Personalized Dual-Hormone Control for Type 1 Diabetes Using Deep Reinforcement Learning

Taiyu Zhu, Kezhi Li, Pantelis Georgiou

Abstract

We propose a dual-hormone control algorithm by exploiting deep reinforcement learning (RL) for people with Type 1 Diabetes (T1D). Specifically, double dilated recurrent neural networks are used to learn the hormone delivery strategy, trained by a variant of Q-learning, whose inputs are raw data of glucose & meal carbohydrate and outputs are the actions to deliver dual-hormone (basal insulin and glucagon). Without prior knowledge of the glucose-insulin metabolism, we develop the data-driven model in the UVA/Padova Simulator. We first pre-train the generalized model in an average T1D environment with a long-term exploration, then adopt importance sampling to train personalized models for each individual. *In-silico*, the proposed algorithm largely reduces adverse glycemic events, and achieves time in range, i.e., the percentage of normoglycemia, 93% for the adults and 83% for the adolescents, which outperforms previous approaches significantly. These results indicate that deep RL has great potential to improve the treatment of chronic illnesses.

Introduction

Diabetes is a lifelong condition that affects an estimated 451 million people worldwide (Cho et al. 2018). Delivering an optimal insulin dose to T1D subjects has been one of the long-standing challenges since the 1970s (Reddy et al. 2016).

In the past, the quality of life of subjects with Type 1 Diabetes (T1D) has relied heavily on the accuracy of human-defined models & features of the delivery strategy. Recently, however, deep learning has provided new ideas and solutions to many healthcare problems (Jiang et al. 2017). This has been empowered by the increased availability of medical data and rapid progress of analytic tools, e.g. deep reinforcement learning (RL). However, several reasons hinder the building of efficient RL models to solve problems in chronic diseases. Firstly, as RL medical data is collected from a dynamic interaction between the human and environment, they are limited and expensive (Artman et al. 2018). In addition, it is different to a scenario such as playing Atari in virtual environment (Mnih et al. 2015); RL costs heavily to 'explore' the possibilities on humans in terms of price and safety. Finally, the variability of physiological responses to the same treatment can be very large for different people with T1D (Vettoretti et al. 2018). These reasons are why there has been so little progress in using RL in chronic illnesses.

To overcome these obstacles, we propose a two-step framework to apply deep RL in chronic illnesses, and use T1D as a case study. T1D is chosen because it is a typical disease that requires dynamic treatment consistently. A generalized deep RL model for a hormone delivery strategy in diabetes is pre-trained using a variant of Q-learning as the first step. Secondly, by prioritizing the transitions in experience memory, importance sampling are implemented to train personalized models with individual data (Schaul et al. 2015). It has been shown that dilated recurrent neural network (DRNN) performs well in processing long-term dependencies and future glucose prediction (Chang et al. 2017; Chen et al. 2018). Thus, we employ DRNN to build a double deep Q-network (DQN) for multi-dimensional medical time series, in which each basal hormone delivery (at five minutes intervals) are considered as an action determined by a stochastic policy, and glucose levels and time in range (TIR) are considered as the reward. TIR presents the time percentage of glucose values within a target range considered to be normoglycemia. It is a key derived metric in glycemia control and preferred in diabetes clinics (Vigersky and McMahon 2018).

We use the UVA/Padova T1D Simulator, a credible glucose-insulin dynamics simulator which has been accepted by the Food and Drug Administration (FDA) (Dalla Man et al. 2014), as the environment. It can generate data from T1D subjects with high variability of meal intake, body conditions and other factors. During the training, the agent interacts with the T1D environment to obtain the optimal policy for the dual-hormone closed-loop delivery, as shown in Figure 1. Then 10 adults and 10 adolescents are tested within a 6 month period of time. The results show that TIRs achieve 93% for adults and 83% for adolescents *in-silico*, which significantly improve the state-of-the-art performances.

Related Work and Preliminaries

The rapid growth of continuous glucose monitoring (CGM) and insulin pump therapy have motivated use of a closed-loop system, known as the artificial pancreas (AP) (Cobelli, Renard, and Kovatchev 2011; Hovorka 2011). Many algorithms are developed and verified as closed-loop single/dual hormone delivery strategies (Bergenstal et al. 2016) that are mostly based on control algorithms (Facchinetti 2016; Haidar 2016). Researchers have investigated a RL routine to update several parameters of glucose controller in gradient (**?**; Herrero et al. 2017). Moreover, we find a RL environment was built in a simulator of the 2008 version (Xie 2018). Based on this simulator, a recent paper has introduced deep RL to improve average risk index (Fox and Wiens 2019); the method uses 1d-CNN or GRU as the DQN to control single insulin delivery. However, we use an updated simulator, the 2013 version (S2013), with many new features such as glucagon kinetics that allows dual-hormone actions (Dalla Man et al. 2014).

We see the problem an infinite-state Markov decision process (MDP) with noise. An MDP can be defined by a tuple $\langle S, A, R, \mathcal{T}, \gamma \rangle$ with state S, action A, reward function $R : S \times A \mapsto [R_{\text{min}}, R_{\text{max}}]$, transition function $\mathcal{T}$, and discount factor $\gamma \in [0, 1]$. At each time period, the agent takes an action $a \in A$, causes the environment from some state $s \in S$ to transit to state s' with probability $T(s', s, a) = P(s'|s, a)$. A policy π specifies the strategy of selecting an action. RL's goal is to find the policy π mapping states to actions that maximizes the expected long-term reward. Thus, we define Q-function $Q^{\pi}(s, a)$ for state-action values and the optimal $Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a) = \mathbb{E}_{s'} [R(s, a) + \gamma \max_{a'} Q^*(s', a')]$.

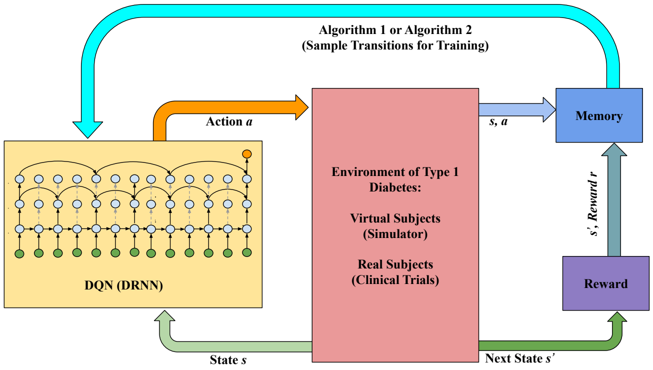

Algorithm 1 or Algorithm 2
(Sample Transitions for Training)
Action *a*
s, *a*
Memory
Environment of Type 1 Diabetes:
Virtual Subjects (Simulator)
Real Subjects (Clinical Trials)
s', Reward *r*
Reward
DQN (DRNN)
State *s*
Next State *s'*

Figure 1: The system architecture to apply deep RL on T1D.

Methods

In the hormone delivery problem, we use a multi-dimensional data as the input D. The data includes blood glucose G (mg/dL), meal data M (g) manually record by individuals, corresponding meal bolus B (U). Dual-hormone (basal insulin and glucagon) delivery is considered as actions A. In this case D can be denoted as $D = \{G, M, I, C\} = [d_1, \cdots, d_L]^{\mathbb{T}} \in \mathbb{R}^{\mathbb{L} \times 4}$, where L is the data length, I is the total insulin including bolus B, basal, and C stands for the glucagon. We use the latest 1 hour data (12 samples) as current state $s_t = [d_{t-11} \cdots, d_{t-1}, d_t]^{\mathbb{T}}$. Here B is computed from M with a standard bolus calculator $B \propto M$ divided by the body weight. Then the problem can be seen as an agent interacting with an environment over time steps sequentially, as depicted in Figure 1. Every five minutes, an observation $o_t = s_t + e_t$ can be obtained, and an action a_t is taken. The action can be chosen from three options: do nothing, deliver basal insulin, or deliver glucagon. The amount of basal insulin and glucagon is a small constant that determined by the subject profile in advance. To maintain the BG in a target range, we define a reward carefully that the agent receives in each time step

$$r_t = \begin{cases} -0.6 + (G_{t+1} - 70)/100 & 30 \leq G_{t+1} < 70 \\ 0 & 70 \leq G_{t+1} < 90 \\ 1 & 90 \leq G_{t+1} < 140 \\ 0 & 140 < G_{t+1} < 180 \\ -0.4 - (G_{t+1} - 180)/200 & 180 < G_{t+1} \leq 300 \\ -1 & \text{else} \end{cases} \tag{1}$$

The goal of the agent is to learn a personalized basal insulin and glucagon delivery strategy with a short period of time (one month) and limited data for each individual. Therefore, we propose a two-step learning approach: A generalized model is pre-trained in average T1D environment with a long-term exploration, and then we obtain personalized models for each subject by fine-tuning the generalized model. From the intrinsic perspective of T1D in a real-life scenario, we cannot use trial and error process with poor initial performance, so the first step needs to be done in the simulator. For the second step, it is possible to be adopted in real clinical trails, with a generalized model as the agent's initial policy. Meanwhile, proper safety constrains, such as insulin/glucagon suspension, are still required during the training process.

Generalized DQN training

With the interactive environment in the simulator, the DRNN is employed as the centerpiece in DQNs, because it has larger receptive field that is crucial for glucose time series processing (Chen et al. 2018). Double DQN weights θ_1, θ_2 in the *simulator* are trained because it has been proved as a robust approach to solve overestimations (Van Hasselt, Guez, and Silver 2016). Action network and value network are trained $J_{DQ}(Q) = (R(o, a) + \gamma Q(o', \arg\max_a Q(o', a; \theta_1); \theta_2) - Q_i(o, a; \theta_1))^2$. The pseudo-code is sketched in Algorithm 1.

The agent explores random hormone delivery actions under policy π that is ε-greedy with respect to Q_{θ_1} in *simulator*. Some human intervention/demonstration at the beginning of the RL process can reduce the training time slightly, but in *simulator* it is not necessary.

Personalized DQN training

In this step we refine the model and customize it for the personal use. Weights and features obtained from the last step

Algorithm 1 Generalized DQN training

```
Inputs: Initializing environment $E$ , historical data $H$, update frequency T, two dilated RNN of random weights $\theta_1$, $\theta_2$, respectively.
**repeat**
    select action from $a \sim \pi(Q_{\theta_1}, \varepsilon)$, observe $(o', r)$ in $E(Is)$
    store $(o, a, r, o')$ into replay buffer $\mathcal{B}$
    sample a mini-bath uniformly from $\mathcal{B}$ and calculate loss $J_{DQ}(Q)$
    perform a gradient descent to update $\theta_1$, $\theta_2$
    **if** $t$ mod $T = 0$ **then** $\theta_2 \leftarrow \theta_1$ **end if**
**until** converge or reach the number of iterations
```

are updated using limited data with an importance sampling (Schaul et al. 2015). Details are shown in Algorithm 2.

Algorithm 2 Personalized DQN training

```
Inputs: Initialized with environment $E$, historical data $H$, generalized Q-function $Q$ with weights $\theta_1$, replay buffer $\mathcal{B}$, target weights $\theta_2$, update frequency $T$
generate $\mathcal{D}$ as a merge of $\mathcal{B}$ and experience collected from $H$
calculate importance probability $Pr$ from $H$
**repeat**
    select action from policy $a \sim \pi(Q_{\theta_1}, \varepsilon)$, observe $(o', r)$
    store $(o, a, r, o')$ in $\mathcal{D}$, overwriting the oldest samples previously merged from $\mathcal{B}$
    sample a mini-batch from $\mathcal{D}$ with importance $Pr$
    calculate loss $J(Q) = J_{DQ}(Q)$
    perform a gradient descent update $\theta$ and the importance sampling to update $Pr$
    **if** $t$ mod $T = 0$ **then** $\theta_2 \leftarrow \theta_1$ **end if**
**until** converge or reach the number of iterations
```

Experiment Results

We compare the results with the following experimental setup (details in supplementary materials): 1. constant basal insulin (CB); 2. insulin suspension and carbohydrate recommendation (ISCR) (Liu et al. 2019); 3. generalized DQN (Algorithm 1); 4. personalized DQN (Algorithm 1, 2). CB is the baseline method in *simulator* as conventional hormone control of T1D, and ISCR is based on proportional-derivative controller and Kalman filter.

In experiments, we used the TIR ($[70, 180]$ mg/dL), the percentage of hypoglycemia ($<$ 70 mg/dL) and hyperglycemia ($>$ 180 mg/dL) as the metrics to measure the performance. In general, either higher TIR or lower Hypo/Hyper indicates better glycemia control. Table 1 presents the overall glycemia performance on the adult subjects. It is noted that the DQN$_{personalized}$ achieves the best performance and increases the mean TIR by 11.21% ($p \leq 0.005$), compared to the CB setup. For the adolescent case in Table 2, the DQN$_{personalized}$ also obtains the best TIR of 83.39%. In both cases, the personalized model outperforms the baseline methods on both TIR and Hypo/Hyper results with considerable improvements.

Table 1: Performance on 10 adult subjects.

Method	Normo (TIR)	Hypo	Hyper
CB	$81.91 \pm 8.66^‡$	$5.29 \pm 3.93^‡$	$12.80 \pm 8.67^‡$
ISCR	$87.62 \pm 7.57^‡$	$2.36 \pm 1.44^‡$	$10.01 \pm 7.35^‡$
DQN$_{Generalized}$	89.16 ± 5.04	1.92 ± 1.36	8.92 ± 5.38
DQN$_{Personalized}$	**93.12±4.48**	**1.25±1.32**	**5.63±3.29**

$^*p \leq 0.05$ $^†p \leq 0.01$ $^‡p \leq 0.005$

Table 2: Performance on 10 adolescent subjects.

Method	Normo (TIR)	Hypo	Hyper
CB	$61.68 \pm 10.95^‡$	$9.04 \pm 7.22^‡$	$29.28 \pm 11.16^‡$
ISCR	$74.55 \pm 9.61^†$	$2.38 \pm 1.82^†$	$23.07 \pm 7.26^‡$
DQN$_{Generalized}$	74.89 ± 8.58	2.36 ± 2.19	22.75 ± 8.63
DQN$_{Personalized}$	**83.39±8.03**	**2.10±1.56**	**14.51±9.98**

$^*p \leq 0.05$ $^†p \leq 0.01$ $^‡p \leq 0.005$

In Figure 2, we visualize the TIR performance through 30-day personalized training, and specific BG values of two subjects, as average cases, over 6-month testing period. We has explored simple fully-connected neural network (NN) as DQNs. Although there are increasing trends of TIR for both DRNN and NN during the training, the TIR performance of the NN is not as stable as DRNN. The performance on the adult is basically in accordance with statistical results in Table 1, and the DQN model avoids many hypoglycemia events during the night. For the adolescents, it is observed that the DQN model also helps avoid adverse glycemic events and improve TIR significantly.

Conclusion

We propose a new dual-hormone delivery algorithm and employ deep RL for glucose management. DRNNs are used in the architecture of double DQN with the 2-step learning framework to develop personalized models. This algorithm has achieved a significant improvement in glycemic control and outperforms existing work.

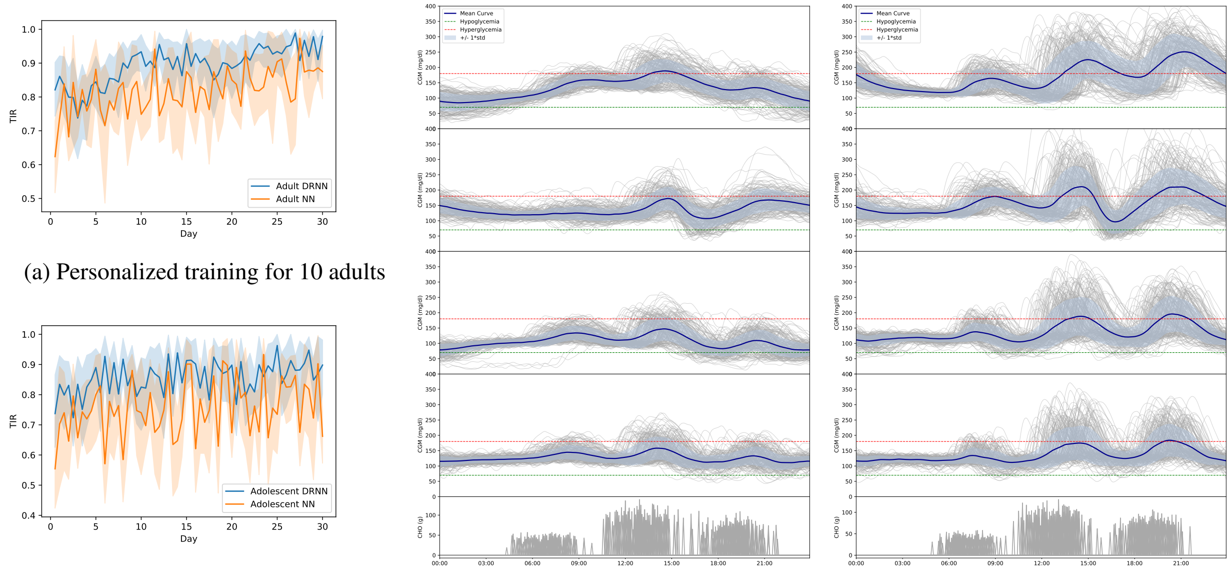
References

Artman, W. J.; Nahum-Shani, I.; Wu, T.; Mckay, J. R.; and Ertefaie, A. 2018. Power analysis in a SMART design: sample size estimation for determining the best embedded dynamic treatment regime. *Biostatistics*.

Bergenstal, R. M.; Garg, S.; Weinzimer, S. A.; Buckingham, B. A.; Bode, B. W.; Tamborlane, W. V.; and Kaufman, F. R. 2016. Safety of a hybrid closed-loop insulin delivery system in patients with type 1 diabetes. *Jama* 316(13):1407–1408.

Chang, S.; Zhang, Y.; Han, W.; Yu, M.; Guo, X.; Tan, W.; Cui, X.; Witbrock, M.; Hasegawa-Johnson, M. A.; and Huang, T. S. 2017. Dilated recurrent neural networks. In *Advances in Neural Information Processing Systems*, 77–87.

Chen, J.; Li, K.; Herrero, P.; Zhu, T.; and Georgiou, P. 2018. Dilated recurrent neural network for short-time prediction of glucose



(a) Personalized training for 10 adults

(b) Personalized training for 10 adolescents

(c) Testing performance on an adult

(d) Testing performance on an adolescent

Figure 2: **Left:** The performance of personalized training with confidence intervals on 10 subjects over a one-month period, using DRNN or NN. **Middle and Right:** The testing performance of four setup over a 6-month period: (Top-to-bottom) CB, ISCR, generalized DQN, personalized DQN, distribution of meal carbohydrate. The average BG levels for 180 days are shown in solid blue lines, and the hypo/hyperglycemia regions are shown in dotted green/red lines. Each gray line stands for glucose trajectory over 1 day (totally 180 ensembles), and the blue shaded regions indicate the standard deviation.

concentration. In *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, 69–73.

Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.; Ohlrogge, A.; and Malanda, B. 2018. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice* 138:271–281.

Cobelli, C.; Renard, E.; and Kovatchev, B. 2011. Artificial pancreas: past, present, future. *Diabetes* 60(11):2672–2682.

Dalla Man, C.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; and Cobelli, C. 2014. The UVA/PADOVA type 1 diabetes simulator. *Journal of diabetes science and technology* 8(1):26–34.

Facchinetti, A. 2016. Continuous glucose monitoring sensors: Past, present and future algorithmic challenges. *Sensors* 16(12).

Fox, I., and Wiens, J. 2019. Reinforcement learning for blood glucose control: Challenges and opportunities. In *Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning (ICML)*.

Haidar, A. 2016. The artificial pancreas: How closed-loop control is revolutionizing diabetes. *IEEE Control Systems Magazine* 36(5):28–47.

Herrero, P.; Bondia, J.; Oliver, N.; and Georgiou, P. 2017. A coordinated control strategy for insulin and glucagon delivery in type 1 diabetes. *Computer methods in biomechanics and biomedical engineering* 20(13):1474–1482.

Hovorka, R. 2011. Closed-loop insulin delivery: from bench to clinical practice. *Nature Reviews Endocrinology* 7(385).

Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; and Wang, Y. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* 2:230–243.

Liu, C.; Avari, P.; Oliver, N.; Georgiou, P.; and Vinas, P. H. 2019. Coordinating low-glucose insulin suspension and carbohydrate recommendation for hypoglycaemia minimization. In *Diabetes technology & therapeutics*, volume 21, A85–A85.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Reddy, M.; Pesl, P.; Xenou, M.; Toumazou, C.; Johnston, D.; Georgiou, P.; Herrero, P.; and Oliver, N. 2016. Clinical safety and feasibility of the advanced bolus calculator for type 1 diabetes based on case-based reasoning: A 6-week nonrandomized single-arm pilot study. *Diabetes Technology & Therapeutics* 18(8):487–493.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2015. Prioritized experience replay. *International Conference on Learning Representations* abs/1511.05952.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double Q-learning. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Vettoretti, M.; Facchinetti, A.; Sparacino, G.; and Cobelli, C. 2018. Type 1 diabetes patient decision simulator for in silico testing safety and effectiveness of insulin treatments. *IEEE Transactions on Biomedical Engineering* 1–1.

Vigersky, R. A., and McMahon, C. 2018. The relationship of hemoglobin a1c to time-in-range in patients with diabetes. *Diabetes technology & therapeutics* 21(2):81–85.

Xie, J. 2018. Simglucose v0.2.1 (2018) [online]. https://github.com/jxx123/simglucose.