Variational Inference for Inverse Reinforcement Learning with Gaussian Processes

Paulius Dilkas (2146879)

7th March 2019

ABSTRACT

1. INTRODUCTION

Inverse reinforcement learning (IRL)—a problem proposed by Russell in 1998 [22]—asks us to find a reward function for a Markov decision process (MDP) that best explains a set of given demonstrations. IRL is important because reward functions can be hard to define manually [1, 2], and rewards are not entirely specific to a given environment, allowing one to reuse the same reward structure in previously unseen environments [2, 9, 14]. Moreover, IRL has seen a wide array of applications in autonomous vehicle control [10, 11] and learning to predict another agent's behaviour [5, 23, 24, 25, 26]. Most approaches in the literature (see Section 2) make a convenient yet unjustified assumption that the reward function can be expressed as a linear combination of features. One proven way to abandon this assumption is by representing the reward function as a Gaussian process (GP) [9, 14, 19]. The original approach used maximum likelihood estimation [14], while the goal of this project is to use variational inference (VI) instead, which learns approximate posterior probability distributions instead of point estimates. This approach can prove useful in three major ways:

1. Modelling full posterior distributions for various parameters can result in more precise reward predictions, as the model simply holds more information.
2. Having variance estimates for rewards can direct our choice in what data should be collected next.
3. An approximate Bayesian treatment of many parameters in the model guards against overfitting [9].

1.1 Statement of the Problem

Definition 1.1 (MDP). *A* Markov decision process *is a set $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, r\}$, 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 (with higher γ values, it makes little difference whether a reward is received now or later, while with lower γ values the future becomes gradually less and less important); and $r : \mathcal{S} \to \mathbb{R}$ is the reward function.*

In *inverse reinforcement learning*, one is presented with an MDP without a reward function $\mathcal{M} \setminus \{r\}$ and a set of expert demonstrations $\mathcal{D} = \{\zeta_i\}_{i=1}^N$, where each demonstration $\zeta_i = \{(s_{i,0}, a_{i,0}), \dots, (s_{i,T}, a_{i,T})\}$ is a multiset of state-action pairs representing the actions taken by the expert during a particular recorded session. Each state is also characterised by a number of features. The goal of IRL is then to find r such that the optimal policy under r

$$\pi^* = \arg\max_\pi \mathbb{E}\left[\left.\sum_{t=0}^\infty \gamma^t r(s_t)\right| \pi\right]$$

matches the actions in $\mathcal{D}$.

Following previous work on GP IRL [14, 9], we use a maximum entropy IRL model [24], under which we have that

$$P(a|s) \propto \exp(Q_\mathbf{r}(s, a)),$$

where

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

and $V_\mathbf{r} : \mathcal{S} \to \mathbb{R}$ is an MDP *value function*, which can be obtained by repeatedly applying the 'soft' version of the Bellman backup operator until convergence: [14, 15]

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

The likelihood of the data can then be expressed as [9, 14]

$$\begin{aligned} p(\mathcal{D}|r) &= \prod_{i=1}^N \prod_{t=1}^T p(a_{i,t}|s_{i,t}) \\ &= \exp\left(\sum_{i=1}^N \sum_{t=1}^T Q_\mathbf{r}(s_{i,t}, a_{i,t}) - V_\mathbf{r}(s_{i,t})\right). \end{aligned} \quad (2)$$

However, a reward function learned by maximising this likelihood is not transferable to new situations [9, 14]. One needs to model the reward structure in a way that would allow reward predictions for previously unseen states.

One way to model rewards without assumptions of linearity is with a *Gaussian process* (GP). A GP is a collection of random variables, any finite combination of which has a joint Gaussian distribution [20]. We write $r \sim \mathcal{GP}(0, k)$ to say that r is a GP with mean 0 and covariance function k, which uses a vector of hyperparameters $\boldsymbol{\lambda}$. Covariance functions take two state feature vectors as input and quantify how similar the two states are, in a sense that we would expect them to have similar rewards.

As training a GP with n data points has a time complexity of $\mathcal{O}(n^3)$ [20], numerous approximation methods have been suggested, many of which select a subset of data called *inducing points* and focus most of the training effort on them [16]. Let $\mathbf{X_u}$ be the matrix of features at inducing states, $\mathbf{u}$ the rewards at those states, and $\mathbf{r}$ a vector with $r(\mathcal{S})$ as elements. Then the full joint probability distribution can be factorised as

$$p(\mathcal{D}, \mathbf{u}, \mathbf{r}) = p(\mathbf{u}) \times p(\mathbf{r}|\mathbf{u}) \times p(\mathcal{D}|r),$$

where

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

is the GP prior [20], where $m \in \mathbb{N}$ is the number of inducing points. The GP posterior is a multivariate Gaussian [14]

$$p(\mathbf{r}|\boldsymbol{\lambda}, \mathbf{X}_\mathbf{u}, \mathbf{u}) = \mathcal{N}(\mathbf{r}; \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}}), \tag{3}$$

and $p(\mathcal{D}|r)$ is as in Equation 2. The matrices such as $\mathbf{K}_{\mathbf{r},\mathbf{u}}$ are called *covariance 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 state in $\mathbf{X}_\mathbf{u}$, respectively [9].

Given this model, data $\mathcal{D}$, and inducing feature matrix $\mathbf{X}_\mathbf{u}$, our goal is then to find optimal values of hyperparameters $\boldsymbol{\lambda}$, inducing rewards $\mathbf{u}$, and the rewards for all relevant states $\mathbf{r}$. While the previous paper to consider this IRL model computed maximum likelihood (ML) estimates for $\boldsymbol{\lambda}$ and $\mathbf{u}$, and made an assumption that $\mathbf{r}$ in Equation 3 has zero variance [14], we aim to avoid this assumption and use VI to approximate the full posterior distribution $p(\mathbf{u}, \mathbf{r} \mid \mathcal{D})$. *Variational inference* is an approximation technique for probability densities [4]. Let $q(\mathbf{u}, \mathbf{r})$ be our approximating family of probability distributions for $p(\mathbf{u}, \mathbf{r} \mid \mathcal{D})$ with its own hyperparameter vector $\boldsymbol{\nu}$. Then the job of VI algorithms is to optimise $\boldsymbol{\nu}$ in order to minimise the *Kullback-Leibler* (KL) divergence between the original probability distribution and our approximation. KL divergence (asymmetrically) measures how different the two distributions are, and in this case can be defined as [4]

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

The last term is both hard to compute and constant w.r.t. $q(\mathbf{u}, \mathbf{r})$ [4], so we can remove it from our optimisation objective. The negation of what remains is often called the *evidence lower bound* (ELBO) and is defined as[1] [3, 4]

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

By considering full probability distributions instead of point estimates—as long as the approximations are able to capture important features of the posterior—our predictions are likely to be more accurate and rely on fewer assumptions. Moreover, we hope to make use of various recent advancements in VI for both time complexity and approximation distribution fit (see Section 2), making the resulting algorithm competitive both in terms of running time and model fit.

The project is primarily concerned with investigating how a VI formulation of the GP IRL model compares against the original ML approach. Most importantly, we aim to compare how the two algorithms converge both over time and as the number of demonstrations increases. It would also be interesting to see how close the approximation of the posterior distribution is to the real thing. Finally, it is reasonable to conjecture that VI can outperform point estimates when dealing with more uncertainty, e.g., when experts make mistakes. We can easily investigate this by adapting how the evaluation data is generated.

2. BACKGROUND

2.1 Linear Algebra and Numerical Analysis

DEFINITION 2.1 (NORMS). *For any finite-dimensional vector* $\mathbf{x} = (x_1, \dots, x_n)^\intercal$*, 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 2.2 (PERTURBATION LEMMA [13]). *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*

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

3. THE WIZWOZ SYSTEM

3.1 Notation

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 $\mathrm{tr}(\mathbf{A})$ to denote its *trace* and $\mathrm{adj}(\mathbf{A})$ 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 $\mathbb{R}$.

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 $s \in \mathcal{S}$.

3.2 Preliminaries

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.

[1] Throughout the proposal, all integrals should be interpreted as definite integrals over the entire sample space.

As recently suggested by Ong et al. [17], we use a decomposition $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\intercal + \mathbf{D}^2$, where $\mathbf{B}$ is a lower triangular $m \times p$ matrix with positive diagonal entries, and $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_m)$. Typically, we would set p so that $p \ll m$ to get an efficient approximation, but it is also worth pointing out that we can retain full accuracy by setting $p = m$ and $\mathbf{D} = \mathbf{O}$. Moreover, we define a few variables that will simplify expressions for the derivatives:

$$\mathbf{U} = (\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})^\intercal$$

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

$$\mathbf{\Gamma} = \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{S}\mathbf{K}_{\mathbf{r},\mathbf{u}},$$

$$\mathbf{R} = \mathbf{S}\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}}{\partial \lambda_i} - \frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{r}}}{\partial \lambda_i} + \left(\frac{\partial \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal}{\partial \lambda_i} - \mathbf{S}\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} \right) \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{r},\mathbf{u}},$$

$$Q = (\mathbf{u} - \boldsymbol{\mu})^\intercal \mathbf{\Sigma}^{-1} (\mathbf{u} - \boldsymbol{\mu}).$$

Derivatives such as $\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i}$ can be expressed as

$$\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} = \frac{1}{\lambda_i} \mathbf{K}_{\mathbf{u},\mathbf{u}}$$

if $i = 0$, and

$$\left[\frac{\partial \mathbf{K}_{\mathbf{u},\mathbf{u}}}{\partial \lambda_i} \right]_{j,k} = k(\mathbf{x}_{\mathbf{u},j}, \mathbf{x}_{\mathbf{u},k}) \left(-\frac{1}{2}(x_{\mathbf{u},j,i} - x_{\mathbf{u},k,i})^2 - \mathbb{1}[j \neq k]\sigma^2 \right)$$

otherwise.

Lemma 3.1 (Derivatives of PDFs).

1. $\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}} = \frac{1}{2} q(\mathbf{u})(\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-\intercal})(\mathbf{u} - \boldsymbol{\mu})$.
2. (a) $\frac{\partial q(\mathbf{u})}{\partial \mathbf{\Sigma}} = \frac{1}{2} q(\mathbf{u})(\mathbf{\Sigma}^{-\intercal}\mathbf{U}\mathbf{\Sigma}^{-\intercal} - \mathbf{\Sigma}^{-\intercal})$.
   (b) $\frac{\partial q(\mathbf{u})}{\partial \mathbf{B}} = q(\mathbf{u})(\mathbf{\Sigma}^{-1}\mathbf{U}\mathbf{\Sigma}^{-1} - |\mathbf{\Sigma}|^{-1} \operatorname{adj}(\mathbf{\Sigma}))\mathbf{B}$.
3. For $i = 0, \dots, d$,
$$\frac{\partial q(\mathbf{r} \mid \mathbf{u})}{\partial \lambda_i} = \frac{1}{2} q(\mathbf{r} \mid \mathbf{u})(|\mathbf{\Gamma}|^{-1} \operatorname{tr}(\mathbf{R} \operatorname{adj}(\mathbf{\Gamma})) - (\mathbf{r} - \mathbf{S}\mathbf{u})^\intercal \mathbf{\Gamma}^{-1} \mathbf{R} \mathbf{\Gamma}^{-1} (\mathbf{r} - \mathbf{S}\mathbf{u})),$$

3.3 Evidence Lower Bound

$$p(\mathcal{D}, \mathbf{u}, \mathbf{r}) = p(\mathbf{u}) \times p(\mathbf{r} \mid \mathbf{u}) \times p(\mathcal{D} \mid \mathbf{r}). \tag{5}$$

$$q(\mathbf{u}, \mathbf{r}) = q(\mathbf{u}) \times q(\mathbf{r} \mid \mathbf{u}). \tag{6}$$

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 Equation 4 and write[2]

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

By substituting in Equations 5 and 6, we get

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

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

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

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

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

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

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

We can simplify $\sum_{i=1}^N \sum_{t=1}^T r(s_{i,t})$ by defining a new vector $\mathbf{t} = (t_1, \dots, t_{|\mathcal{S}|})^\intercal$, 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

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

where

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

3.4 Theoretical Justification

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 3.2. *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

$$V_{\mathbf{r}}(s) = \log \sum_{a \in \mathcal{A}} \exp \left(r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') V_{\mathbf{r}}(s') \right) \tag{7}$$

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

[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'.

for any $\alpha \in \mathbb{R}$, the set

$$\left\{ \mathbf{r} \in \mathbb{R}^{|\mathcal{S}|} \;\middle|\; \begin{array}{l} V_{\mathbf{r}}(s_0) \in (-\infty, \alpha); \\ V_{\mathbf{r}}(s) \in \mathbb{R} \text{ for all } s \in \mathcal{S} \setminus \{s_0\}; \\ \text{Equation 7 is satisfied by all } s \in \mathcal{S} \end{array} \right\}$$

is measurable. Since this set can be constructed in Zermelo-Fraenkel set theory *without* the axiom of choice, it is measurable [8], which proves that $V(s)$ is a measurable function for any $s \in \mathcal{S}$. □

PROPOSITION 3.3. *If the initial values of the MDP value function satisfy the following bound, then the bound remains satisfied throughout value iteration:*

$$|V_{\mathbf{r}}(s)| \le \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}. \tag{8}$$

PROOF. We begin by considering Equation 8 without taking the absolute value of $V_{\mathbf{r}}(s)$, i.e.,

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

and assuming that the initial values of $\{V_{\mathbf{r}}(s) \mid s \in \mathcal{S}\}$ already satisfy Equation 9. Recall that for each $s \in \mathcal{S}$, the value of $V_{\mathbf{r}}(s)$ is updated by applying Equation 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

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

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 Equations 9 and 10 gives Equation 8. □

THEOREM 3.4 (DOMINATED CONVERGENCE THEOREM [21]). *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*

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

Then f is integrable over X and

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

LEMMA 3.5. *Let $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (a, b) \subset \mathbb{R}$ be an arbitrary bounded function. Then, for $i = 0, \dots, d$,*

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

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

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

by Lemma 3.1. 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

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

i.e., without λ_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 $k(\mathbf{x}_j, \mathbf{x}_k)$, for some $\mathbf{x}_j$ and $\mathbf{x}_k$. Since $k(\mathbf{x}_j, \mathbf{x}_k)$ is a decreasing function of λ_i, and $c(\mathbf{r}, \mathbf{u}) > a$,

$$\begin{aligned} k(\mathbf{x}_j, \mathbf{x}_k)|_{\lambda_i = c(\mathbf{r}, \mathbf{u})} = \lambda_0 \exp\Bigg(&-\frac{1}{2} c(\mathbf{r}, \mathbf{u}) (x_{j,i} - x_{k,i})^2 \\ &- \mathbb{1}[j \ne k] \sigma^2 c(\mathbf{r}, \mathbf{u}) - S \Bigg) \\ < \lambda_0 \exp\Bigg(&-\frac{1}{2} a (x_{j,i} - x_{k,i})^2 \\ &- \mathbb{1}[j \ne k] \sigma^2 a - S \Bigg), \end{aligned}$$

where

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

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 3.6. *In order to find $\frac{\partial q(\mathbf{u})}{\partial t}$, where t is the ith element of the vector $\boldsymbol{\mu}$, 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 $\boldsymbol{\Sigma}$.*

LEMMA 3.7. *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 3.1,

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

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 $\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-\intercal}$ 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 3.8. *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

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

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 3.1,

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

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 $\mathbf{C}(\mathbf{r},\mathbf{u}) = \boldsymbol{\Sigma}+\mathbf{E}(\mathbf{r},\mathbf{u})$, 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 2.2 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 $\|\boldsymbol{\Sigma}^{-1}\|_\infty\|\mathbf{E}(\mathbf{r},\mathbf{u})\|_\infty < 1$. Then $\mathbf{C}(\mathbf{r},\mathbf{u})$ is invertible, and

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

which means that

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

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

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

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. □

LEMMA 3.9.

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

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 $\xi_i = \left[\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}\right]_i$, $\sigma_i = \sqrt{[\mathbf{K}_{\mathbf{r},\mathbf{r}}-\mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}}]_{i,i}}$[3], and Φ is the cumulative distribution function of the standard normal distribution. 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

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

[3] 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.

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 Chen [6].

THEOREM 3.10. *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 $t_n \neq t$ for all n. We want to show that

$$F'(t) = \lim_{n \to \infty} \frac{F(t_n) - F(t)}{t_n - t} = \iint \left. \frac{\partial f}{\partial t} \right|_{(\mathbf{r}, \mathbf{u}, t)} d\mathbf{r} \, d\mathbf{u}. \quad (11)$$

We have

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

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

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

Equation 11 follows from Theorem 3.4 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 3.2. 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 3.3, 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

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

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 3.5, 3.7, and 3.8, 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 3.9, which allows us to integrate out $\mathbf{r}$, and we are left with showing the existence of

$$\int \left(a + \|\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}\|_1\right) \max\{|p_1(\mathbf{u})|, |p_2(\mathbf{u})|\} q(\mathbf{u}) \, d\mathbf{u}, \quad (12)$$

where a is a constant. The integral

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

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 Equation 12 as well. □

3.5 Derivatives

3.5.1 $\partial / \partial \boldsymbol{\mu}$

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

$$\frac{\partial}{\partial \boldsymbol{\mu}} \left[\boldsymbol{\mu}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \boldsymbol{\mu}\right] = (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} + \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-\intercal}) \boldsymbol{\mu}$$

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 3.10 and Lemma 3.1. Hence,

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

3.5.2 $\partial / \partial \mathbf{B}$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}} = \frac{1}{2} \left(\frac{\partial}{\partial \mathbf{B}} \log |\boldsymbol{\Sigma}| - \frac{\partial}{\partial \mathbf{B}} \operatorname{tr}\left(\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \boldsymbol{\Sigma}\right) \right) - \frac{\partial}{\partial \mathbf{B}} \mathbb{E}[v].$$

By Theorem 3.10,

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

Then, using the aforementioned tool by Laue et al. [12], 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 3.1 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,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}} = \left(\boldsymbol{\Sigma}^{-1} - \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\right)\mathbf{B} - \mathbb{E}[(\boldsymbol{\Sigma}^{-1}\mathbf{U}\boldsymbol{\Sigma}^{-1} - |\boldsymbol{\Sigma}|^{-1}\operatorname{adj}(\boldsymbol{\Sigma}))\mathbf{B}v].$$

3.5.3 $\partial/\partial\lambda_j$

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

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

by Petersen and Pedersen [18], and

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

by Theorem 3.10 and Lemma 3.1. Thus,

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

where the remaining derivatives can be found in Lemma 3.1.

4. EVALUATION

5. CONCLUSIONS

5.1 Future Work

6. REFERENCES

[1] 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.

[2] S. Arora and P. Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *CoRR*, abs/1806.06877, 2018.

[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] K. D. Bogert and P. Doshi. Multi-robot inverse reinforcement learning under occlusion with estimation of state transitions. *Artif. Intell.*, 263:46–73, 2018.

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

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

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

[9] 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.

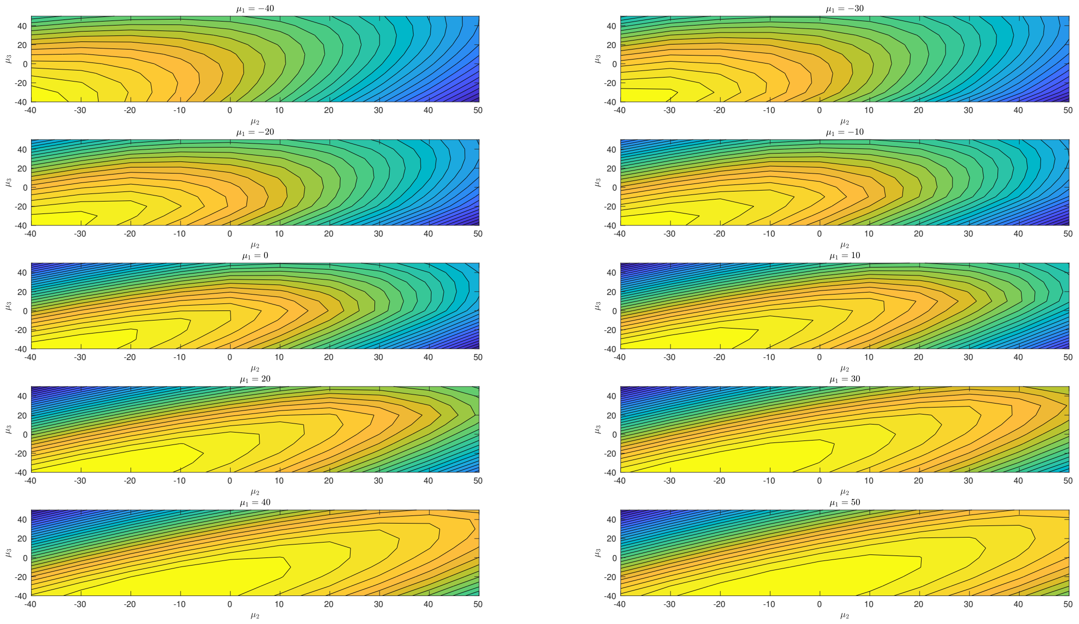
[10] B. Kim and J. Pineau. Socially adaptive path planning in human environments using inverse reinforcement learning. *I. J. Social Robotics*, 8(1):51–66, 2016.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] Q. Qiao and P. A. Beling. Inverse reinforcement learning with Gaussian process. *CoRR*, abs/1208.2112, 2012.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.