Optimal Reinforcement Learning for Gaussian Systems

Philipp Hennig
Max Planck Institute for Intelligent Systems
Department of Empirical Inference
Spemannstraße 38, 72070 Tübingen, Germany
`phennig@tuebingen.mpg.de`

Abstract

The exploration-exploitation trade-off is among the central challenges of reinforcement learning. The optimal Bayesian solution is intractable in general. This paper studies to what extent analytic statements about optimal learning are possible if all beliefs are Gaussian processes. A first order approximation of learning of both loss and dynamics, for nonlinear, time-varying systems in continuous time and space, subject to a relatively weak restriction on the dynamics, is described by an infinite-dimensional partial differential equation. An approximate finite-dimensional projection gives an impression for how this result may be helpful.

1 Introduction – Optimal Reinforcement Learning

Reinforcement learning is about doing two things at once: *Optimizing* a function while *learning* about it. These two objectives must be balanced: Ignorance precludes efficient optimization; time spent hunting after irrelevant knowledge incurs unnecessary loss. This dilemma is famously known as the *exploration exploitation trade-off*. Classic reinforcement learning often considers time cheap; the trade-off then plays a subordinate role to the desire for learning a "correct" model or policy. Many classic reinforcement learning algorithms thus rely on ad-hoc methods to control exploration, such as "ϵ-greedy" [1], or "Thompson sampling" [2]. However, at least since a thesis by Duff [3] it has been known that Bayesian inference allows optimal balance between exploration and exploitation. It requires integration over every possible future trajectory under the current belief about the system's dynamics, all possible new data acquired along those trajectories, and their effect on decisions taken along the way. This amounts to optimization and integration over a tree, of exponential cost in the size of the state space [4]. The situation is particularly dire for continuous space-times, where both depth and branching factor of the "tree" are uncountably infinite. Several authors have proposed approximating this lookahead through samples [5, 6, 7, 8], or ad-hoc estimators that can be shown to be in some sense close to the Bayes-optimal policy [9].

In a parallel development, recent work by Todorov [10], Kappen [11] and others introduced an idea to reinforcement learning long commonplace in other areas of machine learning: Structural assumptions, while restrictive, can greatly simplify inference problems. In particular, a recent paper by Simpkins et al. [12] showed that it is actually possible to solve the exploration exploitation trade-off *locally*, by constructing a *linear* approximation using a Kalman filter. Simpkins and colleagues further assumed to know the loss function, and the dynamics up to Brownian drift. Here, I use their work as inspiration for a study of general optimal reinforcement learning of dynamics *and* loss functions of an unknown, *nonlinear, time-varying* system (note that most reinforcement learning algorithms are restricted to time-invariant systems). The core assumption is that all uncertain variables are known up to Gaussian process uncertainty. The main result is a first-order description of optimal reinforcement learning in form of infinite-dimensional differential statements. This kind of description opens up new approaches to reinforcement learning. As an only initial example of such treatments, Section 4

1

presents an approximate Ansatz that affords an explicit reinforcement learning algorithm; tested in some simple but instructive experiments (Section 5).

An intuitive description of the paper's results is this: From prior and corresponding choice of learning machinery (Section 2), we construct statements about the *dynamics of the learning process* (Section 3). The learning machine itself provides a probabilistic description of the dynamics of the physical system. Combining both dynamics yields a *joint* system, which we aim to control optimally. Doing so amounts to simultaneously controlling exploration (controlling the learning system) and exploitation (controlling the physical system).

Because large parts of the analysis rely on concepts from optimal control theory, this paper will use notation from that field. Readers more familiar with the reinforcement learning literature may wish to mentally replace coordinates x with states s, controls u with actions a, dynamics with transitions $p(s' \mid s, a)$ and utilities q with *losses* (negative rewards) $-r$. The latter is potentially confusing, so note that optimal control in this paper will attempt to *minimize* values, rather than to maximize them, as usual in reinforcement learning (these two descriptions are, of course, equivalent).

2   A Class of Learning Problems

We consider the task of optimally controlling an uncertain system whose states $s \equiv (x, t) \in \mathcal{K} \equiv \mathbb{R}^D \times \mathbb{R}$ lie in a $D+1$ dimensional Euclidean phase space-time: A cost Q (cumulated loss) is acquired at (x, t) with *rate* $\mathrm{d}Q/\mathrm{d}t = q(x, t)$, and the first inference problem is to learn this analytic function q. A second, independent learning problem concerns the dynamics of the system. We assume the dynamics separate into *free* and *controlled* terms affine to the control:

$$\mathrm{d}x(t) = [f(x, t) + g(x, t)u(x, t)]\, \mathrm{d}t \tag{1}$$

where $u(x, t)$ is the control function we seek to optimize, and f, g are analytic functions. To simplify our analysis, we will assume that *either f or g* are known, while the other may be uncertain (or, alternatively, that it is possible to obtain independent samples from both functions). See Section 3 for a note on how this assumption may be relaxed. W.l.o.g., let f be uncertain and g known. Information about both $q(x, t)$ and $f(x, t) = [f_1, \ldots, f_D]$ is acquired stochastically: A Poisson process of constant rate λ produces mutually independent samples

$$y_q(x, t) = q(x, t) + \epsilon_q \text{ and } y_{fd}(x, t) = f_d(x, t) + \epsilon_{fd} \text{ where } \epsilon_q \sim \mathcal{N}(0, \sigma_q^2); \epsilon_{fd} \sim \mathcal{N}(0, \sigma_{fd}^2). \tag{2}$$

The noise levels σ_q and σ_f are presumed known. Let our initial beliefs about q and f be given by Gaussian processes $\mathcal{GP}_{k_q}(q; \mu_q, \Sigma_q)$; and independent Gaussian processes $\prod_d^D \mathcal{GP}_{k_{fd}}(f_d; \mu_{fd}, \Sigma_{fd})$, respectively, with kernels $k_r, k_{f1}, \ldots, k_{fD}$ over $\mathcal{K}$, and mean / covariance functions μ / Σ. In other words, samples over the belief can be drawn using an infinite vector Ω of i.i.d. Gaussian variables, as

$$\tilde{f}_d([x, t]) = \mu_{fd}([x, t]) + \int \Sigma_{fd}^{1/2}([x, t], [x', t'])\Omega(x', t')\mathrm{d}x'\mathrm{d}t = \mu_{fd}([x, t]) + (\Sigma_{fd}^{1/2}\Omega)([x, t]) \tag{3}$$

the second equation demonstrates a compact notation for inner products that will be used throughout. It is important to note that f, q are unknown, but deterministic. At any point during learning, we can use the same samples Ω to describe uncertainty, while μ, Σ change during the learning process.

To ensure continuous trajectories, we also need to regularize the control. Following control custom, we introduce a quadratic control cost $\rho(u) = \frac{1}{2}u^{\mathsf{T}}R^{-1}u$ with control cost scaling matrix R. Its units $[R] = [x/t]/[Q/x]$ relate the cost of changing location to the utility gained by doing so.

The overall task is to find the optimal discounted horizon *value*

$$v(x, t) = \min_u \int_t^\infty e^{-(\tau - t)/\gamma} \left[q[\chi[\tau, u(\chi, \tau)], \tau] + \frac{1}{2}u(\chi, \tau)^{\mathsf{T}}R^{-1}u(\chi, \tau) \right]\, \mathrm{d}\tau \tag{4}$$

where $\chi(\tau, u)$ is the trajectory generated by the dynamics defined in Equation (1), using the control law (policy) $u(x, t)$. The exponential definition of the discount $\gamma > 0$ gives the unit of time to γ.

Before beginning the analysis, consider the relative generality of this definition: We allow for a continuous phase space. Both loss *and* dynamics may be uncertain, of rather general nonlinear form, and may change over time. The specific choice of a Poisson process for the generation of samples is

2

somewhat ad-hoc, but *some* measure is required to quantify the flow of information through time. The Poisson process is in some sense the simplest such measure, assigning uniform probability density. An alternative is to assume that datapoints are acquired at regular intervals of width λ. This results in a quite similar model but, since the system's dynamics still proceed in continuous time, can complicate notation. A downside is that we had to restrict the form of the dynamics. However, Eq. (1) still covers numerous physical systems studied in control, for example many mechanical systems, from classics like cart-and-pole to realistic models for helicopters [13].

3 Optimal Control for the Learning Process

The optimal solution to the exploration exploitation trade-off is formed by the *dual control* [14] of a joint representation of the physical system and the beliefs over it. In reinforcement learning, this idea is known as a belief-augmented POMDP [3, 4], but is not usually construed as a control problem. This section constructs the Hamilton-Jacobi-Bellman (HJB) equation of the joint control problem for the system described in Sec. 2, and analytically solves the equation for the optimal control. This necessitates a description of the learning algorithm's dynamics:

At time $t = \tau$, let the system be at phase space-time $s_\tau = (x(\tau), \tau)$ and have the Gaussian process belief $\mathcal{GP}(q; \mu_\tau(s), \Sigma_\tau(s, s'))$ over the function q (all derivations in this section will focus on q, and we will drop the sub-script q from many quantities for readability. The forms for f, or g, are entirely analogous, with independent Gaussian processes for each dimension $d = 1, \ldots, D$). This belief stems from a finite number N of samples $\boldsymbol{y}_0 = [y_1, \ldots, y_N]^\mathsf{T} \in \mathbb{R}^N$ collected at space-times $\boldsymbol{S}_0 = [(x_1, t_1), \ldots, (x_N, t_N)]^\mathsf{T} \equiv [s_1, \ldots, s_N]^\mathsf{T} \in \mathcal{K}^N$ (note that t_1 to t_N need not be equally spaced, ordered, or $< \tau$). For arbitrary points $s^* = (x^*, t^*) \in \mathcal{K}$, the belief over $q(s^*)$ is a Gaussian with mean function μ_τ, and co-variance function Σ_τ [15]

$$
\begin{aligned}
\mu_\tau(s_i^*) &= \boldsymbol{k}(s_i^*, \boldsymbol{S}_0)[K(\boldsymbol{S}_0, \boldsymbol{S}_0) + \sigma_q^2 \boldsymbol{I}]^{-1} \boldsymbol{y}_0 \\
\Sigma_\tau(s_i^*, s_j^*) &= k(s_i^*, s_j^*) - \boldsymbol{k}(s_i^*, \boldsymbol{S}_0)[K(\boldsymbol{S}_0, \boldsymbol{S}_0) + \sigma_q^2 \boldsymbol{I}]^{-1} \boldsymbol{k}(\boldsymbol{S}_0, s_j^*)
\end{aligned}
\tag{5}
$$

where $K(\boldsymbol{S}_0, \boldsymbol{S}_0)$ is the Gram matrix with elements $K_{ab} = k(s_a, s_b)$. We will abbreviate $K_0 \equiv [K(\boldsymbol{S}_0, \boldsymbol{S}_0) + \sigma_y^2 \boldsymbol{I}]$ from here on. The co-vector $\boldsymbol{k}(s^*, \boldsymbol{S}_0)$ has elements $\boldsymbol{k}_i = k(s^*, s_i)$ and will be shortened to $\boldsymbol{k}_0$. How does this belief *change* as time moves from τ to $\tau + \mathrm{d}t$? If $\mathrm{d}t \to 0$, the chance of acquiring a datapoint y_τ in this time is $\lambda \, \mathrm{d}t$. Marginalising over this Poisson stochasticity, we expect one sample with probability $\lambda \, \mathrm{d}t$, two samples with $(\lambda \, \mathrm{d}t)^2$ and so on. So the mean after $\mathrm{d}t$ is expected to be

$$
\mu_{\tau + \mathrm{d}t} = \lambda \, \mathrm{d}t \, (\boldsymbol{k}_0, k_\tau) \begin{pmatrix} K_0 & \boldsymbol{\xi}_\tau \\ \boldsymbol{\xi}_\tau^\mathsf{T} & \kappa_\tau \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_0 \\ y_\tau \end{pmatrix} + (1 - \lambda \, \mathrm{d}t - \mathcal{O}(\lambda \, \mathrm{d}t)^2) \cdot \boldsymbol{k}_0 K_0^{-1} \boldsymbol{y}_0 + \mathcal{O}(\lambda \, \mathrm{d}t)^2 \tag{6}
$$

where we have defined the map $k_\tau = k(s^*, s_\tau)$, the vector $\boldsymbol{\xi}_\tau$ with elements $\xi_{\tau,i} = k(s_i, s_\tau)$, and the scalar $\kappa_\tau = k(s_\tau, s_\tau) + \sigma_q^2$. Algebraic re-formulation yields

$$
\mu_{\tau + \mathrm{d}t} = \boldsymbol{k}_0 K_0^{-1} \boldsymbol{y}_0 + \lambda(k_t - \boldsymbol{k}_0^\mathsf{T} K_0^{-1} \boldsymbol{\xi}_t)(\kappa_t - \boldsymbol{\xi}_t^\mathsf{T} K_0^{-1} \boldsymbol{\xi}_t)^{-1}(y_t - \boldsymbol{\xi}_t^\mathsf{T} K_0^{-1} \boldsymbol{y}_0) \, \mathrm{d}t. \tag{7}
$$

Note that $\boldsymbol{\xi}_\tau^\mathsf{T} K_0^{-1} \boldsymbol{y}_0 = \mu(s_\tau)$, the mean prediction at s_τ and $(\kappa_\tau - \boldsymbol{\xi}_\tau^\mathsf{T} K_0^{-1} \boldsymbol{\xi}_\tau) = \sigma_q^2 + \Sigma(s_\tau, s_\tau)$, the marginal variance there. Hence, we can define scalars $\bar{\Sigma}, \bar{\sigma}$ and write

$$
\frac{(y_\tau - \boldsymbol{\xi}_\tau^\mathsf{T} K_0^{-1} \boldsymbol{y}_0)}{(\kappa_\tau - \boldsymbol{\xi}_\tau^\mathsf{T} K_0^{-1} \boldsymbol{\xi}_\tau)^{1/2}} = \frac{[\Sigma^{1/2}\Omega](s_\tau) + \sigma\omega}{[\Sigma(s_\tau, s_\tau) + \sigma^2]^{1/2}} \equiv \bar{\Sigma}_\tau^{1/2}\Omega + \bar{\sigma}_\tau\omega \quad \text{with} \quad \omega \sim \mathcal{N}(0, 1). \tag{8}
$$

So the change to the mean consists of a deterministic but uncertain change whose effects accumulate linearly in time, and a *stochastic* change, caused by the independent noise process, whose *variance* accumulates linearly in time (in truth, these two points are considerably subtler, a detailed proof is left out for lack of space). We use the Wiener [16] measure $\mathrm{d}\omega$ to write

$$
\mathrm{d}\mu_{s_\tau}(s^*) = \lambda \frac{k_\tau - \boldsymbol{k}_0^\mathsf{T} K_0^{-1} \boldsymbol{\xi}_\tau}{(\kappa_\tau - \boldsymbol{\xi}_\tau^\mathsf{T} K_0^{-1} \boldsymbol{\xi}_\tau)^{-1/2}} \frac{[\Sigma^{1/2}\Omega](s_\tau) + \sigma\omega}{[\Sigma(s_\tau, s_\tau) + \sigma^2]^{1/2}} \, \mathrm{d}t \equiv \lambda L_{s_\tau}(s^*)[\bar{\Sigma}_\tau^{1/2}\Omega \, \mathrm{d}t + \bar{\sigma}_\tau \, \mathrm{d}\omega]
$$
$$\tag{9}$$

where we have implicitly defined the *innovation function* L. Note that L is a function of both s^* and s_τ. A similar argument finds the change of the covariance function to be the *deterministic* rate

$$
\mathrm{d}\Sigma_{s_\tau}(s_i^*, s_j^*) = -\lambda L_{s_\tau}(s_i^*)L_{s_\tau}^\mathsf{T}(s_j^*) \, \mathrm{d}t. \tag{10}
$$

3

So the dynamics of learning consist of a deterministic change to the covariance, and both deterministic and stochastic changes to the mean, both of which are samples a Gaussian processes with covariance function proportional to LL^T. This separation is a fundamental characteristic of GPs (it is the nonparametric version of a more straightforward notion for finite-dimensional Gaussian beliefs, for data with known noise magnitude).

We introduce the belief-augmented space $\mathcal{H}$ containing states $z(\tau) \equiv [x(\tau), \tau, \mu_q^\tau(s), \mu_{f1}^\tau, \ldots, \mu_{fD}^\tau, \Sigma_q^\tau(s, s'), \Sigma_{f1}^\tau, \ldots, \Sigma_{fD}^\tau]$. Since the means and covariances are functions, $\mathcal{H}$ is infinite-dimensional. Under our beliefs, $z(\tau)$ obeys a stochastic differential equation of the form

$$\mathrm{d}z = [\mathbf{A}(z) + \mathbf{B}(z)u + \mathbf{C}(z)\Omega]\,\mathrm{d}t + \mathbf{D}(z)\,\mathrm{d}\omega \tag{11}$$

with *free dynamics* $\mathbf{A}$, *controlled dynamics* $\mathbf{B}u$, *uncertainty* operator $\mathbf{C}$, and *noise* operator $\mathbf{D}$

$$\mathbf{A} = \left[\mu_f^\tau(z_x, z_t)\,,\; 1\,,\; 0\,,\; 0\,,\; \ldots\,,\; 0\,,\; -\lambda L_q L_q^\mathsf{T}\,,\; -\lambda L_{f1} L_{f1}^\mathsf{T}\,,\; \ldots\,,\; -\lambda L_{fD} L_{fD}^\mathsf{T}\right]; \tag{12}$$

$$\mathbf{B} = [g(s^*), 0, 0, 0, \ldots]; \quad \mathbf{C} = \mathrm{diag}(\Sigma_{f\tau}^{1/2}, 0, \lambda L_q \bar{\Sigma}_q^{1/2}, \lambda L_{f1} \bar{\Sigma}_{f1}^{1/2}, \ldots, \lambda L_{fD} \bar{\Sigma}_{fd}^{1/2}, 0, \ldots, 0);$$

$$\mathbf{D} = \mathrm{diag}(0, 0, \lambda L_q \bar{\sigma}_q, \lambda L_{f1} \bar{\sigma}_{f1}, \ldots, \lambda L_{fD} \bar{\sigma}_{fD}, 0, \ldots, 0) \tag{13}$$

The *value* – the expected cost to go – of any state s^* is given by the Hamilton-Jacobi-Bellman equation, which follows from Bellman's principle and a first-order expansion, using Eq. (4):

$$v(z_\tau) = \min_u \left\{ \iint \left[\left(\mu_q(s_\tau) + \Sigma_{q\tau}^{1/2}\Omega_q + \sigma_q\omega_q + \frac{1}{2}u^\mathsf{T} R^{-1}u\right)\,\mathrm{d}t + v(z_{\tau+\,\mathrm{d}t})\right]\,\mathrm{d}\omega\,\mathrm{d}\Omega \right\} \tag{14}$$

$$= \min_u \left\{ \int \mu_q^\tau + \Sigma_{q\tau}^{1/2}\Omega_q + \frac{1}{2}u^\mathsf{T} R^{-1}u + \frac{v(z_\tau)}{\mathrm{d}t} + \frac{\partial v}{\partial t} + [\mathbf{A} + \mathbf{B}u + \mathbf{C}\Omega]^\mathsf{T}\nabla v + \frac{1}{2}\mathrm{tr}[\mathbf{D}^\mathsf{T}(\nabla^2 v)\mathbf{D}]\mathrm{d}\Omega \right\}\mathrm{d}t$$

Integration over ω can be performed with ease, and removes the *stochasticity* from the problem; The *uncertainty* over Ω is a lot more challenging. Because the distribution over future losses is correlated through space and time, ∇v, $\nabla^2 v$ are functions of Ω, and the integral is nontrivial. But there are some obvious approximate approaches. For example, if we (inexactly) swap integration and minimisation, draw samples Ω^i and solve for the value for each sample, we get an "average optimal controller". This over-estimates the actual sum of future rewards by assuming the controller has access to the true system. It has the potential advantage of considering the actual optimal controller for every possible system, the disadvantage that the average of optima need not be optimal for any actual solution. On the other hand, if we ignore the correlation between Ω and ∇v, we can integrate (17) locally, all terms in Ω drop out and we are left with an "optimal average controller", which assumes that the system locally follows its average (mean) dynamics. This cheaper strategy was adopted in the following. Note that it is myopic, but not greedy in a simplistic sense – it does take the effect of learning into account. It amounts to a "global one-step look-ahead". One could imagine extensions that consider the influence of Ω on ∇v to a higher order, but these will be left for future work. Under this first-order approximation, analytic minimisation over u can be performed in closed form, and bears

$$u(z) = -R\mathbf{B}(z)^\mathsf{T}\nabla v(z) = -Rg(x, t)^\mathsf{T}\nabla_x v(z). \tag{15}$$

The *optimal* Hamilton-Jacobi-Bellman equation is then

$$\gamma^{-1}v(z) = \mu_q^\tau + \mathbf{A}^\mathsf{T}\nabla v - \frac{1}{2}[\nabla v]^\mathsf{T}\mathbf{B}R\mathbf{B}^\mathsf{T}\nabla v + \frac{1}{2}\mathrm{tr}\left[\mathbf{D}^\mathsf{T}(\nabla^2 v)\mathbf{D}\right]. \tag{16}$$

A more explicit form emerges upon re-inserting the definitions of Eq. (12) into Eq. (16):

$$\gamma^{-1}v(z) = [\mu_q^\tau + \underbrace{[\mu_f^\tau(z_x, z_t)\nabla_x + \nabla_t]v(z)}_{\text{free drift cost}} - \underbrace{\frac{1}{2}[\nabla_x v(z)]^\mathsf{T} g^\mathsf{T}(z_x, z_t) R g(z_x, z_t)\nabla_x v(z)}_{\text{control benefit}}$$

$$+ \sum_{c=q,f_1,\ldots,f_D} \underbrace{-\lambda[L_c L_c^\mathsf{T}\nabla_{\Sigma_c}]v(z)}_{\text{exploration bonus}} + \underbrace{\frac{1}{2}\lambda^2\bar{\sigma}_c^2[L_{fd}^\mathsf{T}(\nabla^2_{\mu_{fd}}v(z))L_{fd}]}_{\text{diffusion cost}} \tag{17}$$

Equation (17) is the central result: Given Gaussian priors on nonlinear control-affine dynamic systems, up to a first order approximation, optimal reinforcement learning is described by an infinite-dimensional second-order partial differential equation. It can be interpreted as follows (labels in the

4

equation, note the negative signs of "beneficial" terms): The value of a state comprises the immediate utility rate; the effect of the free drift through space-time and the benefit of optimal control; an *exploration bonus* of learning, and a *diffusion cost* engendered by the measurement noise. The first two lines of the right hand side describe effects from the phase space-time subspace of the augmented space, while the last line describes effects from the belief part of the augmented space. The former will be called *exploitation terms*, the latter *exploration terms*, for the following reason: If the first two lines line dominate the right hand side of Equation (17) in absolute size, then future losses are governed by the physical sub-space – caused by exploiting knowledge to control the physical system. On the other hand, if the last line dominates the value function, exploration is more important than exploitation – the algorithm controls the physical space to increase knowledge. To my knowledge, this is the first differential statement about reinforcement learning's two objectives. Finally, note the role of the sampling rate λ: If λ is very low, exploration is useless over the discount horizon.

Even after these approximations, solving Equation (17) for v remains nontrivial for two reasons: First, although the vector product notation is pleasingly compact, the mean and covariance functions are of course infinite-dimensional, and what looks like straightforward inner vector products are in fact integrals. For example, the average exploration bonus for the loss, writ large, reads

$$-\lambda L_q L_q^\intercal \nabla_{\Sigma_q} v(z) = -\iint_{\mathcal{K}} \lambda L_{s_\tau}^{(q)}(s_i^*) L_{s_\tau}^{(q)}(s_j^*) \frac{\partial v(z)}{\partial \Sigma(s_i^*, s_j^*)} \, \mathrm{d}s_i^* \, \mathrm{d}s_j^*. \tag{18}$$

(note that this object remains a function of the state s_τ). For general kernels k, these integrals may only be solved numerically. However, for at least one specific choice of kernel (square-exponentials) and parametric Ansatz, the required integrals can be solved in closed form. This analytic structure is so interesting, and the square-exponential kernel so widely used that the "numerical" part of the paper (Section 4) will restrict the choice of kernel to this class.

The other problem, of course, is that Equation (17) is a nontrivial differential Equation. Section 4 presents *one, initial* attempt at a numerical solution that should not be mistaken for a definitive answer. Despite all this, Eq. (17) arguably constitutes a useful gain for Bayesian reinforcement learning: It replaces the intractable definition of the value in terms of future trajectories with a differential equation. This raises hope for new approaches to reinforcement learning, based on numerical analysis rather than sampling.

Digression: Relaxing Some Assumptions

This paper only applies to the specific problem class of Section 2. Any generalisations and extensions are future work, and I do not claim to solve them. But it is instructive to consider some easier extensions, and some harder ones: For example, it is intractable to simultaneously learn both g and f *nonparametrically*, if only the actual transitions are observed, because the beliefs over the two functions become infinitely dependent when conditioned on data. But if the belief on either g or f is *parametric* (e.g. a general linear model), a joint belief on g and f is tractable [see 15, §2.7], in fact straightforward. Both the quadratic control cost $\propto u^\intercal R u$ and the control-affine form $(g(x,t)u)$ are relaxable assumptions – other parametric forms are possible, as long as they allow for analytic optimization of Eq. (14). On the question of learning the kernels for Gaussian process regression on q and f or g, it is clear that standard ways of inferring kernels [15, 18] can be used without complication, but that they are not covered by the notion of optimal learning as addressed here.

4   Numerically Solving the Hamilton-Jacobi-Bellman Equation

Solving Equation (16) is principally a problem of numerical analysis, and a battery of numerical methods may be considered. This section reports on one specific Ansatz, a Galerkin-type projection analogous to the one used in [12]. For this we break with the generality of previous sections and assume that the kernels k are given by square exponentials $k(a, b) = k_{\text{SE}}(a, b; \theta, S) = \theta^2 \exp(-\frac{1}{2}(a - b)^\intercal S^{-1}(a - b))$ with parameters θ, S. As discussed above, we approximate by setting $\Omega = 0$. We find an approximate solution through a factorizing parametric Ansatz: Let the value of any point $z \in \mathcal{H}$ in the belief space be given through a set of parameters $\boldsymbol{w}$ and some nonlinear *functionals* $\boldsymbol{\phi}$, such that their contributions separate over phase space, mean, and covariance functions:

$$v(z) = \sum_{e=x,\Sigma_q,\mu_q,\Sigma_f,\mu_f} \boldsymbol{\phi}_e(z_e)^\intercal \boldsymbol{w}_e \qquad \text{with } \boldsymbol{\phi}_e, \boldsymbol{w}_e \in \mathbb{R}^{N_e} \tag{19}$$

5

This projection is obviously restrictive, but it should be compared to the use of radial basis functions for function approximation, a similarly restrictive framework widely used in reinforcement learning. The functionals ϕ have to be chosen conducive to the form of Eq. (17). For square exponential kernels, one convenient choice is

$$\phi_s^a(z_s) = k(s_z, s_a; \theta_a, S_a) \tag{20}$$

$$\phi_\Sigma^b(z_\Sigma) = \iint_\mathcal{K} [\Sigma_z(s_i^*, s_j^*) - k(s_i^*, s_j^*)]k(s_i^*, s_b; \theta_b, S_b)k(s_j^*, s_b; \theta_b, S_b) \, \mathrm{d}s_i^* \, \mathrm{d}s_j^* \quad \text{and} \tag{21}$$

$$\phi_\mu^c(z_\mu) = \iint_\mathcal{K} \mu_z(s_i^*)\mu_z(s_j^*)k(s_i^*, s_c; \theta_c, S_c)k(s_j^*, s_c; \theta_c, S_c) \, \mathrm{d}s_i^* \, \mathrm{d}s_j^* \tag{22}$$

(the subtracted term in the first integral serves only numerical purposes). With this choice, the integrals of Equation (17) can be solved analytically (solutions left out due to space constraints). The approximate Ansatz turns Eq. (17) into an algebraic equation quadratic in $\boldsymbol{w}_x$, linear in all other $\boldsymbol{w}_e$:

$$\frac{1}{2}\boldsymbol{w}_x^\mathsf{T}\boldsymbol{\Psi}(z_x)\boldsymbol{w}_x - q(z_x) + \sum_{e=x,\mu_q,\Sigma_q,\mu_f,\Sigma_f} \boldsymbol{\Xi}^e(z_e)\boldsymbol{w}_e = 0 \tag{23}$$

using co-vectors $\boldsymbol{\Xi}$ and a matrix $\boldsymbol{\Psi}$ with elements

$$\begin{aligned}
\Xi_a^x(z_s) &= \gamma^{-1}\phi_s^a(z_s) - f(z_x)^\mathsf{T}\nabla_x\phi_s^a(z_s) - \nabla_t\phi_s^a(z_s) \\
\Xi_a^\Sigma(z_\Sigma) &= \gamma^{-1}\phi_\Sigma^a(z_\Sigma) + \lambda \iint_\mathcal{K} L_{s_\tau}(s_i^*)L_{s_\tau}(s_j^*)\frac{\partial\phi_\Sigma(z_\Sigma)}{\partial\Sigma_z(s_i^*, s_j^*)} \, \mathrm{d}s_i^* \, \mathrm{d}s_j^* \\
\Xi_a^\mu(z_\mu) &= \gamma^{-1}\phi_\mu^a(z_\mu) - \frac{\lambda^2\bar{\sigma}_{s_\tau}^2}{2} \iint_\mathcal{K} L_{s_\tau}(s_i^*)L_{s_\tau}(s_j^*)\frac{\partial^2\phi_\mu^a(z_\mu)}{\partial\mu_z(s_i^*)\partial\mu_z(s_j^*)} \, \mathrm{d}s_i^* \, \mathrm{d}s_j^* \\
\Psi(z)_{k\ell} &= [\nabla_x\phi_s^k(z)]^\mathsf{T}g(z_x)Rg(z_x)^\mathsf{T}[\nabla_x\phi_s^\ell(z)]
\end{aligned} \tag{24}$$

Note that Ξ_μ and Ξ_Σ are both functions of the physical state, through s_τ. It is through this functional dependency that the value of information is associated with the physical phase space-time. To solve for $\boldsymbol{w}$, we simply choose a number of evaluation points $\boldsymbol{z}_{\text{eval}}$ sufficient to constrain the resulting system of quadratic equations, and then find the least-squares solution $\boldsymbol{w}_{\text{opt}}$ by function minimisation, using standard methods, such as Levenberg-Marquardt [19]. A disadvantage of this approach is that is has a number of degrees of freedom $\boldsymbol{\Theta}$, such as the kernel parameters, and the number and locations $\boldsymbol{x}_a$ of the feature functionals. Our experiments (Section 5) suggest that it is nevertheless possible to get interesting results simply by choosing these parameters heuristically.

5 Experiments

5.1 Illustrative Experiment on an Artificial Environment

As a simple example system with a one-dimensional state space, f, q were sampled from the model described in Section 2, and g set to the unit function. The state space was tiled regularly, in a bounded region, with 231 square exponential ("radial") basis functions (Equation 20), initially all with weight $w_x^i = 0$. For the information terms, only a single basis function was used for each term (i.e. one single $\phi_{\Sigma q}$, one single $\phi_{\mu q}$, and equally for f, all with very large length scales S, covering the entire region of interest). As pointed out above, this does not imply a trivial structure for these terms, because of the functional dependency on L_{s_τ}. Five times the number of parameters, i.e. $N_{\text{eval}} = 1175$ evaluation points $\boldsymbol{z}_{\text{eval}}$ were sampled, at each time step, uniformly over the same region. It is not intuitively clear whether each z_e should have its own belief (i.e. whether the points must cover the belief space as well as the phase space), but anecdotal evidence from the experiments suggests that it suffices to use the current beliefs for all evaluation points. A more comprehensive evaluation of such aspects will be the subject of a future paper. The discount factor was set to $\gamma = 50$s, the sampling rate at $\lambda = 2$/s, the control cost at $10\text{m}^2/(\$\text{s})$. Value and optimal control were evaluated at time steps of $\delta t = 1/\lambda = 0.5$s.

Figure 1 shows the situation 50s after initialisation. The most noteworthy aspect is the nontrivial structure of exploration and exploitation terms. Despite the simplistic parameterisation of the corresponding functionals, their functional dependence on s_τ induces a complex shape. The system

6

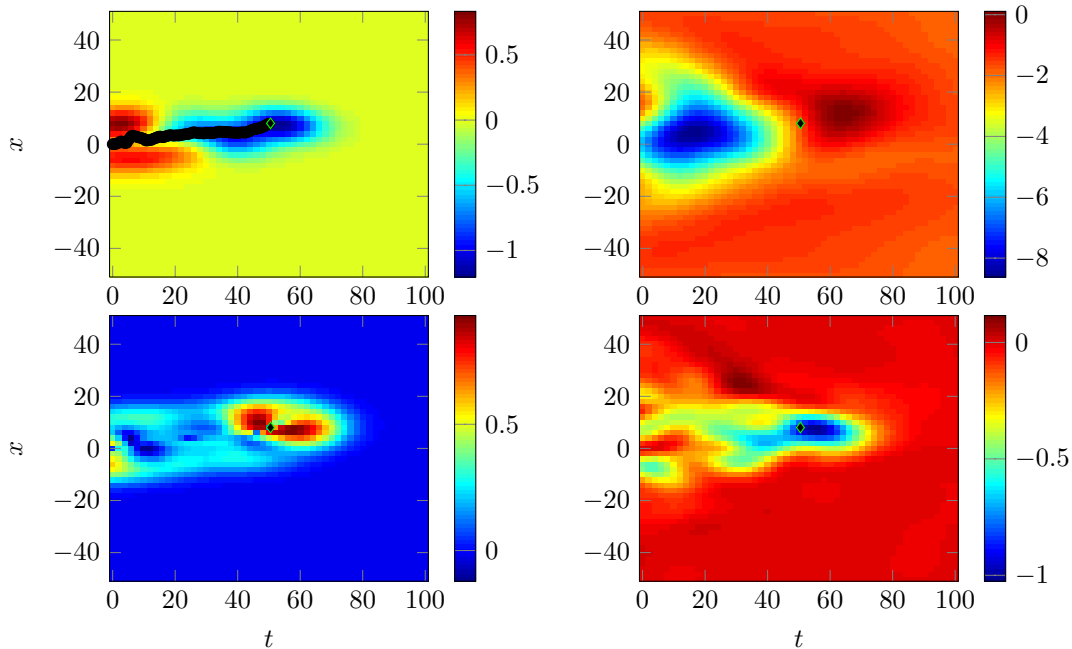



Figure 1: State after 50 time steps, plotted over phase space-time. **top left:** μ_q (blue is good). The belief over f is not shown, but has similar structure. **top right:** value estimate v at current belief: compare to next two panels to note that the approximation is relatively coarse. **bottom left:** exploration terms. **bottom right:** exploitation terms. At its current state (black diamond), the system is in the process of switching from exploitation to exploration (blue region in bottom right panel is roughly cancelled by red, forward cone in bottom left one).

constantly balances exploration and exploitation, and the optimal balance depends nontrivially on location, time, and the actual value (as opposed to only uncertainty) of accumulated knowledge. This is an important insight that casts doubt on the usefulness of simple, local exploration boni, used in many reinforcement learning algorithms.

Secondly, note that the system's trajectory does not necessarily follow what would be the optimal path under full information. The value estimate reflects this, by assigning low (good) value to regions *behind* the system's trajectory. This amounts to a sense of "remorse": If the learner would have known about these regions earlier, it would have strived to reach them. But this is not a sign of sub-optimality: Remember that the value is defined on the augmented space. The plots in Figure 1 are merely a slice through that space at some level set in the belief space.

5.2 Comparative Experiment – The Furuta Pendulum

The cart-and-pole system is an under-actuated problem widely studied in reinforcement learning. For variation, this experiment uses a cylindrical version, the pendulum on the rotating arm [20]. The task is to swing up the pendulum from the lower resting point. The table in Figure 2 compares the average loss of a controller with access to the true f, g, q, but otherwise using Algorithm 1, to that of an ϵ-greedy TD(λ) learner with linear function approximation, Simpkins' et al.'s [12] Kalman method and the Gaussian process learning controller (Fig. 2). The linear function approximation of TD(λ) used the same radial basis functions as the three other methods. None of these methods is free of assumptions: Note that the sampling frequency influences TD in nontrivial ways rarely studied (for example through the coarseness of the ϵ-greedy policy). The parameters were set to $\gamma = 5$s, $\lambda = 50/$s. Note that reinforcement learning experiments often quote total accumulated loss, which differs from the discounted task posed to the learner. Figure 2 reports actual discounted losses. The GP method clearly outperforms the other two learners, which barely explore. Interestingly, none of the tested methods, not even the informed controller, achieve a stable controlled balance, although

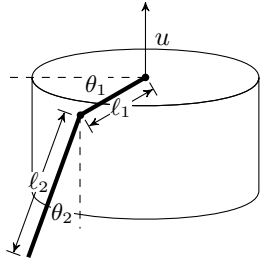
7

Method	cumulative loss
Full Information (baseline)	4.4 ±0.3
TD(λ)	6.401±0.001
Kalman filter Optimal Learner	6.408±0.001
Gaussian process optimal learner	4.6 ±1.4

Figure 2: The Furuta pendulum system: A pendulum of length ℓ_2 is attached to a rotatable arm of length ℓ_1. The control input is the torque applied to the arm. **Right:** cost to go achieved by different methods. Lower is better. Error measures are one standard deviation over five experiments.

the GP learner does swing up the pendulum. This is due to the random, non-optimal location of basis functions, which means resolution is not necessarily available where it is needed (in regions of high curvature of the value function), and demonstrates a need for better solution methods for Eq. (17).

There is of course a large number of other algorithms methods to potentially compare to, and these results are anything but exhaustive. They should not be misunderstood as a critique of any other method. But they highlight the need for units of measure on every quantity, and show how hard optimal exploration and exploitation truly is. Note that, for time-varying or discounted problems, there is no "conservative" option that cold be adopted in place of the Bayesian answer.

6   Conclusion

Gaussian process priors provide a nontrivial class of reinforcement learning problems for which optimal reinforcement learning reduces to solving differential equations. Of course, this fact alone does not make the problem easier, as solving nonlinear differential equations is in general intractable. However, the ubiquity of differential descriptions in other fields raises hope that this insight opens new approaches to reinforcement learning. For intuition on how such solutions might work, one specific approximation was presented, using functionals to reduce the problem to finite least-squares parameter estimation.

The critical reader will have noted how central the prior is for the arguments in Section 3: The dynamics of the learning process are predictions of future data, thus inherently determined exclusively by prior assumptions. One may find this unappealing, but there is no escape from it. Minimizing future loss requires predicting future loss, and predictions are always in danger of falling victim to incorrect assumptions. A finite initial identification phase may mitigate this problem by replacing prior with posterior uncertainty – but even then, predictions and decisions will depend on the model.

The results of this paper raise new questions, theoretical and applied. The most pressing questions concern better solution methods for Eq. (14), in particular better means for taking the expectation over the uncertain dynamics to more than first order. There are also obvious probabilistic issues: Are there other classes of priors that allow similar treatments? (Note some conceptual similarities between this work and the BEETLE algorithm [4]). To what extent can approximate inference methods – widely studied in combination with Gaussian process regression – be used to broaden the utility of these results?

Acknowledgments

The author wishes to express his gratitude to Carl Rasmussen, Jan Peters, Zoubin Ghahramani, Peter Dayan, and an anonymous reviewer, whose thoughtful comments uncovered several errors and crucially improved this paper.

8

References

[1] R.S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, 1998.

[2] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of two samples. *Biometrika*, 25:275–294, 1933.

[3] M.O.G. Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, U of Massachusetts, Amherst, 2002.

[4] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 697–704, 2006.

[5] Richard Dearden, Nir Friedman, and David Andre. Model based Bayesian exploration. In *Uncertainty in Artificial Intelligence*, pages 150–159, 1999.

[6] Malcolm Strens. A Bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, pages 943–950, 2000.

[7] T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *International Conference on Machine Learning*, pages 956–963, 2005.

[8] J. Asmuth, L. Li, M.L. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Uncertainty in Artificial Intelligence*, 2009.

[9] J.Z. Kolter and A.Y. Ng. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th International Conference on Machine Learning*. Morgan Kaufmann, 2009.

[10] E. Todorov. Linearly-solvable Markov decision problems. *Advances in Neural Information Processing Systems*, 19, 2007.

[11] H. J. Kappen. An introduction to stochastic control theory, path integrals and reinforcement learning. In *9th Granada seminar on Computational Physics: Computational and Mathematical Modeling of Cooperative Behavior in Neural Systems.*, pages 149–181, 2007.

[12] A. Simpkins, R. De Callafon, and E. Todorov. Optimal trade-off between exploration and exploitation. In *American Control Conference, 2008*, pages 33–38, 2008.

[13] I. Fantoni and L. Rogelio. *Non-linear Control for Underactuated Mechanical Systems*. Springer, 1973.

[14] A.A. Feldbaum. Dual control theory. *Automation and Remote Control*, 21(9):874–880, April 1961.

[15] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[16] N. Wiener. Differential space. *Journal of Mathematical Physics*, 2:131–174, 1923.

[17] T. Kailath. An innovations approach to least-squares estimation — part I: Linear filtering in additive white noise. *IEEE Transactions on Automatic Control*, 13(6):646–655, 1968.

[18] I. Murray and R.P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. *arXiv:1006.0868*, 2010.

[19] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[20] K. Furuta, M. Yamakita, and S. Kobayashi. Swing-up control of inverted pendulum using pseudo-state feedback. *Journal of Systems and Control Engineering*, 206(6):263–269, 1992.

9