35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Personalized Tuning of a Reinforcement Learning Control Algorithm for Glucose Regulation

Elena Daskalaki, Peter Diem, and Stavroula G. Mougiakakou, *Member, IEEE*

***Abstract*—Artificial pancreas is in the forefront of research towards the automatic insulin infusion for patients with type 1 diabetes. Due to the high inter- and intra-variability of the diabetic population, the need for personalized approaches has been raised. This study presents an adaptive, patient-specific control strategy for glucose regulation based on reinforcement learning and more specifically on the Actor-Critic (AC) learning approach. The control algorithm provides daily updates of the basal rate and insulin-to-carbohydrate (IC) ratio in order to optimize glucose regulation. A method for the automatic and personalized initialization of the control algorithm is designed based on the estimation of the transfer entropy (TE) between insulin and glucose signals. The algorithm has been evaluated *in silico* in adults, adolescents and children for 10 days. Three scenarios of initialization to i) zero values, ii) random values and iii) TE-based values have been comparatively assessed. The results have shown that when the TE-based initialization is used, the algorithm achieves faster learning with 98%, 90% and 73% in the A+B zones of the Control Variability Grid Analysis for adults, adolescents and children respectively after five days compared to 95%, 78%, 41% for random initialization and 93%, 88%, 41% for zero initial values. Furthermore, in the case of children, the daily Low Blood Glucose Index reduces much faster when the TE-based tuning is applied. The results imply that automatic and personalized tuning based on TE reduces the learning period and improves the overall performance of the AC algorithm.**

I. INTRODUCTION

The simultaneous use of Continuous Glucose Monitors (CGMs) for measurement of glucose levels and pumps for the subcutaneous infusion of insulin is one of the fundamental therapeutic schemes for individuals suffering from Type 1 Diabetes (T1D) mellitus. A control algorithm able to estimate the appropriate per patient insulin dose to be infused by the pump based on glucose data provided by the CGM could lead to the development of an Artificial Pancreas (AP). Various approaches for such control algorithms have been proposed including Proportional-Integral-Derivative (PID) control [1]-[4], Model Predictive Control (MPC) [4]-[10], run-to-run algorithms [11]-[14] and MD-Logic (MDL) control [15]-[16].

One of the major challenges in diabetes regulation is the high inter- and intra-population variability. For this purpose, personalized insulin treatment has been recently highlighted as a crucial goal towards efficient glucose control. This study discusses the use of a novel and online adaptive approach for glucose regulation based on the principles of reinforcement learning and optimal control for personalized diabetes treatment. In previous work of the Diabetes Technology Research Group [17], an algorithm based on the Actor-Critic (AC) learning has been designed and developed. The algorithm provides daily updates of the average basal rate (BR) and the insulin-to-carbohydrate (IC) ratio towards minimization of hyper-/hypoglycemia. As an extension to this study, a method for the automatic and personalized tuning of the AC based algorithm is proposed based on the estimation of information transfer (IT) between insulin and glucose signals.

II. METHODS

A. The AC algorithm

AC belongs to the class of reinforcement learning (RL) algorithms. RL involves adaptive agents able to optimize their performance over time through interaction with the environment, which may include partially known or unknown dynamics [18]. AC consists of two complementary adaptive agents: the Critic and the Actor, with the former being responsible for the control policy evaluation and the latter for the control policy optimization. AC implementations may vary in the design of both the Actor and the Critic part. An extensive review can be found in [19]. A schematic view of a system controlled by an AC algorithm is shown in Figure 1.

The system can be modeled as a Markov Decision Process (MDP) with finite state space *X* and action space *U*. The control policy is a deterministic or stochastic function

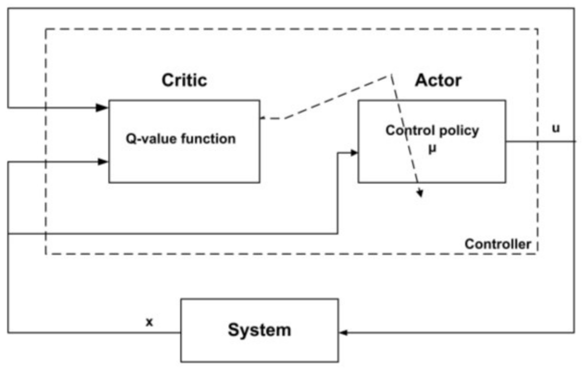


Figure 1: System controlled by an Actor-Critic algorithm

*Research supported by Bern University Hospital "Inselspital".

E. Daskalaki is with the Diabetes Technology Research Group, ARTORG Center for Biomedical Engineering Research, University of Bern, 3010 Bern, Switzerland (e-mail: elena.daskalaki@artorg.unibe.ch).

P. Diem is with the Division of Endocrinology, Diabetes and Clinical Nutrition, Bern University Hospital "Inselspital", 3010 Bern, Switzerland (e-mail: peter.diem@insel.ch).

S.G. Mougiakakou is with the Diabetes Technology Research Group, ARTORG Center for Biomedical Engineering Research, University of Bern, 3010 Bern, Switzerland (e-mail: stavroula.mougiakakou@artorg.unibe.ch: +41 31 632 7592; fax: +41 31 632 7576).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

$\mu(u|x,\theta)$ which maps an action u to a state x based on a policy parameter vector $\theta \in R^K$. A local cost $c(x,u)$ is associated with each state x and action u. Aim of the AC algorithm is to minimize the average expected cost per state defined as:

$$\bar{\alpha}(\theta)=\sum_{x\in X,\ u\in U} c(x,u)\eta_\theta(x,u) \tag{1}$$

where $\eta_\theta(x,u)$ is the stationary probability associated with the Markov chain $\{X_k, U_k\}$ dependent on θ. The Critic estimates the corresponding Q-value function $Q_\theta(x,u)$ which stands for the future expected cost when starting from state x and action u and following control policy $\mu(\cdot\,|\theta)$. Based on the Q-value function, the Actor estimates the gradient $\nabla_\theta \bar{\alpha}(\theta)$ of the average expected cost with respect to θ. The policy parameter vector θ is then updated based on a gradient descent approach:

$$\theta_{k+1}=\theta_k-\beta_k\nabla_\theta\bar{\alpha}(\theta) \tag{2}$$

where β_k a sequence of positive, non-decreasing step sizes and k denotes the iteration counter.

B. Design of an AC-based algorithm for glucose regulation

The algorithm implements a dual control policy for the optimization of the average daily BR and the IC ratio defined as:

$$IC={}^{I_{bolus}}\!/_{CHO} \tag{3}$$

where I_{bolus} is the insulin bolus dose and CHO is the amount of carbohydrates contained in a meal. The Critic estimates the Q-value function based on the Temporal Differences method [20] while the Actor updates the two control policies on a daily basis as follows: At the end of day k, $k=1, \ldots, D$, with D the total number of days of the trial, the daily sensor glucose profile is collected and two features related to hyperglycemia and hypoglycemia are computed as shown in (4) and (5).

$$F_1=max\left(G_{max}\text{-}G_h,0\right) \tag{4}$$

$$F_2=max\left(G_l\text{-}G_{min},0\right) \tag{5}$$

where G_{max}, G_{min} are the measured minimum and maximum glucose concentration and G_h=180 mg/dl, G_l=70 mg/dl are the hyperglycemia and hypoglycemia bounds respectively. Define $F=[F_1\ F_2]$ the feature vector containing the hyper- and hypoglycemia features. The control policy for the average BR and the IC ratio is updated as:

$$\mu\left(u_k\middle|x_k,\theta_k^S\right)=S_k=S_{k-1}+P_k^S{*}S_{k-1} \tag{6}$$

where $S=\{BR, IC\}$, S_k is the control policy for day k and P_k^S is the rate of change of S_k from day k-1 to day k estimated as a linear combination of the features F:

$$P_k^S=F_k\theta_k^S \tag{7}$$

with θ^S being the policy parameter vector of the respective control policy. The policy parameter vectors θ^S are updated based on (2). A detailed description of the design and implementation of the AC-based algorithm for glucose regulation can be found in [17].

A major challenge during the design of adaptive algorithms, especially for applications where safety matters, is to keep the learning period as short as possible. Furthermore, even during learning, the necessary safety constraints should be guaranteed. One way to achieve this goal in designing an AC-based algorithm is through the appropriate initialization of the policy parameter vectors θ^S, which regulate the optimization of the control policy over time. The parameters θ^S can be viewed intuitively as weights that define the percentage of change of the BR and IC ratio according to the daily hypoglycemia and hyperglycemia status. Setting these parameters away from their optimal values results in longer learning period, which can be crucial for the safety of the patient. One would expect that the percentage of change depends on the amount of IT from insulin to glucose in the sense that for high IT small adaptations of the insulin scheme may be sufficient whereas for low IT larger updates may be needed. Based on this reasoning, the IT from insulin to glucose has been estimated and used for the automatic, patient-specific initialization of θ^S.

C. Automatic tuning of the AC-based algorithm

Assessing causality and IT between signals has been extensively studied and various measures have been proposed. A comprehensive review can be found in [21]. Transfer entropy (TE) is a powerful measure of IT, mainly due to its nonlinear and directional structure, and has found promising application in biomedical signal analysis [22]-[24]. TE measures the information flow from a signal Y (source) to a signal X (target) while it excludes redundant effects coming from other signals. Let $X=\{x_i,\ i=1{:}n\}$, $Y=\{y_i,\ i=1{:}n\}$, $Z=\{z_i,\ i=1{:}n\}$ be three observed random processes of length n. TE estimates the IT from process Y to X, which can be also translated as the amount of knowledge we gain about X when we already know Y, based on the following formula:

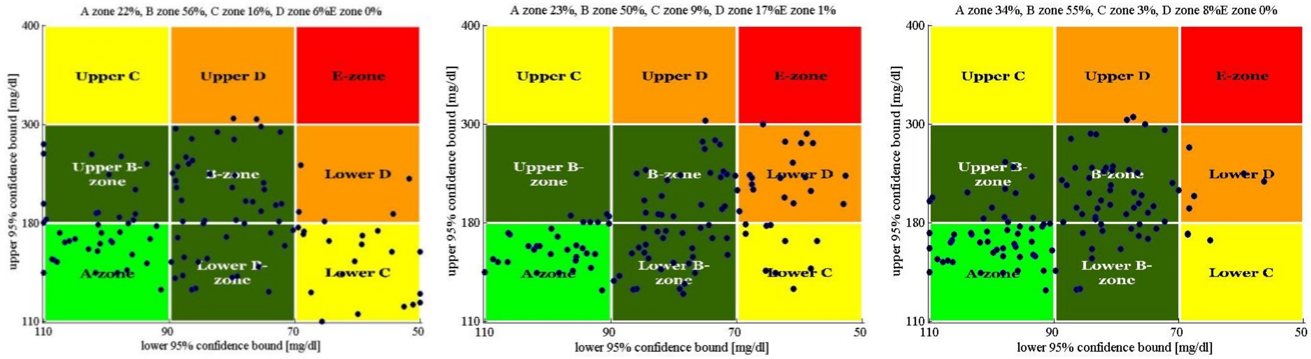
$$T_{Y\rightarrow X}=\sum_i p\left(x_i,y_i,z_i\right) log\frac{p(x_i|y_i,z_i)}{p(x_i|z_i)} \tag{8}$$

where $p(\cdot)$ denotes probability density function (pdf) and log is the basis two logarithm. Division with the conditional probability of X to Z excludes the redundant information coming from both Y and Z without excluding, though, the possible synergistic contribution of the two signals on X [23]. Main challenge in computing (8) is the estimation of the involved pdfs. Several approaches have been proposed for this purpose [21]. One of the most commonly used methods is the fixed data partitioning in which the time-series are partitioned into equi-sized bins and the pdfs are approximated as histograms [25].

Expecting that high TE is related to smaller rates of change in the insulin scheme, the initial values of the policy parameter vectors θ^S are set to be inversely proportional to the estimated TE per patient as:

$$\theta_0^S(p)=W/TE(p) \tag{9}$$

where p denotes a specific patient and W is a constant manually set as W=±1 with +1 for the elements of θ_0^S related to hyperglycemia and -1 for the elements related to hypoglycemia.



(a) (b) (c)

Figure 2: CVGA plots for all patients and the last five trial days when AC initialization is based on scenario (a) S1, (b) S2 and (c) S3

III. Results

The AC-based algorithm has been *in silico* evaluated on 28 virtual T1D patients (10 adults; 10adolescents; 8children) using the educational version of the FDA-accepted University of Virginia (UVa) T1D simulator. Two children have been excluded due to excessive glucose responses. The meal protocol included 3-4 daily meals of random CHO content and timing. A detailed description of the meal scenario is presented in [26]. In order to simulate the errors when real patients estimate the CHO content of their meal, a random meal uncertainty uniformly distributed between -50% and +50% has been introduced. The total trial duration was 10 days. The initial values of BR and IC ratio have been set equal to their optimized values as provided by the UVa simulator. For adults and adolescents, these values are close to the optimal ones, whereas in the case of children they are too high and, when applied in an open-loop scenario, they lead to excessive insulin infusion and frequent hypoglycemic events [17]. Consequently, the AC algorithm must perform significant updates of the BR and IC ratio in order to optimize glucose regulation, a fact that renders the duration of the learning period in children very challenging.

Three different initialization scenarios of the policy parameter vectors θ^S have been investigated:

S1. The policy parameter vectors θ^S are initialized to zero values.

S2. The policy parameter vectors θ^S are initialized to random values with magnitude ranging in (0, 1).

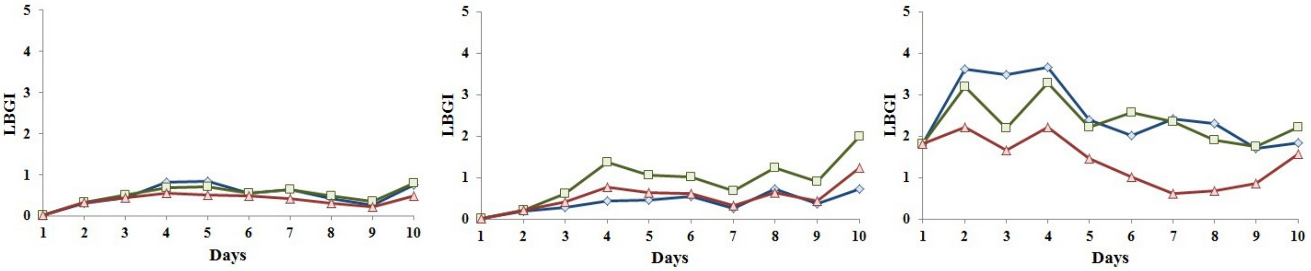
S3. The policy parameter vectors θ^S are initialized based on the estimated insulin-to-glucose transfer entropy per patient as in (9).

The Control Variability Grid Analysis (CVGA) has been used for the evaluation of the AC algorithm, while the risk of hypoglycemia has been estimated based on the Low Blood Glucose Index (LBGI) [27].

TABLE I**:** PERCENTAGES IN THE A+B ZONES OF THE CVGA FOR THE THREE AGE GROUPS AND SCENARIOS S1, S2, S3

Patient age group	**S1**	**S2**	**S3**
adults	93.00	95.00	98.00
adolescents	88.00	78.00	90.00
children	41.00	41.00	73.00

Figure 2 presents the CVGA plots for all patients and scenarios S1-S3. Table 1 presents the percentage of values within the A+B zones of the CVGA separately for each age group and the three scenarios. Setting the maximum acceptable duration of the learning period to five days, these results refer to the last five days of the trial. Finally, the daily evolution of the LBGI for the three groups when following scenarios S1-S3 is presented in Figure 3. This result refers to the total trial duration.



(a) (b) (c)

Figure 3: Evolution of LBGI during the 10 days of the *in silico* trial for adults (a), adolescents (b) and children (c) and scenarios S1 (blue), S2 (green) and S3 (red)

IV. Discussion

Figure 2 shows that, when the AC initialization is based on the patient-specific TE (scenario S3), the general performance of the algorithm after five days of learning is increased with 89% of the values in the A+B zones of the CVGA, compared to 78% for scenario S1 and 73% for S2. The same result can be observed from Table 1, where the values are separately presented for each patient group. Furthermore, from Table 1 it is clear that most of the hypoglycemic events present in Figure 2 belong to the age group of children. This is expected since, as mentioned earlier, the initial simulator-suggested values of BR and IC ratio are far from being optimal. The contribution of the TE-based tuning especially in the case of children is thus critical as it significantly reduces the duration of the learning period and achieves increased overall performance. Figure 3 supports this above remark presenting the evolution of the daily LBGI over the total trial duration. As expected, children start from much higher LBGI values compared to adults and adolescents. It can be further seen that the daily LBGI is kept to low and comparable values among the three scenarios for adults and adolescents, however, in the case of children, LBGI reduces much faster when the AC algorithm is initialized based on the TE compared to the zero or random initialization.

V. Conclusion

An AC-based control algorithm for glucose regulation in T1D has been designed and developed. In order to achieve faster and safer learning process, a novel approach for the automatic and patient-specific initialization of the algorithm, based on the estimation of the TE from insulin to glucose, has been proposed. Significant contribution of this method has been found compared to zero or random initialization especially in the case of children where the initial BR and IC ratio were far from their optimal values. Future work will include investigation of alternative ways for TE estimation and extensive evaluation of the AC control algorithm both *in silico* and in clinical practice.

References

[1] G.M. Steil, A.E. Panteleon and K. Rebrin, “Closed-loop insulin delivery: the path to physiological glucose control,” *Adv. Drug Deliver. Rev.*, vol. 56, pp. 125–144, 2004.

[2] G. Marchetti, M. Barolo, L. Jovanovic, H. Zisser and D. Seborg, “An improved PID switching control strategy for type 1 diabetes,” *IEEE Trans. Biomed. Eng.*, vol. 55, pp. 857-865, 2008.

[3] S.A. Weinzimer, G.M. Steil, K.L. Swan, J. Dziura, N. Kurtz and W.V. Tamborlane, “Fully automated closed-loop insulin delivery versus semi-automated hybrid control in pediatric patients with type 1 diabetes using an artificial pancreas,” *Diabetes Care*, vol. 31, pp. 934-939, 2008.

[4] C.C. Palerm, “Physiologic insulin delivery with insulin feedback: A control systems perspective,” *Comput. Meth. Prog. Biomed.*, vol. 102, pp. 130-137, 2011.

[5] R. Hovorka, V. Canonico, L.J. Chassin, U. Haueter, M. Massi-Benedetti, M.O. Federici, T.R. Pieber, H.C. Schaller, L. Schaupp, T. Vering and M.E. Wilinska, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiol. Meas.*, vol. 25, pp. 905-920, 2004.

[6] L. Magni, D.M. Raimondo, C. Dalla Man, G. De Nicolao, B. Kovatchev and C. Cobelli, “Model predictive control of glucose concentration in type I diabetic patients: An in silico trial,” *Biomed. Signal Proces.*, vol. 4, pp. 338-346, 2009.

[7] H. Lee, B.A. Buckingham, D.M. Wilson and B.W. Bequette, “A closed-loop artificial pancreas using model predictive control and a sliding meal size estimator,” *J. Diabetes Sci. Technol.*, vol. 3, pp. 1082–1090, 2009.

[8] Y. Wang, E. Dassau and F.J. Doyle, “Closed-loop control of artificial pancreatic beta-cell in type 1 diabetes mellitus using model predictive iterative learning control,” *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 211-219, 2010.

[9] B. Grosman, E. Dassau, H.C. Zisser, L. Jovanovic and F.J. Doyle III, “Zone model predictive control: a strategy to minimize hyper-and hypoglycemic events,” *J. Diabetes Sci. Technol.*, vol. 4, pp. 961-975, 2010.

[10] M.W. Percival, Y. Wang, B. Grosman, E. Dassau, H. Zisser, L. Jovanovic L and F.J. Doyle III, “Development of a multi-parametric model predictive control algorithm for insulin delivery in type 1 diabetes mellitus using clinical parameters,” *J. Process Contr.*, vol. 21, pp. 391–404, 2011.

[11] H. Zisser, L. Jovanovic, F.J. Doyle III, P. Ospina and C. Owens, “Run-to-run control of meal-related insulin dosing,” *Diabetes Technol. Ther.*, vol. 7, pp. 48-57, 2005.

[12] C.C. Palerm, H. Zisser, L. Jovanovic L and F.J. Doyle III, “A run-to-run control strategy to adjust basal insulin infusion rates in type 1 diabetes,” *J. Process Contr.*, vol. 18, pp. 258-265, 2008.

[13] H. Zisser, C.C. Palerm, W.C. Bevier, F.J. Doyle III and L. Jovanovic, “Clinical update on optimal prandial insulin dosing using a refined run-to-run control algorithm,” *J. Diabetes Sci. Technol.*, vol. 3, pp. 487-491, 2009.

[14] Y. Wand, E. Dassau, H. Zisser, L. Jovanovic L and F.J. Doyle III, “Automatic bolus and adaptive basal algorithm for the artificial pancreas b-cell,” *Diabetes Technol. Ther.*, vol. 12, pp. 879-887, 2010.

[15] S. Miller, R. Nimri, E. Atlas, E.A. Grunberg, and M. Phillip, "Automatic learning algorithm for the MD-logic artificial pancreas system," *Diabetes Technol. Ther.*, vol. 13, no. 10, pp. 983-990, 2011.

[16] E. Atlas, R. Nimri, S. Miller, E.A. Grunberg, and M. Phillip, “Feasibility Study of Automated Overnight Closed-Loop Glucose Control Under MD-Logic Artificial Pancreas in Patients with Type 1 Diabetes: The DREAM Project,” *Diabetes Technol Ther.*, vol. 14, no. 8, pp. 728-735, 2012.

[17] E. Daskalaki, P. Diem, S. Mougiakakou, “An Actor-Critic Based Controller for Glucose Regulation in Type 1 Diabetes”, DOI: 10.1016/j.cmpb.2012.03.002.

[18] R.S. Sutton and A.G. Barto, “Reinforcement learning,” *MIT Press*, 1998.

[19] C. Szepesvari, “Algorithms for reinforcement learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 4, no.1, 2009.

[20] J.N. Tsitsiklis and B. Van Roy, “An analysis of temporal-difference learning with function approximation,” *IEEE Trans. Aut. Contr.*, vol. 42, no. 5, pp. 674-690, 1997.

[21] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Phys. Rep.*, vol. 441, no. 1 pp. 1-46, 2007.

[22] T. Schreiber, "Measuring information transfer," *Phys. Rev. Let.*, vol. 85, no. 2, 461-464, 2000.

[23] P.L. Williams and R.D. Beer. "Generalized measures of information transfer," *preprint arXiv*, :1102.1507, 2011.

[24] J. Lee, S. Nemati, I. Silva, B.A. Edwards, J.P. Butler, and A. Malhotra, "Transfer Entropy Estimation and Directional Coupling Change Detection in Biomedical Time Series," *Biomed. Eng. OnLine*, vol.11:19, 2012.

[25] A.J. Butte, and I.S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *In Pac. Symp. Biocomput.*, vol. 5, pp. 418-429, 2000.

[26] E. Daskalaki, A. Prountzou, P. Diem, and S.G. Mougiakakou. "Real-Time Adaptive Models for the Personalized Prediction of Glycemic Profile in Type 1 Diabetes Patients." *Diabetes Technol. Ther.,* vol. 14, no. 2, pp. 168-174, 2012.

[27] B.P. Kovatchev, D. J. Cox, L. A. Gonder-Frederick, D. Young-Hyman, D. Schlundt, and W. Clarke. "Assessment of risk for severe hypoglycemia among adults with IDDM: validation of the low blood glucose index." *Diabetes Care,* vol. 21, no. 11, pp. 1870-1875, 1998.