# Optimal exploration as a symmetry breaking phenomenon

**Tonk S.R.H.M.**[1] and **Kappen H.J.** [2]

**Abstract.** We study the problem of joint estimation and control for a small example for which we can numerically compute the exact solution. We demonstrate, that optimal exploration is achieved through symmetry breaking in the Bellman equation; that optimal actions can be discontinuous in the beliefs; and that the optimal value function is typically non-differentiable. This could pose a challenge for the conventional design of value function approximations for solving POMDPs.

## 1 Introduction

The problem of control theory and reinforcement learning is to chose actions that optimize future rewards. When the state transition probabilities and the rewards are known, the optimal action of the agent[3] can be computed using a dynamic programming argument that results in the Bellman equation. However, when parts of this information are not available, a complex situation arises because now the agent can choose between actions that optimize expected reward and actions that are expected to gain information. This presents the agent the problem of finding the right compromise between probing and control, a problem which is known in control theory as dual control and was originally introduced by [3] (see [4] for a recent review).

Any dual control problem can be written as an ordinary control problem by augmenting the state space with parameters that quantify the belief of the agent in the world. This approach is known as POMDP and BAMDP in the AI community and is closely related to model-based reinforcement learning. Augmentation of the state space with a belief parameter provides a natural framework for reasoning about exploration/exploitation trade-offs which are found within many Bayesian reinforcement problems. An optimal balance for this exploration/exploitation trade off can be found by calculating the optimal policy within the POMDP framework. However, the calculation of these optimal policies turns out to be notoriously hard, because policies scale exponentially with the time horizon. Several authors have proposed heuristic methods such as [2, 7], solving the problem of intractability by introducing parameterized value function approximations over the extended state space.

In the control literature, the problem of parameter estimation is also well-known. For linear dynamical systems with known parameters, unobserved states and quadratic costs, the problem is absent. One refers to these systems as *certainty equivalent* [9], which means that the optimal control can be computed as if the state were observed and given by the expected (filtered) estimate of the past observations. However, for the general dual control problem, certainty equivalence does not hold. A particular action selection strategy is called probing, where random, i.e. non-goal directed, actions are proposed with the objective to learn the system parameters. Although it is argued convincingly that probing is an effective way to obtain efficient dual control strategies, the optimal probing strategy is typically not known.

In this paper, we study a very simple dual control problem in continuous space with one binary unknown parameter and a finite time horizon. Variants of this problem have been studied before [5, 11, 10] and various heuristics for this problem have been proposed, but the optimal solution has not been computed before. Because of the simplicity of the control problem, we can gain valuable insight in the optimal exploration/exploitation behavior, necessary to find the optimal solution.

The studied example will show several features that are important for the design of effective exploration strategies for POMDPs, Bayesian RL or dual control.

- We will demonstrate that the optimal exploration in this model is facilitated by a symmetry breaking mechanism. The symmetry breaking separates two phases: a phase where there is sufficient knowledge about the control task to exploit. In this phase the minimization over actions is convex and the optimal action is unique (in this example) and closely resembles the optimal action when there is complete knowledge. In the other phase there is insufficient knowledge to exploit. In this phase many actions are possible, each of them rather unrelated to reaching the goal. The minimization over actions is non-convex and exploratory actions arise as multiple local minima in the Bellman equation. The deepest of these minima is a particular compromise between steering and probing, depending on the state (distance to the goal), the horizon time, the noise and the cost function. The symmetry breaking is not controlled by an external parameter but arises dynamically when the Bellman equation is iterated.
- As a result of the non-uniqueness of the choice of action, the cost-to-go function $J(x, \theta, t)$ is non-smooth. In our example $J(x, \theta, t)$ has discontinuous partial derivatives to $\theta$, with this non-differentiability exactly occurring in the regions of no/low knowledge states. This has important consequences for a function approximation approach to POMDPs. Common function approximators use smooth $C^{\infty}$ functions and these will encounter difficulty to model the non-smooth value function. The example shows that these difficulties can be encountered in the no/low knowledge states, which are of key importance for exploration.
- The model displays probing, which we define as control actions that are stronger than needed if full information were available. However, probing does not only depend on the amount of (lack

[1] Donders Institute for Brain Cognition and Behaviour, Radboud University of Nijmegen, the Netherlands, email: s.r.h.m.tonk@student.ru.nl

[2] Donders Institute for Brain Cognition and Behaviour, Radboud University of Nijmegen, the Netherlands, email: b.kappen@science.ru.nl

[3] We use the term agent throughout this paper, but mean no more by it than simply a control system.

of) knowledge, but is shown to also depend in a complex way on the state of the system and the horizon time.

In section 2 we introduce the dual control problem, we derive the corresponding Bellman equation and describe how to solve it. In section 3 we present the results of our simulations for a task to reach an configuration at the horizon time and for a tracking task where the agent must stay as close as possible to a target location for the period up to the horizon time. Finally we will discuss the results of our analysis and their relevance in section 4.

## 2 The stochastic optimal control problem

We consider the following discrete-time continuous system

$$x_{t+1} = x_t + bu_t + \xi_t \tag{1}$$

where $x_t$ is the location of the agent at time $t$, $u_t$ is the control action, $b = \pm 1$ is an unknown binary parameter and $\xi_t$ is a Gaussian stochastic variable with a mean zero and variance $\nu$. Given a sequence of controls $u_{0:T-1}$ over a future time interval $1, \ldots, T$ and an initial position $x_0$, one can define a probability distribution over future trajectories $p(x_{1:T}|u_{0:T-1}, x_0)$. The control cost is given by the expectation value

$$C(x_0, u_{0:T-1}) = \left\langle Fx_T^2 + \sum_{t=0}^{T-1} Gx_t^2 + Ru_t^2 \right\rangle \tag{2}$$

where $F, G, R$ and constants. We will study the tracking case ($F = 0, G = 1$) and the end cost case ($F = 1, G = 0$) in section 3,

When $b$ is known, the control problem is of the linear quadratic type and the optimal solution can be easily computed. When $b$ is not known (but fixed), the problem is an instance of a POMDP. It is then effective to take a large control step so that $bu_t$ is large compared to the $\xi_t$ so that based on $u_t, x_t$ and $x_{t+1}$ a reliable estimate of $b$ an be computed.

The uncertainty regarding $b$ can be modeled by defining a probability distribution over $p(b|\theta)$ that summarizes our belief about the value of $b$. Since $b$ is binary, we take $p(b|\theta) = \sigma(b\theta)$, with $\sigma(x) = (1 + \exp(-2x))^{-1}$. If $\theta = 0$, we believe that $b = \pm 1$ with equal probability. If $\theta = \pm\infty$, we are certain about the value of $b$. Thus, we trade an unknown parameter $b$ for a known distribution.

By observing the output $x_{t+1}$ after the control action $u_t$ we can update the belief parameter using the posterior obtained from Bayes's rule:

$$\begin{aligned} p_{t+1}(b) &= \sigma(b\theta_{t+1}) = p(b|x_{t+1}, x_t, u_t) \\ &\propto p(x_{t+1}|x_t, u_t) p_t(b) \\ &\propto \exp\left(-\frac{(x_{t+1} - x_t - bu_t)^2}{2\nu}\right) \sigma(b\theta_t) \end{aligned}$$

from which the following update rule for $\theta_{t+1}$ can be derived:

$$\theta_{t+1} = \theta_t + \frac{1}{\nu}(x_{t+1} - x_t) u_t \tag{3}$$

This update ensures that the belief about the control gain is adapted after every control action and gives the system the possibility to 'learn' from its actions. So at any given time $t$ the state of the dynamical system is characterized by $(x_t, \theta_t)$. The initial state is $(x_0, \theta_0)$ and $\theta_0$ defines our initial belief about the value of $b$.

Eq. 2 becomes

$$C(x_0, \theta_0, u_{0:T-1}) = \left\langle Fx_T^2 + \sum_{t=0}^{T-1} Gx_t^2 + Ru_t^2 \right\rangle \tag{4}$$

where the expectation is now over both the noise and the beliefs. In this way, we have converted the dual control problem in $x$ in an ordinary control problem in $(x, \theta)$.

The standard approach to derive the Bellman equation is to define the cost-to-go or value function

$$J_t(x_t, \theta_t) = \min_{u_{t:T-1}} C(x_t, \theta_t, u_{t:T-1}) \tag{5}$$

that solves the control problem from an intermediate time $t$ until $T$. Note, that $J$ in general depends on time. The Bellman equation results from a dynamic programming argument that relates $J_t(x, \theta)$ to $J_{t+1}(x, \theta)$ and is given by [1]

$$J_t(x_t, \theta_t) = \min_{u_t} \left( Ru_t^2 + \left\langle Gx_{t+1}^2 + J_t(x_{t+1}, \theta_{t+1}) \right\rangle_{x_t, \theta_t} \right), \quad t = 0, \ldots, T-1 \tag{6}$$

where the expectation value $\langle\rangle_{x_t, \theta_t}$ denotes that it is evaluated conditioned on the current state $x_t, \theta_t$ using Eqs. 1 and 3. Eq. 6 becomes

$$J_t(x_t, \theta_t) = \min_{u_t} \Bigg( Ru_t^2 + \sum_{b=\pm1} \sigma(b\theta_t) \int d\xi N(\xi|0, \nu) \left( G(x_t + bu_t + \xi)^2 + J_{t+1}(x_t + bu_t + \xi, \theta_t + \frac{1}{\nu}(bu_t + \xi) u_t) \right) \Bigg) \tag{7}$$

Eq. 7 is solved with boundary condition $J_T(x, \theta) = Fx^2$. The numerical details are given in the Appendix.

## 3 Numerical results

Using the theoretical framework derived in the previous section we can now analyse the optimal solutions to a end cost problem and a tracking problem for finite time horizons.

## 3.1 End cost problem

### *3.1.1 Optimal control*

We will first study the case with only end costs, i.e. $F = 1, G = 0$. The end cost $x^2$ forces the agent to reach the end goal $x = 0$ with minimal control cost. In the limit that $\theta \to \infty$, the agent is fully certain that $b = 1$ and the dual control problem reduces to an ordinary control problem for which the optimal control solution can be easily computed and is given by

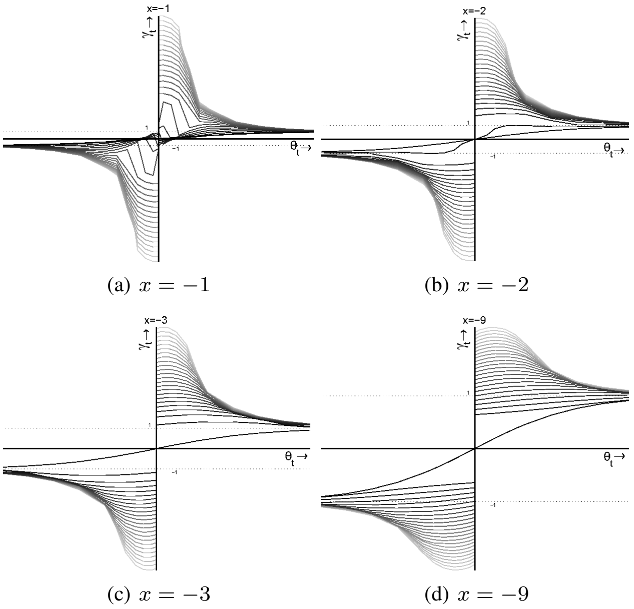
$$u_t^*(x, \theta = \infty) = -\frac{x}{R + T - t}, \quad t = 0, \ldots, T-1 \tag{8}$$

The optimal control obtained from our simulations with arbitrary $\theta$ is analyzed relative to this certain optimal control, by defining the relative control

$$\gamma_t(x, \theta) = \frac{u_t^*(x, \theta)}{u_t^*(x, \infty)} \tag{9}$$

Deviation of the relative control from the value 1, indicates how the optimal dual control differs from the ordinary control solution.

The optimal control is shown in fig. 1 for various $x$ as a function of the $\theta$ and the time-to-go. When the system is certain about the value of $b$ (large $\theta$), the control approaches the control Eq. 8. In these states the agent exploits its beliefs and steer toward the desired end-state in the optimal manner.



(a) $x = -1$ (b) $x = -2$

(c) $x = -3$ (d) $x = -9$

**Figure 1.** Dual control solution for the end cost problem. $\nu = 0.5, R = 1, F = 1, G = 0$. Each plot shows the relative control $\gamma_t(x, \theta)$ as a function of $\theta$ for different values of $t$. The lighter curves with the larger values of $\gamma_t$ are for larger times-to-go, this decreases as the curves get darker. When $\theta$ is large, $\gamma_t(x, \theta)$ approaches one, indicating that the control law when $b$ is known is recovered. The different plots show the same results for different values of $x$.

However, when uncertainty in $b$ increases (smaller absolute values of $\theta$), the optimal behavior starts to differ more and more from the certain case. For most of these situations the relative control becomes large than 1 and the system starts to show probing behavior. The longer the time-to-go ($T - t$), the more aggressive this probing gets.

In addition to probing, there is another interesting phenomenon. For a short time-to-go and large uncertainty in $b$ (small $\theta$), the system will control less than with no uncertainty; the relative optimal control is smaller than one. In this case, the agent accepts that it is in an impossible situation and that there is insufficient time left for probing. Probing would significantly increase the distance from the goal and the remaining time-to-go and the poor estimate of $b$ are likely to be insufficient to steer back to the goal. Instead, the agent becomes risk aversive and reduces its gain in order not to risk to steer even further away from the goal.

In fig. 1a one can see that the optimal control even becomes negative in some case, meaning that based on the agents belief about the most likely sign of $b$, the agent will steer in the direction away from the goal. The reason is that for any $b$ the expected changes in $x$ and $\theta$ are correlated, as shown in fig. 2.

In the optimal control case ($u_-$ in fig. 2) the agent will end up in one of two locations, depending on whether his belief is correct. If the agent is right about the sign of $b$, it will move away from the goal but also increases $\theta$ and thus gains information (the location indicated by (-+) in fig. 2). If the agent is wrong about the sign of $b$, the result of the control step is to move toward the goal (+), but will not gain information (in fact will lose information) (+-).

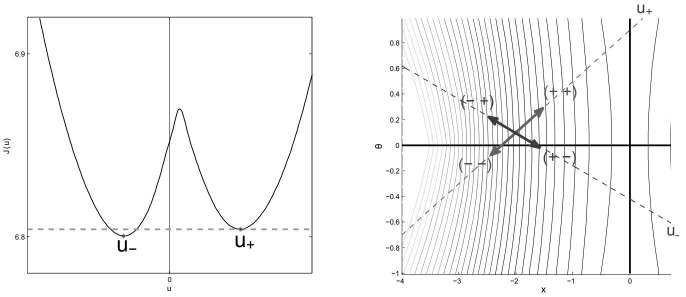


**Figure 2.** Expected next time states $(x_{t+1}, \theta_{t+1})$, given a $(x_t, \theta_t)$ both close to zero for the optimal control direction ($u_-$) and the opposite control ($u_+$). In either case one must distinguish the situation that the agent is right or wrong about his belief.

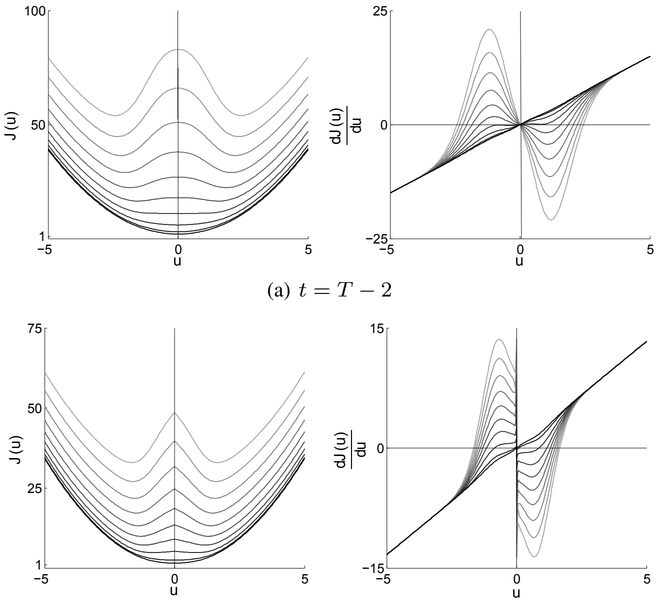
We compare this optimal choice with the suboptimal control $u_+$ in the opposite direction, i.e. the agent steers toward the goal (again based on the agents belief about the most likely sign of $b$). Again, there are two possible outcomes. If the agent is right about the sign of $b$, it will indeed get closer to the goal and also increase $\theta$ (++). However, if the agent is wrong about the sign of $b$, it will steer away from the desired state and it will not learn (- -). This last option is very unfavorable. The reason that the agent decides to steer away from the target is because the average expected cost of the outcomes (+-) and (-+) is less than the average expected cost of the outcomes (++) and (- -).

### *3.1.2 Symmetry breaking and non-differentiability of J*

The observed probing behavior arises as the result of a symmetry breaking in the right hand side of Eq. 7 and is illustrated in fig. 3.

At the final time, $J_T(x, \theta) = x^2$ independent of $\theta$ and the rhs of Eq. 7 is quadratic in $u_{T-1}$, yielding a unique optimum. The value function at $t = T - 1$ depends on both $x$ and $\theta$ and as a result $J_{T-1}$ becomes a complex non-linear function of $u_{T-2}$. At a certain time-to-go, the rhs of Eq. 7 develops multiple minima in $u_t$. Around $\theta = 0$ these minima are equally deep and either solution for $u_t$ is equally good. Both solutions are explorations in either the positive or negative direction, and neither are goal directed. Choosing one of the two will have symmetry breaking in the value function as direct consequence.

The multiple optimal control solutions in the Bellman optimization also give rise to a discontinuity in the optimal control as function of the belief parameter. At a certain time-to-go the optimal control $u^*(\theta)$ becomes discontinuous at $\theta = 0$, as is shown in fig. 1. This gives rise to non-differentiability of the value function from that point on in the Bellmann iteration, this is shown in fig. 4Left, where we plot $J_t(x, \theta)$ for $t = T - 2, x = -2$ and $t = T - 2, x = -6$ as a function of $\theta$. In fig. 4Right we also plot the same $J$ as a function of the belief $p(b = 1|\theta)$ and recover the well-known result that the optimal cost-to-go is convex in the belief [8].



(a) $t = T - 2$

(b) $t = T - 3$

**Figure 3.** The choice of the agent for exploitation or exploration is realized through a symmetry breaking mechanism in the Bellman equation. We plot the rhs of the Bellman equation as a function of $u$ and its derivative for $\theta = 0$. The different curves correspond to different values of $x$, these values increases with the lightness of the curve. Exploratory behavior ($u \neq 0$) arises in the no-knowledge state $\theta = 0$ by proposing non-zero controls. The singularity is absent at $t = T - 2$ and present starting from $t = T - 3$.

## 3.2 The tracking problem

We now look at the problem with $F = 0, G = 1$, where the agent must track a goal location $x = 0$. The main difference between the end cost task and the tracking task is that for the end cost task the optimal control and value function are time dependent for all times, while with increasing time-to-go, the control and value function for the tracking task converge to a time-independent value.

In fig. 5a, we show the optimal control as a function of $\theta$ for $x = -1$. Note, how the solution converges and becomes independent of time as the time-to-go increases. Note that, as in the end cost case, the smooth solution for small time-to-go (exploitation) breaks for larger time-to-go yielding the discontinuity at $\theta = 0$.

The main difference with the end cost task is that probing is much less pronounced and only occurs in states close to the goal. The reason is that in the end cost task probing is not penalized with state costs until the agent reaches the the end time. As a result, the agent can probe to gain knowledge and once $b$ is learned, steer back to reach the goal in time. In the tracking problem, the system pays state costs at each time and thus probings are much more expensive.

The optimal control solutions for the end cost task and tracking task for short time-to-go are very similar, as should be expected. As before, one finds instances where the optimal control steers away from the goal.

We focus on the large time-to-go behavior, when the control solution has converged to a stationary value. In fig. 5b we have redrawn the solution for $x = 1$, where we have also drawn the sub-optimal control solution that arise from the minimization of the rhs of Eq. 7. It indicates the region of symmetry breaking in $u$ around $\theta = 0$ that is caused by the uncertainty in $b$. Similar regions of uncertainty can be defined for the end cost problem in fig. 1.

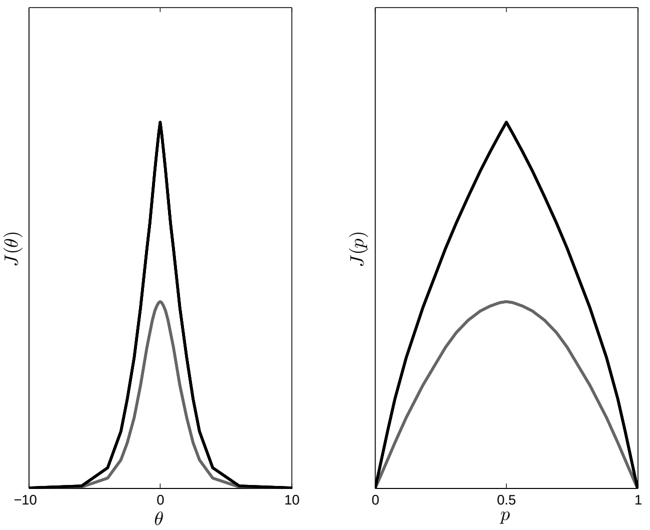


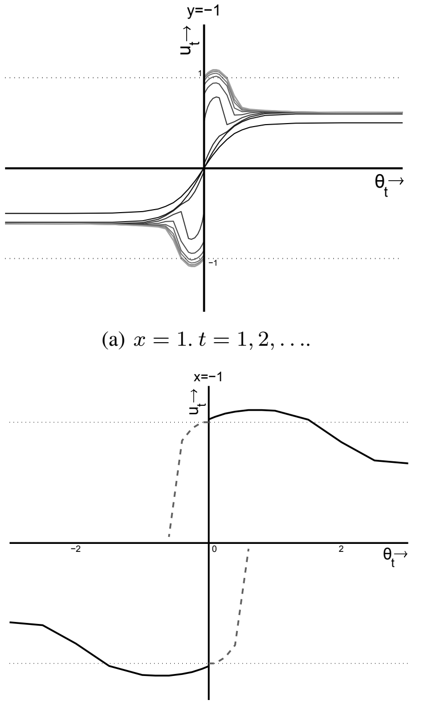
**Figure 4.** Left) The optimal value function $J_t(x, \theta)$ for $t = T - 2, x = -2$ (gray) and $t = T - 2, x = -6$ (black) versus $\theta$ rescaled to fit in the same plot, showing that the value function is smooth for small time-to-go and becomes non-differentiable at $\theta = 0$ for larger time-to-go. Right) Same as left, but as a function of the belief: $J_t(x, p)$, with $p = p(b = 1|\theta)$. The optimal cost-to-go is convex in the belief for any time-to-go.

## 4 Discussion

We have presented a detailed analysis of a simple dual control problem. The analytical solutions to this problem show that optimal exploration occurs through symmetry breaking in Bellman equations, an effect which is accompanied with non-differentiability in the optimal value function. In the example considered here, the optimal exploration that is characterized by a Mexican hat shaped value function with high valued non-exploratory actions on the top of the hat, and the low valued exploratory actions in rim of the hat. The systems must choose an action corresponding to one of the minima in order to advance in its state of knowledge, and by doing so the symmetry of the value function is broken. Alternatively the low knowledge state is a minimum itself, in which case no exploration takes place.

This behavior is not only true and intuitive for a binary system, but also for other types of systems, e.g. with higher dimensionality. In the general case there are by definition many possible control actions that could be useful for exploration. Depending on the belief state, some of these actions will have a lower expected future cost than others. However, at some particular points in the belief state the expected future cost can be the lowest for a subset of different actions. In this case all the actions within this subset are optimal, and by choosing one of the actions as solution the symmetry of the system will be broken. Furthermore, the optimal control as function of the belief parameter will be discontinuous for these particular states, this has non-differentiability in the value function as a direct consequence, as was also shown in the example problem.

We now summarize what we believe that this example says about the type of problems that one may encounter in a POMDP setting and require further investigation.



(a) $x = 1$. $t = 1, 2, \ldots$.

(b) $x = 1$. $t \to \infty$.

**Figure 5.** a). Dual control solution for the tracking problem. $\nu = 0.5, R = 1, F = 0, G = 1$, showing $u_t(x, \theta)$ for $x = 1$ as a function of $\theta$ for different times-to-go $t$ (increasing for lighter curves) The tracking task displays much less pronounced probing compared to the end cost case. As in the end cost case, the optimization for $u$ has local minima resulting in the discontinuity at $\theta = 0$. b). Asymptotic solution at infinite time-to-go.

The tracking problem is similar to an infinite horizon discounted reward RL problem without absorbing state, where the horizon exponentially decays with characteristic length $1/\log\gamma$ [6]. The end cost problem is similar to an RL problem with absorbing state. Thus, we expect that in POMDPs the optimal exploration will display similar features as discussed here: non-convexity in the minimization of the Bellman equation with respect to $u$; optimal actions that are discontinuous in the beliefs; and a non-differentiable (but convex) optimal value function. It is in particular important to realize that the non-convexity in the minimization of the Bellman equation co-exists with the convexity of the optimal value function.

The non-differentiability of the value function poses a problem for function approximation. In our example, we had only one belief variable and the singularity was restricted to the point $\theta = 0$. In this case, we could construct separate solutions for $\theta$ positive and negative. In general, however, with many belief parameters this singular structure may become a quite complex high dimensional object. A smooth function approximation approach that ignores these singularities will not succeed in finding an accurate approximation. One could argue that these difficulties are not of great importance, because they are restricted to some isolated region in parameter space. However, as we also showed in our example, it is exactly in the region of no/low knowledge space were these difficulties are encountered. It is this region that specifically characterizes the initial phase of exploration and therefore is of key importance for optimal exploration. For instance in the method proposed in [2] for finding the optimal policies in POMDPs similar to the analyzed example, the assumption is made that the value function is smooth under the belief state. From the results of our analysis we see that one has to be careful with making these kind of assumptions for these problems.

By analyzing a simple example we were able to get some insight in the intricacies of optimally solving an exploration/exploitation problem. These results provide some hints as to where possible weak spots are in the conventional approach to solving exploration/exploitation in RL. It is hoped that this paper will stimulate further research into these types of problems, so new and better approaches to doing Bayesian RL and solving POMDPs can be developed.

## A Numerical aspects

To numerically solve the Bellman equation, we discretize $x$ uniformly with step size $\Delta x = 0.25$ between $x_- = -9$ and $x_+ = 9$, and $\theta$ non-uniformly with increasing step size for small $\theta$ between $\theta_- = -13$ and $\theta_+ = 13$. For each grid point and each time step, we minimize the rhs of Eq. 7 with respect to $u_t$ using a conjugate gradient descend method.

The integral over $\xi$ is computed by discrete integration in the range $x_{T+1} \in [x_t + bu_t - 10\sqrt{\nu}, x_t + bu_t + 10\sqrt{\nu}]$, with step size $\Delta x_{t+1} = 2^{-4}$ and cubic spline interpolation between the $x, \theta$ grid points. Note, that $\theta_{t+1}$ is given, once $x_t, x_{t+1}, u_t$ and $b$ are given. The range of points $(x_{t+1}, \theta_{t+1})$ thus considered exceeds the above defined grid, meaning that $J_{t+1}(x_{t+1}, \theta_{t+1})$ needs to be extrapolated. This is done in the x-direction by fitting a second order polynomial to $J$ for the large values of $x$ only and in the $\theta$ direction $J$ saturates to a constant value.

The minimization over $u$ is run multiple times with different initializations to ensure that the global minimum is obtained.

## REFERENCES

[1] D.P. Bertsekas. *Dynamic Programming and optimal control.* Athena Scientific, Belmont, Massachusetts, second edition, 2000.

[2] M. Duff. Design for an Optimal Probe. *In Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 131–138, 2003.

[3] A.A. Feldbaum. Automation remote control. *Dual control theory*, I-IV(21-22):874–880, 1033–1039, 1–12, 109–121, 1960.

[4] N.M. Filatov and H. Unbehauen. *Adaptvie dual control.* Springer Verlag, Reading, Massachusetts, 2004.

[5] J.J. Florentin. Optimal, probing, adaptive control of a simple bayesian system. *International Journal of Electronics*, 13:165–177, 1962.

[6] H.J. Kappen. An introduction to stochastic control theory, path integrals and reinforcement learning. *AIP Conference Proceedings*, 887:149–181, 2007.

[7] P. Poupart and N. Vlassis. Model-based bayesian reinforcement learning in partially observable domains. *Proceedings International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2008.

[8] E.J. Sondik. *The optimal control of partially observable Markov processes.* Athena Scientific, PhD thesis, Stanford University, 1971.

[9] H. Theil. A note on certainty equivalence in dynamic planning. *Econometrica*, 25(2):346–349, 1957.

[10] A.M. Thompson and W.R. Cluett. Stochastic iterative dynamic programming: a Monte Carlo approach to dual control. *Automatica*, pages 767–778, 2005.

[11] K.J. Astrom and B. Wittenmark. Problems of identification and control. *Journal of Mathematical analysis and applications*, 34:90–113, 1971.