Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

I See What You See: Inferring Sensor and Policy Models of Human Real-World Motor Behavior

Felix Schmitt, Hans-Joachim Bieg and Michael Herman
Robert Bosch GmbH
70465 Stuttgart, Germany
<first name>.<last name>@de.bosch.com

Constantin A. Rothkopf
Center for Cognitive Science & Dept. of Psychology
Technische Universität Darmstadt
64283 Darmstadt, Germany
rothkopf@psychologie.tu-darmstadt.de

Abstract

Human motor behavior is naturally guided by sensing the environment. To predict such sensori-motor behavior, it is necessary to model what is sensed and how actions are chosen based on the obtained sensory measurements. Although several models of human sensing haven been proposed, rarely data of the assumed sensory measurements is available. This makes statistical estimation of sensor models problematic. To overcome this issue, we propose an abstract structural estimation approach building on the ideas of Herman et al.'s Simultaneous Estimation of Rewards and Dynamics (SERD). Assuming optimal fusion of sensory information and rational choice of actions the proposed method allows to infer sensor models even in absence of data of the sensory measurements. To the best of our knowledge, this work presents the first *general* approach for joint inference of sensor and policy models. Furthermore, we consider its concrete implementation in the important class of sensor scheduling linear quadratic Gaussian problems. Finally, the effectiveness of the approach is demonstrated for prediction of the behavior of automobile drivers. Specifically, we model the glance and steering behavior of driving in the presence of visually demanding secondary tasks. The results show, that prediction benefits from the inference of sensor models. This is the case, especially, if also information is considered, that is contained in gaze switching behavior.

Introduction

Prediction models of human motor behavior are important for several practical applications. For example, an automobile could warn a novice driver of driving errors based on a model of driving strategies of experienced drivers (Shimosaka, Kaneko, and Nishi 2014). As human behavior can be very complex, prediction models are often inferred from data. For prediction in a broad range of situations, as for example encountered in real traffic, models are necessary that generalize well to unseen instances.

Classically, mappings from situational states to likely actions, stochastic policies, are inferred. However, there is evidence, that human motor behavior can also be characterized by a task consisting of situational constraints and an objective (Baron and Kleinman 1969; Todorov and Jordan 2002; Rothkopf, Ballard, and Hayhoe 2007). Objectives often hold globally and therefore allow significantly better generalization than directly learned policies. Given an objective function and an *explicit* model of the situational constraints, the specific stochastic policy can be obtained by probabilistic Optimal Control (OC). Conversely, objectives can be inferred from data using probabilistic Inverse Optimal Control (IOC) (Ramachandran and Amir 2007; Neu and Szepesvári 2007; Ziebart, Bagnell, and Dey 2010; Rothkopf and Dimitrakakis 2011).

Besides task objective, sensory uncertainty strongly shapes human motor behavior. This is because the sensory measurements made by humans are noisy, ambiguous and limited in scope. For example, visual sensing is restricted to stimuli in the field of view, where disparate objects can look the same from a single perspective. According to the theory of *perception by Bayesian inference* (Knill and Richards 1996), humans fuse those measurements with internal models of the sensors and the evolution of the situation for decision on motor actions.

In this work we assume that the human internal model of the situational dynamics equals the actual dynamics that are known at inference time. Instead we focus on the sensor models and their estimation. Specific sensor models have been obtained from signal detection experiments, i.e. artificial tasks, that are e.g. reviewed in (Nash, Cole, and Bigler 2016). For these tasks several inference techniques are available (Acerbi, Ma, and Vijayakumar 2014). However, the validity of the inferred models for *natural* motor tasks is limited: Tasks like vehicle driving involve several sensors, require closed-loop choice of continuous valued actions and often show complex situational dynamics. Hence, methods for inference of sensor models from *real-world* behavioral data are needed, as noted in (Nash, Cole, and Bigler 2016).

Related Work

Data of human sensory measurements is rarely available. Hence, inference approaches commonly exploit knowledge of the dynamics and objectives of the specific motor task. Such tasks are often modeled as Linear Quadratical Gaussian (LQG) problems, rendering them computationally tractable (Todorov and Jordan 2002). *Given* objective or policy in an LQG setting, internal process models can be estimated by maximization of the expected likelihood of the observed actions (Golub, Chase, and Byron 2013). However, in natural tasks normally neither objective nor policy are known beforehand. Hence, (Phatak et al. 1976) already addressed identification of *both* sensor model and optimal policy. In that work, an estimation approach was proposed for optimal behavior in the special case of time-invariant infinite-horizon LQGs.

Similar as (ibid.) we address joint inference of policy and sensor model. However, we relax the assumption of optimal behavior to rational behavior according to the Maximum Causal Entropy (MCE) framework (Ziebart, Bagnell, and Dey 2010). Furthermore, a new abstract approach for inference of sensor models in *arbitrary* Partial Observable Markov Decision Processes (POMDP)s by application of the ideas of Simultaneous Estimation of Rewards and Dynamics (SERD) (Herman et al. 2016) is presented. In addition to that, we derive a concrete implementation of the concept in the problem class of Sensor Scheduling LQGs (SLQG)s. In contrast to ordinary LQGs (Phatak et al. 1976; Golub, Chase, and Byron 2013; Chen and Ziebart 2015) that are characterized by a single static sensor model, SLQGs allow control of the sensor model for active information gathering. Thereby, we both extend the implementation of SERD for small discrete and fully-observable problems (Herman et al. 2016) to SLQGs and previous work on IOC in SLQGs (Schmitt et al. 2016a) to inference of sensor models. Finally, we demonstrate the effectiveness of the proposed method on data of a new driving experiment conducted in real traffic.

Background

In the following section, we briefly review the mathematical background of POMDPs as well as previous results on MCE-IOC and SERD which this work builds on.

POMDPs

The mathematical basis of OC is provided by Markov Decision Processes (MDP). Here, an agent acts in a world defined by states $x_t \in \mathcal{X}$. Applying an action $u_t \in \mathcal{U}$, the world changes according to a stochastic process $\mathcal{P}(x_{t+1}|u_t, x_t)$. The objective of the agent in a finite-horizon MDP is finding a policy π that maximizes expected cumulated reward $\mathbb{E}\big[\sum_{t=0}^{T} r(u_t, x_t)\big|\pi, \mathcal{P}, p_0\big]$ over a horizon T, given a reward model $r(u_t, x_t)$ and an initial state distribution $p_0(x_0)$. Partial observable Markov decision processes further allow for agents that have no direct access to the states x_t but instead obtain sensory measurements $z_t \in \mathcal{Z}$ according to a sensor model $p^z(z_t|x_t)$. Using the Bayes filter every POMDP can be transferred into an equivalent MDP in the belief-states $b(x_t)$ using the process model

$$\begin{aligned}&\mathcal{P}^b(b(x_t)|z_t, u_{t-1}, b(x_{t-1})) \qquad (1)\\ &\quad\propto \mathbb{E}[p^z(z_t|x_t)\mathcal{P}(x_t|u_{t-1}, x_{t-1})|u_{t-1}, b(x_{t-1})]\end{aligned}$$

and the reward model $r^b(u_t, b(x_t)) = \mathbb{E}[r(u_t, x_t)|b(x_t)]$. Hence, POMDPs can be understood as a natural integration of perception by Bayesian inference into optimal control.

MCE-IOC

Given a set of behavioral data $\{x_t, u_t\}^i$, $t = 0, 1, \ldots, T$, $i = 1, 2, \ldots, n$, MCE-IOC infers a policy that could have generated the data. Here, the policy model is recursively defined by the so-called soft Bellman equations:

$$\tilde{Q}^\theta_T(u_T, x_T) = \theta^\top \varphi(u_T, x_T) \qquad (2)$$

$$\tilde{V}^\theta_t(x_t) = \log \int \exp(\tilde{Q}^\theta_t(u_t, x_t))\,\mathrm{d}\,u_t \qquad (3)$$

$$\tilde{Q}^\theta_t(u_t, x_t) = \theta^\top \varphi(u_t, x_t) + \mathbb{E}\big[\tilde{V}^\theta_{t+1}(x_{t+1})\big|\mathcal{P}\big] \qquad (4)$$

$$\tilde{\pi}^\theta_t(u_t|x_t) = \exp(\tilde{Q}^\theta_t(u_t, x_t) - \tilde{V}^\theta_t(x_t)) \qquad (5)$$

with the soft cost-to-go function $\tilde{Q}^\theta$, soft value function $\tilde{V}^\theta$, reward parameters θ and reward features $\varphi(u_t, x_t)$ (Ziebart, Bagnell, and Dey 2010). An important property of the MCE policy $\tilde{\pi}$ is the relationship

$$\log \tilde{\pi}^\theta_t(u_t|x_t) \propto \mathbb{E}\big[\sum_{t'=t}^{T} \theta^\top \varphi(u_{t'}, x_{t'})\big|\tilde{\pi}, \mathcal{P}\big], \qquad (6)$$

i.e. the higher its expected cumulated value of $\theta^\top \varphi(u_t, x_t)$ under $\tilde{\pi}^\theta$ the higher the likelihood of $\tilde{\pi}^\theta_t(u_t|x_t)$, what leverages the interpretation as a stochastic generalization of the OC policy for the reward $\theta^\top \varphi(u_t, x_t)$ (ibid.).

The reward parameters θ can be obtained by minimization of the Lagrangian dual of the MCE problem,

$$\mathcal{D}(\theta) = \mathbb{E}\Big[\tilde{V}^\theta_0(x_0)\Big|p_0\Big] - \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T} \theta^\top \varphi(\{x_t, u_t\}^i) \quad (7)$$

(ibid.) using its gradient $\nabla_\theta \mathcal{D}(\theta)$ which is given by

$$\begin{aligned}&\mathbb{E}\Big[\nabla_\theta \tilde{Q}^\theta_0(u_0, x_0)\Big|\tilde{\pi}^\theta_0, p_0\Big] - \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T} \varphi(\{x_t, u_t\}^i) \qquad (8)\\ &= \mathbb{E}\Big[\sum_{t=0}^{T} \varphi(x_t, u_t)\Big|\tilde{\pi}^\theta, \mathcal{P}, p_0\Big] - \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T} \varphi(\{x_t, u_t\}^i).\end{aligned}$$

The approach is characterized by statistical robustness (ibid.) and can also be extended to LQGs by application to the corresponding belief-MDPs (Chen and Ziebart 2015).

SERD

Consider a process model $\mathcal{P}^\lambda$ differentiable w.r.t. parameters λ. The main idea of simultaneous estimation of rewards and dynamics (Herman et al. 2016) is that observed behavior according to an MCE policy contains information regarding λ. This is because all three terms $\tilde{V}^{\theta,\lambda}, \tilde{Q}^{\theta,\lambda}, \tilde{\pi}^{\theta,\lambda}$ are also differentiable functions of λ. Specifically, (dropping the dependence on x, u) the gradient $\nabla_\lambda \tilde{Q}^{\theta,\lambda}_t$ is given by

$$\mathbb{E}\big[\nabla_\lambda \tilde{Q}^{\theta,\lambda}_{t+1}\big|\tilde{\pi}^{\theta,\lambda}_{t+1}, \mathcal{P}^\lambda\big] + \int \big[\nabla_\lambda \mathcal{P}^\lambda\big]\tilde{V}^{\theta,\lambda}_{t+1}\,\mathrm{d}\,x_{t+1} \qquad (9)$$

(ibid). Hence, λ can also be estimated by minimization of the extended dual function $\mathcal{D}(\theta, \lambda)$ using the gradient

$$\begin{aligned}\nabla_\lambda \mathcal{D}(\theta, \lambda) &= \mathbb{E}\Big[\nabla_\lambda \tilde{Q}^{\theta,\lambda}_0\Big|\tilde{\pi}^{\theta,\lambda}_0, p_0\Big] \qquad (10)\\ &= \mathbb{E}\Big[\sum_{t=0}^{T} \nabla_\lambda \theta^\top \mathbb{E}[\varphi(x_t, u_t)|b_\lambda(x_t)]\\ &\quad + \sum_{t=0}^{T-1} \int \big[\nabla_\lambda \mathcal{P}^\lambda\big]\tilde{V}^{\theta,\lambda}_{t+1}\,\mathrm{d}\,x_{t+1}\ \Big|\tilde{\pi}^{\theta,\lambda}, \mathcal{P}^\lambda, p_0\Big].\end{aligned}$$

Inferring Sensor Models

Using the idea of SERD allows to infer an unknown sensor model p^z_λ of *any* belief process model $\mathcal{P}^b_\lambda$ without knowing the sensory measurements z^i_t: First, the unbiased estimator $\theta^\top\varphi(\{x_t,u_t\}^i)$ can be used instead of the unknown $\theta^\top\mathbb{E}[\varphi(\{x_t,u_t\}^i)|b_\lambda(\{x_t\}^i)]$ for Eq. (7), assuming that the true states $\{x_t\}^i$ are known at inference time. Second, the computation of $\mathcal{D}(\theta,\lambda)$ and $\nabla_{\theta,\lambda}\mathcal{D}(\theta,\lambda)$ requires only integration over *possible* z_t under the current iterate p^z_λ and not the actual z^i_t. Essentially, the sensor model is estimated from its influence on the MCE policy and the corresponding feature expectation.

For practical application computationally tractable Eq. (2)-(10) are required. However, even computation of the soft Bellman equations Eq. (2) is often infeasible for POMDPs, similar to their classic counterparts.

SLQG Implementation

In the remaining part of this work we will address the POMDP class of SLQGs. Building on the work of (Schmitt et al. 2016a), it is shown that most parts of Eq. (2)-(10) allow exact and tractable computation, while there exists an approximation technique and tractable special cases for the remaining ones.

Definition of Class

Sensor scheduling LQGs are characterized by *primary* states x^p_t and actions u^p_t subject to linear-affine dynamics

$$x^\mathrm{p}_{t+1} = A_t x^\mathrm{p}_t + B_t u^\mathrm{p}_t + a_t + \epsilon_t \tag{11}$$

with i.i.d. Gaussian disturbances $\mathcal{N}(\epsilon_t|\mathbf{0},\Sigma)$. The reward function on $x^\mathrm{p}_t, u^\mathrm{p}_t$ is a quadratic form,

$$\theta_p^\top\varphi(u^\mathrm{p}_t,x^\mathrm{p}_t) = x^{\mathrm{p}\,\top}_t\Theta^\mathrm{p}_1 x^\mathrm{p}_t + u^{\mathrm{p}\,\top}_t\Theta^\mathrm{p}_2 u^\mathrm{p}_t, \tag{12}$$

with negative semi-definite matrix Θ^p_1 and negative definite Θ^p_2. This is combined with a set of linear Gaussian sensor models $p^z_\lambda(z_t|x^\mathrm{p}_t;x^z_t) = \mathcal{N}(z_t|H(x^z_t)x^\mathrm{p}_t,\Sigma^z_\lambda(x^z_t))$ parametrized by a sensor state $x^z_t \in \mathcal{X}^z$. The individual sensor models can be switched by means of actions $u^z_t \in \mathcal{U}^z$, $x^z_{t+1} = x^z_{t+1}(u^z_t,x^z_t)$ subject to a reward $\theta_z^\top\varphi(u^z_t,x^z_t)$. Therefore, SLQGs allow to model problems of active information gathering. Their application for modeling human behavior was first proposed in (Baron and Kleinman 1969) and has e.g. been used to predict gaze switching and steering behavior of automobile drivers (Schmitt et al. 2016b).

For the sake of readability, we assume that unknown parameters λ are present in the sensor noise covariances Σ^z_λ only. The matrices $H(x^z_t)$ can be estimated in similar fashion.

SLQGs can be transformed into belief-MDPs in the following way: First, we use $\mathbf{x}^z_t$ to denote the sequence of previous sensor states $(x^z_0,x^z_1,\dots,x^z_t)$, given an initial state x^z_0. Thereafter the reward of the belief b_t, $\mathbb{E}[\theta_\mathrm{p}^\top\varphi_\mathrm{p}(x^\mathrm{p}_t,u^\mathrm{p}_t)|b_t]$ is given by $\mu^{\mathrm{p}\,\top}_t\Theta_1\mu^\mathrm{p}_t + \mathrm{tr}(\Theta_1\Sigma^\mathrm{p}_\lambda(\mathbf{x}^z_t)) + u^{\mathrm{p}\,\top}_t\Theta_2 u^\mathrm{p}_t$, where the variables $\mu^\mathrm{p}_t, \Sigma^\mathrm{p}_\lambda(\mathbf{x}^z_t)$ result from the Kalman filter,

$$\Sigma^\mathrm{p}_\lambda(\mathbf{x}^z_{t+1}) = \hat{\Sigma}^\mathrm{p}_\lambda(\mathbf{x}^z_{t+1}) - \Sigma^\mu_\lambda(\mathbf{x}^z_{t+1}) \tag{13}$$

$$\mu^\mathrm{p}_{t+1} \sim \mathcal{N}(\mu^\mu_{t+1},\Sigma^\mu_\lambda(\mathbf{x}^z_{t+1})) \tag{14}$$

$$\mu^\mu_{t+1} := A_t\mu^\mathrm{p}_t + Bu^\mathrm{p}_t + a_t \tag{15}$$

$$\Sigma^\mu_\lambda(\mathbf{x}^z_{t+1}) := K_\lambda(\mathbf{x}^z_{t+1})H(x^z_{t+1})\hat{\Sigma}^\mathrm{p}_\lambda(\mathbf{x}^z_{t+1}) \tag{16}$$

$$K_\lambda(\mathbf{x}^z_{t+1}) := \hat{\Sigma}^\mathrm{p}_\lambda(\mathbf{x}^z_{t+1})H(x^z_{t+1})^\top S_\lambda(\mathbf{x}^z_{t+1})^+ \tag{17}$$

$$S_\lambda(\mathbf{x}^z_{t+1}) := H(x^z_{t+1})\hat{\Sigma}^\mathrm{p}_\lambda(\mathbf{x}^z_{t+1})H(x^z_{t+1})^\top + \Sigma^z_\lambda(x^z_{t+1})$$

$$\hat{\Sigma}^\mathrm{p}_\lambda(\mathbf{x}^z_{t+1}) := A_t\Sigma^\mathrm{p}_\lambda(\mathbf{x}^z_t)A_t^\top + \Sigma. \tag{18}$$

Here, $\mathrm{tr}(S)$ denotes the sum of diagonal elements and S^+ the Moore-Penrose pseudo-inverse of matrix S.

Policies

Applying the soft Bellman equations, results in functions

$$\tilde{V}^{\theta,\lambda}_t = \mu^{\mathrm{p}\,\top}_t\bar{\Omega}^1_t\mu^\mathrm{p}_t + \mu^{\mathrm{p}\,\top}_t\bar{\omega}^2_t + \bar{\omega}^3_t + \hat{\tau}^\lambda_t(\mathbf{x}^z_t), \tag{19}$$

$$\begin{aligned}\tilde{Q}^{\theta,\lambda}_t = {} & [\mu^\mathrm{p}_t;u^\mathrm{p}_t]^\top\Omega^1_t[\mu^\mathrm{p}_t;u^\mathrm{p}_t] + [\mu^\mathrm{p}_t;u^\mathrm{p}_t]^\top\omega^2_t + \omega^3_t \\ & + \theta_z^\top\varphi(u^z_t,x^z_t) + \hat{\tau}^\lambda_{t+1}(\mathbf{x}^z_{t+1}(u^z_t,\mathbf{x}^z_t)) \\ & + \mathrm{tr}\big(\Theta_1\Sigma^\mathrm{p}_\lambda(\mathbf{x}^z_t) + \hat{\Omega}^1_{t+1}\Sigma^\mu_\lambda(\mathbf{x}^z_{t+1}(u^z_t,\mathbf{x}^z_t))\big),\end{aligned} \tag{20}$$

where the terms $\bar{\Omega}^1_t,\Omega^1_t,\bar{\omega}^2_t,\omega^2_t,\bar{\omega}^3_t,\omega^3_t$ can be obtained in closed-form and independently of states x^z_t (Schmitt et al. 2016a). Here, the expression $[y;x]$ denotes the vertical concatenation of vectors x,y. Hence, the resulting policy can be split into two parts:

$$\tilde{\pi}_t(u^\mathrm{p}_t,u^z_t|\mu^\mathrm{p}_t,\mathbf{x}^z_t) = \mathcal{N}(u^\mathrm{p}_t|F_t\mu^\mathrm{p}_t + f_t,\Sigma^f_t)\tilde{\pi}_t(u^z_t|\mathbf{x}^z_t).$$

Note, that the terms in $\tilde{V}^{\theta,\lambda}_t,\tilde{Q}^{\theta,\lambda}_t$ which depend on $\mathbf{x}^z_t$ can become problematic as the size of its state space is $|\mathcal{X}^z|^t$. Hence, computation of the integral in Eq. (3) is prohibitively expensive even for moderate time horizons and small numbers of sensor models.

There are two options to address this issue. First, there exist special cases were the state space of $\mathbf{x}^z_t$ remains tractable. This is obviously the case if $|\mathcal{X}^z| = 1$, i.e. in ordinary LQGs (Chen and Ziebart 2015). Furthermore, also regularly enforcing switches to a model of perfect sensing of x^p_t, as in (Schmitt et al. 2016a), can reduce the computational burden. This is because shorter sequences $(x^z_{t'},x^z_{t'+1},\dots,x^z_t)$ from the last t' perfect sensing have to be considered. An alternative follows from considering the *maximum likelihood* sequence $\hat{\mathbf{x}}^z_T$. Applying property Eq. (6) of $\tilde{\pi}$ it holds

$$\begin{aligned}\hat{\mathbf{x}}^z_T = \arg\max_{\mathbf{x}^z_T}\sum_{t=0}^{T}\theta_z^\top\varphi(u^z_t,x^z_t) + \mathrm{tr}\big(\Theta_1\Sigma^\mathrm{p}_\lambda(\mathbf{x}^z_t)\big) \\ + \sum_{t=0}^{T-1}\mathrm{tr}\big(\hat{\Omega}^1_{t+1}\Sigma^\mu_\lambda(\mathbf{x}^z_{t+1})\big).\end{aligned} \tag{21}$$

Such *optimal* sensor scheduling problems can be solved by exploiting properties of the Kalman covariance update to prune the search tree (Vitus et al. 2012). Therefore, the feature counts of $\hat{\mathbf{x}}^z_T$ may serve as a tractable approximation of the expectations required for $\nabla_{\theta,\lambda}\mathcal{D}$. A similar maximum likelihood approximation for $\nabla_{\theta,\lambda}\mathcal{D}$ has successfully been used by (Kuderer, Gulati, and Burgard 2015) for inference of rewards of driver trajectory planning.

Reward Gradients

Given the policy $\tilde{\pi}^{\theta,\lambda}$, the reward gradients are obtained by

$$\nabla_{\Theta_1}\tilde{Q}_0^{\theta,\lambda} = \mathbb{E}\Big[\sum_{t=0}^{T}\mu_t^{\mathrm{p}}\mu_t^{\mathrm{p}\,\top} + \Sigma_\lambda^{\mathrm{p}}(\mathbf{x}_t^z)\Big|\tilde{\pi}^{\theta,\lambda},\mathcal{P}^\lambda,p_0\Big] \quad (22)$$

$$\nabla_{\Theta_2}\tilde{Q}_0^{\theta,\lambda} = \mathbb{E}\Big[\sum_{t=0}^{T}u_t^{\mathrm{p}}u_t^{\mathrm{p}\,\top}\Big|\tilde{\pi}^{\theta,\lambda},\mathcal{P}^\lambda,p_0\Big] \quad (23)$$

$$\nabla_{\theta_z}\tilde{Q}_0^{\theta,\lambda} = \mathbb{E}\Big[\sum_{t=0}^{T}\varphi(u_t^z,x_t^z)\Big|\tilde{\pi}^{\theta,\lambda},\mathcal{P}^\lambda,p_0\Big]. \quad (24)$$

Here the expectations can either be approximated using the maximum likelihood sequence $\hat{\mathbf{x}}_T^z$ or exactly be computed by recursion (Schmitt et al. 2016a) in the special cases mentioned before.

Sensor Model Gradients

For computation of $\nabla_\lambda\mathcal{D}$ the gradients $\nabla_\lambda\tilde{Q}_t^{\theta,\lambda}$ are required. Therefore, we consider the terms of $\tilde{Q}_t^{\theta,\lambda}$ that depend on λ:

$$\hat{\tau}_{t+1}^{\lambda}(\mathbf{x}_{t+1}^z) + \mathrm{tr}\big(\Theta_1\Sigma_\lambda^{\mathrm{p}}(\mathbf{x}_t^z) + \hat{\Omega}_{t+1}^1\Sigma_\lambda^{\mu}(\mathbf{x}_{t+1}^z)\big). \quad (25)$$

When forming the gradient $\nabla_\lambda\tilde{Q}_t^{\theta,\lambda}$, the gradient of the first part, $\hat{\tau}_{t+1}^{\lambda}(\mathbf{x}_{t+1}^z)$, is the sum of the gradients $\nabla_\lambda\tilde{Q}_{t+1,\dots,T}^{\theta,\lambda}$ succeeding in time. The gradient of the second part $\mathrm{tr}\big(\Theta_1\Sigma_\lambda^{\mathrm{p}}(\mathbf{x}_t^z) + \hat{\Omega}_{t+1}^1\Sigma_\lambda^{\mu}(\mathbf{x}_{t+1}^z)\big)$ is given by $\partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mathrm{p}}(\mathbf{x}_t^z))^\top\mathrm{vec}(\Theta_1) + \partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mu}(\mathbf{x}_{t+1}^z))^\top\mathrm{vec}(\hat{\Omega}_{t+1}^1)$, where the terms $\partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mathrm{p}}), \partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mu})$ are obtained from derivation of the Kalman filter update Eq.s (13)-(18):

$$\begin{aligned}
\partial_\lambda\mathrm{vec}(\hat{\Sigma}_\lambda^{\mathrm{p}}) =& (A_t\otimes A_t)\partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mathrm{p}}) + \mathrm{vec}(\Sigma^{\mathrm{p}}) && (26)\\
\partial_\lambda\mathrm{vec}(S_\lambda) =& (H\otimes H)\partial_\lambda\mathrm{vec}(\hat{\Sigma}_\lambda^{\mathrm{p}}) + \partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mathrm{o}}) && (27)\\
\partial_\lambda\mathrm{vec}(S_\lambda^+) =& -(S_\lambda^+\otimes S_\lambda^+)\partial_\lambda\mathrm{vec}(S_\lambda) && (28)\\
&+ ((I - S_\lambda S_\lambda^+)\otimes(S_\lambda^+S_\lambda^+))\partial_\lambda\mathrm{vec}(S_\lambda)\\
&+ ((S_\lambda^+S_\lambda^+)\otimes(I - S_\lambda^+S_\lambda))\partial_\lambda\mathrm{vec}(S_\lambda)\\
\partial_\lambda\mathrm{vec}(K_\lambda) =& (S_\lambda^+H\otimes I)\partial_\lambda\mathrm{vec}(\hat{\Sigma}_\lambda^{\mathrm{p}}) && (29)\\
&+ (I\otimes\hat{\Sigma}_\lambda^{\mathrm{p}}H^\top)\partial_\lambda\mathrm{vec}(S_\lambda^+)\\
\partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mu}) =& (\hat{\Sigma}_\lambda^{\mathrm{p}}H^\top\otimes I)\partial_\lambda\mathrm{vec}(K_\lambda) && (30)\\
&+ (I\otimes H^\top K_\lambda^\top)\partial_\lambda\mathrm{vec}(S_\lambda^+)\\
\partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mathrm{p}}) =& \partial_\lambda\mathrm{vec}(\hat{\Sigma}_\lambda^{\mathrm{p}}) - \partial_\lambda\mathrm{vec}(\Sigma_\lambda^{\mu}). && (31)
\end{aligned}$$

Here, I denotes the identity matrix, $\mathrm{vec}(X)$ the vectorization of a matrix by vertical concatenation of columns and $X\otimes Y$ the Kronecker product of matrices X, Y. The derivative of the Moore-Penrose pseudo-inverse X^+ is discussed in (Stewart 1977).

Numerical Evaluation in Application

Our research effort was motivated by the problem of modeling vehicle control and glance behavior of experienced drivers when engaging in additional visually demanding activities. The importance of this motor task arises from the fact that inappropriate gaze behavior while driving, i.e. visual distraction, caused a significant proportion of U.S. road fatalities in 2014, especially among novice drivers (NHTSA 2016). Specifically, we apply the proposed approach to infer models for sensing of position and orientation of the vehicle by its driver. Here, previous work (Summala, Nieminen, and Punto 1996) showed that the amount of glance deviation from the direction of the road scenery correlates with decrease in driving performance.

SLQG model

Similar as (Schmitt et al. 2016b), we consider additional visually demanding activities in the driving task of lane keeping. This can be modeled by the following SLQG: The vehicle states are its position in lane y_t, its lateral velocity $\dot{y}_t$, its orientation w.r.t. the tangent of the lane center line ϕ_t and the angle of the steering wheel α_t that can be controlled by means of the steering angle velocity $\dot{\alpha}_t$. We use the linear-affine kinematic vehicle model (Risack et al. 1998)

$$\begin{bmatrix}\dot{y}_t\\ \dot{\phi}_t\\ \dot{\alpha}_t\end{bmatrix} = \begin{bmatrix}0 & v_t & 0\\ 0 & 0 & c\,v_t\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}y_t\\ \phi_t\\ \alpha_t\end{bmatrix} + \begin{bmatrix}0\\0\\1\end{bmatrix}\dot{\alpha} + \begin{bmatrix}0\\ -v_t\kappa_t\\ 0\end{bmatrix} \quad (32)$$

integrated at 25 Hz and subject to random disturbances on the dynamics of states y_t, ϕ_t. Here, the steering wheel transmission ratio c is a constant, while the vehicle's speed v_t and the lane curvature κ_t parametrize all possible situations. The reward on the primary states was modeled by $\varphi^{\mathrm{p}}(x_t^{\mathrm{p}},u_t^{\mathrm{p}}) = [y_t^2;\dot{y}_t^2;\alpha_t^2;\dot{\alpha}_t^2]$, as first experiments did not show benefits of more complex reward models.

We assume that drivers obtain sensory measurements of states y_t, ϕ_t, α_t only. Thereof, the steering angle α_t is perfectly sensed as it is fully given by its derivative, i.e. actions $\dot{\alpha}_t$. Hence, the sensor model can be formalized by

$$H = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix},\ \Sigma^z(x_t^z) = \begin{bmatrix}\lambda_{x_t^z}^{y} & 0 & 0\\ 0 & \lambda_{x_t^z}^{\phi} & 0\\ 0 & 0 & 0\end{bmatrix}. \quad (33)$$

Four sensor states $x_t^z \in \{\mathrm{R, H, C, N}\}$ are considered that correspond to the driver glancing at the Road (R), the Head-up Display (H), the Combi Instrument (C) and the Navigation System (N), depicted in Fig. 1. In any situation the

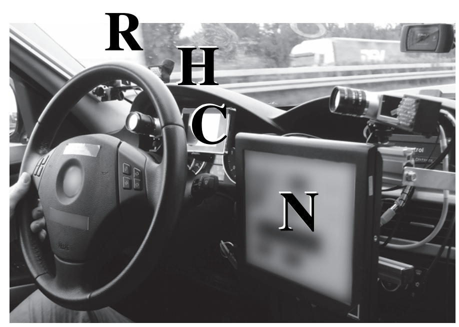

Figure 1: Sensor states x_t^z: Road (R), Head-up Display (H), Combi Instrument (C) and Navigation System (N)

driver can glance at the road, i.e. $x_t^z = \mathrm{R}$. If the driver is

engaging in a secondary task, he/she has to switch gaze between (R) and one of the displays {H, C, N}. We assume a constant cost for the switching effort and constant utility for glancing at the required display. This is formalized by the reward model $\varphi_z(u_t^z, x_t^z) = [\mathbb{I}(x_t^z = \mathrm{R}); u_t^z]$, where $\mathbb{I}$ denotes the indicator function.
To ensure tractable exact solution as in (Schmitt et al. 2016a), we truncated the sequence $\mathbf{x}_t^z$ at the last glance on the road, i.e $t' : x_t^z = \mathrm{R}$. This was done assuming that once the glance returns to the road the belief $b(x_t^\mathrm{p})$ jumps to the state it had when glancing at the road instead of the display $x_k^z = \mathrm{R}$, $k = t', \dots, t$. Additionally, we imposed a maximum length of $\mathbf{x}_t^z$ of $l_{\max} = 187$ elements corresponding to glances lengths of 7.5 s.

Database

Data for inference was obtained by a driving experiment in real traffic similar to that of (Schmitt et al. 2016b). 17 drivers were recruited which all had attended at least two driving safety trainings prior to participation. The experiment consisted of driving at speeds $v_t \in \{80, 90, 110\}$ km/h on a public motorway. Here, speed and distance to preceding vehicles were controlled by the vehicle's Adaptive Cruise Control (ACC), so that the driving task was keeping the vehicle in lane alone. At each speed four experimental conditions were triggered by an instructor: Driving without a secondary task and driving while engaging in a secondary task that required glancing at one of the displays {H, C, N}. The resulting glance statistics are shown in Fig. 2.

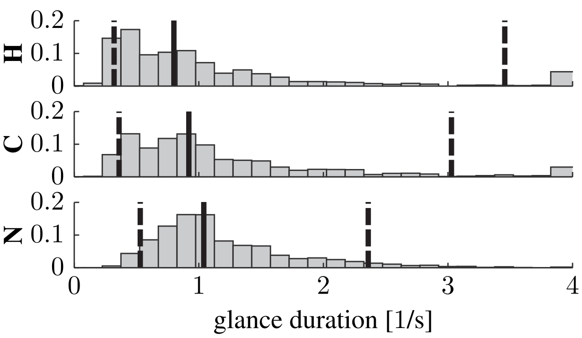

Figure 2: Histograms of durations of glances on displays {H, C, N}. Solid vertical line indicates the median, dashed vertical lines the $0.1, 0.9$ quantiles.

The task consisted of incrementally typing 30 random digits 1 or 2 which were displayed on a small screen (see Fig. 1 at letter C). Typing all digits required 40 s on average. Similar to previous work (ibid.), the participants were instructed to *"perform the secondary task as quickly and correctly as possible while not endangering driving safety"*. All vehicle states x^p were recorded by commercial in-vehicle sensors, while a commercial infra-red camera system (see Fig. 1) was used to track the gaze of the driver and to estimate state x^z. At each of the three driving speeds three repetitions of the task at all three displays and of driving without task were conducted. This resulted in a total amount of four hours of behavioral data for inference.

Numerical Experiment

For inference of sensor models and policies the collected driving periods were subdivided into snippets of exactly 5 s length. Using the states $x_t^\mathrm{p}, \mathbf{x}_t^z$ at the beginning of the snipped as initial state, 100 state/action sequences were sampled based on the vehicle model and the inferred sensor models and policies. Finally, prediction quality was assessed with respect to the following metrics:

1. Kullback Leibler divergence

$$\mathrm{KL} := \sum_{d=0}^{l_{\max}} p(d^i)(\log[p(d^i)] - \log[p(d)]), \tag{34}$$

   between the distributions of observed d^i and predicted off road glance duration $d_t = \min(k : x_{t-k}^z = \mathrm{R}, k \geq 0)$.

2. Average neg. log-likelihood of the observed states $\{x^\mathrm{p}\}^i$

$$\mathrm{NLL} := -\frac{1}{T} \sum_{t=0}^{T} \log p(\{x_t^\mathrm{p}\}^i | \pi, \mathcal{P}, p_0), \tag{35}$$

   based on Gaussian approximation of the predicted state distribution $p(x_t^\mathrm{p})$.

We compared our approach to those previously proposed, as well as MCE SLQG and a direct policy estimation baseline using a "best guess" sensor model:

1 Direct Policy Estimation by linear regression for sub-policy $\pi(u_t^\mathrm{p}|\mu_t^\mathrm{p})$ and logistic regression for sub-policy $\pi(u_t^z|d_t)$ based on the durations as in (Schmitt et al. 2016a). Here, λ_R^y, λ_R^ϕ were chosen to a signal-to-noise ratio of $10^{1.33}$. This corresponds to a 96% confidence interval for estimation of the lane position of 0.3 m at a constant speed of $v = 80$ km/h. The remaining parameters $\lambda_{\mathrm{H,C,N}}^y$, $\lambda_{\mathrm{H,C,N}}^\phi$ were set to ∞, what implements a driver who does not sense y, ϕ when not glancing at the road.

2 MCE SLQG using the same parameters λ as in 1.

3 Estimation of λ by maximization of the expected log-likelihood of the observed actions $\{u_t^\mathrm{p}\}^i$ $\mathbb{E}\Big[\log[\pi^\theta(\{u_t^\mathrm{p}\}^i|\mu_t^\mathrm{p})]\Big|\{x_t^\mathrm{p}\}^i, \Sigma_\lambda^\mathrm{p}(\{\mathbf{x}_t^z\}^i)\Big]$ (Golub, Chase, and Byron 2013). Here, the MCE policy $\pi^\theta(u_t^\mathrm{p}|x_t^\mathrm{p})$ for randomly sampled parameters θ was used in the first step. This was followed by MCE SLQG to infer all reward parameters using the obtained parameters $\lambda_{\mathrm{R,H,C,N}}^{y,\phi}$.

4 Estimation of sensor models *and* rewards by maximization of the expected log-likelihood of the observed actions $\{u_t^\mathrm{p}\}^i$ under the MCE policy model Eq. (5), interpretable as an MCE version of (Phatak et al. 1976). This was followed by MCE SLQG to infer all reward parameters using the obtained parameters $\lambda_{\mathrm{R,H,C,N}}^{y,\phi}$.

5 Our approach for joint inference of sensor model and rewards.

For numerical optimization problem **1** was cast as inference in a generalized linear model (Nelder and Baker 1972), while we employed an interior-point method to solve **2-5** subject to $\theta \leq 0$ and $\lambda_{\mathrm{R,H,C,N}}^{y,\phi} \geq 0$.

Results

We report on the metrics on the tests sets of 5 Monte Carlo cross-validations (50%/50% split). Here, table 1 presents the medians of the skewed-distributed values of both metrics, where the least prediction errors are in **bold** digits.

Table 1: Results of the numerical experiment

	Methods				
	1	**2**	**3**	**4**	**5**
KL	$+0.64$	$+0.55$	$+0.28$	$+0.27$	$\mathbf{+0.25}$
NLL	-7.00	$\mathbf{-8.58}$	-8.52	-8.56	-8.58

In the numerical experiment, the optimization problems **4-5** turned out to be notoriously unbounded if the sensor model for driving while glancing at the road, $\lambda_{\mathrm{R}}^{y,\phi}$ was inferred. Problem **3** was not affected by unboundedness. Therefore, the results are reported for parameters $\lambda_{\mathrm{R}}^{y,\phi}$ fixed to a signal-to-noise ratio of $10^{1.33}$ similar to **1**,**2**. Interestingly, this did not have any negative impact on the approach **3**.
In the numerical experiments approaches **2-5** had a significantly (signed-rank $p < 0.01$) lower prediction error than the behavior cloning baseline with respect to both metrics. Although **2** had the lowest neg. log-likelihood the differences to approaches **3-5** were not statistically significant (signed-rank $p > 0.01$). In contrast, both the differences in KL between **2** and **3-5** as well as those between **3-4** and **5** were statistically significant (signed-rank $p < 0.01$).

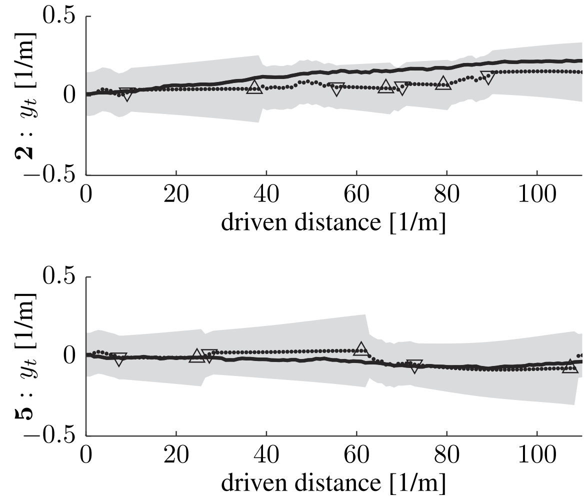

Figure 3: Sample for display H using methods **2** and **5**. Thick line $(-)$ is the actual vehicle's trajectory, dotted line $(\cdots)$ the estimated trajectory, shaded area the 96% confidence interval of the estimation. Triangles (Δ/∇) indicate switches to $\mathrm{R/H}$.

Discussion

The results of the numerical evaluation, show a clear benefit of inference of sensor models in addition to the estimation of policy models. The higher KL of MCE SLQG using the best guess sensor model could mainly be attributed to shorter glances off the road and more frequent switches for the displays H,C. Fig. 3 gives an exemplary comparison between samples from the MCE SLQG policies with the best guess sensor model and with the estimated sensor model.
Although we expected differences in the neg. log-likelihood between approaches **2-5** the results may be explained by the characteristics of the collected data. First, measuring the lane position y_t and orientation ϕ_t is limited in accuracy and therefore small changes cannot be detected. Second, the choice of the lane position in real-world driving is also influenced by aspects that are not considered in the model SLQG, such as the presence of vehicles in the neighboring lane. Hence, the prediction errors of MCE SLQG based approaches may rather be dominated by unmodelled influences than by the used sensor model.
The observation of unbounded optimization problems **4-5** in the case of estimation of $\lambda_{\mathrm{R}}^{y,\phi}$ is in line with the theoretical analysis of over-parametrization in (Phatak et al. 1976): If no data of the sensory measurements obtained by the human is available, deviation from the optimal linear-affine policy $u_t^{\mathrm{p}} = \hat{\pi}(x_t^{\mathrm{p}})$ can be explained by a stochastic policy $\pi^1(u_t^{\mathrm{p}}|x_t^{\mathrm{p}}) = \hat{\pi}(x_t^{\mathrm{p}}) + \epsilon_1$ (as the MCE policy model), the optimal policy acting on noisy estimates of the state $\pi^2(u_t^{\mathrm{p}}|x_t^{\mathrm{p}}) = \hat{\pi}(\mu_t^{\mathrm{p}}) = \hat{\pi}(x_t^{\mathrm{p}} + \epsilon_2)$ or a combination of both.

In the approach **3** this issue is avoided, as the policy and sensor model are not estimated simultaneously. However, it comes at the cost of higher KL in the numerical experiment of this work. In contrast to $\lambda_{\mathrm{R}}^{y,\phi}$, sensor model parameters $\lambda_{\mathrm{H,C,N}}^{y,\phi}$ could be inferred under the MCE policy model in the numerical experiment. This is because the noise covariance $\Sigma_{\lambda}^{\mathrm{p}}(\mathbf{x}_t^z)$ of the state estimate μ_t^{p} in Eq. (13) changed over time if $x_t^z \in \{\mathrm{H}, \mathrm{C}, \mathrm{R}\}$ (see also Fig. 3) which allows to separate the influences of the stochastic policy and the noisy sensory measurements.

Conclusion and Future Work

We presented a general method for inference of sensor and policy models of motor behavior. Here, the key elements are the assumption of *perception by Bayesian inference* and rational choice of actions. Specifically, exploiting the differentiability of the maximum causal entropy dual $\mathcal{D}(\theta, \lambda)$ allows to infer both reward parameters θ and sensor model parameters λ. We considered the concrete implementation for sensor scheduling LQGs that generalize LQGs addressed in most of the previous work. Finally, the approach was evaluated in the important application domain of modeling driver behavior. The results show that prediction of the driver's glance behavior in the presence of different visually demanding additional activities can be improved using the presented methodology.
Despite the promising first results, some issues with the proposed approach need to be addressed. As discussed problems of over-parametrization were present in the considered application. Hence, future work should investigate if and how sensor models obtained in *laboratory* can be used to define priors for inference from *real-world* data. Addition-

ally, criteria for a-priory detection of over-parametrization as (Acerbi, Ma, and Vijayakumar 2014) are relevant. This would allow to find an appropriate parametrization and an experimental design for data collection. Although SLQGs can be used to model real-world active information gathering as demonstrated, the class poses strong restriction on reward models and dynamics. Recently, human motor behavior has successfully been modeled by approximate solution of more complex POMDPs (Belousov et al. 2016). Those approximation techniques may also be applicable for approximate inference of sensor models in general POMDPs.

Acknowledgments

This work was part of the public project UR:BAN which was co-funded by the Federal Ministry for Economic Affairs and Energy on basis of a decision by the German Bundestag.
F.S. wishes to thank all volunteers for their participation in the experiment.

References

Acerbi, L.; Ma, W. J.; and Vijayakumar, S. 2014. A framework for testing identifiability of bayesian models of perception. In *Advances in Neural Information Processing Systems (NIPS)*, 1026–1034.

Baron, S., and Kleinman, D. L. 1969. The human as an optimal controller and information processor. *IEEE Transactions on Man-Machine Systems* 10(1):9–17.

Belousov, B.; Neumann, G.; Rothkopf, C.; and Peters, J. 2016. Catching heuristics are optimal control policies. In *Advances in Neural Information Processing Systems (NIPS)*.

Chen, X., and Ziebart, B. D. 2015. Predictive inverse optimal control for linear-quadratic-gaussian systems. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 165–173.

Golub, M.; Chase, S.; and Byron, M. Y. 2013. Learning an internal dynamics model from control demonstration. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 606–614.

Herman, M.; Gindele, T.; Wagner, J.; Schmitt, F.; and Burgard, W. 2016. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Knill, D. C., and Richards, W. 1996. *Perception as Bayesian inference*. Cambridge University Press.

Kuderer, M.; Gulati, S.; and Burgard, W. 2015. Learning driving styles for autonomous vehicles from demonstration. In *Proceedings of the IEEE International Conference on Robotics & Automation (ICRA)*, volume 134.

Nash, C. J.; Cole, D. J.; and Bigler, R. S. 2016. A review of human sensory dynamics for application to models of driver steering and speed control. *Biological Cybernetics* 1–26.

Nelder, J. A., and Baker, R. J. 1972. Generalized linear models. *Encyclopedia of Statistical Sciences*.

Neu, G., and Szepesvári, C. 2007. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI'07)*, 295–302.

NHTSA. 2016. Traffic safety facts: Distracted driving 2014. Technical Report DOT HS 812 260, U.S. National Highway Traffic Safety Administration.

Phatak, A.; Weinert, H.; Segall, I.; and Day, C. N. 1976. Identification of a modified optimal control model for the human operator. *Automatica* 12(1):31–41.

Ramachandran, D., and Amir, E. 2007. Bayesian inverse reinforcement learning. In *Proceedings of the International Joint Conference on Artifical Intelligence*.

Risack, R.; Klausmann, P.; Krüger, W.; and Enkelmann, W. 1998. Robust lane recognition embedded in a real-time driver assistance system. In *Proceedings of the IEEE Intelligent Vehicles Symposium*.

Rothkopf, C. A., and Dimitrakakis, C. 2011. Preference elicitation and inverse reinforcement learning. In *Proceedings of the European Conference in Machine Learning and Knowledge Discovery in Databases (ECML)*, 34–48. Springer.

Rothkopf, C. A.; Ballard, D. H.; and Hayhoe, M. M. 2007. Task and context determine where you look. *Journal of vision* 7(14):16–16.

Schmitt, F.; Bieg, H.-J.; Manstetten, D.; Herman, M.; and Stiefelhagen, R. 2016a. Exact maximum entropy inverse optimal control for modeling human attention scheduling and control. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.

Schmitt, F.; Bieg, H.-J.; Manstetten, D.; Herman, M.; and Stiefelhagen, R. 2016b. Predicting lane keeping behavior of visually distracted drivers using inverse suboptimal control. In *Proceedings of the IEEE Intelligent Vehicles Symposium*.

Shimosaka, M.; Kaneko, T.; and Nishi, K. 2014. Modeling risk anticipation and defensive driving on residential roads with inverse reinforcement learning. In *Proceedings of the IEEE 17th Conference on Intelligent Transport Systems (ITSC)*.

Stewart, G. 1977. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review* 19(4):634–662.

Summala, H.; Nieminen, T.; and Punto, M. 1996. Maintaining lane position with peripheral vision during in-vehicle tasks. *Human Factors* 38(3):442–451.

Todorov, E., and Jordan, M. I. 2002. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience* 5(11):1226–1235.

Vitus, M. P.; Zhang, W.; Abate, A.; Hu, J.; and Tomlin, C. J. 2012. On efficient sensor scheduling for linear dynamical systems. *Automatica* 48(10):2482–2493.

Ziebart, B. D.; Bagnell, J.; and Dey, A. K. 2010. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 1255–1262.