Agent Behavior Monitoring using Optimal Action Selection and Twin Gaussian Processes

Luis Avila, Ernesto Martinez

INGAR (CONICET-UTN), Avellaneda 3657, Santa Fe S3002 GJC, Argentina
{avilaol, ecmarti}@santafe-conicet.gob.ar

Abstract. The increasing trend towards delegating complex tasks to autonomous artificial agents in safety-critical socio-technical systems makes agent behavior monitoring of paramount importance. In this work, a probabilistic approach for on-line monitoring using optimal action selection and twin Gaussian processes (TGP) is proposed. A Kullback-Leibler (*KL*) based metric is proposed to characterize the deviation of an agent behavior (modeled as a controlled stochastic process) to its specification. The optimal behavior specification is obtained using Linearly Solvable Markov Decision Processes (LSMDP) whereby the Bellman equation is made linear through an exponential transformation such that the optimal control policy is obtained in an explicit form.

Keywords: Agent monitoring, Gaussian processes, optimal action selection.

1 Introduction

Safety-critical systems increasingly integrate autonomous software agents in an increasing number of applications which requires new monitoring tools. As an example, consider the case of collision avoidance in driving systems [1] where the monitoring task involves a number of autonomous vehicles interacting with each other in a high-speed highway. Any monitoring system aimed to warn or prevent collisions and dangerous circumstances must contemplate the expected behavior of nearby cars to detect quickly a collision scenario. Also, automating the task of gazing in video surveillance to find suspicious behaviors [2] highlights the importance of agent behavior monitoring. Detecting dangerous objects and intruders is essential for safety in crowded environments, but monitoring human behaviors and reporting about potential threats is a complex task to be automated.

The novelty and relevance of information contained in a data stream, can be correlated with the effect such data has on the observer (monitor) [3]. The amount of information can be measured in a natural way by the Kullback-Leibler (KL) distance -or *relative entropy*- between the prior and posterior distributions in the monitor beliefs, i.e. regarding the available space of hypotheses about the state of a controlled system. In this work, the novelty of information in a data stream regarding deviations from the specified agent behavior is measured using twin Gaussian Processes [4]. For behavior

specification, optimal choice of actions under uncertainty is used here to characterize the desired behavior of an intelligent agent.

2 Optimal Action Selection

To characterize the expected behavior of an agent, the novel approach of LSMDP is used [5]. This is 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. Consider an agent that by perceiving an environmental state $\mathbf{x} \in \mathbb{R}^{n_x}$, chooses the scalar action $u \in \mathbb{R}^{1_u}$, which causes the system to evolve to the state $\mathbf{x}_{k+1}$ and receives an immediate cost $\ell(\mathbf{x},u)$. The state transition function obey to a controlled Ito's diffusion process of the form

$$d\mathbf{x} = \mathrm{a}(\mathbf{x})dt + B(\mathbf{x})(udt + \sigma d\omega) \tag{1}$$

where $\omega \in \mathbb{R}^{1_u}$ (same space as actions) and σ denote Brownian noise and its scaling parameter, respectively. The term $\mathrm{a}(\mathbf{x})$ describes the so called passive dynamics and $B(\mathbf{x})$ is the input-gain matrix. The passive dynamics represents the behavior of the stochastic dynamics in absence of control actions, it is defined as a diffusion process in continuous domains [5]. To put it in a more convenient form, the *h*-step transition probability for the uncontrolled dynamics is expressed as

$$p^0(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \mathcal{N}\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k + h\mathrm{a}(\mathbf{x}) + hB(\mathbf{x}), h\sigma B(\mathbf{x})^T B(\mathbf{x})\right) \tag{2}$$

In turn, the controlled diffusion is approximated as a deterministic function expressed as a Gaussian distribution with mean and covariance given as

$$p^{\mathbf{u}_k}(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = \mathcal{N}\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k + h\left(\mathrm{a}(\mathbf{x}) + hB(\mathbf{x})u\right), h\sigma B(\mathbf{x})^T B(\mathbf{x})\right) \tag{3}$$

The controller shifts the probability mass from one region of the state space to another by acting on the system dynamics. A control policy $\pi(\mathbf{x})$ is thus defined as a probability of selecting the action u_k at state $\mathbf{x}_k$. For any optimal control application, the main objective is to find an optimal control policy $\pi^*(\mathbf{x})$ which minimizes the expected cumulative cost function $v(\mathbf{x})$ as

$$v^*(\mathbf{x}) = \min_u \left\{ \ell(\mathbf{x}, \pi(\mathbf{x})) + \mathbf{E}_{\mathbf{x}' \sim p^u(\cdot|\mathbf{x})}\left[v(\mathbf{x}')\right] \right\} \tag{4}$$

where $\mathbf{x}'$ denotes the next state for a given action u. The minimum cumulative cost for starting at state $\mathbf{x}$ and acting optimally thereafter enables greedy computation of optimal actions. Eq. (4) is fundamental to optimal control theory and is called the Bellman fundamental equation. The Bellman equation can be simplified by assuming the immediate cost function is given as

$$\ell(\mathbf{x},u) = hq(\mathbf{x}) + KL\left(p^u(\mathbf{x}'\mid\mathbf{x}) \,\|\, p^0(\mathbf{x}'\mid\mathbf{x})\right) \tag{5}$$

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 from the optimally-controlled dynamics to the passive one. The distance can be understood as the price to pay for the optimal shift of the passive dynamics by action u. The *KL* divergence between the above Gaussians can be proven to be $h/2\sigma^2 u^2$ which is the quadratic energy cost accumulated over interval h. Defining a desirability function as

$$z(\mathbf{x}) = \exp\left(-v^*(\mathbf{x})\right) \tag{6}$$

the Bellman equation can be now expressed linearly as

$$z(\mathbf{x}) = \exp\left(-hq(\mathbf{x})\right)\mathcal{G}[z](\mathbf{x}) \tag{7}$$

where $\mathcal{G}[z](\mathbf{x})$ is an integral operator given by

$$\mathcal{G}[f](\mathbf{x}) = \int p^0(\mathbf{x}'\mid\mathbf{x})\, f(\mathbf{x}')d\mathbf{x}' \tag{8}$$

The optimal control policy is computed analytically and expressed as

$$u^*(\mathbf{x}) = -\sigma^2 B(\mathbf{x})^{\mathrm{T}} v_{\mathbf{x}}(\mathbf{x}) \tag{9}$$

Case study: Car-on-a-hill

The car-on-a-hill continuous control problem illustrated in Fig. 1 and details follows. The admissible range of forces is not sufficient to drive up the car greedily from the initial state. The 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 dynamics is given by

$$\begin{aligned} dx_1 &= x_2 \cos\left(\operatorname{atan}\left(s'(x_1)\right)\right)dt \\ dx_2 &= -g\, x_2 \operatorname{sgn}(x_1) \sin\left(\operatorname{atan}\left(s'(x_1)\right)\right)dt - \beta x_2 dt + u dt + d\omega \end{aligned} \tag{10}$$

where $s'(x_1) = 2x_1 \exp\left(-x_1^2/2\right)$ is the slope over the horizontal plane, g = 9.8 is the gravitational constant and β = 0.5 is the damping coefficient. The goal states for the driving agent are 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 control energy. This is a first-exit setting, since costs are accumulated from time 0 to infinity but accumulation stops when the system reaches a terminal state. Error tolerance is needed because the dynamics is stochastic.

As the continuous problem is in the form given in Eq. (1) it can be approximated with a LSMDP. The approximation is constructed 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. (2). Defining the vector $\mathbf{z}$ with elements $z(\mathbf{x}_n)$ and the matrix Q with elements $\exp(-hq(\mathbf{x}_n))$ on its main diagonal Eq. (7) becomes

$$\lambda \mathbf{z} = QP\mathbf{z} \tag{11}$$

where λ is an eigenvalue; Eq. (11) is solved by an iteration method in exponentiated form. The approximation uses a state space discretization: a 101-by-101 grid spanning $x_1 \in [-3, +3]$, $x_2 \in [-9, +9]$. The passive dynamics is constructed by discretizing 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 in the x_2 direction and using a noise scale parameter σ=0.1. This parameterization corresponds to the agent dynamics used as the specification, i.e. the manner the agent is expected to behave when it is optimally controlled. Once the desirability function was optimized using the estimated passive dynamics, the control policy was subsequently derived from the obtained desirability function. The results are shown in Fig. 1, where b) corresponds to the state cost function $q(\mathbf{x})$ of the first exit cost formulation; c) is the optimal cost-to-go function obtained. Solid lines show stochastic trajectories resulting from the optimal behavior with different scales of noise and initial states; d) is the obtained optimal control policy used as the specification for agent behavior monitoring. In all plots blue correspond to smaller values and red to larger values.

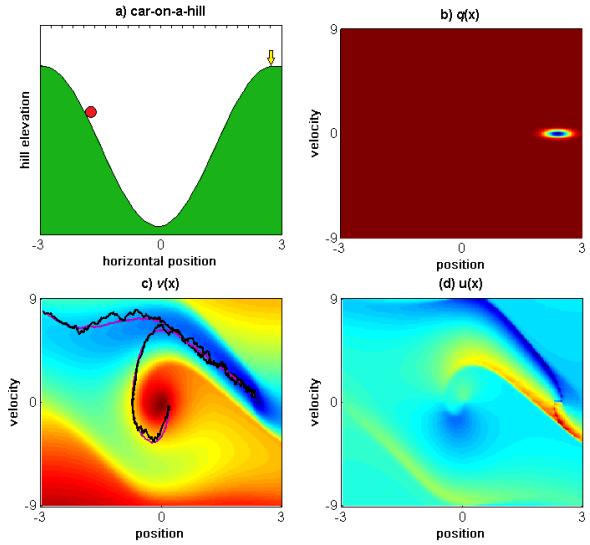

Fig. 1. The car moves along a curved road in the presence of gravity.

3 Behavior Monitoring

3.1 Learning Transition Probabilities

A Gaussian process (GP) is a collection of random variables, any finite number of which has a joint Gaussian distribution. For GP regression (GPR), random variables represent the value of the function $f(\mathbf{x})$ for inputs $\mathbf{x}$. GPR assumes $f(\mathbf{x})$ is a zero mean stationary GP with covariance function $\mathrm{k}(\mathbf{x}_i,\mathbf{x}_j)$, encoding correlations between pairs of random variables

$$\mathrm{k}(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\gamma_r \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right) + \lambda_r \delta_{ij} \tag{12}$$

with $\gamma_r \geq 0$ the kernel width parameter, $\lambda_r \geq 0$ 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.

Short-term transition dynamics are modeled based on interactions between the agent and its environment [6]. Given a state vector $\mathbf{x}$ a separate GP model is trained for each state dimension x, in such a way the effect of uncertainty about its change due to a control action is modeled statistically as

$$\Delta x_k \sim GP(m, \mathrm{k}) \tag{13}$$

where m is the mean function and k is the covariance function. The training inputs to the Gaussian model GP are the states, whereas the targets are the differences between the successor state and the state in which the action is applied. For an input $\mathbf{x}_k$, the predictive distribution $p(\mathbf{x}_{k+1} \mid \mathbf{x}_k)$ is Gaussian distributed. We can build the optimal transition probability $p^*(\mathbf{x}_{k+1} \mid \mathbf{x}_k)$ as a GP^* model which describes the stochastic specification of the agent behavior. The superscript indicates the transition probability is shifted by the optimal control policy $\pi^*(\mathbf{x})$. On the other hand, GP^g describes the observed agent behavior modeled as the transition probability $p^g(\mathbf{x}_{k+1} \mid \mathbf{x}_k)$ which may deviate from optimal action selection. Noteworthy, the GP^g model that characterize the current system functioning requires to be updated on-line to accommodate the arriving data stream.

3.2 Twin Gaussian Processes

Agent behavior monitoring must quantify how observing new data affects the internal beliefs that the agent may have over a set of hypotheses or models $\mathfrak{M}$ of the world. Since the agent acts over an uncertain environment, the monitoring approach should be probabilistic using distributions to capture subjective expectations or beliefs over the current space. Agent's beliefs must be updated on-line, as data is acquired, transforming prior belief distributions $P(M)$ into posterior ones $P(M \mid D)$ and computing the distance between them, which is best done using the *KL* divergence

$$KL\big(P(M|D)\,|\,P(M)\big)=\int_{\mathfrak{M}}P(M|D)\log\frac{P(M|D)}{P(M)}d\mathfrak{M} \tag{14}$$

To provide a sensitive on-line estimation of the divergence from optimal behavior with small samples TGP are used. TGP can be described as a structured prediction method that employs GPs to estimate outputs, by minimizing the *KL* distance between a given system implementation and its specification. Both processes are modeled as normal distributions over a finite set of *L* training examples [4]. Through TGP, we obtain a powerful monitoring tool by comparing the two GPs.

Suppose a particular set containing a sequence of the last *W* state differences $\boldsymbol{X}^g=\{\Delta x_k^g\}_{k-W}^k$ for the stochastic process that results from applying an arbitrary control policy π^g. Then, the joint distribution of observed state differences can be modeled using a zero mean multivariate Gaussian distribution as

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

whose covariance $\boldsymbol{K}^g$ is given by the kernel matrix $\boldsymbol{R}_{ij}=\mathrm{k}\left(\Delta x_i^g,\Delta x_j^g\right)$, the kernel vector $\boldsymbol{r}_i=\mathrm{k}\left(\Delta x_i^g,\Delta x_k^g\right)$ and the kernel value $r=\mathrm{k}\left(\Delta x_k^g,\Delta x_k^g\right)$. For the given Gaussian dynamics $\mathcal{N}^g(\mathbf{0},\boldsymbol{K}^g)$ of a sampled sequence of state transitions, the offset or distance to a specified distribution $\mathcal{N}^*(\mathbf{0},\boldsymbol{K}^*)$ is key to calculate a robust measure of the twinned Kullback-Leibler divergence (T_{KL}). Considering the *KL* divergence as a measure of the alignment between two kernels, the divergence between Gaussian (stochastic) processes is defined as

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

The *KL* divergence is non-negative and zero if and only if the two multivariate Gaussian distributions have the same covariance. In Fig. 2, T_{KL} is used to compute the performance of the controlled dynamics GP^g against the specified dynamics GP^p for the optimal agent behavior. Notice that instead of computing the *KL* distance

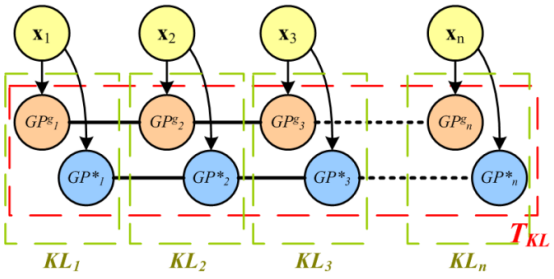

Fig. 2. The data ordinates are **x** and *GP* are the GP distributions for the implemented and the specification. The horizontal black lines indicate fully-connected sets.

pointwise (green rectangles) for each new estimation Δx_k , TGP uses a sequence of observed state transitions $\{\Delta x_k\}_{k-W}^{k}$ over a finite horizon (red rectangle), which gives a more robust description of a deviant behavior.

Case study (Cont'd).

It is clear the importance of describing any change in the system dynamics using a reduced and relevant data set for modeling the corresponding GPs. Hence, a data set containing the last L=50 state-transition pairs is used to train the GP in order to capture any deviant behavior from the specification. T_{KL} is computed using a moving window strategy over the last W=30 estimations of state transitions. 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. T_{KL} is measured for the implemented dynamics GP^g to its specification GP^* in Fig. 3. This measure emphasizes the fact that similar states should produce similar estimates of

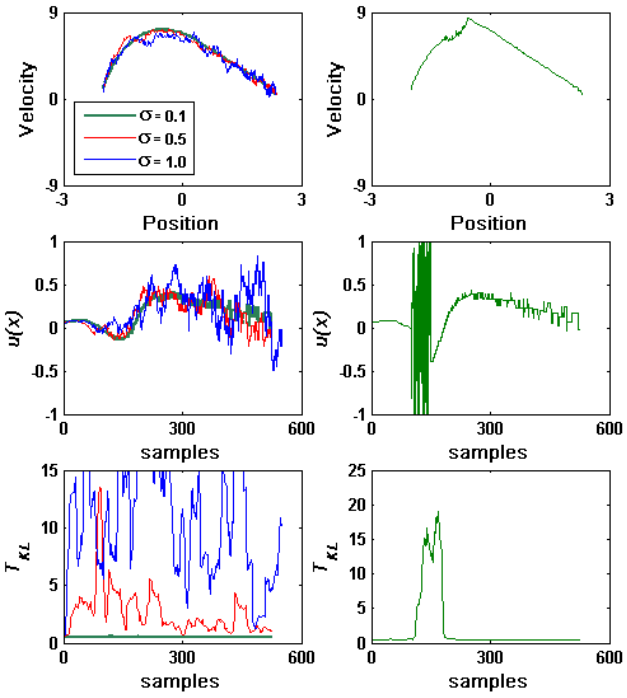

Fig. 3. a) Sample trajectories are generated using different values of the noise scale parameter σ. The performance degradation in the agent behavior is clearly revealed below by T_{KL} . b) Simulation outcome for random actions between samples 100 to 150.

both covariates and responses. In a), 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 parameter σ certainly change the agent behavior. In b), a failure in the car actuator is simulated. During samples 100-150 the car dynamics is governed by a random policy with $u \in [-1, +1]$.

Discussion

This paper presents a probabilistic approach for on-line to monitoring of an agent behavior under uncertainty based on Bayesian surprise and optimal action selection. The desired behavior is modeled by a prior Gaussian distribution for state transitions, in order to assess if an observed control policy respects its specification. An analytical specification of the desired optimal behavior is obtained using a class of Markov decision processes which are linearly solvable. Twin Gaussian processes are used to compare on-line observe state transitions due to the actual agent behavior with the stochastic specification.

References

1. Broggi, A., Medici, P., Zani, P., Coati, A., Panciroli, M.: Autonomous vehicles control in the VisLab intercontinental autonomous challenge. Annu. Rev. Control. 36, 161–171 (2012).
2. Fernández-Caballero, A., Castillo, J.C., Rodríguez-Sánchez, J.M.: Human activity monitoring by local and global finite state machines. Expert Syst. Appl. 39, 6982–6993 (2012).
3. Hasanbelliu, E., Kampa, K., Principe, J.C., Cobb, J.T.: Online learning using a Bayesian surprise metric. Neural Networks (IJCNN), The 2012 International Joint Conference on. pp. 1–8. IEEE (2012).
4. Bo, L., Sminchisescu, C.: Twin gaussian processes for structured prediction. Int. J. Comput. Vis. 87, 28–52 (2010).
5. Todorov, E.: Efficient computation of optimal actions. Proc. Natl. Acad. Sci. U. S. A. 106, 11478–11483 (2009).
6. Deisenroth, M.P., Rasmussen, C.E., Peters, J.: Gaussian process dynamic programming. Neurocomputing. 72, 1508–1524 (2009).