COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

ORCS E4529 Reinforcement Learning

Final Project

Reinforcement Learning on Type 1 Diabetes Control

Yecheng Ma *Zunke Ma*

Supervised by Professor Shipra Agrawal

Fall 2023

Introduction

Type 1 Diabetes (T1D) is a chronic condition characterized by the pancreas's inadequate insulin production, essential for regulating blood glucose levels [1]. This deficiency leads to hyperglycemia, a critical rise in blood glucose. Currently, there is no cure for T1D, necessitating lifelong insulin management to maintain normal glucose levels [2]. Treatments range from basal-bolus therapy, combining long-acting and short-acting insulin, to insulin pumps, small devices attached to the body, monitor blood glucose and provide a continuous insulin supply as needed. These pumps, while regulating immediate glucose fluctuations, can struggle with delayed post-meal spikes and rely on potentially inaccurate patient carbohydrate estimates [3].

In the insulin pump (artificial pancreas) systems two main control algorithms are prevalent: Proportional Integral Derivative (PID) control [4-5] and Model Predictive Control (MPC) [6-9]. MPC, using dynamic patient-specific models, aims to maintain target glucose levels but is limited by its inability to fully integrate external factors [9].

Reinforcement learning, increasingly utilized in healthcare, offers new perspectives in T1D management [10-13]. Various approaches range from employing PID as a baseline to utilizing neural networks for deep Q learning. The effectiveness of these methods varies, often based on the action space and the time frame of observational data used, such as the incorporation of 24-hour CGM samples and insulin doses in prior studies [11]. However, enhancements over traditional PID models have been modest.

This paper explores training a Deep Reinforcement Learning algorithm, specifically an Actor Critic model, in a simulated environment to control blood glucose in T1D patients across three age groups: adults, adolescents, and children.

The code can be found in:
https://drive.google.com/drive/folders/1bfZHMc-4LqeJxYGWZOljcKasxXnG2hqh?usp=sharing

Problem Formulation

Data from real patients is extremely hard to obtain. We utilized the *simglucose* simulator environment [17], which implemented the UVA/Padova glucose model [14] and was widely used for research purposes in T1D studies. We choose to use the OpenAI gym environment of this simulator.

State

We used only the current blood glucose value as the observation variable defined by the gym environment. Initially, we tried to include previous blood glucose values and insulin data into the states, similar to [11]. However, the simulator in the gym environment provides the single blood glucose at the current time as the state. Thus, we choose to adopt the single blood glucose level as the state, which is measured every 3 minutes.

Action

The action is the units of insulin administered to the patient. According to the specifications of the insulin pump incorporated in the simulator, this amount varies from 0 to 30 units. Action is also taken every 3 minutes.

Reward

In determining the reward function, we examined the risk function associated with blood glucose levels. Given the paramount importance of safety in healthcare applications, our objective is to enable our controller to sustain a healthy blood glucose level for extended periods, thereby minimizing health risks to patients. A key metric in assessing blood glucose-related risks is the blood glucose risk index (BGRI), depicted in Figure 1. BG represents blood glucose level.

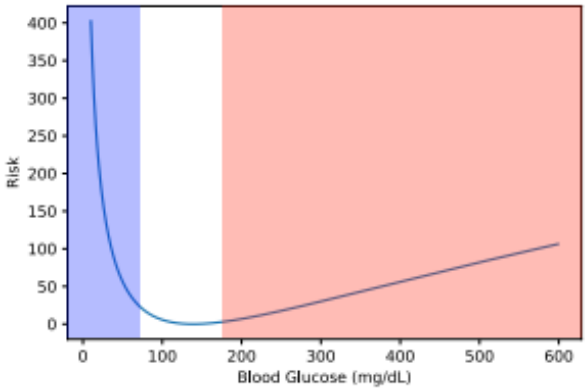


Figure 1: Continuous Risk Index Function

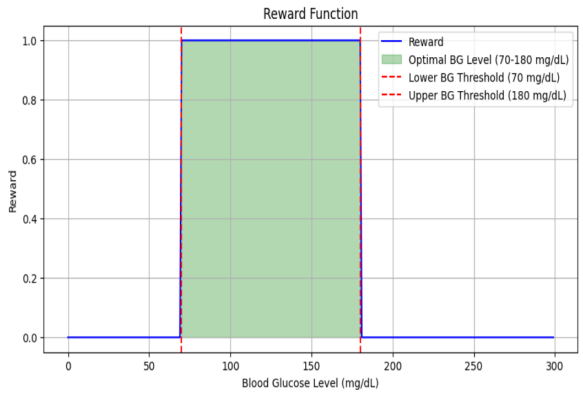


Figure 2: Our reward Function

Figure 1 illustrates the correlation between blood glucose levels (mg/dL) and the risk index value. The blue region signifies hypoglycemia (low blood glucose level), and the red region indicates the risk associated with hyperglycemia (high blood glucose level). The established target range for T1D patients is [70, 180] mg/dL [15].

Previous research has suggested using the inverse of this curve as a reward function [11]. However, this approach, which typically involves solely negative rewards, may prompt premature termination of the agent's activity, as noted in other studies [13].

To address this, we propose a simplified reward function similar to [13] designed to encourage the agent to maintain blood glucose within the healthy range, while also averting premature termination due to excessive negative rewards:

$$reward = \begin{matrix} 1,\ if\ BG \in [70,\ 180]\ mg/dL \\ 0, otherwise \\ -100,\ if\ terminate \end{matrix}$$

Our reward function is visualized in Figure 2.The termination criteria are set such that 1) if the patient's BG level falls to approximately 39 mg/dL for 20 to 30 minutes 2) rises to around 500 mg/dL for about an hour, the simulation ends. These parameters are predefined in the gym environment. The rationale behind the -100 penalty is to impose a substantial penalty for situations where BG levels reach critically low or high values, directly impacting patient safety.

Additionally, the architecture of this reward function is constructed so that, except in extreme cases where a -100 penalty is incurred, the cumulative reward effectively reflects the total time the patient's blood glucose level remains within the healthy range. This duration serves as an essential indicator for assessing the efficacy of our approach.

Policy Design

Our project aimed to precisely administer insulin units to maintain a patient's glucose levels within a normal range, leveraging the Actor-Critic method for this purpose. By innovatively combining policy and value networks, we effectively trained our agent to respond appropriately to various patient states.

The cornerstone of our approach was the Policy class, realized using PyTorch. The actor network's policy was parameterized by a Gaussian distribution, where the mean was determined by the neural network and the standard deviation was a learnable parameter. The network's architecture consisted of three layers, transitioning dimensions from 1 to 64, then 32, and finally back to 1. The patient's current BG level was input to the network, which then

computed the mean action corresponding to the insulin dosage. We applied a sigmoid function at the output to scale and constrain these values between 0 and 1, subsequently adjusting them to represent insulin dosages in the range of 0 to 30 units.

Our ValueFunction class, also designed in PyTorch, mirrored the architectural design of the policy network. Its main function was to estimate the value of each state the agent encountered, providing a benchmark for evaluating the decisions made by the policy network and thus enhancing the learning process.

We trained our agent using simulated patient scenarios within the SimGlucose framework, employing an epsilon-greedy strategy for exploration and exploitation. Initially, we introduced high randomness by generating insulin doses between 0 and 5 units, based on observations that insulin requirements are generally less than 1 unit, even when blood glucose levels are high. As training progressed, we reduced the randomness through an exponential decay factor of 0.995, gradually shifting from random actions to more deterministic decisions guided by the policy network.

A critical aspect of our training methodology was the computation of advantages and the calculation of loss. The advantage function was instrumental in assessing the relative benefits of actions taken by the policy network. By juxtaposing the predicted values against the actual returns, we directed the network towards actions yielding higher returns than those anticipated by the critic. The loss functions were tailored for each network component. For the actor, we used the negative log probability of the actions, weighted by the advantage, to promote actions leading to higher returns. For the critic, we employed the mean squared error between the predicted values and the actual returns, aiming to refine the value predictions. This dual approach to loss calculation served as a robust feedback mechanism.

Through backpropagation, adjustments were continually made to the policy network, minimizing the loss and enhancing the agent's decision-making capabilities. This iterative process of learning and refinement was central to achieving our goal of maintaining the patient's blood glucose levels within the desired range.

Evaluation

Our evaluation process involved training the model separately on three patient profiles: Adult#008, Adolescent#008, and Child#008. And then, we tested the adult model on adult#008 himself/herself, adult#006, and adult#002. We evaluated the similar pattern in the other two age groups as well. This was done 20 simulation rounds on each patient to determine the 3-day survival rate and the percentage of time patients remained in a healthy blood glucose (BG) range (70-180).

As shown in figure 3, the 3-day survival rate data, we observed that training on a specific patient did not necessarily translate to better performance on that individual compared to others in the same age group. For example, after training on Adult#008, the model only had a 55% success rate in ensuring Adult#008’s survival over three days, while other adults in the same group had a 100% survival rate.

A concerning finding is the 0% survival rate for Child#008 across all evaluations. This could indicate a higher volatility in children's responses to insulin, suggesting that the current training regimen of 1000 iterations may be insufficient for this age group. Adjusting the model or increasing the training iterations could be necessary to address this issue.

Regarding the figure 4, healthy percentage, adults generally maintained a healthy BG level more consistently than adolescents and children. This could imply that adults have a more stable response to insulin. One critical goal for future improvements is to ensure that all patients, regardless of age, can maintain their BG levels within the healthy range with minimal fluctuations.

In light of these findings, further research could focus on tailoring the training process to better accommodate the unique physiological responses of different age groups.

3-day Survival Rate

	Adult #008	Adolescent#008	Child #008
Eval 1	55%	100%	0%
Eval 2	100%	100%	0%
Eval 3	100%	55%	0%

Figure 3: 3-day Survival Rate

Healthy Percentage (70-180)

	Adult #008	Adolescent#008	Child #008
Eval 1	52.7%	37.2%	56.1%
Eval 2	47.5%	59.2%	47.5%
Eval 3	83.1%	34.9%	46.2%

Figure 4: Healthy Percentage

Insights

We assessed each model with three patients from distinct age groups, with visualizations for evaluation in the Google Drive folders we shared at the beginning. Generally, the agent demonstrated effective control strategies, yet there is room for improvement, particularly across different age demographics.

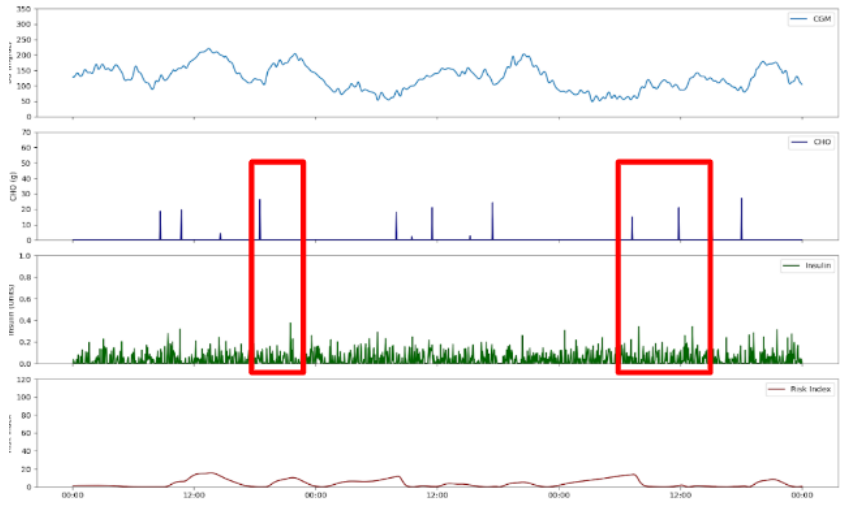


Figure 5: Evaluation on Adult 2 Simulation 5

The Figure 5 sequentially presents blood glucose levels, carbohydrate intake, insulin dosages, and risk index. Notably, the agent administers insulin after high carbohydrate consumption, indicated by red boxes. However, this response is reactive rather than proactive, as insulin is administered only after significant glucose elevations due to carbohydrate consumption. This approach could be refined for real-world application, considering the health risks associated with delayed responses to fluctuating blood glucose levels. Enhancing the

model to include historical blood glucose and carbohydrate intake data could possibly help address this issue.

The graph also reveals that blood glucose levels frequently deviate from the ideal range of 70-180 mg/dL. This may stem from an inadequate penalty for levels outside this range. Additionally, our model often terminates early in hypoglycemia scenarios due to frequent insulin injections at three-minute intervals. This kind of frequent dosing, despite being minimal, results in a gradual decrease in blood glucose, leaving high blood glucose states relatively underexplored. These states, currently without significant penalties, may cause the agent to result in higher-than-optimal glucose levels. Implementing a tiered negative reward [16] for deviations from the target range could mitigate this issue.

$$f_R(G_k) = \begin{cases} 0.5, & 70 \leq G_k \leq 180, \\ -0.8, & 180 < G_k \leq 300, \\ -1, & 300 < G_k \leq 350, \\ -1.5, & 30 \leq G_k < 70 \\ -2, & \text{else.} \end{cases}$$

Example reward function from [16]

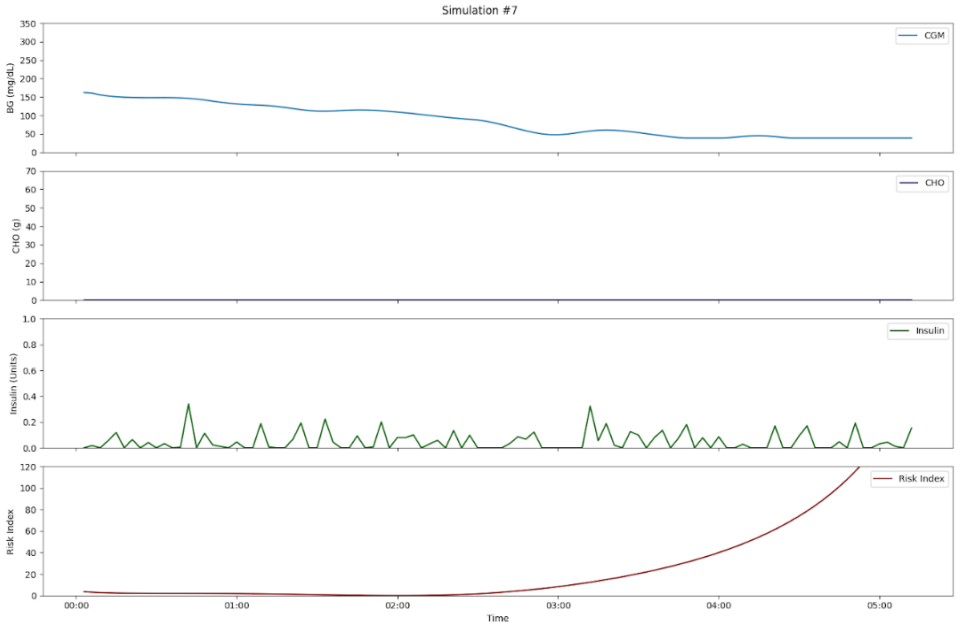


Figure 6: Evaluation on Children 8 Simulation 7

Children's evaluations, in particular, showed suboptimal outcomes, as illustrated in Figure 6. Our findings suggest that children respond more rapidly and intensely to insulin, leading to quicker declines into hypoglycemia, often terminating the evaluation within six hours. Given children's heightened sensitivity to insulin, extending action intervals from three to fifteen minutes or more could more effectively regulate blood glucose within a safe range [13].

To further improve outcomes in children, we implemented two key strategies. Firstly, we largely reduced the initial standard deviation of the action. Secondly, we introduced a buffer to record blood glucose levels from the past hour, which then informed the computation of subsequent actions. These combined modifications led to a substantial 75% improvement in training rewards. The primary factor contributing to this enhancement was the reduction in the initial standard deviation. While this is a learnable parameter, its reduction was subtle over 1000 iterations. In the original model, the standard deviation was initialized at e^-2, and after 1000 iterations, it remained around 0.1, substantially higher than the mean action value, thereby introducing excessive noise. In the revised model, we initialized the standard deviation at e^-18, effectively mitigating this issue. This approach should be beneficial not only for children but also for other age groups.

Limitation & Future Work

In our study, we identified several key limitations and directions for future enhancements. The current state representation in our model is solely based on single blood glucose level, omitting critical factors like historical BG level, insulin quantity, insulin sensitivity and hormonal levels that significantly influence physiological responses. This limitation raises concerns about the potential overfitting of our neural network, given the relatively small state space. Exploring simpler models or incorporating time-series analysis with Recurrent Neural Networks (RNNs) with more state variables could mitigate this risk [7]. Additionally, while we employed Actor-Critic methods, experimenting with more algorithms such as Proximal Policy Optimization (PPO) or Soft Actor-Critic (SAC) might enhance performance for this complex task.

Looking ahead, there is a compelling opportunity to evolve our models for proactive intervention. Future iterations could aim to predict and manage insulin doses preemptively by learning patterns in carbohydrate intake of each patient. Additionally, we are currently training the model on a single patient at a time. To enhance its generalizability, we may employ population-based training or transfer learning techniques using data from multiple patients.

Furthermore, ensuring that our models are designed and validated in compliance with healthcare regulations is crucial for their readiness for clinical trials and real-world applications. Finally, integrating techniques to increase the explainability of our AI models will be essential to build trust among healthcare practitioners and patients, an important step towards the practical deployment of such systems in medical settings.

Conclusion

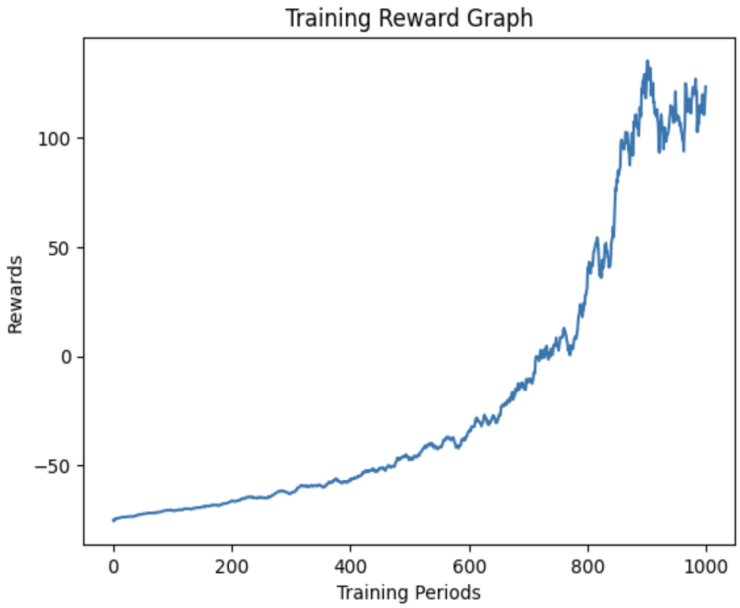
In conclusion, this study presents our experimentations in the management of Type 1 Diabetes using a Deep Reinforcement Learning approach, particularly an Actor-Critic model. Our research demonstrates the potential of this method to maintain blood glucose levels within a healthy range across different patient profiles, including adults, adolescents, and children. While our model showed promising results in terms of improving training rewards and managing blood glucose levels, it also highlighted the unique challenges presented by different age groups, especially in children. However, limitations in state representation and the need for more advanced algorithms and proactive intervention strategies were identified. Future work will focus on expanding the state variables, exploring more algorithms, and enhancing the model's generalizability through population-based training and transfer learning. Ensuring regulatory compliance and increasing model explainability will be crucial for clinical application.

References

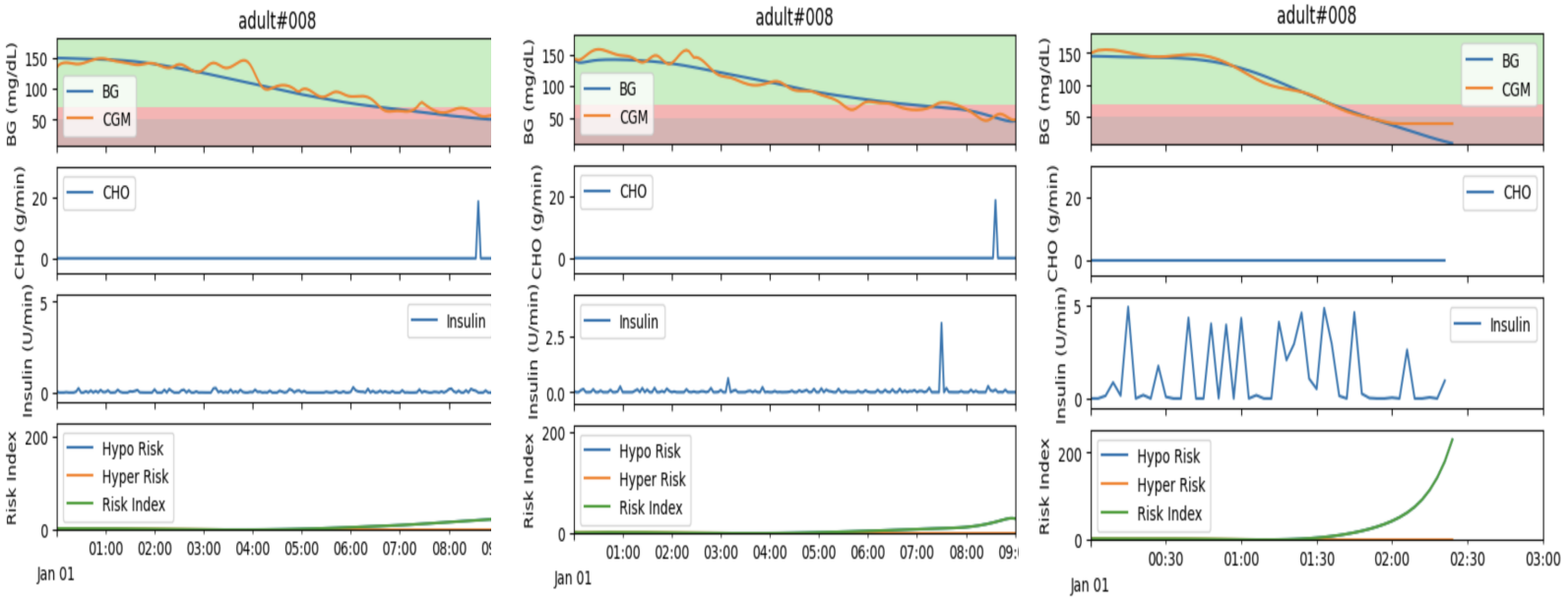
1. “Diabetes - Symptoms and causes.” *Mayo Clinic*, 15 September 2023, https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444. Accessed 12 December 2023.
2. Misso, Marie L et al. “Continuous subcutaneous insulin infusion (CSII) versus multiple insulin injections for type 1 diabetes mellitus.” The Cochrane database of systematic reviews ,1 CD005103. 20 Jan. 2010, doi:10.1002/14651858.CD005103.pub2
3. Reiterer, F.; Freckmann, G.; del Re, L. Impact of Carbohydrate Counting Errors on Glycemic Control in Type 1 Diabetes. IFAC-PapersOnLine 2018, 51, 186–191
4. Messer, L.H.; Forlenza, G.P.; Sherr, J.L.; Wadwa, R.P.; Buckingham, B.A.; Weinzimer, S.A.; Maahs, D.M.; Slover, R.H. Optimizing hybrid closed-loop therapy in adolescents and emerging adults using the MiniMed 670G system. Diabetes Care 2018, 41, 789–796.
5. Renard E, Place J, Cantwell M, Chevassus H, Palerm CC. Closed-loop insulin delivery using a subcutaneous glucose sensor and intraperitoneal insulin delivery: feasibility study testing a new model for the artificial pancreas. Diabetes care. 2010;33(1):121–127. doi: 10.2337/dc09-1080
6. Harvey, R.A.; Dassau, E.; Bevier, W.C.; Seborg, D.E.; Jovanovič, L.; Doyle, F.J., III; Zisser, H.C. Clinical evaluation of an automated artificial pancreas using zone-model predictive control and health monitoring system. Diabetes Technol. Ther. 2014, 16, 348–357.
7. Magni L, Raimondo DM, Bossi L, Man CD, Nicolao GD, Kovatchev B, et al. Model Predictive Control of Type 1 Diabetes: An in Silico Trial. Journal of Diabetes Science and Technology. 2007;1(6):804–812. doi: 10.1177/193229680700100603
8. Yamagata T, Ayobi A, O’Kane A, Katz D, Stawarz K, Marshall P, et al. Model-Based Reinforcement Learning for Type 1 Diabetes Blood Glucose Control. In: Singular Problems for Healthcare Workshop at ECAI 2020; Conference date: 29-08-2020 Through 08-09-2020; 2020. p. 1–14.
9. Ferdinando MD, Pepe P, Gennaro SD, Palumbo P. Sampled-Data Static Output Feedback Control of the Glucose-Insulin System. IFAC-PapersOnLine. 2020;53(2):3626–3631. doi: 10.1016/j.ifacol.2020.12.2044
10. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: A systematic review. Artificial Intelligence in Medicine. 2020;104:101836. doi: 10.1016/j.artmed.2020.101836
11. Fox I, Lee J, Pop-Busui R, Wiens J. Deep reinforcement learning for closed-loop blood glucose control. In: Machine Learning for Healthcare Conference. PMLR; 2020. p. 508–536.
12. Ngo PD, Wei S, Holubová A, Muzik J, Godtliebsen F. Control of Blood Glucose for Type-1 Diabetes by Using Reinforcement Learning with Feedforward Algorithm. Computational and Mathematical Methods in Medicine. 2018;2018:1–8. doi: 10.1155/2018/4091497
13. Viroonluecha P, Egea-Lopez E, Santa J (2022) Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning. PLoS ONE 17(9): e0274608. https://doi.org/10.1371/journal.pone.0274608
14. Man, Chiara Dalla et al. “The UVA/PADOVA Type 1 Diabetes Simulator: New Features.” Journal of diabetes science and technology vol. 8,1 (2014): 26-34. doi:10.1177/1932296813514502
15. Battelino, Tadej et al. “Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range.” Diabetes care vol. 42,8 (2019): 1593-1603. doi:10.2337/dci19-0028

16. Zhu, Taiyu et al. “An Insulin Bolus Advisor for Type 1 Diabetes Using Deep Reinforcement Learning.” Sensors (Basel, Switzerland) vol. 20,18 5058. 6 Sep. 2020, doi:10.3390/s20185058
17. Jinyu Xie. Simglucose v0.2.1 (2018) [Online]. Available: https://github.com/jxx123/simglucose. Accessed on: 12-14-2023

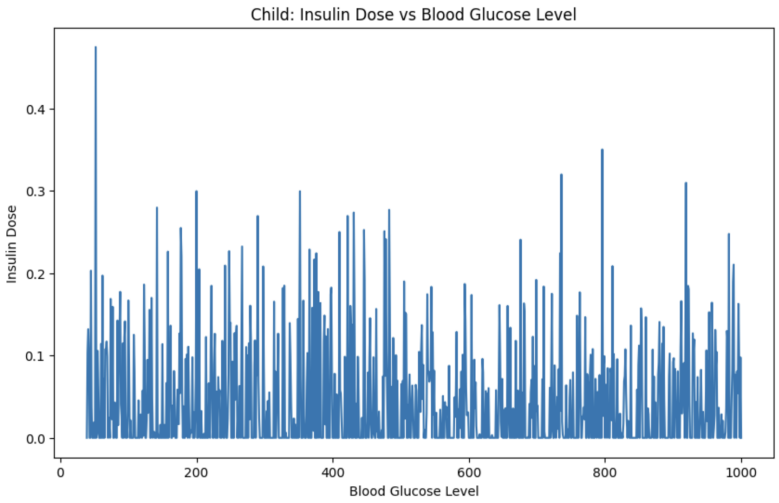
Appendix



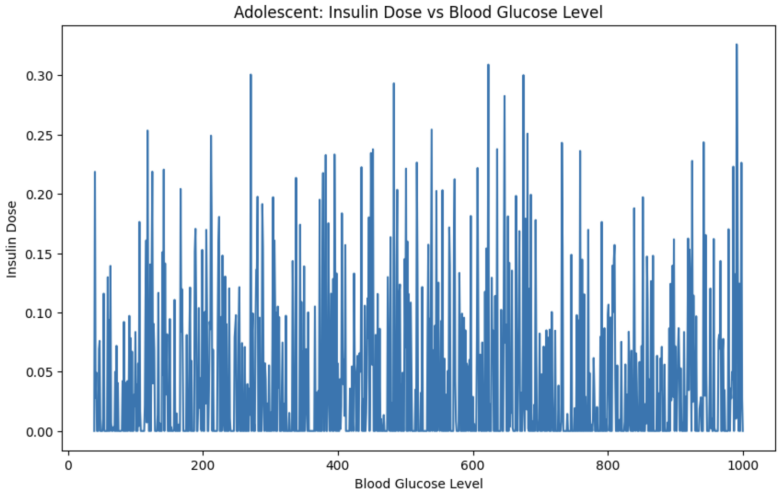
Training Reward Example: Adolescent#008



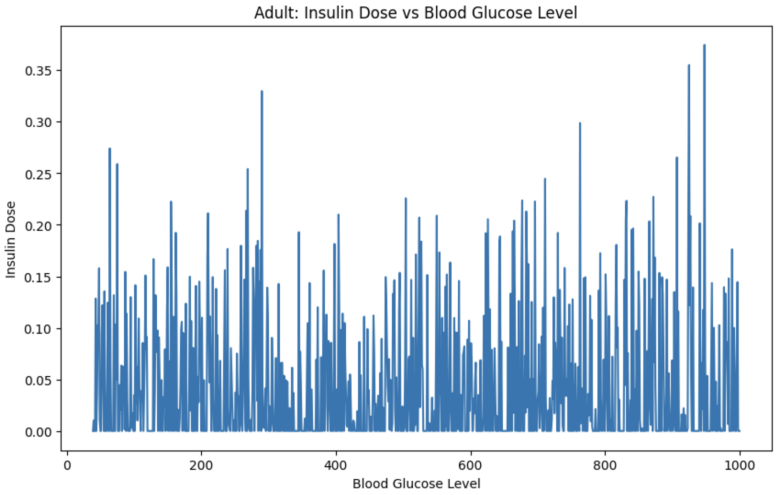
1000 iteration Training Progress Example: Adult#008 (progress from left to right)



Policy Visualization: Mean of Action at different BG for child



Policy Visualization: Mean of Action at different BG for adolescent



Policy Visualization: Mean of Action at different BG for adult