Advanced Engineering Informatics 29 (2015) 1083–1095

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

An active inference approach to on-line agent monitoring in safety–critical systems

Luis Avila, Ernesto Martínez *

INGAR (CONICET-UTN), Avellaneda 3657, Santa Fe S3002 GJC, Argentina

ARTICLE INFO

Article history:
Received 1 April 2015
Received in revised form 30 July 2015
Accepted 31 July 2015
Available online 13 August 2015

Keywords:
Active inference
Bayesian surprise
On-line monitoring
Twin Gaussian processes

ABSTRACT

The current trend towards integrating software agents in safety–critical systems such as drones, autonomous cars and medical devices, which must operate in uncertain environments, gives rise to the need of on-line detection of an unexpected behavior. In this work, on-line monitoring is carried out by comparing environmental state transitions with prior beliefs descriptive of optimal behavior. The agent policy is computed analytically using linearly solvable Markov decision processes. Active inference using prior beliefs allows a monitor proactively rehearsing on-line future agent actions over a rolling horizon so as to generate expectations to discover surprising behaviors. A Bayesian surprise metric is proposed based on twin Gaussian processes to measure the difference between prior and posterior beliefs about state transitions in the agent environment. Using a sliding window of sampled data, beliefs are updated a posteriori by comparing a sequence of state transitions with the ones predicted using the optimal policy. An artificial pancreas for diabetic patients is used as a representative example.

1. Introduction

Anomaly or novelty detection refers to the process of pinpointing unusual and unexpected events, behaviors or patterns that give rise to concerns regarding system safety or performance. This is especially important for monitoring safety–critical systems in which faulty conditions need to be fast accounted for [1]. The issue of anomaly detection has generated substantial research over past years. Complete reviews of the novelty detection literature during the last decade can be found in [2,3]. It is rather clear from the existing works that the most common form of anomaly detection in use today is based on thresholds, fixed limits or crisp boundaries between what would be considered "normal" or expected and something unexpected or abnormal. Typically, when the value of a key variable is above or below of a pre-specified bound or limit, the system is considered to be in an anomalous state. For example, a performance monitor for model-based controllers was proposed using a minimum variance benchmark for a model residuals obtainable from closed-loop data [4]. Similarly in [5], the temperature measured across bridge girders was correlated to a certain state of structural performance degradation. Patterns of physiological deterioration in hospital patients (evident in the vital signs, such as heart and respiratory rate) have been classified using support vector machines [6]. Automatic detection of landmarks in the environment for topological mapping has been associated with surprising measurements, where a location is classified to be a landmark if its surprise value exceeds a given threshold [7]. In a method for detecting lane deviation of a vehicle [8], the difference between a center point of lane markers and the center point of the vehicle is computed in order to alert the driver that a dangerous deviation exists.

In system monitoring, the most active research area is control loop performance monitoring based on statistics computed from plant data. The most significant development was due to the work of Harris [9], who proposed the novel concept of a minimum variance controller as the characteristic behavior of an optimal regulator. Thus, performance monitoring can be based on comparing the observed mean squared error of the controlled output against its corresponding minimum variance. This optimal behavior is based on the theoretical framework of minimum variance control previously developed [10] and has already been implemented with some success in monitoring the performance of a control loop for a diabetic patient [11]. More recently, Harrison and Qin [12] developed a minimum variance performance map that address the impact of constraints on a predictive controller operation, which highlights the inadequacy of the minimum variance criterion for more complex control tasks such as real-time optimization and intelligent control with economic objectives. Unlike the well-known minimum variance index for SISO loops, a recent approach

* Corresponding author. Tel.: +54 (342) 4534451; fax: +54 (342) 4553439.
E-mail address: ecmarti@santafe-conicet.gov.ar (E. Martínez).

http://dx.doi.org/10.1016/j.aei.2015.07.008
1474-0346/

Nomenclature

$a(\mathbf{x})$	uncontrolled or passive dynamics
$B(\mathbf{x})$	input-gain matrix
$\mathcal{D}$	dictionary of training examples
$\mathcal{G}[z](\mathbf{x})$	integral operator of the desirability function
$GP(m,k)$	Gaussian process with mean m and covariance k
$k(\mathbf{x}_i,\mathbf{x}_j)$	covariance function
$KL(\|\|)$	Kullback–Leibler distance
p^0	transition probability distribution for the passive dynamics
$p^{\mathbf{u}}$	transition probability distribution for the controlled dynamics
$p^*(\|)$	state transition distribution under optimal control
$p^g(\|)$	state transition distribution under any implemented control
$P_{k,k+1}$	matrix of transition probabilities for the passive dynamics
$q(\mathbf{x})$	cost function
$\mathbf{s}$	environment internal or hidden state
SI	surprise index
$T_{KL}(\|\|)$	Twin Kullback-Leibler distance
$\mathbf{u}$	current control action of agent
$v(\mathbf{x})$	optimal cost-to-go function
$\mathbf{x}$	observable environmental state
x	each state dimension
$\boldsymbol{X}^g$	sequence of state transition given system observations
$\boldsymbol{X}^*$	sequence of state transition given the specification
Δx	transition for each state dimension.
$z(\mathbf{x})$	desirability function

Greek symbols

$\boldsymbol{\mu}$	monitor's beliefs
$\boldsymbol{\Psi}(\mathbf{x})$	a given state transition distribution in the agent environment
$\boldsymbol{\Psi}(\boldsymbol{\mu})$	monitor's belief distributions over future state transitions
$\xi = D(\boldsymbol{\Psi}(\boldsymbol{\mu})\|\|\boldsymbol{\Psi}(\mathbf{x}))$	distance between belief and state transition distributions
γ_r	kernel width parameter
λ_r	noise variance
δ_{ij}	Kronecker delta function
$\pi^*(\mathbf{x})$	optimal control policy
$\pi^g(\mathbf{x})$	an implemented control policy
σ	scaling parameter for Brownian noise
$d\omega$	Brownian noise

Glycemic symbols

BG	blood glucose level
Sh	hepatic sensitivity
Sp	insulin sensitivity
I	insulin infusion level
S	glycemic sensor output
τ	sensor time-lag parameter
ξ	sensor calibration parameter

was proposed by Srinivasan et al. [13], which does not require for performance monitoring any knowledge about time delays or other plant parameters. However, due to unavoidable uncertainties about environmental conditions, none of the existing performance monitoring algorithms can give a conclusive answer regarding normality or abnormality in a control loop operation [14,15].

It is worth noting that most of the existing literature about online monitoring assumes that the observed system is a "passive" entity not situated in an environment. Therefore, a passive monitoring approach assumes that certain environmental variables are simply correlated to a certain system state or condition according to an immutable relationship; e.g., a temperature distribution across a structure have a strong correlation with structural response. However, agents are "situated" entities and usually perform in time-varying environments that maintain a feedback loop with them. By responding to an external stimulus, the agent carry out actions whose effects give rise to a sequence of state transitions in its environment, which is in turn affected by other agent's actions, setting a dynamic structure as depicted in Fig. 1. An agent behavior is thus an emergence of its control policy for responding to different environmental stimulus and hence it cannot be regarded to a fixed reference, threshold or bound. Unfortunately, an agent policy cannot be directly monitored since it remains hidden to any external observer.

Behavior monitoring aims at characterizing the control policy that an agent uses to react to external stimuli by inferring the state transition distribution $\boldsymbol{\Psi}(\mathbf{x}_k)$ in the agent environment as a result of its actions. Besides noise and variability, uncertainty arises as a consequence of hidden states either because some variables are only partially observable or because they are related to the perception an agent has of its environment. As an example consider highway monitoring where a driver's actions (e.g. to change lane) can only be inferred through observable changes in its immediate environment (the car steering), as well as the effect on it of other nearby affected agents (drivers). Clearly, the monitor perception change over time as sensory information from the agent environment is updated [16]. Thus, for behavior monitoring, the main consequence of the assumption of a situated agent is that any change in the environmental dynamics is mostly a consequence of the agent's policy. Furthermore, notice that the raw readings from sensors do not usually correspond to the agent actions nor the states it perceives or the rewards it obtains, which are difficult if not impossible to measure directly by the monitor. It is thus reasonable to assume that only changes occurring in the nearby environment in which the agent performs are part of the monitor's perception. For example, a highway monitor may recognize if a driver increases the speed by considering temporal variations in the vehicle position, as opposed to predicting car velocity based on the specific force applied on the car's throttle.

By ascribing the evolution of environmental transitions to a certain agent policy, a representation that captures much of the relevant information of an agent behavior is obtained. Since behavior in uncertain environments is a phenomenon that unfolds over

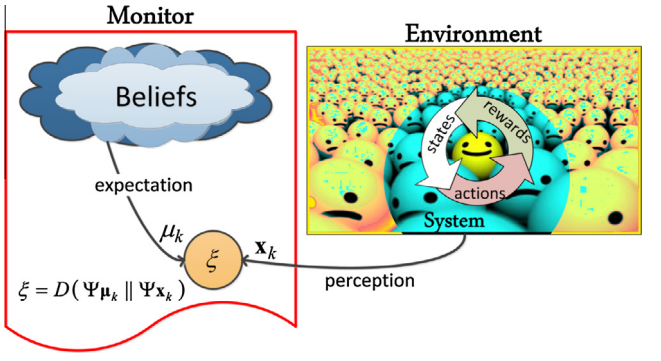

Fig. 1. Active inference approach to agent behavior monitoring. The probabilistic process involved by the real environmental dynamics is repetitively contrasted to the expectation of such dynamics that is modeled by the monitor.

time, the monitor needs to constantly infer potential outcomes of future agent actions in response to sensory information ($\mathbf{x}_k$) of the environmental state. To achieve this, the monitor continuously revises its hypotheses in order to update prior beliefs ($\boldsymbol{\mu}_k$) that include predictive distributions $\boldsymbol{\Psi}(\boldsymbol{\mu}_k)$ over future state transitions. This strategy is characterized in this work in terms of an active inference approach [17], in which the monitor perceives the nearby environment of an agent in order to contrast its prior beliefs about expected state transitions. These prior beliefs are built around a generative model of an optimally controlled stochastic process for state transitions, which involves predictions of what should be sampled in order to validate prior beliefs. Since this stochastic process has been a priori optimized, some of the state transitions are expected to be observed more frequently than others and labeled accordingly as "desirable" [18]. In particular, an optimal behavior is one that reduces the divergence between desirable state transitions and the observed ones. Therefore, any deviation between the agent behavior and the desired one can be detected in a natural way through the Kullback–Leibler distance between belief distributions and posterior distributions for state transitions perceived through sensory data $\xi = D(\boldsymbol{\Psi}(\boldsymbol{\mu}_k)||\boldsymbol{\Psi}(\mathbf{x}_k))$.

For on-line monitoring, an indication of surprise should arise from a mismatch between expectations and what is actually perceived. Surprise is thus an information measure that requires an active inference of the environmental effect of future agent actions, which may change an observer beliefs as the environmental dynamics unfolds. However, it is important to figure out expectations about the different environmental situations the agent may face and how each situation may be characterized. In general, expectations are representations of the values that some perceptual features are likely to assume in the future. Because of uncertainty, expectations are naturally expressed in probabilistic terms such that a probability distribution over the range of possible observations can be considered to be a "belief state," usually conditioned on a particular unobservable state or hidden context. This approach has been applied in [19] to support patients with dementia during hand-washing tasks. The position of certain objects (e.g., hands and towel) is evaluated by the monitor to estimate the progress of the user in the task as well as its current mental condition. This is expressed as a belief state over a range of possible sub-tasks and mental states. Therefore, if an estimated probability of an observation is available, then the certainty of a sequence of perceptions can be compared to its probability of occurrence yielding a measure of surprise.

This article is structured as follows. Section 2 describes agent behavior monitoring as an active inference problem based on an optimal control policy. Section 3 briefly introduces the principles of optimal action selection to characterize the expected behavior under uncertainty. Expected behavior is formalized as a controlled stochastic process that makes the agent policy as close as possible to the desired one by describing environmental transitions in probabilistic terms. In Section 4, surprise is quantified as the Kullback–Leibler divergence between distributions about the desired and the actual state transitions in order to pinpoint any deviation from the expected agent behavior. In Section 5, an artificial pancreas is used as case study in which sensor and actuator faults which may endanger safety or optimal operation are considered. Finally, in Section 6 some remarks and future works are discussed.

2. Active inference

2.1. Action and perception

Perception is essential to intelligent systems since it helps cognitive agents to build their hypotheses and select an optimal course of action in the face of a changing environment. A rational agent chooses actions seeking to maximize the utility obtained despite uncertainty and unplanned events. As it is shown in Fig. 1, the agent influences the nearby environment which is mostly affected by its actions whereas the environment responds by means of rewards and state transitions. As a result, the sequence of observable state transitions in the agent environment is important for monitoring its behavior. This is fundamental, since the agent's policy usually remains unobservable to an external observer in most cases. Bearing this in mind, and through a probabilistic characterization of the agent policy under uncertainty, it is possible to indirectly monitor the agent policy by accounting for changes in the environmental dynamics. These changes are considered as the observable responses to the agent actions, but could be influenced by actions taken by other agents and uncertainty about internal agent states.

Our monitoring approach starts by differentiating the probabilistic process descriptive of the real environmental dynamics that generates new observations, from the expectation of such dynamics that is modeled by the monitor. As an example, Fig. 2 highlights that by perceiving an internal state $\mathbf{s}_k$, the driver chooses the action $\mathbf{u}_k$, which causes the environment to evolve to the state $\mathbf{s}_{k+1}$, while the driver obtains a (unobserved) reward r_k. The Markovian dependences between the driver actions affecting internal states, which subsequently yield observable responses $\mathbf{x}_k$ to the monitor, are highlighted in Fig. 2. Notice how the given sequence of actions performed by the agent does not have a direct influence on the monitor's posterior beliefs $\boldsymbol{\mu}_{k+1}$; the agent policy just becomes apparent when it manipulates the sequence of environmental transitions. Therefore, the monitor perception cannot be directly related to the agent actions, but rather to the observable changes in the environmental dynamics which is affected as a result. Thus, active inference necessarily involves on-line rehearsing of future actions of agents over a rolling horizon so as to generate expectations that help discovering surprising agent behaviors.

From the monitor's perspective, each environmental state $\mathbf{x}_k$ perceived is able to modify or reinforce the prior beliefs $\boldsymbol{\mu}_k$, continuously transforming them into a posterior distribution $\boldsymbol{\mu}_{k+1}$ according to the Bayes rule. This constitutes a behavior monitoring approach that employs, from a Bayesian point of view, the probability distribution for the next environmental response as a reference to the monitor. It is important to remark this is a moving horizon method, since the posterior state transition density obtained corresponds to the a priori distribution for the next time interval.

The active inference monitoring approach is based on an optimal control policy for the agent under uncertainty. This policy allows simulating fictitious state transitions that serve as reference to the monitor. In this sense, prior beliefs correspond to a description of how the agent is expected to act, and hence beliefs are

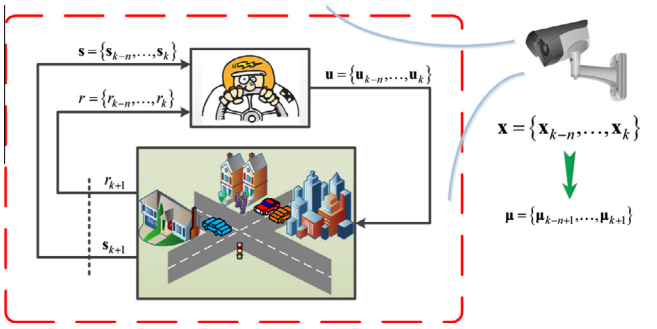

Fig. 2. Markovian dependencies among hidden states generating a sequence of sensory data, which in turn modify or reinforce prior beliefs.

described as a probabilistic model of expected state transitions that the monitor uses to assess the agent behavior. It is important to differentiate between the probabilistic process over which the monitor makes predictions, and the actual environmental dynamics that generates the sensor signals perceived by the monitor. Predictions about state transitions are the result of a generative model, which has been trained to infer a sequence of fictive state transitions and is described in the next section. Predicted transitions are optimal in the sense they are the result of an optimally acting agent in the face of uncertainty. However, predicted transitions are nothing more than an internal representation used by the monitor that may or may not match the future evolution of the environmental dynamics.

The active inference approach to agent monitoring implies that based on prior beliefs, new predictions about environmental state transitions can be made. The monitor's specification is built upon Gaussian distributions that embody a sequence of state transitions, which take place when the environmental dynamics is controlled by an optimal policy under uncertainty. The optimal policy is obtained using a type of Markov decision process whose control law can be framed as a linear problem with an analytical solution (see Section 3). This linearization is accomplished through an exponential transformation of the Bellman's equation, which leads to more efficient numerical methods. Since the sequence of state transitions simulated by the generative model embeds an optimal policy, any agent that behaves properly is one that reinforces the monitor beliefs over time.

Consider for instance the problem of monitoring a vehicle on a road, as in Fig. 3. The image describes the virtual environment perceived by a monitor that is observing the vehicle while the driver is about to make a left turn. There exist pedestrians, cyclists and other vehicles setting a dynamical scene which assists the driver and the monitor to characterize the situation by means of only observable information. Detection dropouts caused by noise, occlusions, and other artifacts are assumed to be hidden states since they can influence sensory data but cannot be measured directly. Based on an optimal behavior specification, the monitor infers a hypothetical trajectory for the vehicle, depending on its prior knowledge and the environment state [20]. After observing the agent environment, the monitor contrasts incoming sensory information with prior beliefs that describe the expected behavior of a proficient driver, such that the car efficiently avoids static and dynamic obstacles while obeying traffic rules. Any deviation from the expected behavior can be detected by quantifying the distance between the observed state transitions and the predicted ones.

It is worth noting that the environmental states perceived by the driver are not necessarily the same states perceived by the monitor, since the driver may discern a number of activities that remain unobservable to the monitor. For instance, a mobile phone vibrating may alter the driver attention, but the monitor is clearly unaware of such situation. As a result, the monitor must infer the agent (driver) behavior just through the evolution of the state transitions in the system composed by the driver and its vehicle.

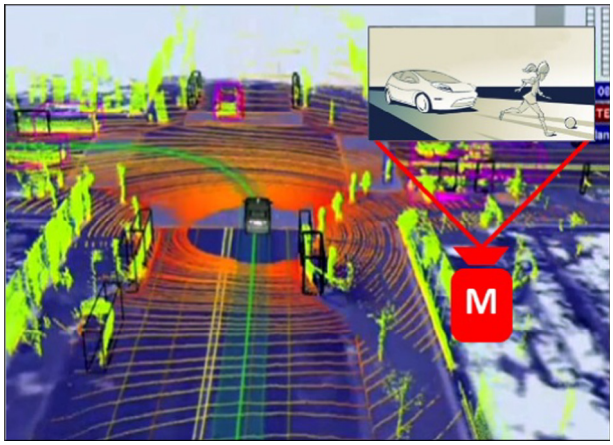

Fig. 3. Virtual environment perceived by a monitor that is observing the vehicle while the driver is about to make a left turn.

As another example (left side of Fig. 3), a rolling ball on the lane should prompt -for both the driver and the monitor- the possibility of a child running out from the sidewalk. From the point of view of the monitor, a proficient driver should react by slowing down the car to a halt and looking out for someone. This is because the monitor beliefs should consider as prior knowledge both normal and abnormal situations and specify the optimal agent behavior in each scenario. In turn, a rolling ball may be an unexpected or surprising event for the driver considering its prior beliefs, yet the driver behavior to handle it should not to be surprising to the monitor. This example shows that for on-line monitoring of an agent behavior, an event is surprising not because its probability is small in an absolute sense, but rather because its probability is relatively small given the prior belief distribution of the observer (monitor) [21].

2.2. Prior beliefs specification

A distinctive problem in monitoring situated agents is that the monitor's perception is partially blinded, which creates uncertainty about the agent policy that explains the observed state transitions. Prior beliefs about the optimal policy are instrumental in order to predict expected responses given the history of observed state transitions over a rolling horizon. In this work, the monitor beliefs are transformed into a generative (probabilistic) model which makes possible to infer future state transitions resulting from an optimally behaving agent.

In the generative model, state transition probabilities are modeled using Gaussian processes (GPs) [22] that provide information about confidence intervals for the predicted next state. A GP is a collection of random variables, any finite number of which has a joint Gaussian distribution. For Gaussian process regression, those random variables represent the value of the function $f(x)$ for inputs x. It is assumed that $f(x)$ is a zero mean stationary Gaussian process with covariance function $k(x_i, x_j)$, encoding correlations between pairs of random variables

$$\mathrm{cov}(f(x_i, x_j)) = k(x_i, x_j) \tag{1}$$

One covariance function particularly useful is the Gaussian

$$k(x_i, x_j) = \exp(-\gamma_r \|x_i - x_j\|^2) + \lambda_r \delta_{ij} \tag{2}$$

with $\gamma_r \geqslant 0$ the kernel width parameter, $\lambda_r \geqslant 0$ the noise variance and δ_{ij} the Kronecker delta function, which is 1 if $i = j$ and 0 otherwise. This prior for the kernel function constrains input samples that are nearby to have highly correlated outputs.

GPs are parameterized based on a sequence of observations resulting from on-going interactions between the agent and its environment. Given a state vector $\mathbf{x}_k$, a separate GP model is trained for each state dimension x_k, in such a way the uncertainty about its change due to a (hidden) agent action is modeled statistically as

$$\Delta x_k \sim GP(m, k) \tag{3}$$

where m is the mean function and k is the covariance function. The training inputs for a Gaussian model GP are the environmental states, whereas the targets are the differences between the successor state and the state in which the action was applied. For an input $\mathbf{x}_k$, the multivariate predictive distribution $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$ is Gaussian distributed.

To describe the monitor's prior beliefs, the transition probability $p^*(\mathbf{x}_{k+1}|\mathbf{x}_k)$ should correspond to an optimally controlled system dynamics. The superscript indicates that the transition probability is shifted by the optimal control policy $\pi^*(\mathbf{x})$. As a result, the monitor is allowed to ascribe any change in the environmental dynamics to the agent policy. On the other hand, another Gaussian model would describe the current observed system behavior in terms of a Gaussian process $p^g(\mathbf{x}_{k+1}|\mathbf{x}_k)$ which may deviate from the corresponding Gaussian process for optimal action selection. The superscript indicates that the transition probability is shifted by any implemented control policy $\pi^g(\mathbf{x})$. It is worth noting that the Gaussian model that characterizes the current implementation needs to be updated on-line as new data is available. This vis-à-vis comparison between $p^g(\mathbf{x}_{k+1}|\mathbf{x}_k)$ and $p^*(\mathbf{x}_{k+1}|\mathbf{x}_k)$ allows the monitor to contrast the characterization of the observed agent behavior to its specification. As it will be discussed later, this comparison can be performed through Bayesian methods that can measure the divergence between two stochastic processes.

3. Optimal control under uncertainty

3.1. Linearly-solvable control problems

As expressed above, the main challenge for monitoring an agent behavior is how the environmental response to an optimal agent policy can be characterized in the face of uncertainty. To this aim, a probabilistic representation of the desired behavior has to be built upon a stochastic process of the optimally controlled system dynamics. To obtain the expected agent policy under uncertainty a class of Markov decision process is used. Linearly-solvable Markov decision processes (LSMDPs) [23] correspond to a class of optimal control problems in which the Bellman's equation can be converted into a linear equation by an exponential transformation of the state value function. The Bellman's equation is fundamental to optimal control theory. This equation was first introduced as the cornerstone of the Dynamic Programming framework (DP) [24]. Later on, the Bellman's optimality condition was instrumental for reinforcement Learning (RL) [25] algorithms, which are very general but can be inefficient [23]. The RL problem consists of learning iteratively to achieve a goal from ongoing interactions with a real or simulated system, while DP is a general method of solving sequential optimization problems using a probabilistic model of state transitions. In both cases, the idea is to predict the value or utility of future actions. Accordingly, optimal actions cannot be found by greedy optimization of the immediate cost, but instead all future costs must be taken into account. To this aim, the optimal cost-to-go function $v(\mathbf{x})$ is defined as the expected cumulative cost for starting at state $\mathbf{x}$ and acting optimally thereafter. Indeed, the Bellman equation characterizes $v(\mathbf{x})$ only implicitly, as the solution to a dynamic optimization problem. A major advantage of using this optimization method is that it provides an explicit optimal decision policy for the agent. Therefore, based on LSMDP it is possible to build a specification of the system dynamics over time when the agent acts according to the obtained optimal policy. The latter is instrumental to detect on-line discrepancies between the beliefs and the environmental dynamics affected by the agent behavior.

Stochastic processes do not have time derivatives in the conventional sense and, as a result, they cannot be manipulated using the ordinary rules of calculus. Ito [26] provided a way around this problem by defining a particular kind of uncertainty representation based on the Wiener process as a building block. An Ito process is thus a stochastic process whose state transition function is represented by the equation

$$d\mathbf{x} = a(\mathbf{x})dt + B(\mathbf{x})(\mathbf{u}dt + \sigma d\omega) \tag{4}$$

where $\omega \in \mathbb{R}^{n_u}$ (same space as actions) and σ denote Brownian noise and its scaling parameter, respectively. The expression $a(\mathbf{x})$ describes the uncontrolled or passive dynamics [23] and $B(\mathbf{x})$ is the input-gain matrix. It is important to remark that, since noise and control signals act over the same space, any state can be reached by the effect of either inherent system variability or control actions.

In order to express Eq. (4) in a more convenient form, the h-step transition probability for the passive dynamics p^0 is expressed as a Gaussian distribution $\mathcal{N}$ as

$$p^0(\mathbf{x}_{k+1}|\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + ha(\mathbf{x}) + hB(\mathbf{x}), h\sigma B(\mathbf{x})^T B(\mathbf{x})) \tag{5}$$

The controlled diffusion process $p^{\mathbf{u}}$ is approximated as a deterministic function expressed as a Gaussian distribution whose mean and covariance are given as

$$p^{\mathbf{u}_k}(\mathbf{x}_{k+1}|\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_{k+1}|\mathbf{x}_k + h(a(\mathbf{x}) + hB(\mathbf{x})\mathbf{u}), \Sigma) \tag{6}$$

One way of thinking of the net effect of control actions is noting how they change the distribution of the next state from $(\mathbf{x}_k + ha(\mathbf{x}) + hB(\mathbf{x}), \Sigma)$ to $(\mathbf{x}_k + h(a(\mathbf{x}) + hB(\mathbf{x})\mathbf{u}), \Sigma)$, where $\sum = h\sigma B(\mathbf{x})^T B(\mathbf{x})$ is the covariance. In other words, the controller shifts the probability distribution from one region of the state space to another [27]. More generally, we can think of the system under study as having a passive dynamics with a distribution p^0 over future states, thus the controller acts by modifying this distribution to obtain a new dynamics $p^{\mathbf{u}_k}$.

Thereby, the situated agent shifts the probability distribution for state transitions from one region of its state space to another by acting on its environment. A control policy $\pi(\mathbf{x})$ is thus defined as the probability of choosing the control action $\mathbf{u}_k$ at state $\mathbf{x}_k$. For any optimal control application, the main objective is to find an optimal policy $\pi^*(\mathbf{x})$ which minimizes the expected cumulative cost function $v(\mathbf{x})$ as

$$v^*(\mathbf{x}) = \min_u \{\ell(\mathbf{x}, \pi(\mathbf{x})) + \mathbf{E}_{x' \sim p^{\mathbf{u}}(x'|x)}[v(\mathbf{x}')]\} \tag{7}$$

where $\mathbf{x}'$ denotes the next state for a given control action $\mathbf{u}$. The minimum cumulative cost for starting at state $\mathbf{x}_k$ and acting optimally thereafter enables greedy computation of optimal actions. Notice that Eq. (7) corresponds to the Bellman fundamental equation, which can be simplified by assuming that the immediate cost function is

$$\ell(\mathbf{x}, u) = hq(\mathbf{x}) + KL(p^{\mathbf{u}}(\mathbf{x}_{k+1}|\mathbf{x}_k)||p^0(\mathbf{x}_{k+1}|\mathbf{x}_k)) \tag{8}$$

The state cost $q(\mathbf{x})$ is an arbitrary function encoding how (un) desirable different states are, and KL is the Kullback–Leibler divergence that measures the distance between the optimally-controlled dynamics and the passive one. This distance can be understood as the price to pay for the optimal shift of the passive dynamics by action $\mathbf{u}$. The KL divergence between the above Gaussian processes can be proven to be $h/2\sigma^2\|\mathbf{u}\|^2$ which is the quadratic energy cost accumulated over interval h. By introducing the exponential transformation $z = \exp(-v)$, the Bellman's equation can be conveniently re-written as [23]:

$$z(\mathbf{x}) = \exp(-hp(\mathbf{x}))\mathcal{G}[z](\mathbf{x}) \tag{9}$$

where $z(\mathbf{x})$ is the desirability function defined as

$$z(\mathbf{x}) = \exp(-v^*(\mathbf{x})) \tag{10}$$

and $\mathcal{G}[z](\mathbf{x})$ is an integral operator defined as

$$\mathcal{G}[f](\mathbf{x}) = \int p^0(\mathbf{x}'|\mathbf{x}) f(\mathbf{x}') d\mathbf{x}' \tag{11}$$

In contrast to the cost function $v(\mathbf{x})$, the negative exponential portrays which states are more desirable. Once the desirability function is found, the optimal control policy is computed analytically and expressed in closed form as

$$\mathbf{u}^*(\mathbf{x}) = -\sigma^2 B(\mathbf{x})^T v_{\mathbf{x}}(\mathbf{x}) \tag{12}$$

In this manner, optimal actions can be expressed analytically given the optimal cost-to-go. Thus, instead of finding a trajectory-based optimal solution, the goal is to find a globally optimal policy over the entire state space.

The continuous problem given in Eq. (4) is solved by choosing a set of states $\{\mathbf{x}_n\}$ and adjusting the matrix $P_{k,k+1}$ of transition probabilities from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$ given by the passive dynamics described in Eq. (5). Since the operator $\mathcal{G}[z](\mathbf{x})$ is linear, Eq. (9) is also linear and can be expressed in vector notation. Defining the vector z with elements $z(\mathbf{x}_n)$ and the matrix Q with elements $\exp(-hq(\mathbf{x}_n))$ on its main diagonal Eq. (9) becomes

$$z = QPz \tag{13}$$

for the first exit formulation, namely goal-direct problems. Noteworthy, Eq. (13) can be solved by an iteration method in exponential form [18].

3.2. Example: car-on-a-hill problem

The mountain car problem provides a challenging problem, particularly when random fluctuations affect the observable state transitions. The model dynamics simulates a point mass moving along a convex surface in the presence of gravity, see Fig. 4. The admissible range of forces is not sufficient to drive up the car greedily from the initial state to the goal position. The observable state vector is $\mathbf{x} = [x_1, x_2]$, where x_1 and x_2 denote horizontal position and the tangential velocity of the car, respectively. The car dynamics is given by

$$\begin{aligned} dx_1 &= x_2 \cos(\operatorname{atan}(s'(x_1)))dt \\ dx_2 &= -gx_2 sgn(x_1) sin(\operatorname{atan}(s'(x_1)))dt - \beta x_2 dt + udt + \sigma d\omega \end{aligned} \tag{14}$$

where $g = 9.8\ \mathrm{m/s^2}$ is the gravitational constant and $\beta = 0.5$ is the damping coefficient. The goal for the driving agent is defined by all states such that $|x_1 - 2.5| < 0.2$ and $|x_2| < 0.5$. The cost model thus encodes the task of parking the car at the horizontal position 2.5 in minimal time and with minimal terminal velocity. Costs are accumulated from time 0 to infinity but accumulation stops when the car achieves the goal. Error tolerance is needed because the system dynamics is stochastic.

To approximate the desirability function in Eq. (13), we use a discretization of the environment using a 101-by-101 grid spanning $x_1 \in \pm 3$ and $x_2 \in \pm 9$ for the car problem. The passive dynamics is constructed by a discretization of the time axis (with time step $h = 0.05$) and defining probabilistic transitions among discrete states. The noise distribution is discretized at 9 points spanning ± 3 standard deviations in the x_2 direction and using a noise scale parameter $\sigma = 0.1$.

Once the desirability function is optimized using the passive dynamics, the control policy is derived from the computed desirability function using Eq. (12). The results are shown in Fig. 4; where (b) corresponds to the state cost function $q(\mathbf{x})$ bearing that the agent is goal-driven; (c) is the optimal cost-to-go function obtained (solid lines show stochastic trajectories resulting from the optimal behavior with different levels of variability and initial states); (d) depicts the obtained optimal control policy regarding prior beliefs about the agent behavior. In all plots blue correspond to smaller values and red to larger values.

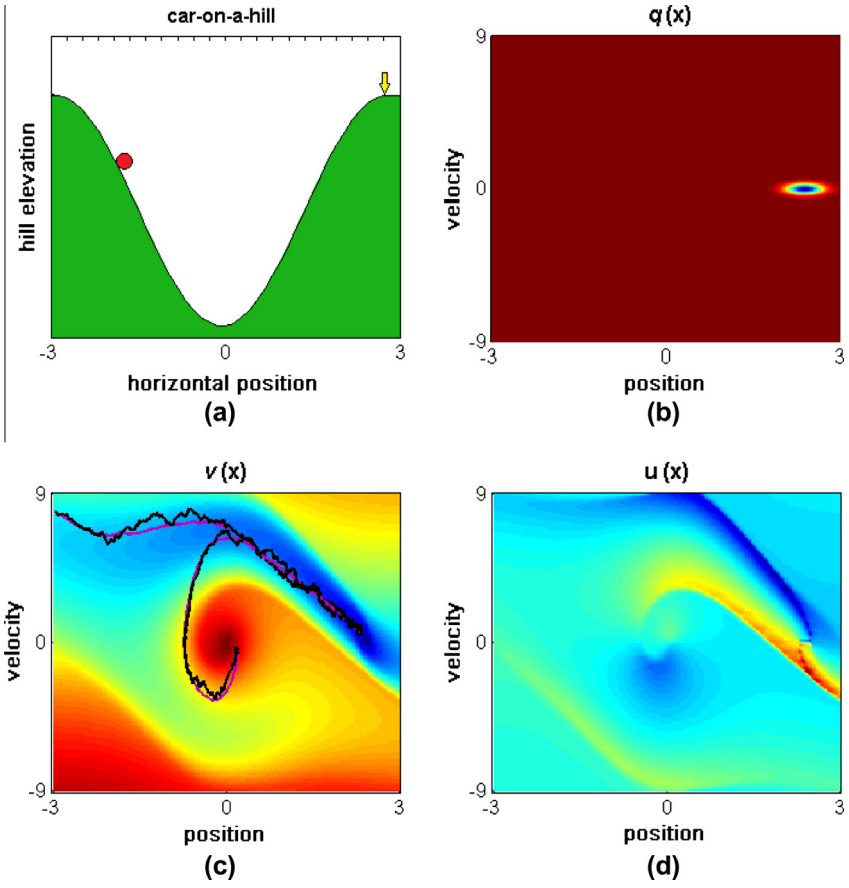

Fig. 4. (a) The car moves along a curved road in the presence of gravity. The control signal is the tangential acceleration. (b) State cost function $q(x)$ of the first exit cost formulation. (c) The optimal cost-to-go function. Solid lines show stochastic trajectories resulting from the optimal behavior with different scales of noise and initial states. (d) The obtained optimal control policy used as the specification for agent behavior monitoring.

4. Bayesian surprise

4.1. Twin Gaussian processes

Surprise quantifies how observing new data affects the internal beliefs a monitor may have about an agent behavior and its control policy. Observations that leave the prior beliefs unaffected are not surprising and -revealing that the monitor hypotheses are confirmed by data-, whereas data observations that cause the monitor to significantly revise their prior beliefs give rise to a surprising condition. Thus, a surprising agent behavior is related to a sequence of state transitions which become suspicious whenever the stochastic process describing the agent policy through environmental state transitions deviates with respect to the monitor's beliefs based on an optimal policy.

The monitor's beliefs are updated on-line as new sensory information arrives, transforming prior belief distributions into posterior ones. According to this, the fundamental effect of data D on the monitor is to change its prior distribution $P(M)$ into a posterior distribution $P(M|D)$ via the Bayes theorem. Conveniently, Bayesian surprise is measured using the distance between the posterior and prior distributions based on the Kullback–Leibler divergence $T(D,M) = KL(P(M|D)||P(M))$. The KL divergence, or relative entropy, should be understood as a measure of the difficulty of discriminating between two distributions.

Since the environmental dynamics is affected by uncertainty, a robust metric of Bayesian surprise must be considered. To quantify the similarity between the optimal behavior and a given stochastic implementation of the agent policy, twin Gaussian processes (TGP) proposed by Bo and Sminchisescu [28] are used (see Fig. 5). A TGP provides a powerful strategy for structured prediction using GP priors on both covariates and responses [29]. Hyper-parameters for both multivariate inputs and estimated outputs are obtained through minimization of the Kullback–Leibler divergence between two GPs modeled over finite input sets of training examples, emphasizing the goal that similar inputs should produce similar responses. Instead of computing surprise point-wise for each new estimation using the GPs models, two data sets containing a sequence of the last W state transitions, $\boldsymbol{X}^g = \{\Delta x_k^g\}_{k-W}^k$ and $\boldsymbol{X}^* = \{\Delta x_k^*\}_{k-W}^k$, are used to characterize the observed and the specified environmental dynamics, respectively. Then, the joint distribution of observed state differences can be modeled using a zero-mean multivariate Gaussian distribution as

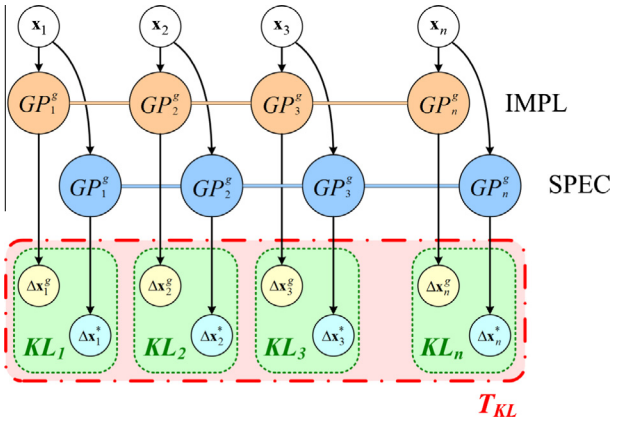

Fig. 5. The data ordinates are x and GP are the distributions for the implementation and the specification. The horizontal lines indicate fully-connected GP sets. While green connectors describe pointwise KL distance, red connectors illustrates twinned distance between the two stochastic processes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$(\boldsymbol{X}^g)^T \sim \mathcal{N}^g\left(0, \begin{bmatrix} \boldsymbol{R}_{ij} & \boldsymbol{r}_i \\ \boldsymbol{r}_i^T & r \end{bmatrix}\right) \tag{15}$$

whose covariance $\mathbf{K}^g$ is given by the kernel matrix $\boldsymbol{R}_{ij} = k(\Delta x_i^g, \Delta x_j^g)$, the kernel vector $\boldsymbol{r}_i = k(\Delta x_i^g, \Delta x_j^g)$ and the kernel value $r = k(\Delta x_k^g, \Delta x_k^g)$. The kernel used here, is the one given in Eq. (2). For the implementation modeled as $\mathcal{N}^g(\mathbf{0}, \mathbf{K}^g)$ based on a sampled sequence of state transitions, the offset or distance to the prior beliefs distribution $\mathcal{N}^*(\mathbf{0}, \mathbf{K}^*)$ is key to calculate a robust measure of surprise. This is achieved by computing the Kullback–Leibler divergence between Gaussian processes as

$$T_{KL}(\mathcal{N}^g||\mathcal{N}^*) = -\frac{N}{2} - \frac{1}{2}\log|\boldsymbol{K}^g| + \frac{1}{2}\mathrm{Tr}\{\boldsymbol{K}^g(\boldsymbol{K}^*)^{-1}\} + \frac{1}{2}\log|\boldsymbol{K}^*| \tag{16}$$

The Kullback–Leibler divergence is therefore non-negative, and zero if and only if the two multivariate Gaussian distributions have the same covariance. In the latter case, the sequence of state transitions caused by a given agent policy has no surprise regarding observed environmental transitions, i.e. they can be associated to an optimal behavior. In Fig. 5, T_{KL} is used to compute the performance of the implemented dynamics GP^g against the dynamics GP^p for the optimal agent behavior. Notice that instead of computing the KL distance point-wise for each new estimation Δx_k, T_{KL} uses a sequence of the observed state transitions $\{\Delta x_k\}_{k-W}^k$ over a rolling horizon, which gives a more robust description of any deviant behavior.

4.2. The surprise index

To quantify the impact of changes in the agent policy, it is necessary to introduce metrics aimed to describe in relative terms the resulting performance under uncertainty. Here, a ratio between the computed surprise value and the maximum value of surprise observed when the agent performs optimally is proposed. This is a good indicator of a deviant agent behavior, since meaningful peaks in the T_{KL} index produced by unexpected state transitions are many orders of magnitude larger than "background" stochastic values associated with nominal data. The surprise index (SI) is then expressed as

$$SI = \frac{T_{KL}}{T^*_{KL\ max}} \tag{17}$$

where T_{KL} is the surprise value for the observed transitions (subject to anomalies) and $T^*_{KL\ max}$ is the maximum surprise value faced when the agent behaves optimally. From Eq. (17) we can presume that any SI value bigger than the unit might be a symptom of an anomaly (or anomalies) affecting the agent behavior or its nearby environment.

A general description of the approach for on-line monitoring using Bayesian surprise is shown in the flowchart in Fig. 6. To start, the monitor observes the environmental state $\mathbf{x}_k$. Based on the current state, the posterior transition distributions for prior beliefs and the implementation are computed. The associated estimations Δx_k^g and Δx_k^* are included in their respective data sets in order to estimate T_{KL}. This is achieved through a moving window strategy which adds new estimations whereas the oldest one is removed. This allows maintaining the monitor's beliefs updated so as to fast detect changes in the agent behavior. At every moment the T_{KL} value obtained is compared against its maximum value allowed through the SI index. If at any instant the index is bigger than 1, a warning signal should be activated.

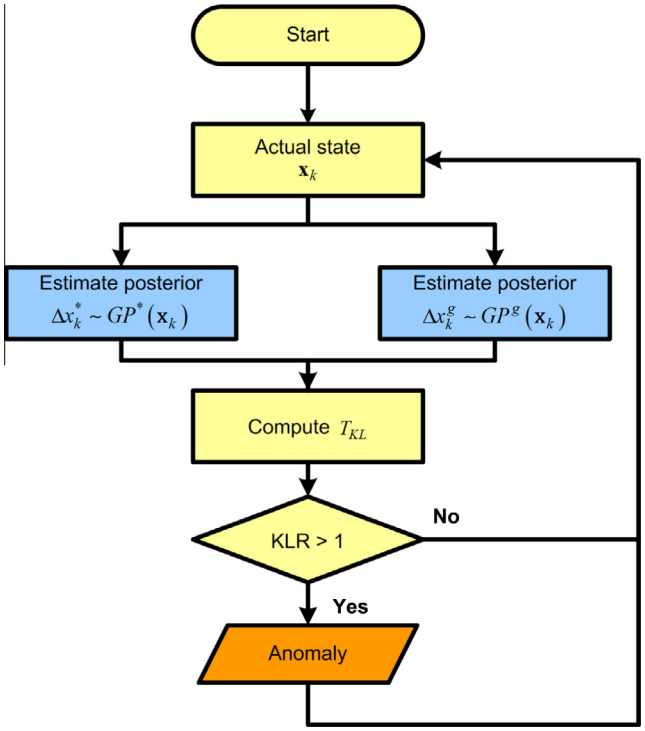

Fig. 6. Flowchart for the on-line monitoring approach. T_{KL} is used to compare the statistical distribution of a test data set against that of a reference.

4.3. Example: car-on-a-hill problem (cont'd)

To fast detect any change in the agent behavior, it is clear the importance of training the corresponding GPs using a reduced and relevant set of state transitions. Hence, the last $N_{\max} = 50$ transitions are used as the input set to train the GP model for the optimally controlled stochastic process (see Section 2.2). Bayesian surprise is later computed using a moving window strategy over the last $W = 30$ state transition estimations given by Eq. (15). The number of estimations used is a tradeoff between the speed of detection of any event or disturbance and the proper characterization of the degradation in the agent behavior, and may change depending on the dynamics of each implementation.

The *SI* index is measured by comparing the implemented dynamics GP^g to the desired one GP^* in Fig. 7. This measure emphasizes the fact that similar states should produce equivalent estimates of both covariates and responses, and so, the *SI* value should be less than 1 if no anomalies or changes in the agent policy exist. In Fig. 7a, the noise scale parameter σ is increased from 0.1 to 0.5 and 1.0, to simulate different degrees of variability. It is quite clear that an increase of the parameter σ might deteriorate the agent behavior. In Fig. 7b, a temporary failure in the car actuator is simulated. In the set of samples from #100 to #150, the car dynamics is governed by a random policy with control actions sampled from the interval $u \in [1, +1]$. This type of actuator malfunctioning which alters the driver behavior is notably captured by the proposed Bayesian surprise index.

5. Case study: artificial pancreas

5.1. Agents in medical systems

Agent based applications have definitely entered into the challenging fields of medical decision making, where faults or errors prompt serious consequences to patients. The importance of intelligent agents in this area lays on a number of properties such as autonomy in operation, communication and cooperation with other agents or systems and their perception capabilities. An agent in a medical systems interacts with the environment by executing different actions autonomously and communicating with other agents and humans, which may allow cooperative problem solving [30]. In doing so, agent based medical systems can support physicians and patients in different medical decisions and monitoring tasks. Technological devices for supporting safety–critical systems and safe patient management require of systematic methods aimed to continuously evaluate their performance and fast pinpointing any life-threaten situation. However, effective monitoring of health care situations is complex as involves interpretation of many variables and evaluation of many patient relative parameters, a great number of which cannot be measured directly.

The potential of agent based technologies can be successfully applied to aid diagnosis, treatment and prediction of many clinical problems. In this sense, recent technology breakthroughs towards a fully automated artificial pancreas (AP) give rise to the need of improved monitoring tools aimed to increase both reliability and performance of closed-loop glycemic control. Briefly, the goal of diabetes management is to mimic the basal and postprandial patterns of insulin secretion produced by normal pancreatic function [31], i.e., maintaining blood glucose (BG) levels between 80 and 140 mg/dl. Successful glucose regulation in the face of unknown changes in diet, exercise, stress, medications and most important of all diminishing the risks of hypoglycemia or hyperglycemia events is a challenging problem.

Poor predictability of BG dynamics is a key issue that both patients and doctors must deal with, mainly because the glucose-insulin dynamics shows great variability among different patients according to the carbohydrate content of meals, exercise level, age and stress. Because of this variability even the same insulin dose with the same meal routine may give rise to different blood glucose responses to insulin boluses on consecutive days. Closing the glycemic loop with a fully automated AP will certainly improve the quality of life for insulin-dependent patients. Such a device is made up of (i) a glucose measuring device, (ii) an automated insulin infusion pump, and (iii) an intelligent control algorithm, which calculates the optimal insulin bolus to be delivered based upon glycemic data and facilitates the communication between the components and other systems. However, besides the dynamics of human physiology, errors in glucose sensors and failures in insulin infusion pumps give rise to a number of challenges for implementing an artificial pancreas. The safety–critical condition of such an automated device makes its performance monitoring task of paramount importance in order to reduce glucose level variability and to minimize the risks of dangerous excursions outside the euglycemic range.

5.2. Modeling variability in glucose dynamics

In this work, glycemic variability and other sources of uncertainty in a diabetic patient are simulated using a stochastic process superimposed on an otherwise deterministic model of the glucose–insulin dynamics. To this aim, the Lehmann and Deutsch model [32] parameterized as described in Acikgoz and Diwekar [33] is used as the basis to describe the deterministic dynamics in a diabetic patient. All sources of variability are accounted for by adding an Ito's stochastic process to a deterministic model of the glucose dynamics. Introducing a stochastic process ensures a heterogeneous cohort of *in silico* subjects that accounts sufficiently well for the observed inter- and intra-subject variability, a key aspect in characterizing an optimal control policy under uncertainty. The glycemic variability in a diabetic patient is thus modeled as

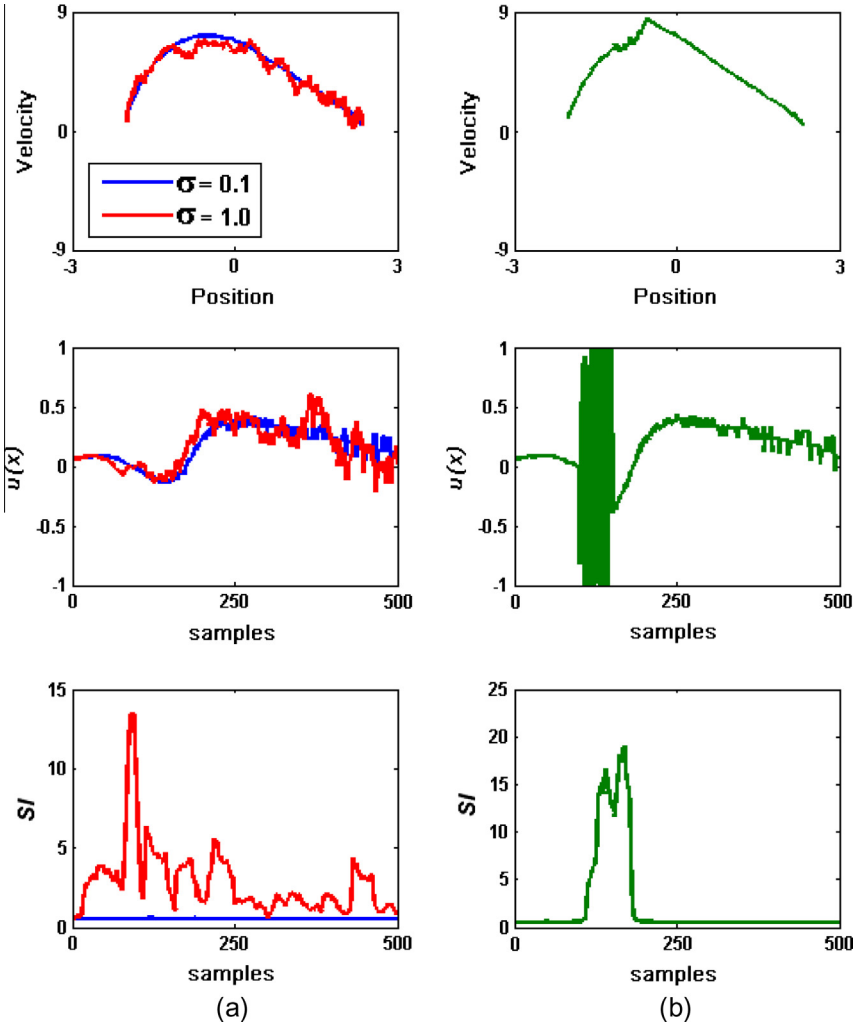

Fig. 7. (a) Sample trajectories are generated using different values of the noise scale parameter σ. The performance degradation in the agent behavior is clearly revealed below by T_{KL}. (b) Simulation outcome for random actions between samples 100–150.

$$\frac{dBG}{dt} = \frac{G_{in} + NHGB - G_{out} - G_{ren}}{V_G} + \frac{\sigma\varepsilon}{\sqrt{dt}} \tag{18}$$

where BG is the blood glucose concentration, G_{in} is the systemic appearance of glucose via glucose absorption from the gut, $NHGB$ is net hepatic glucose balance, G_{out} is the overall rate of peripheral and insulin dependent glucose utilization, G_{ren} is the renal excretion of glucose and V_G is the volume of distribution of glucose. The meal routine in Table 1 described in terms of the carbohydrate intakes is used.

Notice that the only available measurement of glucose concentration in blood is the one obtained by the subcutaneous sensor which is an estimation of the real plasmatic concentration -and corresponds to a hidden quantity to be inferred-. As the sensor needle is placed in the subcutaneous tissue, it measures interstitial fluid glucose concentration (IG) rather than the plasma glucose concentration. In the simulation model, each IG value can be estimated through integrating a BG–IG dynamics, where δ is the static gain (considered equal to 1) and τ is the time-lag constant

Table 1
Carbohydrate intake schedule.

Carbohydrate content (g)	47	16	63	31	63	31
Meal times (h)	8.00	10.00	12.30	16.00	19.30	22.00

$$dIG(k) = -\frac{1}{\tau}IG(k) + \frac{\delta}{\tau}BG(k) \tag{19}$$

Furthermore, sensor readings are corrupted by a random time-varying calibration error $\xi(k)$ and a white Gaussian noise process $v(k)$ so that

$$\begin{aligned} S(k) &= (1 + \xi(k))IG(k) + v(k) \\ \xi(k+1) &= 3\xi(k) - 3\xi(k-1) + \xi(k-2) + w(k) \end{aligned} \tag{20}$$

where $S(k)$ is the glycemic sensor output and the calibration error $\xi(k)$ has been created using a triple integrator of a zero mean white noise $w(k)$. For simplicity, BG is used instead of IG hereafter but recall it refers to the outcome of a continuous glucose sensor.

5.3. Glycemic control

The specification of the expected behavior when monitoring an artificial pancreas is here built upon the state transition distribution for an optimally controlled glucose dynamics. To obtain the optimal control policy using the LSMDP technique, it is required in advance a model of the passive (uncontrolled) dynamics of the glucose regulatory system. Even if different methods exist to model the passive dynamics according the each problem, most of them require of previous knowledge of the transition matrix given in Eq. (13) and immediate costs over the entire state-space [34,35].

However, since the passive dynamics describes nothing but the space of likely state transitions, it would be a valid to use an approximation of the matrix. This approximation can be obtained by means of a reduced model of the glycemic system. In this sense, the Lehmann and Deutsch model above described corresponds to an augmented representation of the two-compartments model described in Bergman et al. [36]

$$\begin{aligned}\frac{dBG}{dt} &= (p_1 - Ia)BG + p_1 G_b \\ \frac{dIa}{dt} &= p_2 Ia + p_3 I(t)\end{aligned} \tag{21}$$

where Ia represents the time course of insulin infusion and G_b is the basal glucose level. In this minimal model of glucose kinetics, plasma insulin $I(t)$ enters a remote compartment where it acts by accelerating the glucose disappearance into the periphery and liver, and inhibiting hepatic glucose production. While some of the parameters p_i given in Table 2, describe important physiological responses, as the insulin sensitivity $S_I = -(p_3/p_2)$, others lack a meaningful interpretation. Beyond any limitation of this minimal model, it is still useful to describe the glucose–insulin state transitions in the passive dynamics.

The control task corresponds to maintaining acceptable glycemic variability. For this purpose, the cost function is conveniently designed in such a way that allows guaranteeing an acceptable behavior of blood glucose dynamics within a target band (80–140 mg/dl). Thereby, the state cost function $q(\mathbf{x})$ is represented by a square exponential form which saturates for great deviations from a chosen reference blood glucose value (BG = 110 mg/dl) and basal insulin (I_b = 30 mU/l). In this way, the control system will aim to keep the patient glucose level as close as possible to the reference. Note that the glycemic control problem has a two-dimensional state $\mathbf{x} = [BG, Ia]$ with only one control input corresponding to the insulin bolus, i.e. $u = I(t)$. Consequently, Eq. (13) can be now solved by an iteration method in exponential form. The approximation uses a state space discretization with a 151-by-151 grid spanning $BG \in [0,220]$ mg/dl and $Ia \in [0,60]$ mU/l. The passive dynamics is constructed by a discretization of the time axis (with time step h=0.05) and defining probabilistic transitions among discrete states so that the mean and variance of the continuous state dynamic are preserved. The noise distribution is discretized at 9 points spanning ±3 standard deviations and using a noise scale parameter σ = 0.1. Thus, the real system dynamics has to be inferred by the controller. The desirability function was computed using the estimated passive dynamics matrix whereas the control policy was subsequently derived from the obtained expression. Results are displayed in Fig. 8, where (a) depicts a scheme of the Bergman's minimal model and (b) is the state cost function $q(\mathbf{x})$ -in all plots blue corresponds to small values and red to high values-; (c) is the optimal cost-to-go function obtained -the two small paths correspond to stochastic trajectories generated by the optimal policy-. More specifically, the red one was obtained using a noise scale σ = 0.1 whereas the black with σ = 0.25. Higher insulin levels are desirable (lower costs) when BG is high, whereas smaller insulin levels are needed when BG is lower. In (d) the obtained optimal control policy for agent behavior monitoring is shown.

Table 2
Parameters of the Bergman's minimal model.

Parameter	Value
p_1	2.96×10^{-2} [min^{-1}]
p_2	1.86×10^{-2} [min^{-1}]
p_3	6.51×10^{-6} [min^{-2}/μU/ml]
G_b	97 [mg/dl]

5.4. On-line monitoring

The controlled dynamics of the implemented AP is represented using a suboptimal insulin policy, assuming a certain degree of glycemic variability σ and multiple meal consumption as in Table 1. The desired dynamics -the one controlled by the optimal insulin policy- is parameterized with: variability parameter $\sigma = 0.10$; calibration error ξ = 2% and time-lag τ = 5 min. In this way, a typical level of uncertainty in the system is considered. A set of $N_{\max}$ = 50 training examples are selected to set up the sequence $\mathcal{D}$ of training examples, while W = 30 observed state transitions are used to compute the surprise with T_{KL}. At the 12th hour, sensor parameters are varied to simulate performance degradation in the AP components, whereas σ is increased to augment patient glycemic variability. To sum up, this allows us to evaluate the clinical impact of real-life glycemic control that is affected by sensor errors as well as time-lags compounded with the effect of patient variability in diabetes management. In Fig. 9, a realization (obtained by applying the optimal control policy for a scale of variability σ = 0.10) of the glucose stochastic process is shown. From the 12th hour onwards, glycemic variability is increased by changing the corresponding parameter to σ = 0.50. The predicted state transitions distributions are presented in the lower part of Fig. 9; shaded areas describe the uncertainty in predictions whereas solid lines correspond to prediction means. An advantage of using GPs is that they also provide information about confidence intervals for each prediction. It is worth noting that increasing variability not only affects the predicted means but also the prediction errors, which reveals a significant degradation in the behavior of the AP controller.

5.4.1. Glycemic variability

In Fig. 10, the ability of the optimal control policy to mitigate excessive glycemic variability is evaluated. The scale of the noise parameter of the glucose-insulin model is increased from σ = 0.25 to σ = 0.50 from the 12 h onwards, giving rise to a larger glycemic variability. It is noticeable how the performance of the closed-loop quickly degrades. The computed T_{KL} metric is irregular at the beginning while the training set $\mathcal{D}$ is still capturing enough information to properly describe the current system behavior. Moreover, stochastic behavior in the glycemic model gives rise to T_{KL} values that are not strictly equal to 0, even if the observed and expected behaviors include the same degree of variability over the interval [0,12] hours. It is worth noting that performance degradation is quickly revealed when parameters are varied from the 12 h onwards in the simulation model of a diabetic patient.

5.4.2. Sensor miscalibration

Any failure resulting in an sudden change in glycemic levels should be followed by a similar increment in the surprise index. Glucose sensor readings affected by miscalibration due ξ = 10% are depicted in Fig. 11 for optimal control. Since the control algorithm responds to glycemic states acquired by sensory devices susceptible to noise and disturbances, the loop performance degrades if a failure in the device occurs. Similarly, a stuck fault in the sensor starting from the 12 h and onwards it is also considered. Because of this, the sensor outcome freezes at the 120 mg/dl reading giving rise to a significant deviation from the expected performance, even when BG levels is maintained in the desired euglycemia range. However, since the performance of the AP implementation is contrasted with its specified behavior this anomaly is fast detected using the T_{KL} metric.

5.4.3. Insulin pump failures

To end the analysis for this case study, Fig. 12 simulates a likely scenario in which a catheter blockage occurs during the use of a

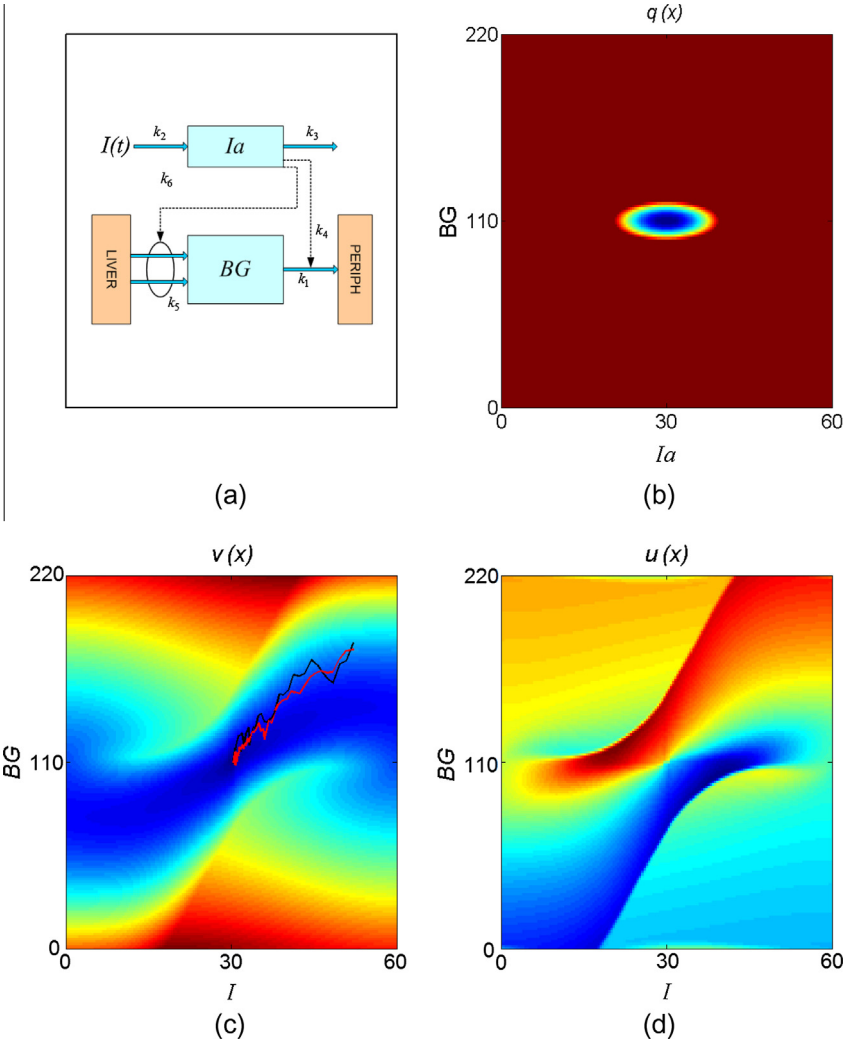

Fig. 8. (a) Reduced model of the glucose-insulin dynamics. (b) State cost function $q(x)$. (c) The optimal cost-to-go function. Sample trajectories are generated using different values of the noise scale parameter σ. (d) The optimal control policy used as the specification for agent behavior monitoring.

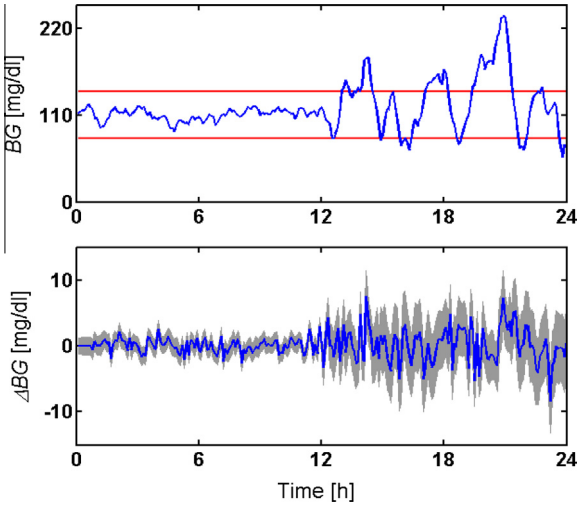

Fig. 9. Increase in glycemic variability using stochastic optimal control. Mean and standard deviation values are generated using Ito's parameter σ.

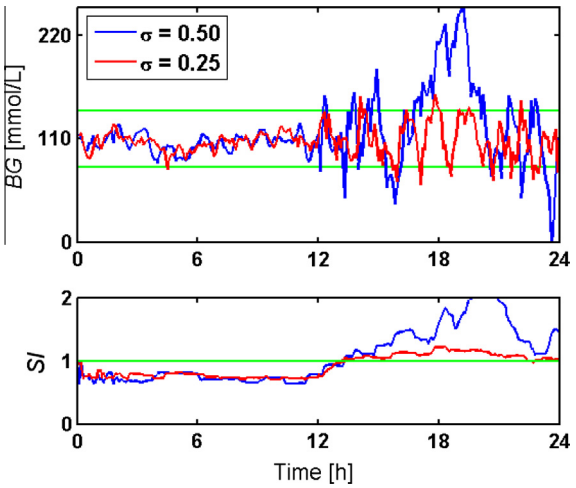

Fig. 10. Assessment of the effect of high glycemic variability using parameter $\sigma = 0.25$ and 0.50.

continuous infusion pump. As a consequence, the insulin dose is considerably reduced such that only 80% of the level prescribed by the optimal control algorithm is in fact administrated. This mitigates the effect of the insulin bolus and leads to a poorly controlled glycemic variability. In a different, a chaotic controller is simulated by considering $u(t) = u^* + \beta dw$, where u^* corresponds

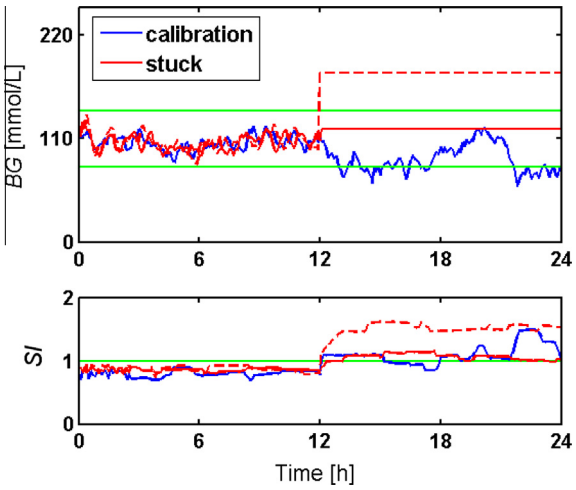

Fig. 11. Performance loss due to glucose sensor miscalibration using $\xi = 10\%$ and a stuck sensor in two different BG levels.

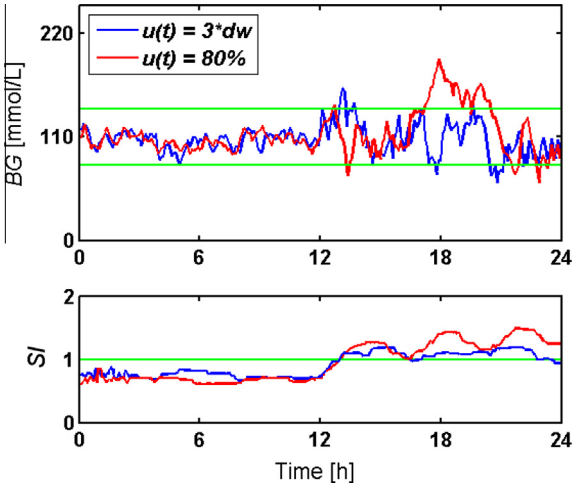

Fig. 12. Performance degradation due to a catheter blockage and a chaotic controller.

to the value given by the optimal policy, dw is the differential of a Brownian random noise and $\beta = 3$ is the standard deviation of the added noise. As a result, even if the control algorithm can still compute the optimal action, it is unable to properly adjust the amount of insulin administrated to the patient. Since a zero mean noise is used, the path still converges to the vicinity of the target level $BG = 110$ mg/dl, despite the significant increase in the variability of the glucose dynamics.

6. Final discussion

This work presents a novel probabilistic approach built upon an optimally controlled stochastic system for on-line monitoring of an agent behavior under uncertainty, which has been conceived in terms of active inference and optimal action selection. Checking if an autonomous agent behavior fulfills expectations is a key issue to guarantee safety and performance of an increasing number of autonomous agent applications such as driverless cars, drones and biomedical systems. The main problem for behavior monitoring is generating prior beliefs under the uncertainty the agent should face in its own environment. In this work, the desired behavior is modeled by a prior Gaussian distribution for state transitions, in order to verify if a given agent control policy respects its specification. The desired optimal behavior is obtained analytically using a class of Markov decision processes which are linearly solvable. Through an exponential transformation, the Bellman equation for such problems can be made linear, despite nonlinearity in the stochastic dynamical models, which facilitates applying efficient numerical methods.

The availability of an optimal control policy allows simulating the desired behavior over time and comparing it with the current system performance in order to identify deviations from the desired behavior. To favor on-line monitoring, a robust metric based on surprise and twin Gaussian processes is introduced to characterize the progressive degradation in the agent behavior by quantifying the distance between the implementation and prior beliefs. A distinctive advantage of computing surprise using Gaussian processes is that the divergence from prior beliefs can be estimated not only using the expected value of state transitions but also the corresponding prediction uncertainty for optimal action selection.

The proposed active inference approach to on-line monitoring of an agent behavior allows its incorporation to a number of applications in diverse domains. Typical examples include the detection of unauthorized access to computer systems [37], irregularities in vital signs and other variables in intensive care patients [38], fraud in financial services [39], detection of saccadic objects for visual applications [40] and detection of path deviation in autonomous vehicles [41].

Future work, aims to extend the approach to multi-agent system monitoring by characterizing the desired collective behavior using a game-theoretic perspective. In turn, despite the efficiency of the LSMDP scheme, it is yet necessary to previously know the passive dynamics of the system, which probably is the most crucial and complex task in applying the proposed methodology. Instead of using a simplistic model of the system under study, current research work aims to include the estimation of the passive dynamics while finding the optimal cost-to-go, by using a temporal difference algorithm called *z-learning*.

References

[1] L. Avila, E. Martínez, Behavior monitoring under uncertainty using Bayesian surprise and optimal action selection, Expert Syst. Appl. 41 (2014) 6327–6345.
[2] M.A. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249.
[3] X. Ding, Y. Li, A. Belatreche, L.P. Maguire, An experimental evaluation of novelty detection methods, Neurocomputing 135 (2014) 313–327.
[4] Z. Sun, S.J. Qin, A. Singhal, L. Megan, Performance monitoring of model-predictive controllers via model residual assessment, J. Process Control. 23 (2013) 473–482.
[5] R. Kromanis, P. Kripakaran, Support vector regression for anomaly detection from measurement histories, Adv. Eng. Inform. 27 (2013) 486–495.
[6] L. Clifton, D.A. Clifton, P.J. Watkinson, L. Tarassenko, Identification of patient deterioration in vital-sign data using one-class support vector machines, in: Comput. Sci. Inf. Syst. FedCSIS 2011 Fed. Conf. On, IEEE, 2011, pp. 125–131.
[7] A. Ranganathan, F. Dellaert, Bayesian surprise and landmark detection, in: Robot. Autom. 2009 ICRA09 IEEE Int. Conf. On, IEEE, 2009, pp. 2017–2023.
[8] D.-M. Lee, Method for detecting lane deviation of vehicle, U.S. Patent No. 6,317,057, U.S. Patent and Trademark Office., Washington D.C., (2001).
[9] T.J. Harris, Assessment of control loop performance, Can. J. Chem. Eng. 67 (1989) 856–861.
[10] K.J. Åström, Introduction to Stochastic Control Theory, Academic Press, 1970.
[11] C.L. Owens, F.J. Doyle III, Performance monitoring of diabetic patient systems, in: Eng. Med. Biol. Soc. 2001 Proc. 23rd Annu. Int. Conf. IEEE, IEEE, 2001, pp. 2047–2050.
[12] C.A. Harrison, S.J. Qin, Minimum variance performance map for constrained model predictive control, J. Process Control 19 (2009) 1199–1204.
[13] B. Srinivasan, T. Spinner, R. Rengaswamy, Control loop performance assessment using detrended fluctuation analysis (DFA), Automatica 48 (2012) 1359–1363.
[14] B. Huang, Bayesian methods for control loop monitoring and diagnosis, J. Process Control 18 (2008) 829–838.
[15] M. Qing-wei, Z. Zhen-fang, L. Ji-zhen, A practical approach of online control performance monitoring, Chemom. Intell. Lab. Syst. 142 (2015) 107–116.

[16] Jeff Hawkings, The Science of Anomaly Detection, 2014. <http://numenta.com/blog/science-of-anomaly-detection.html>.
[17] K. Friston, Active inference and agency, Cogn. Neurosci. 5 (2014) 119–121.
[18] E. Todorov, Eigenfunction approximation methods for linearly-solvable optimal control problems, in: Adapt. Dyn. Program. Reinf. Learn. ADPRL09 IEEE Symp. On, IEEE, 2009, pp. 161–168.
[19] J. Hoey, P. Poupart, A. von Bertoldi, T. Craig, C. Boutilier, A. Mihailidis, Automated handwashing assistance for persons with dementia using video and a partially observable markov decision process, Comput. Vis. Image Underst. 114 (2010) 503–519.
[20] A. Broggi, P. Medici, P. Zani, A. Coati, M. Panciroli, Autonomous vehicles control in the VisLab intercontinental autonomous challenge, Annu. Rev. Control 36 (2012) 161–171.
[21] W. Weaver, Probability, rarity, interest, and surprise, Pediatrics 38 (1966) 667–670.
[22] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[23] E. Todorov, Efficient computation of optimal actions, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 11478–11483.
[24] D.P. Bertsekas, Dynamic Programming and Optimal Control, 2nd ed., vol. I, Athena Scientific, 2000.
[25] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.
[26] K. Ito, Stochastic differentials, Appl. Math. Optim. 1 (1975) 374–381.
[27] K. Dvijotham, E. Todorov, Linearly solvable optimal control, Reinf. Learn. Approx. Dyn. Program. Feedback Control (2012) 119–141.
[28] L. Bo, C. Sminchisescu, Twin gaussian processes for structured prediction, Int. J. Comput. Vis. 87 (2010) 28–52.
[29] A. Naish-Guzman, S.B. Holden, Robust regression with twinned gaussian processes, in: Advances in Neural Information Processing Systems. 2008, pp. 1065–1072.
[30] B.L. Iantovics, Agent-based medical diagnosis systems, Comput. Inform. 27 (2012) 593–625.
[31] M.E. Alexander, R. Mathur, S.M. Moghadas, P.N. Shivakumar, Modelling the effect of CSII on the control of glucose concentration in type 1 diabetes, Appl. Math. Comput. 187 (2007) 1476–1483.
[32] E.D. Lehmann, T. Deutsch, A physiological model of glucose-insulin interaction in type 1 diabetes mellitus, J. Biomed. Eng. 14 (1992) 235–242.
[33] U. Acikgoz, U.M. Diwekar, Blood glucose regulation with stochastic optimal control for insulin-dependent diabetic patients, Chem. Eng. Sci. 65 (2010) 1227–1236.
[34] M. Burdelis, K. Ikeda, Estimating passive dynamics distributions and state costs in linearly solvable markov decision processes during Z learning execution, SICE J. Control Meas. Syst. Integr. 7 (2014) 48–54.
[35] A. Li, P.R. Schrater, Efficient learning in linearly solvable MDP models, in: Proc. Twenty-Third Int. Jt. Conf. Artif. Intell., AAAI Press, Beijing, China, 2013, pp. 248–253.
[36] R.N. Bergman, L.S. Phillips, C. Cobelli, Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose, J. Clin. Invest. 68 (1981) 1456–1467.
[37] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, E. Vázquez, Anomaly-based network intrusion detection: techniques, systems and challenges, Comput. Secur. 28 (2009) 18–28.
[38] D. Li, X. Li, Z. Liang, L.J. Voss, J.W. Sleigh, Multiscale permutation entropy analysis of EEG recordings during sevoflurane anesthesia, J. Neural Eng. 7 (2010) 046010.
[39] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decis. Support Syst. 50 (2011) 559–569.
[40] E. Vig, M. Dorr, T. Martinetz, E. Barth, Intrinsic dimensionality predicts the saliency of natural dynamic scenes, Pattern Anal. Mach. Intell. IEEE Trans. On 34 (2012) 1080–1091.
[41] A.B. Curtis, Path planning for unmanned air and ground vehicles in urban environments, Doctoral dissertation, Brigham Young University, 2008.