A MINIMUM RELATIVE ENTROPY PRINCIPLE FOR ADAPTIVE CONTROL IN LINEAR QUADRATIC REGULATORS

Daniel A. Braun and Pedro A. Ortega
University of Cambridge, Dept. of Engineering, CB2 1PZ Cambridge, U.K.

Keywords: Minimum relative entropy principle, Adaptive control, Bayesian control rule, Linear quadratic regulator.

Abstract: The design of optimal adaptive controllers is usually based on heuristics, because solving Bellman's equations over information states is notoriously intractable. Approximate adaptive controllers often rely on the principle of certainty-equivalence where the control process deals with parameter point estimates as if they represented "true" parameter values. Here we present a stochastic control rule instead where controls are sampled from a posterior distribution over a set of probabilistic input-output models and the true model is identified by Bayesian inference. This allows reformulating the adaptive control problem as an inference and sampling problem derived from a minimum relative entropy principle. Importantly, inference and action sampling both work forward in time and hence such a Bayesian adaptive controller is applicable on-line. We demonstrate the improved performance that can be achieved by such an approach for linear quadratic regulator examples.

1 INTRODUCTION

Learning how to act in an unknown environment poses the problem of adaptive control (Åström and Wittenmark, 1995). Solving adaptive control problems optimally is a notoriously hard problem because it requires the solution of Bellman's optimality equations over large trees of information states, which becomes quickly intractable. Therefore, a number of approximate adaptive control methods have been devised in the literature (Åström and Wittenmark, 1995). Most heuristics for adaptive control are based on the certainty-equivalence principle, i.e. when they estimate the unknown plant parameters, the uncertainty of these estimates has no impact on the pertinent control strategies. Instead, a point estimate of the system parameters is treated as if it represented the "true" system parameters.

It is well known in optimal control theory that the certainty-equivalence principle holds exactly for linear quadratic systems with known dynamics (Åström and Wittenmark, 1995). In case of adaptive control, however, the certainty-equivalence principle breaks down in general and is only used as a heuristic. In fact, previous studies have shown that even for the linear quadratic controller correct closed-loop system identification cannot be guaranteed under certainty-equivalence, which has led to the proposal of cost-biased estimators (Campi and Kumar, 1996). Non-certainty-equivalent controllers are usually designed as extensions of a certainty-equivalent solution, such as *cautious* or *dual* controllers that reduce the control gain in the face of high parameter uncertainty or actively probe the environment by random excitation (Wittenmark, 1975). Here we propose a non-certainty equivalent approach to adaptive control based on a Bayesian control rule derived from a minimum relative entropy principle. We demonstrate how such an approach can be employed to solve adaptive control problems with linear dynamics and quadratic cost.

2 A BAYESIAN RULE FOR ADAPTIVE CONTROL

In the following we assume that the observations of our controller are given by a state variable x_t and the possible actions of our controller are u_t. The controller can then be defined as an input-output system that is characterized by the conditional probabilities

$$P(x_{t+1}|x_{\leq t}, u_{\leq t}) \quad \text{and} \quad P(u_{t+1}|x_{\leq t+1}, u_{\leq t})$$

where $x_{\leq t} = x_1, x_2, \ldots, x_t$ and $u_{\leq t} = u_1, u_2, \ldots, u_t$ denote concatenations of past states and actions respectively. Analogous to the controller, the plant can be thought of as an input-output system with conditional probabilities

$$Q(u_{t+1}|x_{\leq t+1}, u_{\leq t}) \quad \text{and} \quad Q(x_{t+1}|x_{\leq t}, u_{\leq t}).$$

A. Braun D. and A. Ortega P. (2010).
A MINIMUM RELATIVE ENTROPY PRINCIPLE FOR ADAPTIVE CONTROL IN LINEAR QUADRATIC REGULATORS.
In *Proceedings of the 7th International Conference on Informatics in Control, Automation and Robotics*, pages 103-108
DOI: 10.5220/0002938801030108
Copyright © SciTePress

If the controller can perfectly predict the plant for all histories $x_{\leq t}, u_{\leq t}$ then

$$P(x_{t+1}|x_{\leq t}, u_{\leq t}) = Q(x_{t+1}|x_{\leq t}, u_{\leq t}).$$

In this case the plant equation is perfectly known and the controller P can be tailored to the particular plant Q. Especially, the control law $P(u_{t+1}|x_{\leq t+1}, u_{\leq t})$ can be chosen in such a way that it maximizes some optimality criterion given full knowledge of the plant Q.

Consider now the case when the controller does not know the plant dynamics, but assume we know that the plant has dynamics Q_m drawn randomly from a set $\mathcal{Q}$ of possible dynamics indexed by m. Assume further we have available a set of tailored controllers P_m, where each P_m is tailor-made for one of the possible plants Q_m. The set of possible plant dynamics and tailored controllers can then be expressed as conditional probabilities given by the following likelihood and intervention models

$$P(x_{t+1}|m, x_{\leq t}, u_{\leq t}) \quad \text{and} \quad P(u_{t+1}|m, x_{\leq t+1}, u_{\leq t})$$

with $m \in \mathcal{M}$ indexing the different plant dynamics Q_m and the different tailored controllers P_m. How can we now construct a controller P such that its behavior is as close as possible to the tailored controller P_m under any realization of $Q_m \in \mathcal{Q}$?

A convenient measure of how much P deviates from P_m is given by the relative entropy. In particular, we can quantify the average deviation of a control law $P(u_{t+1}|x_{\leq t+1}, \bar{u}_{\leq t})$ from the tailored control law $P(u_{t+1}|m, x_{\leq t+1}, \bar{u}_{\leq t})$ of P_m by computing

$$\left\langle D_{KL}\left(P(u_{t+1}|m, x_{\leq t+1}, \bar{u}_{\leq t}) || P(u_{t+1}|x_{\leq t+1}, \bar{u}_{\leq t})\right)\right\rangle$$

where the average is taken with respect to a prior $P(m)$ and all possible input-output sequences with probabilities $P(x_{\leq t+1}\bar{u}_{\leq t}|m)$. The bar symbol $\bar{u}_{\leq t}$ indicates that past actions have been set by the controller and therefore have to be formalized as interventions (Pearl, 2000; Ortega and Braun, 2010). One can then show that the above quantity is minimized by the following control rule.

Theorem 1 (**Bayesian Control Rule**).

$$P(u_{t+1}|x_{\leq t+1}, \bar{u}_{\leq t}) = \sum_m P(u_{t+1}|m, x_{\leq t+1}, u_{\leq t}) P(m|x_{\leq t+1}, \bar{u}_{\leq t})$$

where $P(m|x_{\leq t+1}, \bar{u}_{\leq t})$ *is given by the recursive expression*

$$P(m|x_{\leq t+1}, \bar{u}_{\leq t}) = \frac{P(x_{t+1}|m, x_{\leq t}, u_{\leq t}) P(m|x_{\leq t}, \bar{u}_{<t})}{\sum_{m'} P(x_t|m', x_{\leq t}, u_{\leq t}) P(m'|x_{\leq t}, \bar{u}_{<t})} \quad (1)$$

The proof can be found in (Ortega and Braun, 2010). Here we apply the Bayesian control rule to adaptive control. It describes a mixture distribution over different tailored controllers indexed by m, each of them suggesting the next control signal u_{t+1} with probability $P(u_{t+1}|m, x_{\leq t+1}, u_{\leq t})$. The mixture weights are given by the posterior probability $P(m|x_{\leq t+1}, \bar{u}_{\leq t})$. It resembles Bayesian inference in that it starts out with a prior distribution over input-output models index by m and computes a posterior distribution after experiencing an interaction. Actions can then be sampled from this posterior distribution.

3 LINEAR QUADRATIC REGULATOR

A linear quadratic regulator is characterized by a linear dynamical system and a quadratic cost function. In the following we will deal with the time-discrete case. Formally, let $\mathbf{x}_t \in \mathbb{R}^N$ be the state vector of the plant at time t, $\mathbf{u}_t \in \mathbb{R}^M$ be the action of the controller, and $\mathbf{F} \in \mathbb{R}^{N\times N}$ and $\mathbf{G} \in \mathbb{R}^{N\times M}$ the time-invariant system matrices describing the dynamics of the plant such that

$$\mathbf{x}_{t+1} = \mathbf{F}\mathbf{x}_t + \mathbf{G}\mathbf{u}_t + \xi_t$$

where $\xi_t \in \mathbb{R}^N$ is a Gaussian random variable with known covariance matrix Ω_ξ. Furthermore, let c_t be the scalar instantaneous cost

$$c_t(\mathbf{x}_t, \mathbf{u}_t) = \mathbf{x}_t^T \mathbf{Q}\mathbf{x}_t + \mathbf{u}_t^T \mathbf{R}\mathbf{u}_t$$

where $\mathbf{R} \in \mathbb{R}^{M\times M}$ is positive definite and $\mathbf{Q} \in \mathbb{R}^{N\times N}$ is positive semi-definite. Thus, the time-average cost J is given by

$$J(\mathbf{x}_t, \mathbf{u}_t) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c_t(\mathbf{x}_t, \mathbf{u}_t).$$

If the matrices $\mathbf{F}, \mathbf{G}, \mathbf{Q}$ and $\mathbf{R}$ are all known the optimal controller has a well-known solution that is a simple state-feedback law

$$\mathbf{u}_t^* = -\mathbf{L}^* \mathbf{x}_t$$

where $\mathbf{L}^*$ can be computed from the algebraic Riccati equation (Stengel, 1993).

3.1 Indirect Adaptive Bayesian Control

In this section we will assume that we know the cost matrices $\mathbf{Q}$ and $\mathbf{R}$, but have to estimate $\mathbf{F}$ and $\mathbf{G}$ during the control process. Since we have to estimate them explicitly in order to compute the optimal policy

$\mathbf{L}^*$ this is often called *model-based* or *indirect* adaptive control. This means we have to deal with an inference problem—estimating $\mathbf{F}$ and $\mathbf{G}$—and an optimal control problem—generating control commands given the estimates $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$.

In order to solve the estimation problem we use an Unscented Kalman Filter (UKF) in our simulation experiments because it can estimate Gaussian random variables both under linear and nonlinear circumstances (Julier and Durrant-Whyte, 1995; Haykin, 2001). The parameter vector we want to estimate is given by the vectorized system matrices $\hat{\mathbf{w}} = \mathrm{vec}([\hat{\mathbf{F}}; \hat{\mathbf{G}}])$. Initially, we assume a Gaussian prior over $\hat{\mathbf{w}}_0$. We model the evolution of the parameter estimate as a Brownian diffusion process given by

$$\hat{\mathbf{w}}_{t+1} = \hat{\mathbf{w}}_t + \omega_t \tag{2}$$

where $\omega \in \mathbb{R}^{N(N+M)}$ is a Gaussian random variable with covariance matrix Ω_ω. The covariance matrix determines the step size of the adaptation process. The likelihood model needed for the inference process is provided by

$$P(\mathbf{x}_{t+1}|\hat{\mathbf{w}}, \mathbf{x}_t, \mathbf{u}_t) \propto e^{-\frac{1}{2}\left(\mathbf{x}_{t+1}-\hat{\mathbf{F}}\mathbf{x}_t-\hat{\mathbf{G}}\mathbf{u}_t\right)^T \Omega_\xi^{-1}\left(\mathbf{x}_{t+1}-\hat{\mathbf{F}}\mathbf{x}_t-\hat{\mathbf{G}}\mathbf{u}_t\right)} \tag{3}$$

The adaptation rate Ω_ω can be adjusted dynamically depending on how well the current parameter estimates fit the observations. In case of poor predictions this should lead to high variability and fast adaptation in big steps, in case of very good predictions this should implies only small adaptation steps. This can be implemented using a Robbins-Monroe innovation update

$$\begin{aligned} \Omega_\omega^{(t+1)} &= (1-\alpha)\Omega_\omega^{(t)} + \alpha \mathbf{I}_t \\ \mathbf{I}_t &= \mathbf{K}_t^{\hat{\mathbf{w}}}\left[\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\right]\left[\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\right]^T (\mathbf{K}_t^{\hat{\mathbf{w}}})^T \end{aligned}$$

where $\mathbf{K}_t^{\hat{\mathbf{w}}}$ is the Kalman gain as used in the UKF and $\hat{\mathbf{x}}_{t+1}$ stems from the prediction step of the UKF (Haykin, 2001).

In order to solve the control problem we have to use the current estimate $\hat{\mathbf{w}} = \mathrm{vec}([\hat{\mathbf{F}}; \hat{\mathbf{G}}])$ to compute the optimal control commands. A certainty-equivalent self-tuning regulator would simply take the mean estimate $\mathbb{E}[\hat{\mathbf{w}}]$ and use this estimate in the algebraic Riccati equation at every point in time as if it was the true parameter vector. While this often works fine if only a few parameters of the matrix are unknown, in general this can lead to suboptimal solutions. Instead, we propose to use the Bayesian control rule as laid down in equation (1). This means we have to specify a likelihood and an intervention model. The likelihood model $P(\mathbf{x}_{t+1}|\hat{\mathbf{w}}, x_{\leq t}, u_{\leq t})$ is given by equation (3). The intervention model is deterministic and given by

$$P(\mathbf{u}_{t+1}|\hat{\mathbf{w}}, \mathbf{x}_{\leq t+1}, \mathbf{u}_{\leq t}) \propto \delta\left(\mathbf{u}_{t+1} + \mathbf{L}_{\hat{\mathbf{w}}}\mathbf{x}_{t+1}\right)$$

It might seem that this would imply taking the entire probability distribution over $\hat{\mathbf{w}}$ and propagating it through the Riccati equation. Then we would sample an $\mathbf{L}$ at each point in time to determine $\mathbf{u}_{t+1}$. Fortunately, an explicit computation of the posterior is not necessary. We can simply sample from the distribution over $\hat{\mathbf{w}}$, propagate this sampled value through the Riccati equation and obtain a sampled policy $\mathbf{L}$. The more precise the estimates over $\hat{\mathbf{w}}$ are going to be, the more precise the sampled policies $\mathbf{L}$ will get.

Example. In many motor control studies the hand is modeled as a point mass, where the state vector $\mathbf{x}_t$ comprises position and velocity in the plane (Todorov and Jordan, 2002). In a discrete state space this yields the following equation:

$$\mathbf{x}_{t+1} = \begin{pmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{x}_t + \begin{pmatrix} 0 & 0 \\ \Delta t/m & 0 \\ 0 & 0 \\ 0 & \Delta t/m \end{pmatrix} \mathbf{u}_t + \xi_t$$

where we chose ξ_t to be distributed according to

$$\xi_t \propto \mathcal{N}\left[\mathbf{0}, \sqrt{\Delta t}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}\right]$$

The noise ξ_t models uncertainty in the force production when controlling the mass point. In our simulation of a reaching task with unknown system dynamics the controller had to learn to bring the mass point from the periphery to the center of the coordinate system trying to find the optimal feedback gains. This requires estimating a 24-dimensional parameter vector $\mathbf{w}$ and sampling a 2×8-dimensional feedback gain. We chose the following parameter settings: $\Delta t = 0.01$, $m = 1$, $R = [[0.001, 0]; [0, 0.001]]$, $Q = [[1,0,0,0]; [0,0.01,0,0]; [0,0,1,0]; [0,0,0,0.01]]$ and $\alpha = 0.05$ for the UKF. The results can be seen in figure 1. The first entry of the parameter vector $\hat{\mathbf{w}}$ is depicted in figure 1a, the first entry of the correspondingly sampled $\mathbf{L}$ is depicted in figure 1b. After an initial exploration phase in which $\mathbf{L}$ is sampled from a broad distribution the controller settles down and only samples from a very narrow distribution centered at the optimal value. Figure 1c,d shows initial and final trajectories and speed profiles: initially amorphous, a straight-line movement is learned with a bell-shaped speed profile. Importantly, the Bayesian controller converges much faster to the correct feedback gain than the certainty-equivalent controller which never

fully reached the optimal value in our simulation—compare figure 1e,f. In the following table the mean absolute feedback gain error—the difference between optimal feedback gain and actually executed feedback gain—is shown averaged over the last 3000 time steps of 100 runs. We have also averaged over all 2×8 feedback gains.

	Abs. Error
Certainty Equivalent Controller	8.26 ± 0.01
Bayesian Control Rule	2.085 ± 0.002

The results show that the Bayesian control rule incurs approximately 4 times less error on average than the certainty-equivalent controller in this example. To ensure that this result does not depend on the particular system we chose we ran the same simulation but all the entries of the true F and G were drawn randomly from a uniform distribution $[0;1]$ in each run, with 100 runs in total. However, these random draws were "frozen" such that both controllers faced the same random variables and differences cannot be attributed to different random draws. Each run of this simulation had 500 time steps and we compared the feedback gain error in the last 100 time steps.

	Abs. Error
Certainty Equivalent Controller	0.536 ± 0.002
Bayesian Control Rule	0.111 ± 0.001

On average the Bayesian control rule incurred approximately 5 times less error than the certainty-equivalent controller.

3.2 Direct Adaptive Bayesian Control

The adaptive linear quadratic control problem can be reformulated in a way that does not require estimating the system matrices $\mathbf{F}$ and $\mathbf{G}$ explicitly (Bradtke, 1993). Instead we can work directly on the policy space and assign a Q *value* to each policy such that the Q value of policy $\mathbf{L}$ is given by

$$Q_{\mathbf{L}}(\mathbf{x}_t, \mathbf{u}_t) = c_t(\mathbf{x}_t, \mathbf{u}_t) + \big(\mathbf{F}\mathbf{x}_t + \mathbf{G}\mathbf{u}_t\big)^T \mathbf{V}_{\mathbf{L}} \big(\mathbf{F}\mathbf{x}_t + \mathbf{G}\mathbf{u}_t\big)$$

where $\mathbf{V}_{\mathbf{L}}$ corresponds to the cost-to-go function. Thus, $Q_{\mathbf{L}}(\mathbf{x}_t, \mathbf{u}_t)$ can be expressed as a quadratic form

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{pmatrix}^T \underbrace{\begin{pmatrix} \mathbf{Q} + \mathbf{F}^T \mathbf{V}_{\mathbf{L}} \mathbf{F} & \mathbf{F}^T \mathbf{V}_{\mathbf{L}} \mathbf{G} \\ \mathbf{G}^T \mathbf{V}_{\mathbf{L}} \mathbf{F} & \mathbf{R} + \mathbf{G}^T \mathbf{V}_{\mathbf{L}} \mathbf{G} \end{pmatrix}}_{= \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}} \begin{pmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{pmatrix}$$

The matrix $\mathbf{M} \in \mathbb{R}^{(N+M)(N+M)}$ is positive definite and represents the Q value of policy $\mathbf{L}$. The relationship between $\mathbf{M}$ and $\mathbf{L}$ is given by

$$\mathbf{L} = -\mathbf{M}_{22}^{-1} \mathbf{M}_{21} \tag{4}$$

as can be readily seen when computing $\partial_{\mathbf{u}_t} Q_{\mathbf{L}}(\mathbf{x}_t, \mathbf{u}_t) = 0$. Previous studies have applied Q-learning to solve this *direct* adaptive control problem by reinforcement learning methods (Bradtke, 1993). Here we want to transform it into an inference problem. To this end, we need to relate $\mathbf{M}$ to an observable quantity in a way that is independent of the policy that is currently executed by the controller. We can achieve this by noting that Bellman's optimality equation imposes a recurrent relationship between consecutive Q values, namely

$$Q_{\mathbf{L}}(\mathbf{x}_t, \mathbf{u}_t) = c_t(\mathbf{x}_t, \mathbf{u}_t) + Q_{\mathbf{L}}(\mathbf{x}_{t+1}, -\mathbf{L}\mathbf{x}_{t+1}) \tag{5}$$

Since c_t is an observable quantity we can take it on one side of the equation and put all Q-quantities of equation (5) on the other side. Only the "true" Q-function can predict all c_t for all data points $\{\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1}\}$. Thus, we can use this relationship to do inference over $\mathbf{M}$ where

$$\begin{aligned} \hat{c}_t \;=\; & \begin{pmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{pmatrix}^T \mathbf{M} \begin{pmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{pmatrix} \\ & - \begin{pmatrix} \mathbf{x}_{t+1} \\ -\mathbf{M}_{22}^{-1}\mathbf{M}_{21}\mathbf{x}_{t+1} \end{pmatrix}^T \mathbf{M} \begin{pmatrix} \mathbf{x}_{t+1} \\ -\mathbf{M}_{22}^{-1}\mathbf{M}_{21}\mathbf{x}_{t+1} \end{pmatrix} \end{aligned}$$

Assuming Gaussian noise with known variance σ^2 for the cost observations we obtain the following likelihood model for our Bayesian controller:

$$P(c_t | \mathbf{M}, \mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{u}_t) \propto \exp\left[-\frac{1}{2\sigma^2} \Big(\hat{c}_t - c_t \Big)^2 \right]$$

The intervention model is again deterministic:

$$P(\mathbf{u}_{t+1} | \mathbf{M}, \mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{u}_t) \propto \delta\Big(\mathbf{u}_{t+1} + \mathbf{M}_{22}^{-1}\mathbf{M}_{21}\mathbf{x}_{t+1} \Big)$$

Doing inference over $\mathbf{M}$ is complicated by three facts: (i) the likelihood model is highly nonlinear in the parameters, (ii) $\mathbf{M}$ must be constrained to the set of positive definite matrices and (iii) $\mathbf{M}$ will be ill-conditioned in many examples because the different parts of the matrix differ usually by various orders of magnitude, as for example the unknown cost matrices $\mathbf{Q}$ and $\mathbf{R}$ are often of different orders of magnitude. Here we can only address problem (i) and (ii), i.e. the examples to demonstrate the Bayesian controller have to be well-conditioned—which is, for instance not true for the previous simulation. With regard to (i) we found that for this inference process the UKF only works robustly when the propagated means are simply computed as an un-weighted average over sigma points instead of the more common weighted average. With regard to (ii) we note that any positive definite matrix can be expressed as a product of its unique Cholesky factors: $\mathbf{M} = \mathbf{m}^T\mathbf{m}$ where $\mathbf{m}$ is upper triangular with diagonal elements strictly positive. Then we can do inference over $\mathbf{m}$ with the simpler

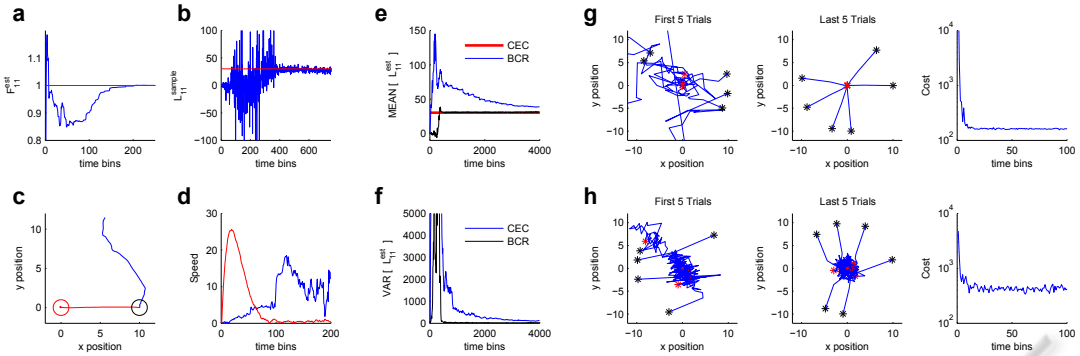

Figure 1: Results. (a-d) Learning to move a mass point when the system dynamics matrices **F** and **G** are unknown. A single run of the control process is shown. (a) Temporal evolution of estimate of the first entry of $\hat{\mathbf{F}}$ and the respective uncertainty as represented by the Kalman filter. (b) Sampled feedback gain—only the first entry of **L** is shown. The initial exploration phase is followed by a stable performance after 400 time steps. The thin red line indicates the optimal feedback gain. (c,d) Trajectories and speed profiles. Initially, the trajectory takes a random direction with an amorphous speed profile (blue curves). Later movement trajectories are straight and speed profiles bell-shaped (black curves). Panels (e-f) show sampled feedback gains over 100 runs. (e) Mean executed feedback gain. The certainty-equivalent controller (CEC) slowly converges to the region of optimal feedback gains. The exact optimal value was not reached in this simulation. The Bayesian control rule (BCR) converges very fast to the optimal feedback gain. (f) Variance of executed feedback gain. The Bayesian controller that used sampled feedback gains converges much faster than the certainty-equivalent controller. (g,h) Learning to move a mass-less point when both the system dynamics matrices **F** and **G** and the cost matrices **Q** and **R** are unknown. (g) Bayesian Control Rule. Trajectories of the first and last 5 trials. Initially, movements are undirected but later converge to straight line movements. The pertinent cost converges to the optimum. (h) Policy Iteration. Trajectories of the first and last 5 trials. The trajectories are wiggly because noise has to be added to the controller for exploration. Due to this extra noise the controller cannot converge to the optimal cost.

constraint that the diagonal elements must be positive. In our simulation we implemented this constraint by simply discarding any Kalman filter updates that would violate it. In general, such constraints can be easily implemented using particle filters.

Example. A simple well-conditioned example is a mass-less particle that moves around in the plane. The system dynamics can be formalized as:

$$\mathbf{x}_{t+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x}_t + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{u}_t$$

The observations are noisy observations of the cost

$$c_t = \mathbf{x}_t^T \mathbf{Q} \mathbf{x}_t + \mathbf{u}_t^T \mathbf{R} \mathbf{u}_t + \xi_t$$

where ξ_t is a normally distributed scalar variable with variance $\sigma_{obs} = 0.1$. Both Q and R were assumed to be identity matrices and $\alpha = 0.5$ as previously. This is a 10-dimensional estimation problem. Figure 1g shows that the Bayesian controller managed to find the optimal control solution only relying on inference and sampling. We compared against a policy iteration algorithm for linear quadratic controllers as proposed in (Bradtke, 1993) – compare Figure 1h. In the latter exploration can only be achieved by adding extra noise to the control command. Note that the Bayesian control rule incurs this noise automatically by sampling from the posterior. We simulated 100 trials with 50 time steps each.

To ensure again that this result does not depend on the particular system we chose we ran another simulation where each entry of **F** and **G** were drawn from the uniform distribution $[0;1]$ and **Q** and **R** were drawn from an inverse Wishart distribution with identity covariance matrix and degree of freedom 2. The noise was again "frozen" for comparison between the two algorithms. We compared the absolute error between the optimal and the actually executed feedback gain over the last 20 trials. The Bayesian control rule outperformed the policy iteration algorithm roughly by factor 5.

	Abs. Error
Policy Iteration (Bradtke, 1993)	2.5 ± 0.1
Bayesian Control Rule	0.55 ± 0.01

4 CONCLUSIONS

In this paper we suggest a minimum relative entropy formulation of adaptive control problems when the plant dynamics are unknown but known to belong to a pre-defined set of possible dynamics. This formulation has an explicit solution given by the Bayesian control rule, a stochastic rule for adaptive control. We have presented two example classes that show how adaptive linear quadratic control problems can

be tackled using this problem formulation. Usually, adaptive controllers rely on the certainty equivalence principle and ignore parameter uncertainty in the control process (Åström and Wittenmark, 1995). In contrast, a controller based on the Bayesian control rule considers this uncertainty for balancing exploration and exploitation in a way that minimizes the expected relative entropy with regard to the true control law.

In particular, indirect control methods provide an interesting perspective here, because they allow solving the adaptive control problem purely based on inference and sampling methods that can be recruited from a rich arsenal in machine learning. Both inference and action sampling work forward in time and are therefore applicable online. Also they do not require different phases of policy evaluation and policy improvement as some of the previous reinforcement learning methods. Inference can be done online independent of the sampled policy. Several other studies have previously proposed to solve adaptive control problems based on inference methods (Toussaint et al., 2006; Engel et al., 2005; Haruno et al., 2001). Crucially, however, these studies have concentrated on the observation part of the learning problem with no principled solution for the action selection problem. Usually, exploration noise has to be introduced in an *ad hoc* fashion in order to avoid suboptimal performance. In contrast, the minimum relative entropy cost function naturally leads to stochastic policies.

The main contribution of this study is to illustrate how a relative entropy formulation can be applied to solve an adaptive control problem. This is done by deriving a stochastic controller based on the Bayesian control rule for the LQR problem with unknown system and cost matrices. Similar minimum relative entropy formulations have recently also been proposed to solve optimal control problems with known system dynamics (Todorov, 2009; Kappen et al., 2009). How these two approaches for adaptive and optimal control relate is an interesting question for future research. Also, the Bayesian control rule suggested here could in principle be employed to solve more general adaptive control problems with possibly nonlinear dynamics. However, finding optimal tailored controllers for complex sub-environments can in general be highly non-trivial. Therefore, finding inference and sampling methods that work for more general classes of adaptive control problems poses a future challenge.

REFERENCES

Åström, K. and Wittenmark, B. (1995). *Adaptive Control*. Prentice Hall, 2nd edition.

Bradtke, S. (1993). Reinforcement learning applied to linear quadratic control. *Advances in Neural Information Processing Systems 5*.

Campi, M. and Kumar, P. (1996). Optimal adaptive control of an lqg system. *Proc. 35th Conf. on Decision and Control*, pages 349–353.

Engel, Y., Mannor, S., and Meir, R. (2005). Reinforcement learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 201–208.

Haruno, M., Wolpert, D., and Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, 13:2201–2220.

Haykin, S. (2001). *Kalman filtering and neural networks*. John Wiley and Sons.

Julier, S.J., U. J. and Durrant-Whyte, H. (1995). A new approach for filtering nonlinear systems. *Proc. Am. Control Conference*, pages 1628–1632.

Kappen, B., Gomez, V., and Opper, M. (2009). Optimal control as a graphical model inference problem. *arXiv:0901.0633*.

Ortega, P. and Braun, D. (2010). A bayesian rule for adaptive control based on causal interventions. In *Proceedings of the third conference on artificial general intelligence*, pages 121–126. Atlantis Press.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.

Stengel, R. (1993). *Optimal control and estimation*. Dover Publications.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences U.S.A.*, 106:11478–11483.

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

Toussaint, M., Harmeling, S., and Storkey, A. (2006). Probabilistic inference for solving (po)mdps. Technical report, EDI-INF-RR-0934, University of Edinburgh, School of Informatics.

Wittenmark, B. (1975). Stochastic adaptive control methods: a survey. *International Journal of Control*, 21:705–730.