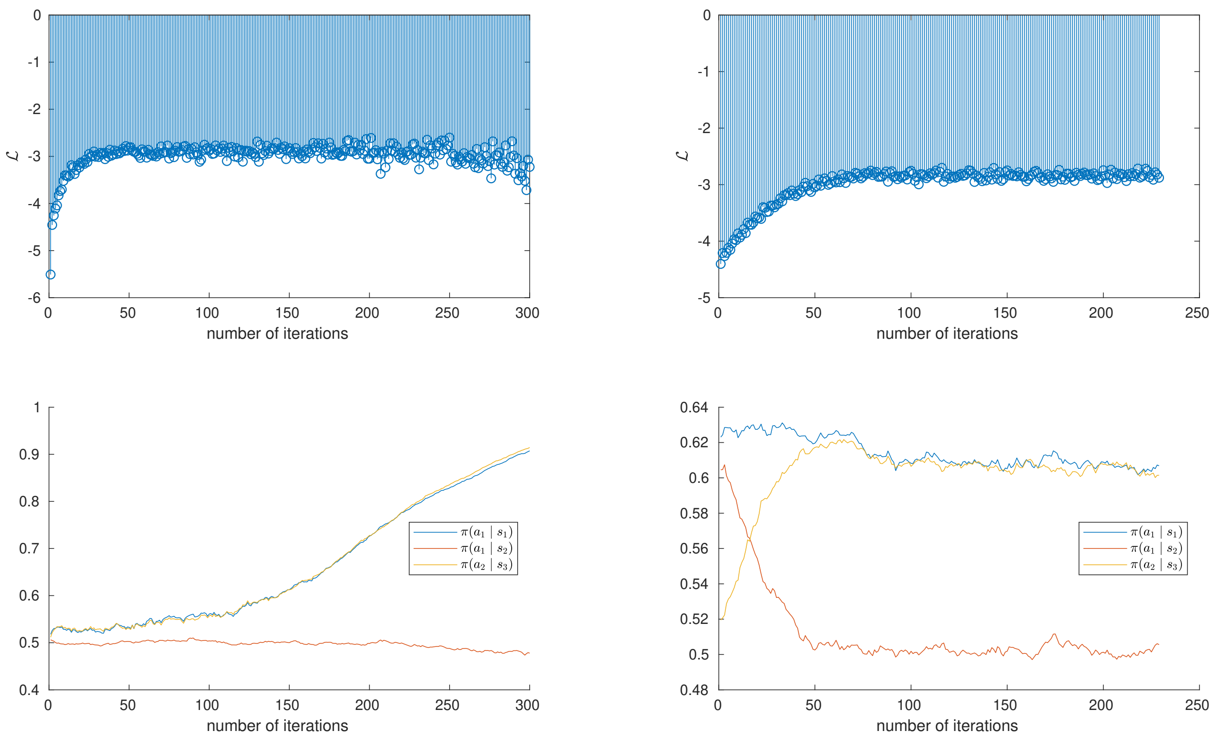

Figure 4: The convergence of $\mathcal{L}$ (at the top) and several example policies (at the bottom) over a number of iterations for Scenario 1 on the left and Scenario 2 on the right.

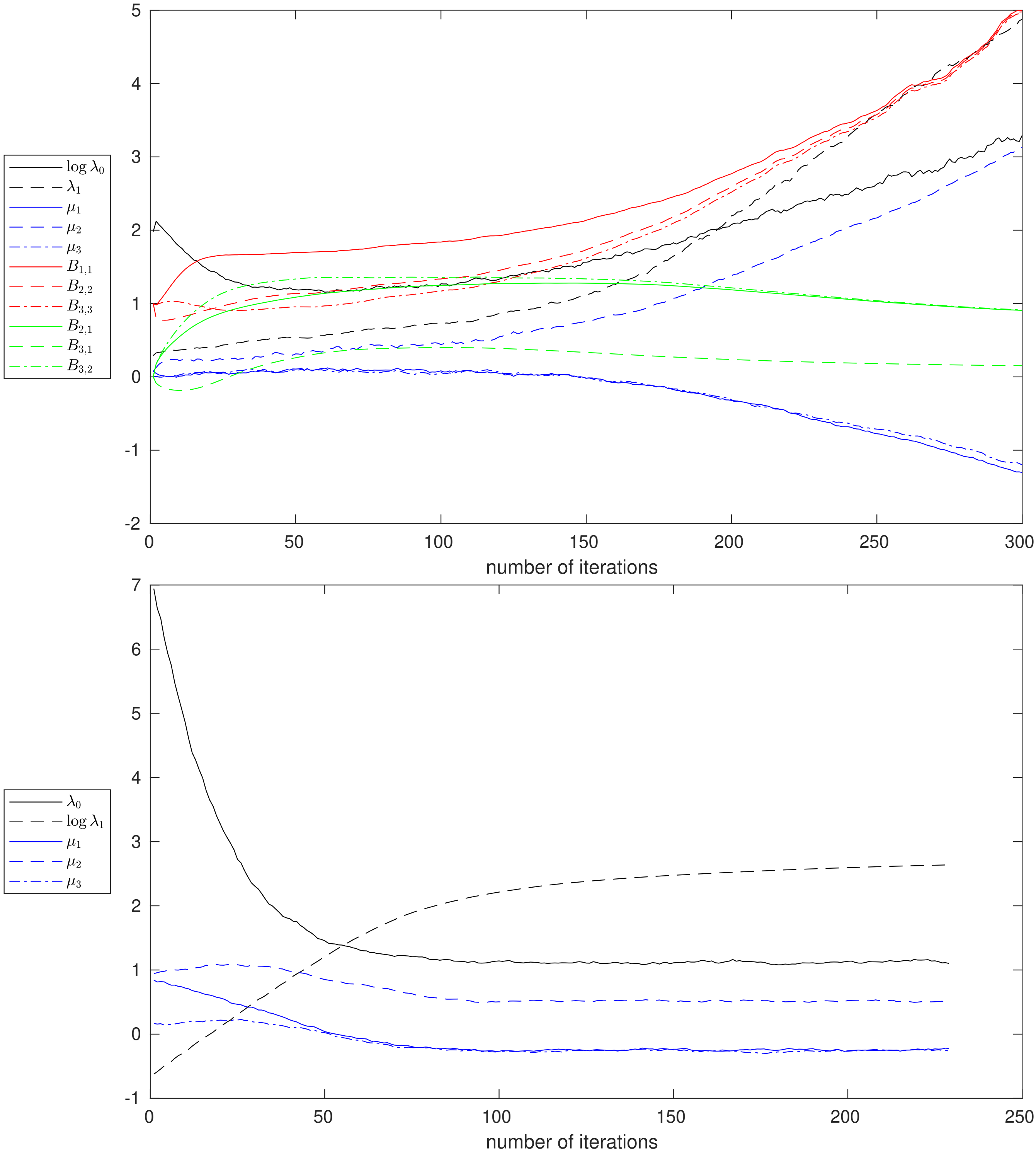

Figure 5: Convergence of all optimised parameters for Scenario 1 at the top and Scenario 2 at the bottom. In order to represent different variables on the same scale, some variables have been log-transformed. Colours denote which vector or matrix each scalar comes from: black for $\boldsymbol{\lambda}$, blue for $\boldsymbol{\mu}$, red for diagonal elements of B, and green for its non-diagonal elements.

Another approach to flexibility in modelling could come from considering different GP kernels. For instance, Wilson and Adams [29] show how all *stationary* (i.e., invariant to translations) kernels can be generated (or at least approximated) from a mixture of Gaussians in their spectral representation using Bochner's theorem [5, 30]. It looks promising to combine these kernels with the variational Fourier features approach by Hensman et al. [8] that leverages the same spectral representations for efficient VI.

8. CONCLUSIONS

Reasonable results in convergence. More variables/information. Fewer assumptions.

We show how to avoid the deterministic training conditional assumption.

8.1 Further Work

An interesting extension to our work would be to consider IRL in the context of a reinforcement learning (RL) agent. Suppose we have an agent whose purpose is to learn optimal behaviour from observing other agents using IRL. It could then take reward variance estimates into account when choosing what states to visit next. It would have to handle the balance between exploration and exploitation similarly to many RL agents, but the information about rewards would come from observing (presumably near-optimal) behaviour exhibited by other agents rather than directly from the environment.

It is also worth noting the approach presented in this paper requires solving S MDPs for every iteration of optimising the parameters (where S is the number of samples drawn from $q(\mathbf{u}, \mathbf{r})$). There are at least two ways to reduce or eliminate this performance bottleneck:

- The MDP value function could be approximated, allowing for some minor mistakes in the resulting policy.
- Perhaps there is a good way to use information about previously computed value functions for similar rewards to hasten the current computation. One simple way to do this would be by initialising the current values to the optimal values of the previous MDP value function computation.

Finally, building on the idea that variance estimates can be used to judge whether the model has learned optimal policy, an interesting question for MDP (or, perhaps, dynamical systems) research would be: how much does a reward have to change in order to affect the deterministic policy? A simple answer to this question would allow us to use variance estimates in order to quantify the model's confidence regarding optimal behaviour.

9. REFERENCES

[1] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1–8. MIT Press, 2006.

[2] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In C. E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

[3] C. M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] S. Bochner. *Lectures on Fourier integrals*. Princeton University Press, 1959.

[6] C. Cheng and B. Boots. Variational inference for Gaussian process models with linear complexity. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5190–5200, 2017.

[7] J. Duchi. Derivations for linear algebra and optimization. Stanford University.

[8] J. Hensman, N. Durrande, and A. Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18:151:1–151:52, 2017.

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

[10] M. Jin, A. C. Damianou, P. Abbeel, and C. J. Spanos. Inverse reinforcement learning via deep Gaussian process. In G. Elidan, K. Kersting, and A. T. Ihler, editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.

[11] B. Kim and J. Pineau. Socially adaptive path planning in human environments using inverse reinforcement learning. *I. J. Social Robotics*, 8(1):51–66, 2016.

[12] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *I. J. Robotics Res.*, 35(11):1289–1307, 2016.

[13] S. Laue, M. Mitterreiter, and J. Giesen. Computing higher order derivatives of matrix and tensor expressions. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2755–2764, 2018.

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

[15] S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with Gaussian processes. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a*

meeting held 12-14 December 2011, Granada, Spain., pages 19–27, 2011.
[16] S. Levine, Z. Popovic, and V. Koltun. Supplementary material: Nonlinear inverse reinforcement learning with Gaussian processes. `http://graphics.stanford.edu/projects/gpirl/gpirl_supplement.pdf`, December 2011.
[17] H. Liu, Y. Ong, X. Shen, and J. Cai. When Gaussian process meets big data: A review of scalable GPs. *CoRR*, abs/1807.01065, 2018.
[18] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 663–670. Morgan Kaufmann, 2000.
[19] V. M.-H. Ong, D. J. Nott, and M. S. Smith. Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27(3):465–478, 2018.
[20] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.
[21] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 814–822. JMLR.org, 2014.
[22] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press, 2006.
[23] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1530–1538. JMLR.org, 2015.
[24] H. Royden and P. Fitzpatrick. *Real Analysis.* Prentice Hall, 2010.
[25] S. J. Russell. Learning agents for uncertain environments (extended abstract). In P. L. Bartlett and Y. Mansour, editors, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998.*, pages 101–103. ACM, 1998.
[26] S. J. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach (3. internat. ed.).* Pearson Education, 2010.
[27] R. M. Timoney. The dominated convergence theorem and applications. Trinity College Dublin, `https://www.maths.tcd.ie/~richardt/MA2224/MA2224-ch4.pdf`, March 2018.
[28] A. Vogel, D. Ramachandran, R. Gupta, and A. Raux. Improving hybrid vehicle fuel efficiency using inverse reinforcement learning. In J. Hoffmann and B. Selman, editors, *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.* AAAI Press, 2012.
[29] A. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1067–1075, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
[30] R. Woodard. Interpolation of spatial data: Some theory for kriging. *Technometrics*, 42(4):436–437, 2000.
[31] M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.
[32] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.
[33] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell. Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In H. Y. Youn and W. Cho, editors, *UbiComp 2008: Ubiquitous Computing, 10th International Conference, UbiComp 2008, Seoul, Korea, September 21-24, 2008, Proceedings*, volume 344 of *ACM International Conference Proceeding Series*, pages 322–331. ACM, 2008.
[34] B. D. Ziebart, N. D. Ratliff, G. Gallagher, C. Mertz, K. M. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. S. Srinivasa. Planning-based prediction for pedestrians. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 11-15, 2009, St. Louis, MO, USA*, pages 3931–3936. IEEE, 2009.

APPENDIX

A. PROOFS

LEMMA 5.1 (DERIVATIVES OF PDFS).

1. $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}} = \frac{1}{2} q(\mathbf{u})(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\intercal})(\mathbf{u} - \boldsymbol{\mu})$.
2. (a) $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}} = \frac{1}{2} q(\mathbf{u})(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1})$.
   (b) $\frac{\partial q(\mathbf{u})}{\partial \mathbf{B}} = q(\mathbf{u})(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{B}$.
3. *For $i = 0, \dots, d$,*
   (a)
$$\begin{aligned}\frac{\partial q(\mathbf{r} \mid \mathbf{u})}{\partial \lambda_i} &= \frac{1}{2} q(\mathbf{r} \mid \mathbf{u})(|\boldsymbol{\Gamma}|^{-1} \operatorname{tr}(\mathbf{R} \operatorname{adj}(\boldsymbol{\Gamma})) \\ &\quad - (\mathbf{r} - \mathbf{S}\mathbf{u})^\intercal \boldsymbol{\Gamma}^{-1} \mathbf{R} \boldsymbol{\Gamma}^{-1} (\mathbf{r} - \mathbf{S}\mathbf{u})).\end{aligned}$$
   (b) *For any covariance matrix* $\mathbf{K}$,
$$\frac{\partial \mathbf{K}}{\partial \lambda_i} = \begin{cases} \frac{1}{\lambda_i}\mathbf{K} & \text{if } i = 0, \\ \mathbf{L} & \text{otherwise,} \end{cases}$$
   where
$$L_{j,k} = k(\mathbf{x}_j, \mathbf{x}_k) \left(-\frac{1}{2}(x_{j,i} - x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \right).$$

PROOF.

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

2. An online tool by Laue et al.[5] [13] can be used to find both derivatives.

3. (a) Since
$$\begin{aligned}q(\mathbf{r} \mid \mathbf{u}) &= \mathcal{N}(\mathbf{r}; \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}}) \\ &= \mathcal{N}(\mathbf{r}; \mathbf{S}\mathbf{u}, \boldsymbol{\Gamma}),\end{aligned}$$
we have
$$\frac{\partial q(\mathbf{r} \mid \mathbf{u})}{\partial \lambda_i} = -\frac{1}{2}q(\mathbf{r} \mid \mathbf{u})\frac{\partial}{\partial \lambda_i}[(\mathbf{r}-\mathbf{S}\mathbf{u})^\intercal \boldsymbol{\Gamma}^{-1}(\mathbf{r}-\mathbf{S}\mathbf{u})+\log|\boldsymbol{\Gamma}|].$$
The same online tool can be used to show that
$$\frac{\partial}{\partial \lambda_i}\log|\boldsymbol{\Gamma}| = -|\boldsymbol{\Gamma}|^{-1}\operatorname{tr}(\mathbf{R}\operatorname{adj}(\boldsymbol{\Gamma})),$$
and
$$\frac{\partial}{\partial \lambda_i}\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{-1}\mathbf{R}\boldsymbol{\Gamma}^{-1}.$$
(b) If $i = 0$, then
$$\frac{\partial \mathbf{K}}{\partial \lambda_i} = \frac{1}{\lambda_i}\mathbf{K}$$
by the structure of each element of $\mathbf{K}$. If $i \neq 0$, then each element of $\frac{\partial \mathbf{K}}{\partial \lambda_i}$ is
$$\begin{aligned}L_{j,k} &= \frac{\partial k(\mathbf{x}_j, \mathbf{x}_k)}{\partial \lambda_i} \\ &= k(\mathbf{x}_j, \mathbf{x}_k)\frac{\partial}{\partial \lambda_i}\Big[-\frac{1}{2}(\mathbf{x}_j-\mathbf{x}_k)^\intercal \boldsymbol{\Lambda}(\mathbf{x}_j-\mathbf{x}_k) \\ &\qquad - \mathbb{1}[j \neq k]\sigma^2\operatorname{tr}(\boldsymbol{\Lambda})\Big] \\ &= k(\mathbf{x}_j, \mathbf{x}_k)\frac{\partial}{\partial \lambda_i}\Big[-\frac{1}{2}\sum_{l=1}^{d}\lambda_l(x_{j,l}-x_{k,l})^2 \\ &\qquad - \mathbb{1}[j \neq k]\sigma^2\sum_{l=1}^{d}\lambda_l\Big] \\ &= k(\mathbf{x}_j, \mathbf{x}_k)\left(-\frac{1}{2}(x_{j,i}-x_{k,i})^2 - \mathbb{1}[j \neq k]\sigma^2\right).\end{aligned}$$

□

B. DERIVATIVES OF THE ELBO

B.1 $\partial/\partial\boldsymbol{\mu}$

We begin by removing terms independent of $\boldsymbol{\mu}$:
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}}[\mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}] - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}}\left[\boldsymbol{\mu}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}\right] - \frac{\partial}{\partial \boldsymbol{\mu}}\mathbb{E}[v].$$
Here
$$\frac{\partial}{\partial \boldsymbol{\mu}}\left[\boldsymbol{\mu}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}\right] = (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}+\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal})\boldsymbol{\mu}$$
by Petersen and Pedersen [20], and
$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\mathbb{E}[V_\mathbf{r}(s)] &= \frac{\partial}{\partial \boldsymbol{\mu}}\iint V_\mathbf{r}(s)q(\mathbf{r} \mid \mathbf{u})q(\mathbf{u})\,d\mathbf{r}\,d\mathbf{u} \\ &= \iint V_\mathbf{r}(s)q(\mathbf{r} \mid \mathbf{u})\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}\,d\mathbf{r}\,d\mathbf{u} \\ &= \frac{1}{2}\mathbb{E}[V_\mathbf{r}(s)(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\intercal})(\mathbf{u}-\boldsymbol{\mu})]\end{aligned}$$
by Theorem 5.10 and Lemma 5.1. Hence,
$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} &= \mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} - \frac{1}{2}(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}+\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal})\boldsymbol{\mu} \\ &\quad - \frac{1}{2}\mathbb{E}\left[(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\intercal})(\mathbf{u}-\boldsymbol{\mu})v\right].\end{aligned}$$

B.2 $\partial/\partial\mathbf{B}$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}} = \frac{1}{2}\left(\frac{\partial}{\partial \mathbf{B}}\log|\boldsymbol{\Sigma}| - \frac{\partial}{\partial \mathbf{B}}\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right)\right) - \frac{\partial}{\partial \mathbf{B}}\mathbb{E}[v].$$
By Theorem 5.10,
$$\frac{\partial}{\partial \mathbf{B}}\mathbb{E}[V_\mathbf{r}(s)] = \iint V_\mathbf{r}(s)q(\mathbf{r} \mid \mathbf{u})\frac{\partial q(\mathbf{u})}{\partial \mathbf{B}}\,d\mathbf{r}\,d\mathbf{u}.$$
Then, using the aforementioned tool by Laue et al. [13], we get
$$\frac{\partial}{\partial \mathbf{B}}\log|\boldsymbol{\Sigma}| = 2\boldsymbol{\Sigma}^{-1}\mathbf{B}, \quad \frac{\partial}{\partial \mathbf{B}}\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right) = 2\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{B},$$
and Lemma 5.1 gives
$$\frac{\partial q(\mathbf{u})}{\partial \mathbf{B}} = q(\mathbf{u})(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - |\boldsymbol{\Sigma}|^{-1}\operatorname{adj}(\boldsymbol{\Sigma}))\mathbf{B}.$$
Therefore,
$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}} = \left(\boldsymbol{\Sigma}^{-1} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\right)\mathbf{B} - \mathbb{E}[(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - |\boldsymbol{\Sigma}|^{-1}\operatorname{adj}(\boldsymbol{\Sigma}))\mathbf{B}v].$$

B.3 $\partial/\partial\lambda_j$

For $j = 0, \dots, d$,
$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda_j} &= \mathbf{t}^\intercal \frac{\partial}{\partial \lambda_j}\left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\right]\boldsymbol{\mu} - \frac{\partial}{\partial \lambda_j}\mathbb{E}[v] \\ &\quad - \frac{1}{2}\left(\frac{\partial}{\partial \lambda_j}\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^\intercal \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}}{\partial \lambda_j}\boldsymbol{\mu} + \frac{\partial}{\partial \lambda_j}\log|\mathbf{K}_{\mathbf{u},\mathbf{u}}|\right),\end{aligned}$$
where
$$\begin{aligned}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}}{\partial \lambda_j} &= -\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}, \\ \frac{\partial}{\partial \lambda_j}\left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\right] &= \frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} + \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}}{\partial \lambda_j} \\ &= \left(\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal}{\partial \lambda_j} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right)\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}, \\ \frac{\partial}{\partial \lambda_j}\operatorname{tr}(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}) &= \operatorname{tr}\left(\frac{\partial}{\partial \lambda_j}\left[\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right]\right) = \operatorname{tr}\left(\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}}{\partial \lambda_j}\boldsymbol{\Sigma}\right) \\ &= -\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right), \\ \frac{\partial}{\partial \lambda_j}\log|\mathbf{K}_{\mathbf{u},\mathbf{u}}| &= \operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right)\end{aligned}$$

[5] `http://www.matrixcalculus.org/`

by Petersen and Pedersen [20], and

$$\frac{\partial}{\partial \lambda_j}\mathbb{E}[V_{\mathbf{r}}(s)] = \iint V_{\mathbf{r}}(s)\frac{\partial q(\mathbf{r} \mid \mathbf{u})}{\partial \lambda_j} q(\mathbf{u})\, d\mathbf{r}\, d\mathbf{u}$$
$$= \frac{1}{2}\mathbb{E}[V_{\mathbf{r}}(s)(|\boldsymbol{\Gamma}|^{-1}\operatorname{tr}(\mathbf{R}\operatorname{adj}(\boldsymbol{\Gamma}))$$
$$- (\mathbf{r}-\mathbf{S}\mathbf{u})^{\intercal}\boldsymbol{\Gamma}^{-1}\mathbf{R}\boldsymbol{\Gamma}^{-1}(\mathbf{r}-\mathbf{S}\mathbf{u}))]$$

by Theorem 5.10 and Lemma 5.1. Thus,

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = \mathbf{t}^{\intercal}\left(\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}}{\partial \lambda_j} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right)\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}$$
$$+\frac{1}{2}\left[\operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu}\right.$$
$$\left. - \operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_j}\right)\right]$$
$$-\frac{1}{2}\mathbb{E}[(|\boldsymbol{\Gamma}|^{-1}\operatorname{tr}(\mathbf{R}\operatorname{adj}(\boldsymbol{\Gamma}))$$
$$- (\mathbf{r}-\mathbf{S}\mathbf{u})^{\intercal}\boldsymbol{\Gamma}^{-1}\mathbf{R}\boldsymbol{\Gamma}^{-1}(\mathbf{r}-\mathbf{S}\mathbf{u}))v],$$

where the remaining derivatives can be found in Lemma 5.1.