Variational Inference for Inverse Reinforcement Learning with Gaussian Processes

Paulius Dilkas

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Masters project proposal

Date of submission placed here

Contents

1 Introduction **2**

2 Statement of the Problem **2**

3 Literature Survey **4**

4 Proposed Approach **4**

5 Work Plan **4**

6 Notes on papers (to be removed) **5**

6.1 Miscellaneous 5

6.2 Gaussian Processes 5

6.3 Interpretability 5

6.4 Inverse Reinforcement Learning 5

6.4.1 Multiple Strategies 6

6.5 Variational Inference 6

1 Introduction

Inverse reinforcement learning (IRL)—a problem proposed by Russell in 1998 [40]—asks us to find a reward function for a Markov decision process 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, 22, 25]. Moreover, IRL has seen a wide array of applications in autonomous vehicle control [23, 24] and learning to predict another agent's behaviour [7, 43, 47, 48, 49]. Most approaches in the literature (see Section 3) 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 [22, 25, 34].

2 Statement of the Problem

Definition 1. A *Markov decision process* (MDP) 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 $\zeta_i = \{(s_{i,0}, a_{i,0}), \ldots, (s_{i,T}, a_{i,T})\}$ 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

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

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

The likelihood of the data can be written down as [22, 25]

$$p(\mathcal{D}|r) = \prod_{i=1}^{N}\prod_{t=1}^{T} p(a_{i,t}|s_{i,t}) = \exp\left(\sum_{i=1}^{N}\sum_{t=1}^{T} Q(s_{i,t}, a_{i,t}; r) - V(s_{i,t}; r)\right), \tag{1}$$

where

$$Q(s_{i,t}, a_{i,t}; r) = r(s_{i,t}) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s_{i,t}, a_{i,t}, s')V(s'; r),$$

and $V(s; r)$ can be obtained by repeatedly applying the equation [26]

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

However, a reward function learned by maximising this likelihood is not transferable to new situations [22, 25]. 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 [37]. We write $r \sim \mathcal{GP}(0, k_\Theta)$ to say that r is a GP with mean 0 and covariance function k_Θ, which uses a set of hyperparameters Θ. 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 $\mathcal{O}(n^3)$ [37], 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 [28]. 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

$$p(\mathcal{D}, \Theta, \mathbf{X_u}, \mathbf{u}, \mathbf{r}) = p(\mathbf{X_u}) \times p(\Theta|\mathbf{X_u}) \times p(\mathbf{u}|\Theta, \mathbf{X_u}) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}) \times p(\mathcal{D}|r). \quad (2)$$

Here $p(\mathbf{X_u})$ and $p(\Theta|\mathbf{X_u})$ are freely chosen priors,

$$\begin{aligned} p(\mathbf{u}|\Theta, \mathbf{X_u}) &= \mathcal{N}(\mathbf{0}, \mathbf{K_{u,u}}) \\ &= \frac{1}{(2\pi)^{n/2}|\mathbf{K_{u,u}}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{K_{u,u}^{-1}}\mathbf{u}\right) \\ &= \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{K_{u,u}^{-1}}\mathbf{u} - \frac{1}{2}\log|\mathbf{K_{u,u}}| - \frac{n}{2}\log 2\pi\right) \end{aligned}$$

is the GP prior [37], the GP posterior is a multivariate Gaussian [25]

$$p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}) = \mathcal{N}(\mathbf{K_{r,u}^T}\mathbf{K_{u,u}^{-1}}\mathbf{u}, \mathbf{K_{r,r}} - \mathbf{K_{r,u}^T}\mathbf{K_{u,u}^{-1}}\mathbf{K_{r,u}}), \quad (3)$$

and $p(\mathcal{D}|r)$ is as in Equation 1. The matrices such as $\mathbf{K_{r,u}}$ are called *covariance matrices* and are defined as $[\mathbf{K_{r,u}}]_{i,j} = k_\Theta(\mathbf{x_{r},}_i, \mathbf{x_{u},}_j)$, where $\mathbf{x_{r},}_i$ and $\mathbf{x_{u},}_j$ denote feature vectors for the ith state in $\mathcal{S}$ and the jth state in $\mathbf{X_u}$, respectively.

Given this model, data $\mathcal{D}$, and inducing feature matrix $\mathbf{X_u}$, our goal is then to find optimal values of hyperparameters Θ, inducing rewards $\mathbf{u}$, and the reward function r. While the previous paper that considered this IRL model computed maximum likelihood estimates for Θ and $\mathbf{u}$, and made an assumption that $\mathbf{r}$ in Equation 3 has 0 variance [25], we aim to avoid this assumption and use variational inference to approximate the full posterior distribution $p(\Theta, \mathbf{u}, \mathbf{r}|\mathcal{D}, \mathbf{X_u})$. *Variational inference* (VI) is an approximation technique for probability densities [6]. Let $q(\Theta, \mathbf{u}, \mathbf{r}; \Lambda)$ be our approximating family of probability distributions for $p(\Theta, \mathbf{u}, \mathbf{r}|\mathcal{D}, \mathbf{X_u})$ with its own set of hyperparameters Λ. Then it is up to VI algorithms to optimise Λ 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 follows [6]:

$$D_{\mathrm{KL}}(q(\Theta, \mathbf{u}, \mathbf{r}; \Lambda)||p(\Theta, \mathbf{u}, \mathbf{r}|\mathcal{D}, \mathbf{X_u})) = \mathbb{E}_q[\log q(\Theta, \mathbf{u}, \mathbf{r}; \Lambda)] - \mathbb{E}_q[\log p(\Theta, \mathbf{u}, \mathbf{r}|\mathcal{D}, \mathbf{X_u})].$$

Since KL divergence is typically hard to compute, instead of minimising it, VI typically tries to maximise the *evidence lower bound* (ELBO) defined as

$$\begin{aligned}\mathcal{L}(\Lambda) &= \mathbb{E}_q[\log p(\mathcal{D}, \Theta, \mathbf{X_u}, \mathbf{u}, \mathbf{r})] - \mathbb{E}_q[\log q(\Theta, \mathbf{u}, \mathbf{r}; \Lambda)] \\ &= \iiint q(\Theta, \mathbf{u}, \mathbf{r}; \Lambda) \log \frac{p(\mathcal{D}, \Theta, \mathbf{X_u}, \mathbf{u}, \mathbf{r})}{q(\Theta, \mathbf{u}, \mathbf{r}; \Lambda)} \, d\Theta \, d\mathbf{u} \, d\mathbf{r}.\end{aligned}$$

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 3), making the resulting algorithm competitive both in terms of speed and model fit.

3 Literature Survey

present an overview of relevant previous work including articles, books, and existing software products. Critically evaluate the strengths and weaknesses of the previous work.

4 Proposed Approach

We begin by rewriting the posterior by using the chain rule and Bayes' theorem, trying to extract parts of the distribution we know how to compute, namely those in Equation 2:

$$\begin{aligned}p(\Theta, \mathbf{u}, \mathbf{r}|\mathcal{D}, \mathbf{X_u}) &= p(\Theta|\mathbf{X_u}, \mathcal{D}) \times p(\mathbf{u}|\Theta, \mathbf{X_u}, \mathcal{D}) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}, \mathcal{D}) \\ &\propto p(\Theta|\mathbf{X_u}, \mathcal{D}) \times p(\mathbf{u}|\Theta, \mathbf{X_u}, \mathcal{D}) \times p(\mathcal{D}|r) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}) \\ &\propto p(\Theta|\mathbf{X_u}, \mathcal{D}) \times p(\mathcal{D}|\Theta, \mathbf{X_u}, \mathbf{u}) \times p(\mathbf{u}|\Theta, \mathbf{X_u}) \times p(\mathcal{D}|r) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}) \\ &\propto p(\mathcal{D}|\Theta, \mathbf{X_u}) \times p(\Theta|\mathbf{X_u}) \times p(\mathcal{D}|\Theta, \mathbf{X_u}, \mathbf{u}) \times p(\mathbf{u}|\Theta, \mathbf{X_u}) \times p(\mathcal{D}|r) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u})\end{aligned}$$

Note that now there are only two unknown probability distributions: $p(\mathcal{D}|\Theta, \mathbf{X_u})$ and $p(\mathcal{D}|\Theta, \mathbf{X_u}, \mathbf{u})$. They can be computed as follows:

$$\begin{aligned}p(\mathcal{D}|\Theta, \mathbf{X_u}, \mathbf{u}) &= \int p(\mathcal{D}|r) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}) \, d\mathbf{r}, \\ p(\mathcal{D}|\Theta, \mathbf{X_u}) &= \iint p(\mathcal{D}|r) \times p(\mathbf{r}|\Theta, \mathbf{X_u}, \mathbf{u}) \times p(\mathbf{u}|\Theta, \mathbf{X_u}) \, d\mathbf{u} \, d\mathbf{r}.\end{aligned}$$

5 Work Plan

show how you plan to organize your work, identifying intermediate deliverables and dates.

6 Notes on papers (to be removed)

6.1 Miscellaneous

(Directed) similarity between MDPs using restricted Boltzmann machines [8]

Chapter 6 on distance measures [30]

The PhD thesis behind maximum causal entropy [46]

6.2 Gaussian Processes

Simple introduction to GPs for time-series modelling [39]

Spectral kernels [44]

GPs over graphs instead of vectors (haven't actually read) [42]

Another introduction from physics (skimmed through) [21]

Learning a GP from very little data [33]

One GP for multiple correlated output variables [4]

Kernels for categorical and count data [41]

Scalability/Approximations thesis [20]

6.3 Interpretability

Learning latent factors [27]

The behaviour of Reddit users [14]

6.4 Inverse Reinforcement Learning

One of the first papers on the topic [32]

Bayesian setting [36]

Learning optimal composite features [13]

A different take on IRL with GPs [34]

IRL for large state spaces (haven't read) [9]

Multiple reward functions [12]

A recent survey [2]

Some not-very-successful method [31]

6.4.1 Multiple Strategies

EM clustering [3]

Structured priors [15]

There are more, but I haven't gotten to them yet.

6.5 Variational Inference

Part IV on probabilities and inference [29]

Normalizing flows [38]

Linear VI for GPs [11]

Stochastic VI [19]

Structured stochastic VI (haven't read) [18]

Another review of recent advances [45]

Sparse VI for GP [17]

Sparse GPs [10].

Tighter ELBOs are not necessarily better [35].

[5]

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In Carla 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. Two early algorithms based on max margin optimisation.

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

[3] Monica Babes, Vukosi N. Marivate, Kaushik Subramanian, and Michael L. Littman. Apprenticeship learning about multiple intentions. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 897–904. Omnipress, 2011.

[4] Ilias Bilionis, Nicholas Zabaras, Bledar A. Konomi, and Guang Lin. Multi-output separable Gaussian process: Towards an efficient, fully Bayesian paradigm for uncertainty quantification. *J. Comput. Physics*, 241:212–239, 2013.

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

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

[7] Kenneth D. Bogert and Prashant Doshi. Multi-robot inverse reinforcement learning under occlusion with estimation of state transitions. *Artif. Intell.*, 263:46–73, 2018. Using IRL to predict positions of patrolling robots.

[8] H. Bou Ammar, E. Eaton, M.E. Taylor, D.C. Mocanu, K. Driessens, G. Weiss, and K.P. Tuyls. An automated measure of MDP similarity for transfer in reinforcement learning. In *Proceedings of the MLIS-2014 collocated with The Twenty-Eighth AAAI Conference on Artificial Intelligence, 27-28 July 2014, Quebec City, Canada*, pages 31–37, 2014.

[9] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 182–189. JMLR.org, 2011.

[10] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[11] Ching-An Cheng and Byron Boots. Variational inference for Gaussian process models with linear complexity. In Guyon et al. [16], pages 5190–5200.

[12] Jaedeug Choi and Kee-Eung Kim. Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 314–322, 2012.

[13] Jaedeug Choi and Kee-Eung Kim. Bayesian nonparametric feature construction for inverse reinforcement learning. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1287–1293. IJCAI/AAAI, 2013.

[14] Sanmay Das and Allen Lavoie. The effects of feedback on human behavior in social media: an inverse reinforcement learning model. In Ana L. C. Bazzan, Michael N. Huhns, Alessio Lomuscio, and Paul Scerri, editors, *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014*, pages 653–660. IFAAMAS/ACM, 2014.

[15] Christos Dimitrakakis and Constantin A. Rothkopf. Bayesian multitask inverse reinforcement learning. In Scott Sanner and Marcus Hutter, editors, *Recent Advances in Reinforcement Learning - 9th European Workshop, EWRL 2011, Athens, Greece, September 9-11, 2011, Revised Selected Papers*, volume 7188 of *Lecture Notes in Computer Science*, pages 273–284. Springer, 2011.

[16] Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman 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*, 2017.

[17] James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18:151:1–151:52, 2017.

[18] Matthew D. Hoffman and David M. Blei. Structured stochastic variational inference. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

[19] Matthew D. Hoffman, David M. Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[20] Hanna Hultin. Evaluation of massively scalable Gaussian processes. Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden, June 2017.

[21] David J.C. MacKay. Introduction to Gaussian processes. 168, 01 1998.

[22] Ming Jin, Andreas C. Damianou, Pieter Abbeel, and Costas J. Spanos. Inverse reinforcement learning via deep Gaussian process. In Gal Elidan, Kristian Kersting, and Alexander 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. IRL with an extra GP layer to represent features and variational approximation.

[23] Beomjoon Kim and Joelle Pineau. Socially adaptive path planning in human environments using inverse reinforcement learning. *I. J. Social Robotics*, 8(1):51–66, 2016. An example of how IRL can be used in socially adaptive robot path planning.

[24] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *I. J. Robotics Res.*, 35(11):1289–1307, 2016. An example of how IRL can be used in socially adaptive robot path planning.

[25] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with Gaussian processes. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian 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. First paper to tackle rewards that cannot be expressed as a linear combination of features.

[26] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Supplementary material: Nonlinear inverse reinforcement learning with Gaussian processes. `http://graphics.stanford.edu/projects/gpirl/gpirl_supplement.pdf`, December 2011. Contains derivations of likelihood partial derivatives and additional details about the implementation.

[27] Yunzhu Li, Jiaming Song, and Stefano Ermon. InfoGAIL: Interpretable imitation learning from visual demonstrations. In Guyon et al. [16], pages 3815–3825.

[28] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When Gaussian process meets big data: A review of scalable GPs. *CoRR*, abs/1807.01065, 2018. A recent review of scalable GPs.

[29] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2003.

[30] B. McCune, J.B. Grace, and D.L. Urban. *Analysis of Ecological Communities*. MjM Software Design, 2002.

[31] Gergely Neu and Csaba Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In Ronald Parr and Linda C. van der Gaag, editors, *UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, Canada, July 19-22, 2007*, pages 295–302. AUAI Press, 2007.

[32] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In Pat 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.

[33] John C. Platt, Christopher J. C. Burges, S. Swenson, C. Weare, and A. Zheng. Learning a Gaussian process prior for automatically generating music playlists. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 1425–1432. MIT Press, 2001.

[34] Qifeng Qiao and Peter A. Beling. Inverse reinforcement learning with Gaussian process. *CoRR*, abs/1208.2112, 2012. An alternative formulation of IRL with GPs.

[35] Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 4274–4282. JMLR.org, 2018.

[36] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In Manuela M. Veloso, editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2586–2591, 2007.

[37] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. The main book on GPs.

[38] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis R. Bach and David 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.

[39] Stephen J. Roberts, Matt Osborne, Mark Ebden, Steve Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 371 1984:20110550, 2013.

[40] Stuart J. Russell. Learning agents for uncertain environments (extended abstract). In Peter L. Bartlett and Yishay 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. The first paper (talk) that defines inverse reinforcement learning.

[41] Terrance Savitsky, Marina Vannucci, and Naijun Sha. Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1):130, 2011.

[42] Arun Venkitaraman, Saikat Chatterjee, and Peter Händel. Gaussian processes over graphs. *CoRR*, abs/1803.05776, 2018.

[43] Adam Vogel, Deepak Ramachandran, Rakesh Gupta, and Antoine Raux. Improving hybrid vehicle fuel efficiency using inverse reinforcement learning. In Jörg Hoffmann and Bart Selman, editors, *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*. AAAI Press, 2012. Using IRL to predict where the driver is going.

[44] Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In Sanjoy Dasgupta and David 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.

[45] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *CoRR*, abs/1711.05597, 2017.

[46] Brian D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Pittsburgh, PA, USA, 2010. AAI3438449.

[47] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008. An influential idea: maximum entropy IRL.

[48] Brian D. Ziebart, Andrew L. Maas, Anind K. Dey, and J. Andrew Bagnell. Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In Hee Yong Youn and We-Duke 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. Using IRL to predict the behaviour of taxi drivers.

[49] Brian D. Ziebart, Nathan D. Ratliff, Garratt Gallagher, Christoph Mertz, Kevin M. Peterson, James A. Bagnell, Martial Hebert, Anind K. Dey, and Siddhartha 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. Using IRL to predict the movement of pedestrians.