Non-linear Continuous Action Spaces for Reinforcement Learning in Type 1 Diabetes

Chirath Hettiarachchi[1(✉)], Nicolo Malagutti[1], Christopher J. Nolan[1], Hanna Suominen[1,2], and Elena Daskalaki[1]

[1] Australian National University, Canberra, Australia
{chirath.hettiarachchi,nicolo.malagutti,christopher.nolan,
hanna.suominen,eleni.daskalaki}@anu.edu.au
[2] University of Turku, Turku, Finland

Abstract. Artificial Pancreas Systems (APS) aim to improve glucose regulation and relieve people with Type 1 Diabetes (T1D) from the cognitive burden of ongoing disease management. They combine continuous glucose monitoring and control algorithms for automatic insulin administration to maintain glucose homeostasis. The estimation of an appropriate control action—or—insulin infusion rate is a complex optimisation problem for which Reinforcement Learning (RL) algorithms are currently being explored due to their performance capabilities in complex, uncertain environments. However, insulin requirements vary markedly according to sleep patterns, meal and exercise events. Hence, a large dynamic range of insulin infusion rates is required necessitating a large continuous action space which is challenging for RL algorithms. In this study, we introduced the use of non-linear continuous action spaces as a method to tackle the problem of efficiently exploring the large dynamic range of insulin towards learning effective control policies. Three non-linear action space formulations inspired by clinical patterns of insulin delivery were explored and analysed based on their impact to performance and efficiency in learning. We implemented a state-of-the-art RL algorithm and evaluated the performance of the proposed action spaces *in-silico* using an open-source T1D simulator based on the UVA/Padova 2008 model. The proposed exponential action space achieved a 24% performance improvement over the linear action space commonly used in practice, while portraying fast and steady learning. The proposed action space formulation has the potential to enhance the performance of RL algorithms for APS.

Keywords: Reinforcement learning · Glucose regulation · Continuous action space

1 Introduction

Reinforcement Learning (RL) is a class of machine learning algorithms where an intelligent agent learns to act in an underlying environment to maximise

H. Aziz et al. (Eds.): AI 2022, LNAI 13728, pp. 557–570, 2022.
https://doi.org/10.1007/978-3-031-22695-3_39

a cumulative reward [25]. The reward is formulated to reflect a desired objective. RL algorithms have been successfully applied in games, where they have demonstrated superhuman performance capability/potential [20]. However, the application of RL to real-world problems is challenging due to complexities and constraints such as critical safety requirements, lack of knowledge for the formulation of reward functions, delays in sensors or actuators, partial observability, and high-dimensional continuous state or action spaces [6].

The problem of glucose regulation in Type 1 Diabetes (T1D) features all of the above challenges in RL. In healthy individuals, insulin secretion is performed by the islet β-cells of the pancreas. During periods of fasting (e.g., during sleep), a low basal rate of insulin secretion is required, whereas after meals surges in insulin secretion superimposed on basal secretion are necessary to maintain normal blood glucose concentrations [19]. In people with T1D, the autoimmune destruction of the β-cells of the pancreas results in complete insulin deficiency [5]. As a result, external insulin administration is vital to maintain glucose homeostasis [17]. Efforts to estimate the right rates of insulin infusion in T1D are challenged by delays in continuous glucose monitoring (CGM) sensing and insulin action; high inter- and intra-population variability; and critical safety constraints that cannot be compromized. Current advancements in T1D management methods include insulin administration by a continuous subcutaneous insulin infusion (CSII) pump alongside CGM in open-loop or hybrid closed-loop systems. In both cases, insulin delivery is divided into two distinct infusion patterns: low-range, almost-continuous basal pattern of delivery, and a pattern of intermittent high-range (or bolus) delivery of insulin used mainly to counter the glucose elevation due to meals [17]. In an open-loop setting, both basal and bolus insulin rates are calculated based on patient-specific characteristics (e.g., total daily insulin requirement, carbohydrate ratio) combined with the estimated amount of carbohydrate (CHO) content of a meal [17], as well as on insulin pharmacokinetic and pharmacodynamic properties [23]. In hybrid closed-loop schemes, basal insulin infusion rates are automatically estimated by a control algorithm according to CGM inputs, while insulin bolus dosing is manually calculated and administered by the user prior to meals. For decades, research interest has been on developing a fully automated Artificial Pancreas System (APS) (Fig. 1A) [4]. An APS consists of a CGM, a CSII pump, and a control algorithm to calculate automatically the insulin infusion rate for all circumstances in an effort to improve the total time in the normoglycemic range and relieve the people living with T1D from the heavy cognitive burden included in the manual calculation of meal CHO and insulin bolus doses [2].

Current APS research is investigating the use of RL algorithms due to their capability to perform well in uncertain and complex dynamic environments with disturbances [1]. However, one of the main challenges faced by RL algorithms in the APS context is the large and continuous insulin action space, which differs from the discrete actions present in game environments. According to the basal-bolus scheme, the low-range basal insulin actions account for the vast majority of the total insulin actions, while the large bolus actions are intermittent.

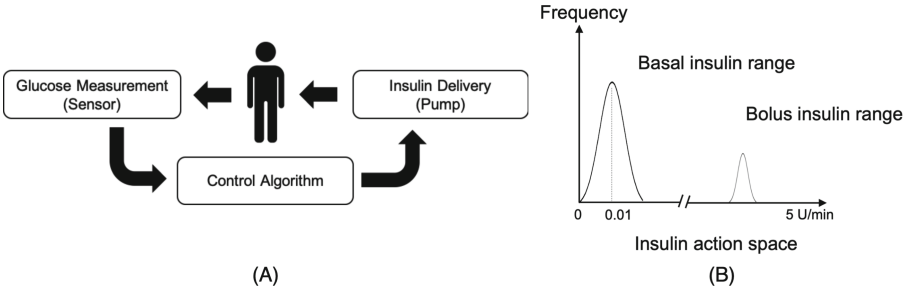

Fig. 1. (A) Artificial Pancreas System, (B) Frequency distribution of insulin action based on a clinical perspective (not to scale).

Typical clinical patterns of insulin delivery can hence be interpreted as bi-modal in the frequency (Fig. 1B). This challenges the RL algorithm by requiring efficient exploring of the entire continuous insulin action space to learn suitable control strategies for different situations which necessitates varying amounts of insulin out of a large dynamic range. The application of RL algorithms to continuous action spaces is not straightforward compared to low-dimensional discrete action spaces, while high-dimensional actions further increase the learning difficulty [13]. The complexity of continuous action spaces can be sub-optimally solved by discretizing the action space. However, this may not be suitable for high-precision control problems such as glucose regulation, as it could eliminate required information regarding the structure of the action space. Current available CSII pumps are discretized with fine resolution (e.g., Medtronic Minimed Pump basal increment of 0.025U/hr [18]), hence allowing the assumption of a continuous insulin action space. An alternative approach could be the introduction of two separate actions for the RL algorithm to focus on the clinical conventions of basal and bolus insulin separately. However, this increases the degrees of freedom in the algorithm and, due to the very sparse use of large insulin doses, could add complexity to learning.

In this study, we introduced the use of non-linear continuous action spaces as a method to overcome the challenges associated with efficiently exploring the dynamic range of insulin to learn effective glucose regulation strategies. Three non-linear translation functions were designed to map the RL action to the insulin infusion rate, inspired by the basal-bolus pattern of clinical insulin treatment practice. We implemented a state-of-the-art RL algorithm used in continuous control (e.g., 3D humanoid motion problems, physics simulations [22]) to evaluate the learning performance and efficiency of the proposed non-linear continuous action spaces. We evaluated our approach *in-silico* using an open-source T1D simulator based on the FDA approved UVA/Padova 2008 model [9]. We demonstrated that a linear action space is not suitable for the problem of glucose regulation in T1D and show that the proposed non-linear continuous action spaces improve the performance while portraying fast and steady learning.

2 Related Work

The use of RL for continuous control has gained much attention in the recent past due to applications such as locomotion, self-driving, and dexterous manipulation tasks [13]. In the problem of glucose regulation in T1D, the majority of the RL-based approaches focus on hybrid systems, which only control basal insulin levels, while bolus insulin infusion is carried out manually by the user [7,14,31,32]. These studies use a small set of handcrafted discrete actions to represent basal insulin [31,32]. A handful of studies have sought to control both basal and bolus insulin without any user input [7,8,12]. In particular, [7,12] used a discrete action distribution to control insulin. However, the action space discretization could lead to a loss of information, while a continuous action space is expected to enable more flexible RL agents that can learn more robust control strategies. [8] is the only research which focused on a continuous insulin action space. They divided the action space to two equal regions, where one represented no insulin administration and the other mapping the action linearly to the insulin pump. This strategy encouraged sparse insulin dosing and was evaluated on reasonably low CHO meals which required insulin rates <0.5 U/min. However, in real-life scenarios the presence of meals with large CHO content results in a large insulin action space ([0, 5] U/min).

According to the RL-related literature, applications with a large continuous actions space can present challenges related to its efficient exploration [11] and the existence of redundant or irrelevant actions [30]. [11] used a novel actor-critic algorithm based on a Sequential Monte Carlo approach to improve the exploration, while [30] proposed an approach which combined the RL algorithm with an action elimination network to eliminate sub-optimal actions.

In the present study, we design and develop a fully automated APS based on a RL algorithm. We introduce a challenging meal protocol including meals of large CHO content, which translates to the need of a large insulin action space, reflective of a real-life scenario. To address this challenge, we propose the use of non-linear representations of the RL algorithm action spaces. This results in a non-uniform resolution across the action space which guides the RL algorithm to explore insulin delivery patterns observed in clinical treatment. To the best of our knowledge this is the first attempt to explore action space representations to tackle complexities in large continuous action spaces associated with glucose regulation in T1D.

3 Method

3.1 Problem Formulation

The glucose control problem can be formulated as a Partially Observable Markov Decision Process (POMDP), where perfect state information is unavailable and limited to noisy sensor measurements. This POMDP can be defined as a 6-tuple $(S^\star, S, O, A, P, R)$, where $s^\star \in S^\star$ denotes the true states, $s \in S$ the noisy states observed by an observation function O, $a \in A$ the actions, $P : (s^\star, a) \rightarrow s'$ the

transition function where s' denotes the next state, and $R : (s, a) \rightarrow r \in \mathbb{R}$ the reward function. The reward function is designed based on glucose risk indices proposed in [10], where dangerous glucose levels are penalised and normal glucose levels encouraged. We define an observation function $O : s_t^\star \rightarrow g_{t-k:t}, i_{t-k:t}$ which maps the true state $s_t^\star$ at current time t to glucose sensor observation g_t and administered insulin i_t augmented by their past k historical values. Hence, the observed state space is formulated as $s_t = (g_{t-k:t}, i_{t-k:t})$ where past k samples encompass the information related to glucose dynamics and the effect of insulin.

3.2 Action Space

We take a policy gradient approach for designing the RL algorithm since it is more suited for continuous action spaces and for learning stochastic policies [25]. In this formulation, the RL algorithm is required to predict a distribution over the actions ($\pi(a|s)$) for a given state (s). We use a normal distribution ($\mathcal{N}(\mu, \sigma)$) where the RL algorithm learns both μ & σ parameters. The final predicted action is bounded to the range $[-1, 1]$ which is then mapped to the insulin infusion rate of the insulin pump ($I_{pump} \in [0, 5]$ U/min) based on a translation function T. As discussed earlier, the common practice in RL is to map the predicted action linearly to the underlying actuator [3,26,27] as shown in Eq. 1, where I_{max} corresponds to the maximum insulin.

$$I_{pump} = I_{max} \cdot \frac{(a+1)}{2}, a \in [-1, 1]. \tag{1}$$

3.3 Proposed Translation Functions

Glucose regulation requires frequent use of very small insulin doses for basal insulin compared to the less frequent larger doses, resulting in a skewed concentration in the action space, as opposed to the uniform resolution provided by the linear mapping (Eq. 1). In order to capture this property, we explore three non-linear translation functions; (1) quadratic, (2) proportional-quadratic, and (3) exponential to formulate non-linear action spaces in order to provide better resolution to the important target insulin ranges (Fig. 2).

Quadratic. The translation function T is a quadratic function of the RL action a. This formulation integrates the two distinct actions (basal-bolus) used in typical insulin treatment to a single continuous action space avoiding the complexity of using multiple actions. The action space is divided into two segments, where actions in $[-1, 0]$ are translated to a basal dose range with a maximum basal insulin of $\delta_1(0.05)$ and actions in (0,1] considered as the bolus range with a maximum bolus insulin of I_{max}. This results in a duplication of the basal range [0, 0.05] in the bolus range [0, 5], which could be considered negligible due to the low resolution of the bolus range.

$$I_{pump} = \begin{cases} \delta_1 \cdot a^2 & -1 \leq a \leq 0 \\ I_{max} \cdot a^2 & 0 < a \leq 1 \end{cases}. \tag{2}$$

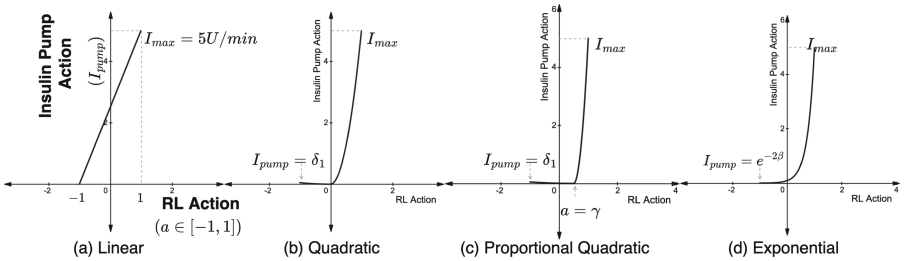

Fig. 2. Translation functions used to map RL action to the insulin infusion rate: (a) linear, (b) quadratic, (c) proportional-quadratic, and (d) exponential.

Proportional-Quadratic. This function is a modification of the Quadratic function where the parameter $\gamma(0.5)$ is introduced to adjust the resolution of the two dose ranges and δ_1 is set to 0.5.

$$I_{pump} = \begin{cases} \frac{\delta_1}{(\gamma+1)^2} \cdot (a-\gamma)^2 & -1 \leq a \leq \gamma \\ \frac{I_{max}}{(\gamma-1)^2} \cdot (a-\gamma)^2 & \gamma < a \leq 1 \end{cases}. \tag{3}$$

Exponential. The translation function T is an exponential function of the RL action $a \in [-1, 1]$ with a tuneable parameter $\beta(4.0)$ which ensures $I_{pump} \in (0, 5]$. This increases the resolution of the basal dose range while ensuring the action space is continuous without any duplication of actions. This formulation provides more flexibility for the RL algorithm to use the fully continuous structure of the action space and avoids any instabilities in learning, which might be caused by action duplication.

$$I_{pump} = I_{max} \cdot e^{\beta(a-1)}, a \in [-1, 1]. \tag{4}$$

3.4 Algorithm

The RL algorithm was designed based on PPO [22], which is one of the state-of-the-art on-policy RL methods used in continuous control problems. We formulate the glucose control problem as a continuing (not episodic) task, where the goal of the RL algorithm is to maximise the average reward r [16,25] while following a control policy π defined as,

$$r(\pi) \doteq \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_t | S_0, A_{0:t-1} \sim \pi]. \tag{5}$$

The PPO algorithm consists of a policy network (π_θ) and a value network (V_ϕ) which we have implemented using recurrent and dense neural network layers. The main objective of the policy network is to learn a suitable policy while the value network learns the n-step expected return being in a given state (s_t).

The PPO algorithm imposes constraints on policy updates to avoid excessive changes between the old policy ($\pi_{\theta_{old}}$) and the new policy (π_θ) by clipping the probability ratios of the new and old policies at $1-\epsilon$ or $1+\epsilon$ as shown in the policy objective below:

$$L^{policy}(\theta) = \hat{\mathbb{E}}_t \Bigg[min(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t, \\ clip(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon)\hat{A}_t) + \beta_s H(\pi(\cdot|s_t)) \Bigg], \tag{6}$$

where $\hat{A}_t$ is the advantage function [21] estimate at timestep t. The entropy term $H(\pi(\cdot|s_t))$ facilitates exploration, where β_s is a hyperparameter. We used n-step returns to compute the advantage function and value function targets (V_t^{target}). The value network is optimised using the objective:

$$L^{value}(\phi) = \hat{\mathbb{E}}_t \left[\frac{1}{2}(V_\phi(s_t) - \hat{V}_t^{target})^2 \right]. \tag{7}$$

3.5 Simulation Protocol

The UVA/Padova T1D simulator was used for the conduction of our study [9]. This is the only FDA-approved T1D simulator and can be used as a replacement of animal studies prior to clinical evaluation in humans. The simulator comprises a cohort of 30 *in-silico* subjects of three age categories (adults, adolescents and children) as well as models of different CGM and CSII pumps available in the market. In order to allow for reproducability of our results by the community, we used an open-source Python implementation of this simulator [29]. We conducted the evaluation using the adolescent cohort (10 subjects) due to their highly complex individual dynamics and glucose variability which create a very challenging glucose control environment. The Guardian RT glucose sensor and the Insulet pump with a sampling time of 5 min was used for the experiments. A challenging meal scenario was defined for the training and testing of the RL algorithm. For the training phase, the meal scenario consisted of three random meals (breakfast, lunch, and dinner) which were randomised based on the amount of CHO, time, and probability of occurrence (Table 1). The testing scenario spanned 24 h starting at 00:00 hrs and was fixed with three meals: 40 g of CHO for breakfast at 8:00 h, 80 g of CHO for lunch at 13:00 h, and 60 g of CHO for dinner at 20:00 h. Simulations which recorded glucose levels that exceeded the detectable range (39–600 mg/dL) of the glucose sensor were terminated and considered as a *catastrophic failure*.

3.6 Implementation Details and Data Analysis

The simulations were carried out on a workstation machine with 2 × NVIDIA 3090 GPUs. Each action space representation was evaluated for three random

Table 1. Training meal protocol.

Meal type	Time (hours)	Probability	Carbohydrates (g)
Breakfast	7.00–9.00	0.95	30–60
Lunch	12.00–14.00	0.95	70–100
Dinner	19.00–21.00	0.95	50–110

seeds per subject, where all other hyperparameters were kept fixed. The RL algorithms were trained for 500,000 interactions (1,736 human days, 1 interaction = insulin action taken every 5 min), which was identified as sufficient to reach convergence for the above proposed meal protocol. Upon the conclusion of training, 1,500 testing simulations were also conducted for each subject. The best performing action space representation was compared against the benchmarking linear action space by conducting statistical significance tests for each individual subject. A Shapiro-Wilk Test [24] was performed to check the normality and a Mann-Whitney U Test [15] was conducted to evaluate significance using a confidence level of 0.05.

3.7 Evaluation Metrics

The evaluation was twofold, and included analysis of the final performance after training using the results of the testing simulations and analysis of the learning efficiency derived from the training phase. For the final performance assessment, we used as metrics the total reward achieved by the RL algorithm as a percentage of the maximum achievable reward (PR) and the Time In Range (TIR) calculated as the average percentage of time that the glucose levels were maintained in the normoglycemic range (70–180 mg/dL) during a simulation. In addition we calculated the Failure Rate (FR) as the percentage of simulations which resulted in catastrophic failures over the total testing simulations. The clinical objective is to increase TIR and reduce the FR.

The aim of the RL algorithm during training is to iteratively improve a control policy reasonably fast (relative to the application's time scale), while avoiding excessive changes (smooth learning) which could lead to sudden unanticipated behaviour and even result in catastrophic failures. To evaluate the learning efficiency, we first defined reward thresholds of 25%, 50%, 70%, and 80% of the maximum achievable reward. The average number of interactions required to reach the threshold as a percentage of total interactions (PI) was used to compare the learning efficiency between the candidate action spaces. Furthermore, we qualitatively assessed the learning smoothness through visual inspection of the shape and fluctuations of the reward curves during training.

4 Results

The exponential action space improved the performance in terms of PR by 24%, while reducing the FR by 42% compared to the benchmark linear action space, on average across the subjects. The improvement was statistically significant ($p < .001$) for all the subjects. The performance of the candidate action spaces for all subjects, based on PR and FR metrics is summarised in Table 2. An inter-subject variability in performance improvements was identified, with substantial performance improvements for some subjects (Adolescent8). All the proposed non-linear action spaces were able to outperform the linear action space. The proposed non-linear functions all performed similarly in terms of TIR and FR (Fig. 3). Figure 3 also highlighted the variability in glucose control present among the subjects.

The exponential action space was the most efficient in reaching the 80% reward threshold in 34.90% PI, for all 10 subjects. Table 3 summarises the PI required for identified reward thresholds and the number of subjects reaching the target threshold. The linear action space was unable to reach the 80% reward threshold and also took more PI to cross lower reward thresholds. The quadratic and proportional-quadratic functions also performed better compared to the linear action space. The structure of the reward graphs (Fig. 4) for the exponential and quadratic action spaces gave evidence of steady learning and better convergence for all subjects. The reward graphs of adolescent 3 and 5 clearly indicated unsteady learning for the linear action space where sudden large reward fluctuations are observed during training. The linear action space failed to achieve convergence in some subjects (Adolescent 0, 8) and convergence to sub-optimal reward levels was observed in some subjects (Adolescent 1, 6, 9).

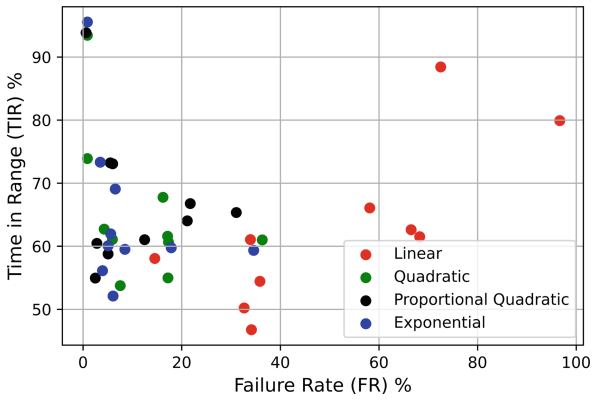

Fig. 3. The percentage time in normoglycemic range (TIR) and Failure Rate (FR) for candidate action spaces (Each dot represents an adolescent subject and each color a candidate action space.) (Color figure online)

Table 2. Adolescent Cohort Summary Results—Total Reward Achieved As A Percentage of Maximum Achievable Reward (PR) & Failure Rate (FR) for the candidate functions: Linear(**L**), Quadratic(**Q**), Proportional-Quadratic(**PQ**), Exponential(**E**).

Adolescent	Reward (PR)				Failure rate (FR)			
ID	L	Q	PQ	E	L	Q	PQ	E
0	52.55%	97.08%	97.33%	97.52%	72.53%	0.87%	0.60%	0.87%
1	66.70%	79.23%	81.75%	79.36%	35.87%	7.53%	2.47%	6.07%
2	76.09%	80.82%	77.92%	87.71%	32.67%	17.33%	21.73%	5.60%
3	58.48%	88.27%	75.20%	86.95%	34.13%	4.27%	31.07%	5.07%
4	68.21%	85.74%	76.81%	84.57%	58.13%	5.93%	21.13%	8.47%
5	70.20%	81.93%	88.96%	87.89%	33.93%	16.20%	6.00%	6.53%
6	54.96%	71.19%	82.78%	76.20%	68.27%	36.33%	2.80%	17.87%
7	75.94%	74.63%	81.40%	80.75%	14.53%	17.20%	5.07%	3.93%
8	25.17%	93.43%	90.72%	91.97%	96.67%	0.87%	5.53%	3.47%
9	59.65%	78.29%	80.90%	74.05%	66.53%	17.13%	12.47%	34.60%
Average (mean ± std)	60.79% ± 14.98%	83.06% ± 8.12%	83.38% ± 6.93%	84.70% ± 7.24%	51.33% ± 24.97%	12.37% ± 10.82%	10.89% ± 10.33%	9.25% ± 9.99%

Table 3. Efficiency analysis—Average Number of Interactions Required to reach the reward threshold as a percentage of total interactions (PI) and the number of adolescents achieving identified reward thresholds.

Translation function	Reward threshold			
	25%	50%	70%	80%
Linear	64.55% (10)	71.91% (9)	80.28% (5)	None
Quadratic	15.65% (10)	19.91% (10)	26.71% (10)	41.37% (8)
Proportional quadratic	38.75% (10)	45.79% (10)	52.51% (10)	63.44% (9)
Exponential	16.96% (10)	21.22% (10)	25.97% (10)	34.90% (10)

5 Discussion

The application of RL algorithms to problems with large continuous action spaces is challenging and currently being tackled through the design of efficient exploration algorithms [11] and irrelevant/redundant action elimination [30]. The common practice in continuous control RL tasks present in OpenAI Gym [3], DeepMind Control Suite [26], and MuJoCo physics environments [27] is to use a linear action space. Inspired by clinical treatment methods for T1D, in this study we introduced non-linear continuous action space representations to tackle the challenge of the large, continuous, and non-uniform insulin action space. To the best of our knowledge this is the first study to explore non-linear action space formulations to compensate the challenges present in continuous action spaces associated to glucose regulation in T1D.

The proposed exponential action space outperformed the linear action space, with statistically significant ($p < .001$) improvements in PR and FR metrics for

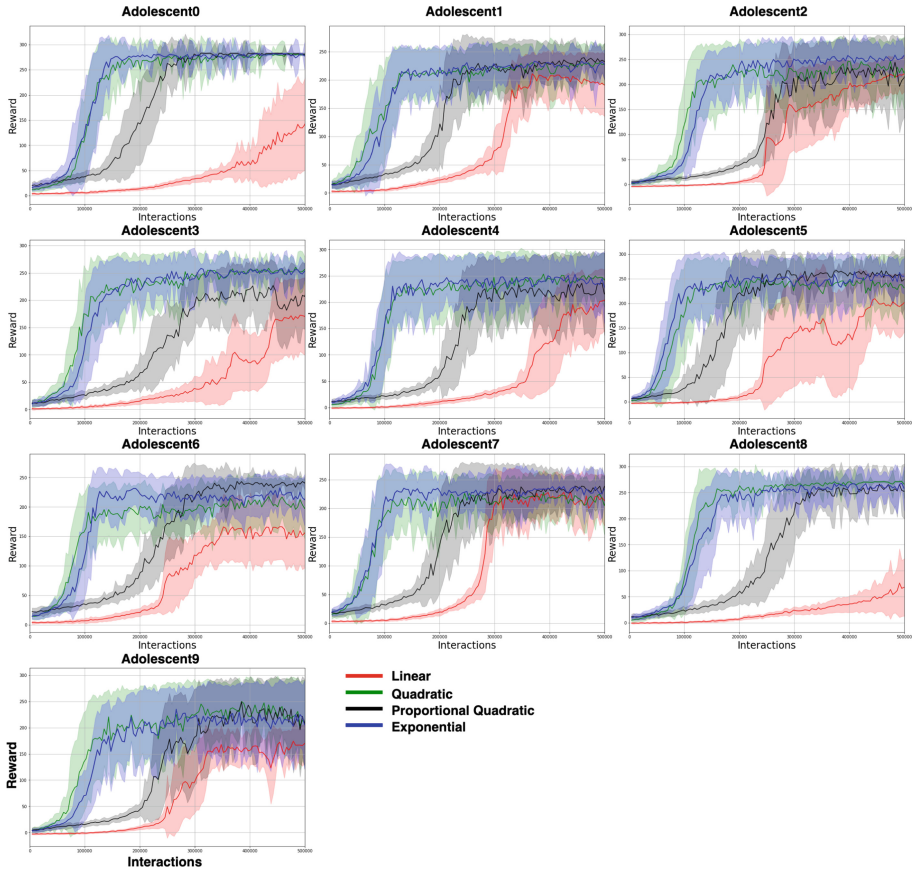

Fig. 4. Comparison of candidate action spaces during training. The mean and standard deviation of the total testing reward (3 random seeds × 20 testing scenarios) achieved for each candidate action space is presented against 500,000 learning interactions.

all subjects. It also exhibited steady convergence properties and was the most efficient in reaching the 80% reward threshold out of all the candidate action spaces. Unsteady learning (Adolescent 3, 5) and convergence to sub-optimal reward levels (Adolescent 1, 6, 9) was observed for subjects under a linear action space. The linear action space showed very poor performance for some subjects (Adolescent 0, 8) as it was unable to converge within the target number of training interactions. However, it is expected that the linear action space based RL algorithm might converge if the number of training interactions are increased. Meanwhile, all the proposed non-linear action spaces were able to converge within the target training interactions. The subjects who achieved convergence under the linear action space only converged to a sub-optimal level which indicates that increasing training interactions might not be beneficial for these subjects. The learning efficiency achieved in our proposed approach indicates efficient exploration by the

RL algorithm. This also reduces the computational time requirements for training, which is very valuable in the design and development phase of RL algorithms for glucose control, as often multiple iterations of designs are explored and tested. Increasing the complexity in the RL algorithmic architecture or the simulation protocol is expected to result in increased compute times until convergence is achieved. Hence, faster learning can become not only desirable but also vital for the experimental design of future RL algorithms for glucose regulation.

The successful real-world application of a RL-based APS would require online continual learning to adapt the control strategy based on biological variability (e.g., ageing, hormonal disturbances) of the user. Hence, the steady learning observed in the exponential and quadratic action spaces is vital to ensure safety by avoiding sudden excessive changes. The proposed approach illustrated favourable characteristics in this regard, while further future research is required. Our approach can also be applied to other medical applications with similar action space properties. The application of propofol dosing in general anaesthesia is such an application where a non-uniform action distribution is observed and RL currently being explored [28].

We selected the PPO algorithm due to its suitability towards the glucose regulation application which requires continuous control and steady learning. Hence, the performance improvement through non-linear action spaces were only analysed based on PPO. It is expected that the identified benefits would also be applicable to other similar on-policy RL algorithms. The results draw a promising line of research to explore the contribution of non-linear action spaces on other on-policy and off-policy RL algorithms. The use of a linear action space does not impose any bias in the learning process as all actions are equally probable. In contrast, our approach adds prior knowledge about the insulin action distribution to facilitate better learning which imposes the bias of the current clinical practice in T1D insulin treatment. This introduced bias was demonstrated as necessary for the glucose regulation task to achieve effective and efficient control. However, the effect of the bias in the proposed approach could be detrimental in problem domains with limited expert knowledge, and future research could focus on methods to identify the most suitable translation functions to design the target non-linear action spaces. The inter-subject variability in performance observed in the analysis highlights that the design of personalised action spaces using clinically recognised parameters of individuals may be beneficial, thus a potential area of future research. The designed RL-based system can be further improved to increase the TIR and reduce the FR. In future work, we aim to explore reward function formulations and algorithmic improvements to enhance the performance, while focussing on aspects such as safety, explainability, and transferability to real-life, which are vital for a robust APS.

6 Conclusions

This study proposed the use of non-linear action space representations for a RL-based APS with the aim to address the challenge of the large, continuous,

and non-uniform insulin action space and enhance the learning efficiency and performance of glucose control strategies. Our results demonstrated superior performance of the non-linear action spaces compared to the standard linear one with faster and smoother convergence and higher final reward. This research is expected to contribute to the development of RL-based fully-autonomous APS.

Acknowledgments. This research was funded in part by the Australian National University and the Our Health in Our Hands initiative.

Code Availability. A repository of code used in this study, and further supplementary material, is available at https://github.com/chirathyh/G2P2C.

References

1. Bothe, M.K., Dickens, L., et al.: The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Rev. Med. Devices **10**(5), 661–673 (2013)
2. Brew-Sam, N., Chhabra, M., et al.: Experiences of young people and their caregivers of using technology to manage type 1 diabetes mellitus: systematic literature review and narrative synthesis. JMIR Diabetes **6**(1), e20973 (2021)
3. Brockman, G., et al.: OpenAI gym. arXiv Eprint arXiv:1606.01540 (2016)
4. Cobelli, C., Renard, E., Kovatchev, B.: Artificial pancreas: past, present, future. Diabetes **60**(11), 2672–2682 (2011)
5. DiMeglio, L.A., Evans-Molina, C., Oram, R.A.: Type 1 diabetes. Lancet **391**(10138), 2449–2462 (2018)
6. Dulac-Arnold, G., Mankowitz, D., Hester, T.: Challenges of real-world reinforcement learning. arXiv preprint arXiv:1904.12901 (2019)
7. Fox, I., Wiens, J.: 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 (2019)
8. Fox, I., et al.: Deep reinforcement learning for closed-loop blood glucose control. In: Machine Learning for Healthcare Conference, pp. 508–536. PMLR (2020)
9. Kovatchev, B.P., Breton, M., et al.: In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. J. Diabetes Sci. Technol. **3**(1), 44–55 (2009)
10. Kovatchev, B.P., Clarke, W.L., et al.: Quantifying temporal glucose variability in diabetes via continuous glucose monitoring: mathematical methods and clinical application. Diabetes Technol. Ther. **7**(6), 849–862 (2005)
11. Lazaric, A., Restelli, M., Bonarini, A.: Reinforcement learning in continuous action spaces through sequential monte carlo methods. In: Advances in Neural Information Processing Systems, vol. 20 (2007)
12. Lee, S., Kim, J., et al.: Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: in silico validation. IEEE J. Biomed. Health Inform. **25**(2), 536–546 (2020)
13. Lillicrap, T.P., Hunt, J.J., Pritzel, A., et al.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)
14. Lim, M.H., Lee, W.H., et al.: A blood glucose control framework based on reinforcement learning with safety and interpretability: in silico validation. IEEE Access **9**, 105756–105775 (2021)

15. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**, 50–60 (1947)
16. Naik, A., Shariff, R., et al.: Discounted reinforcement learning is not an optimization problem. arXiv preprint arXiv:1910.02140 (2019)
17. Nathan, D., Genuth, S., et al.: The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N. Engl. J. Med. **329**(14), 977–986 (1993)
18. Online: Insulin pump comparison. http://www.betterlivingnow.com/forms/Insulin-Pump-Comparison.pdf. Accessed 24 Mar 2022
19. Rorsman, P., Eliasson, L., Renstrom, E., Gromada, J., Barg, S., Gopel, S.: The cell physiology of biphasic insulin secretion. Physiology **15**(2), 72–77 (2000)
20. Schrittwieser, J., Antonoglou, I., et al.: Mastering Atari, go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020)
21. Schulman, J., Moritz, P., Levine, S., et al.: High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015)
22. Schulman, J., Wolski, F., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
23. Shah, R.B., Patel, M., et al.: Insulin delivery methods: past, present and future. Int. J. Pharm. Investig. **6**(1), 1–9 (2016)
24. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)
25. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
26. Tassa, Y., Doron, Y., Muldal, A., Erez, T., et al.: DeepMind control suite. arXiv preprint arXiv:1801.00690 (2018)
27. Todorov, E., Erez, T., Tassa, Y.: MuJoCo: a physics engine for model-based control. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5026–5033. IEEE (2012)
28. Vajapey, A.: Predicting optimal sedation control with reinforcement learning. Ph.D. thesis, Massachusetts Institute of Technology (2019)
29. Xie, J.: Simglucose v0. 2.1 (2018). https://github.com/jxx123/simglucose. Accessed 13 Jan 2022
30. Zahavy, T., et al.: Learn what not to learn: action elimination with deep reinforcement learning. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
31. Zhu, T., Li, K., Georgiou, P.: A dual-hormone closed-loop delivery system for type 1 diabetes using deep reinforcement learning. arXiv preprint arXiv:1910.04059 (2019)
32. Zhu, T., Li, K., Herrero, P., Georgiou, P.: Basal glucose control in type 1 diabetes using deep reinforcement learning: an in silico validation. IEEE J. Biomed. Health Inform. **25**(4), 1223–1232 (2020)