


Data-Driven Robust Control Using Reinforcement Learning

Phuong D. Ngo ¹, Miguel Tejedor ^{1,*} and Fred Godtlibsen ^{2,*}¹ Norwegian Centre for E-Health Research, 9019 Tromsø, Norway; phuong.dinh.ngo@ehealthresearch.no² Department of Mathematics and Statistics, Faculty of Science and Technology, UiT The Arctic University of Norway, 9019 Tromsø, Norway

* Correspondence: miguel.tejedor@ehealthresearch.no (M.T.); fred.godtlibsen@uit.no (F.G.)

Abstract: This paper proposes a robust control design method using reinforcement learning for controlling partially-unknown dynamical systems under uncertain conditions. The method extends the optimal reinforcement learning algorithm with a new learning technique based on the robust control theory. By learning from the data, the algorithm proposes actions that guarantee the stability of the closed-loop system within the uncertainties estimated also from the data. Control policies are calculated by solving a set of linear matrix inequalities. The controller was evaluated using simulations on a blood glucose model for patients with Type 1 diabetes. Simulation results show that the proposed methodology is capable of safely regulating the blood glucose within a healthy level under the influence of measurement and process noises. The controller has also significantly reduced the post-meal fluctuation of the blood glucose. A comparison between the proposed algorithm and the existing optimal reinforcement learning algorithm shows the improved robustness of the closed-loop system using our method.

Keywords: reinforcement learning; robust control; data-driven



Citation: Ngo, P.D.; Tejedor, M.; Godtlibsen, F. Data-Driven Robust Control Using Reinforcement Learning. *Appl. Sci.* **2022**, *12*, 2262. <https://doi.org/10.3390/app12042262>

Academic Editors: Wen-June Wang, Chung-Hsun Sun and Luigi Fortuna

Received: 11 January 2022

Accepted: 18 February 2022

Published: 21 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Control of unknown dynamic systems with uncertainties is a challenge because exact mathematical models are often required. Since many processes are complicated, nonlinear, and varying with time, a control algorithm that does not depend on a mathematical model and can adapt to time-varying conditions is required. A popular approach is to develop a universal approximator for predicting the output of unknown systems [1]. Control algorithms can then be designed based on the parameters of the approximator. Based on this approach, many control techniques have been proposed using machine learning models such as neural networks and fuzzy logic. For example, Goyal et al. [2] proposed a robust sliding mode controller which can be designed from Chebyshev neural networks. Chadli and Guerra [3] introduced a robust static output feedback controller for Takagi Sugeno fuzzy models. Ngo and Shin [4] proposed a method to model unstructured uncertainties and a new Takagi Sugeno fuzzy controller using type-2 fuzzy neural networks.

However, obtaining a good approximator requires a significant amount of training data, especially for a complicated model with high-dimensional state spaces or with many inputs and outputs. The data-driven model must also be updated frequently for time-varying systems. In addition, many control design techniques assume uncertainties as functions of system parameters. However, in many cases, the causes of uncertainties are unknown and unstructured. With the development of data science and machine learning, model-free approaches such as reinforcement learning (RL) have emerged as an effective method to control unknown nonlinear systems [5–8]. The principle of RL is based on the interaction between a decision-making agent and its environment [9], and the actor–critic method is often used as the RL framework for many control algorithms. In the actor–critic framework, the critic agent uses current state information of the environment in order to

update the value or action value function. The actor agent then uses the value or action value function to calculate the optimal action.

It can be seen that many data-driven algorithms lack stability analysis of the closed-loop systems. Among recent techniques focusing on the robustness of control algorithms, Yang et al. [10] presented an off-policy reinforcement learning (RL) solution to solve robust control problems for a certain class of unknown systems with structured uncertainties. In [11], a robust data-driven controller was proposed based on the frequency response of multivariable systems and convex optimization. Based on data-driven tuning, Takabe et al. [12] introduced a detection algorithm suitable for massive overloaded multiple-input multiple-output systems. In more recent works, Na et al. [13] proposed an approach to address the output-feedback robust control for continuous-time uncertain systems using online data-driven learning, while Makarem et al. [14] used data-driven techniques for iterative feedback tuning of a proportional-integral-derivative controller's parameters. However, in many cases, stability can only be ensured for specific systems where uncertainties are structured. In addition, the value function must be estimated accurately, which is difficult to achieve, especially at the beginning of the control process when the agent has just started interacting with the environment. Additionally, in many applications, the state space is either continuous or high-dimensional. In these cases, the value function approximation is often inaccurate, potentially leading to instability. Therefore, new RL approaches for which stability can be guaranteed under uncertain conditions are essential if algorithms are to be used in critical and safety-demanding systems.

Type 1 diabetes is a disease caused by the lack of insulin secretion. The condition results in uncontrolled increase of blood glucose level if the patients are not provided with insulin doses. High blood glucose level can lead to both acute and chronic complications, and eventually result in failure of various organs. One of the major challenges in controlling the blood glucose is that the biochemical and physiologic kinetics of insulin and glucose is complicated, nonlinear, and only approximately known [15]. Additionally, the stability of the control system is essential in this case since unstable control effort will lead to life-threatening condition for the patients.

This paper proposes a novel method to capture uncertainty in estimating the value function in reinforcement learning based on observation data. Using the uncertainty information, the paper also presents a new technique to improve the policy while guaranteeing the stability of the closed-loop system under uncertainty conditions for partially-unknown dynamical systems. The proposed methodology is applied to a blood glucose model for testing its effectiveness in controlling the blood glucose level in patients with Type 1 diabetes.

Structure of Paper

The content of the paper is organized as follows. Section 2 describes the proposed robust RL algorithm. Section 3 shows the simulation results of the methodology. The conclusions are given in Section 4.

2. Materials and Methods

In this section we present the robust RL method and the simulation setup used for evaluation of the algorithm.

2.1. Robust Control Using Reinforcement Learning

In this paper, a class of dynamical systems is considered, which can be described by the following linear state-space equation:

$$\dot{x}(t) = Ax(t) + Bu(t), \quad (1)$$

where $x \in \mathbb{R}^n$ is the vector of n state variables, $u \in \mathbb{R}^m$ is the vector of m control inputs, $A \in \mathbb{R}^{n \times n}$ is the state matrix, and $B \in \mathbb{R}^{n \times m}$ is the input matrix. It is assumed that matrix A is a squared $n \times n$ unknown matrix and the system (A and B) is stabilized. Our target

is to derive a control algorithm $u(t)$ that can regulate the state variables contained in $x(t)$ based on input and output data without knowing matrix A .

As an RL framework, the proposed robust control algorithm consists of an agent that takes actions and learns the consequences of its actions in an unknown environment. The environment is defined by a state vector $x(t)$ that describes its states at time t . The action at time t is represented by $u(t)$. As a consequence of the action, a cost $r(t)$ is incurred and accumulated. The cost function $r(t)$ is assumed to be known and predefined as a function of the current state and action. The objective of the learning process is to minimize the total cost accumulation in the future.

At each decision time point, the agent receives information about the state of the environment and chooses an action. The environment reacts to this action and transitions to a new state, which determines whether the agent receives a positive or negative reinforcement. Current RL techniques propose optimal actions by minimizing the predicted cost accumulation. However, uncertainties due to noises in the data or inaccurate estimation of the cost accumulation can lead to suboptimal actions and even unstable responses. Our target is to provide the agent with a robust and safe action that can guarantee the reduction of the future cost accumulation in the presence of uncertainties. The action calculated by the proposed algorithm may not be the optimal action that reduces the cost in the fastest way, but it can always guarantee the stability of the system, which is imperative in many critical applications.

2.1.1. Estimation of the Value Function by the Critics

In the RL context, the accumulation of cost over time, when starting in the state $x(t)$ and following policy π , is defined as the value function of policy π , i.e.,

$$V^\pi(x(t)) = E_\pi \left\{ \int_t^\infty \gamma^{\tau-t} r(\tau) d\tau \right\}, \tag{2}$$

where γ is the discount factor. The cost $r(t)$ is assumed to be a quadratic function of the states:

$$r(t) = x^T(t)Qx(t), \tag{3}$$

where the positive definite matrix $Q \in \mathbb{R}^{n \times n}$ is symmetric, positive semidefinite (since the cost is assumed to be non-negative), and contains the weighting factors of the variables that are minimized.

In order to facilitate the formulation of the stability condition in the form of linear matrix inequalities (LMI), the value function $V(x(t))$ is approximated by a quadratic function of the states:

$$V^\pi(x(t)) \approx x^T(t)Px(t), \tag{4}$$

where the kernel matrix $P \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite (since matrix Q in the cost function is symmetric and positive semidefinite).

By using the Kronecker operation, the approximated value function can be expressed as a linear combination of the basis function $\phi(x(t)) = (x(t) \otimes x(t))$:

$$\begin{aligned} V^\pi(x(t)) &\approx x^T(t)Px(t) = \text{vec}(P)^T(x(t) \otimes x(t)) \\ &= w^T(x(t) \otimes x(t)) = w^T\phi(x(t)), \end{aligned} \tag{5}$$

where w is the parameter vector, $\phi(x(t))$ is the vector of basis functions, and \otimes is the Kronecker product. The transformation between w and P can be performed as follows:

$$w = \text{vec}(P) = [P_{11}, P_{21}, \dots, P_{n1}, P_{12}, \dots, P_{1n}, P_{nn}]^T, \tag{6}$$

where $P_{i,j}$ is the element of matrix P in the i th row and j th column. With T as the interval time for data sampling, the integral RL Bellman equation can be used to update the value function [8]:

$$V^\pi(x(t)) = \int_t^{t+T} \gamma^{\tau-t} r(\tau) d\tau + V^\pi(x(t+T)). \tag{7}$$

By using the quadratic cost function (Equation (3)) and the approximated value function (Equation (5)), the integral RL Bellman equation can be written as follows:

$$x^T(t)Px(t) = \int_t^{t+T} x(\tau)^T Qx(\tau) d\tau + x^T(t+T)Px(t+T) \tag{8}$$

or

$$w^T \phi(x(t)) = \int_t^{t+T} x(\tau)^T Qx(\tau) d\tau + w^T \phi(x(t+T)). \tag{9}$$

At each iteration, n samples along the state trajectory are collected ($x^{(1)}(t), x^{(2)}(t), \dots, x^{(n)}(t)$). The mean value of w can be obtained by using least-square technique:

$$\hat{w} = (XX^T)XY, \tag{10}$$

where

$$X = [\phi_\Delta^1 \quad \phi_\Delta^2 \quad \dots \quad \phi_\Delta^N]^T, \tag{11}$$

$$\phi_\Delta^i = \phi(x^i(t)) - \phi(x^i(t+T)), \tag{12}$$

$$Y = [d(x^1(t)) \quad d(x^2(t)) \quad \dots \quad d(x^n(t))]^T \tag{13}$$

and

$$d(x^i(t)) = \int_t^{t+T} x^i(\tau)^T Qx^i(\tau) d\tau \tag{14}$$

with $i = 1, 2, \dots, N$.

The confidence interval for the coefficient $w^{(j)}$ is given by

$$w^{(j)} \in [\hat{w}^{(j)} - q_{1-\frac{\theta}{2}} \sqrt{\tau_j \hat{\sigma}^2}, \hat{w}^{(j)} + q_{1-\frac{\theta}{2}} \sqrt{\tau_j \hat{\sigma}^2}], \tag{15}$$

where $1 - \theta$ is the confidence level, $q_{1-\frac{\theta}{2}}$ is the quantile function of standard normal distribution, τ_j is the j th element on the diagonal of $(XX^T)^{-1}$, and $\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n-p}$, with $\epsilon = Y - \hat{w}X$. From that, the uncertainty Δw is defined as the deviation interval around the nominal value:

$$\Delta w = \left[-q_{1-\frac{\theta}{2}} \sqrt{\tau_j \hat{\sigma}^2}, -q_{1-\frac{\theta}{2}} \sqrt{\tau_j \hat{\sigma}^2} \right]. \tag{16}$$

Matrices \hat{P} and ΔP can be obtained by placing elements of \hat{w} and Δw into columns.

2.1.2. Policy Improvement by the Actor

Linear feedback controllers have been widely used as a stabilization tool for nonlinear systems where dynamic behavior is considered approximately linear around the operating condition [16–18]. Hence, in this paper, we use linear functions of the states with gain K_i as the control policy at iteration i :

$$u(t) = \pi(x(t)) = -K_i x(t), \tag{17}$$

and the level of uncertainty is constant during the controlling process. The task of the actor is to robustly improve the current policy such that the value function is guaranteed to be reduced during the next policy implementation. If the following differential inequality is satisfied:

$$\dot{V}_i(x(t)) + \alpha V_i(x(t)) \leq 0 \tag{18}$$

with some positive constant α , then by using the comparison lemma (Lemma 3.4 in [19]), the derivative of function $\dot{V}_i(x(t))$ can be bounded by

$$\dot{V}_i(x(t)) \leq V_i(x(t_0))e^{-\alpha(t-t_0)}. \tag{19}$$

Therefore, maximizing the rate α will ensure a maximum exponential decrease in the value of $\dot{V}_i(x(t))$.

The following part shows the main results of the paper, which describe how the policy gain can be improved during the learning process. Derivations of the results are provided in the stability analysis (Section 2.1.3).

Definition 1. Assume A is a square matrix with dimension $n \times n$ and x is a vector with dimension $n \times 1$. The maximize operation on matrix A and vector x is defined as follows:

$$\text{maximize}(A, x) = C, \tag{20}$$

where

$$C_{ij} = \begin{cases} \max(A_{ij}) & \text{if } x_i x_j \geq 0 \\ \min(A_{ij}) & \text{if } x_i x_j < 0 \end{cases} \text{ with } i, j = 1 \dots n. \tag{21}$$

Assuming that the sign of all state variables cannot be changed between each policy update interval, the improved policy K_{i+1} can be obtained by minimizing α subject to

$$\begin{bmatrix} V & K_{i+1}^T B^T \\ BK_{i+1} & -\gamma I \end{bmatrix} \leq 0 \tag{22}$$

and

$$\begin{bmatrix} \zeta & K_{i+1} \\ K_{i+1}^T & -I \end{bmatrix} \leq 0, \tag{23}$$

where:

$$V = M + \Delta P_{i,\max}^T \Delta P_{i,\max} \gamma - \hat{P}_i B K_{i+1} - K_{i+1}^T B^T \hat{P}_i + \alpha (\hat{P}_i + \frac{1}{2} \Delta P_{i,\max}^T \Delta P_{i,\max} + I) \tag{24}$$

and

$$M = -Q - K_i^T R K_i + \hat{P}_i B K_i + K_i^T B^T \hat{P}_i + H_i \tag{25}$$

with $\Delta P_{i,\max} = \text{maximize}(\Delta P_i, x)$ and $H_i = \text{maximize}(\Delta P_i B K_i + K_i^T B^T \Delta P_i, x)$, utilizing the maximize operation defined in Definition 1. Inequality (22) provides the stable condition and its derivation is provided in Section 2.1.3. Inequality (23) provides the upper bound for the updated gain K_{i+1} through the user-defined parameter ζ . The value of $-\zeta$ limits the maximum L^2 gain of K_{i+1} since inequality (23) is equivalent to $K_{i+1} K_{i+1}^T \leq -\zeta$.

2.1.3. Stability Analysis

With the control policy as described in Equation (17), the equation for the closed-loop system can be derived as follows:

$$\dot{x}(t) = Ax(t) - BKx(t) = (A - BK)x(t). \tag{26}$$

Lemma 1. Assuming that the closed-loop system described by Equation (26) is stable, solving for P in Equation (8) is equivalent to finding the solution of the underlying Lyapunov equation [8]:

$$P(A - BK) + (A - BK)^T P = -Q. \tag{27}$$

Proof of Lemma 1. We start with Equation (27) and try to prove that matrix P is also the solution of Equation (8). Consider $V(x(t)) = x^T(t)Px(t)$, where P is the solution of Equation (27):

$$\begin{aligned} \dot{V}(x(t)) &= \frac{d(x^T(t)Px(t))}{dt} \\ &= \dot{x}^T(t)Px(t) + x^T(t)P\dot{x}(t) \\ &= x^T(t) \left[(A - BK)^T P + P(A - BK) \right] x(t) \\ &= -x^T(t)Qx(t) \quad (\text{using Equation (27)}). \end{aligned} \tag{28}$$

Since the closed-loop system is stable, the Lyapunov Equation (27) has a unique solution, $P_i > 0$. From (28), this solution will satisfy

$$\frac{d(x^T(t)P_i x(t))}{dt} = -x^T(t)Qx(t), \tag{29}$$

which is equivalent to

$$x^T(t+T)Px(t+T) - x^T(t)Px(t) = \int_t^{t+T} -x^T(\tau)Qx(\tau)d\tau. \tag{30}$$

Therefore, P is also the solution of Equation (8). \square

Lemma 2. Given matrices E and F with appropriate dimensions, the following LMI can be obtained:

$$EF^T + FE^T \leq EE^T + FF^T. \tag{31}$$

Proof of Lemma 2. From the properties of matrix norm, we have

$$(E - F)(E - F)^T \geq 0, \tag{32}$$

which is equivalent to

$$EE^T + FF^T - EF^T - FE^T \geq 0 \tag{33}$$

or

$$EF^T + FE^T \leq EE^T + FF^T. \tag{34}$$

\square

Lemma 3. Given A as a square matrix with dimension $n \times n$ and x as a vector with dimension $n \times 1$, the following LMI can be obtained:

$$x^T Ax \leq x^T Cx, \tag{35}$$

where $C = \text{maximize}(A, x)$ as in Definition 1.

Proof of Lemma 3. We have

$$\begin{aligned} x^T Ax &= \sum_{i,j=1,2,\dots,n} a_{ij}x_i x_j \leq \sum_{i,j=1,2,\dots,n} |a_{ij}x_i x_j| \\ &= \sum_{i,j=1,2,\dots,n} c_{ij}x_i x_j = x^T Cx, \end{aligned} \tag{36}$$

where $c_{ij} = \begin{cases} \max(a_{ij}) & \text{if } x_i x_j \geq 0 \\ \min(a_{ij}) & \text{if } x_i x_j < 0 \end{cases}$ with $i, j = 1 \dots n$. \square

Theorem 1. Consider a dynamic system that can be represented by Equation (1) with the state matrix A unknown. Assume that the sign of all state variables cannot be changed between each

policy update interval and the estimated value function at iteration i is $V_i(x(t)) = x^T(t)P_i x(t)$ with $P_i = \hat{P}_i + \Delta P_i$. If

- The current control policy $u(t) = \pi_i(x(t)) = -K_i x(t)$ is stabilizing;
- The LMI given in (22) is satisfied with some positive constant γ ;

then the closed-loop system with the control policy $u(t) = -K_{i+1}x(t)$ is quadratic stable with convergence rate α .

Proof of Theorem 1. Since the current control policy is stable, the estimated parameter matrix P_i is positive definite. Hence, $V_i(x(t)) = x_t^T P_i x_t > 0$. Here, $V_i(x(t))$ is used as the Lyapunov function for the updated control policy $u(t) = \pi_{i+1}(x(t)) = -K_{i+1}x(t)$. For notation convenience, the state vector $x(t)$ and input vector $u(t)$ are denoted as x_t and u_t , respectively. By using Equation (27) in Lemma 1 and the representation $P_i = \hat{P}_i + \Delta P_i$, we can calculate the left side of Equation (18) as follows:

$$\begin{aligned} & \dot{V}_i(x(t)) + \alpha V_i(x(t)) \\ &= \dot{x}_t^T P_i x_t + x_t P_i \dot{x}_t + \alpha x_t^T P_i x_t \\ &= (Ax_t + Bu_t)^T P_i x_t + x_t P_i (Ax_t + Bu_t)^T + \alpha x_t^T P_i x_t \\ &= x_t^T [P_i(A - BK_{i+1}) + (A - BK_{i+1})^T P_i + \alpha P_i] x_t \\ &= x_t^T [P_i(A - BK_i) + (A - BK_i)^T P_i + \alpha P_i] x_t + x_t^T [P_i B(K_i - K_{i+1}) + (K_i - K_{i+1})^T B^T P_i + \alpha P_i] x_t \\ &= -x_t^T [Q + K_i^T R K_i] x_t + x_t^T [(\bar{P}_i + \Delta P_i) B(K_i - K_{i+1}) + (K_i - K_{i+1})^T B^T (\bar{P}_i + \Delta P_i) + \alpha \bar{P}_i + \alpha \Delta P_i] x_t \\ &= x_t^T [-Q - K_i^T R K_i + \bar{P}_i B K_i + K_i^T B^T \bar{P}_i + \alpha \bar{P}_i + \Delta P_i B K_i + K_i^T B^T \Delta P_i - \Delta P_i B K_{i+1} \\ &\quad - K_{i+1}^T B^T \Delta P_i - \bar{P}_i B K_{i+1} - K_{i+1}^T B^T \bar{P}_i + \alpha \Delta P_i] x_t. \end{aligned}$$

By using Lemma 3, we have the following inequality:

$$\Delta P_i B K_i + K_i^T B^T \Delta P_i \leq H_i, \tag{37}$$

and the following inequality can be obtained by Lemma 2:

$$\begin{aligned} -\Delta P_i B K_{i+1} - K_{i+1}^T B^T \Delta P_i &\leq \gamma \Delta P_i \Delta P_i^T + \frac{1}{\gamma} (B K_{i+1})^T (B K_{i+1}) \\ &\leq \gamma \Delta P_{i,\max} \Delta P_{i,\max}^T + \frac{1}{\gamma} K_{i+1}^T B^T B K_{i+1} \end{aligned} \tag{38}$$

Additionally,

$$\alpha \Delta P_i \leq \alpha \left(\frac{1}{2} \Delta P_i \Delta P_i^T + I \right) \leq \alpha \left(\frac{1}{2} \Delta P_{i,\max} \Delta P_{i,\max}^T + I \right), \tag{39}$$

where $H_i = \text{maximize}(\Delta P_i B K_i + K_i^T B^T \Delta P_i, x)$, and $\Delta P_{i,\max} = \text{maximize}(\Delta P_i, x)$, utilizing the maximize operator defined in Definition 1.

Hence, $\dot{V}_i(x(t)) + \alpha V_i(x(t))$ can be bounded by

$$\begin{aligned} \dot{V}_i(x(t)) + \alpha V_i(x(t)) &\leq x_t^T [-Q - K_i^T R K_i + \bar{P}_i B K_i + K_i^T B^T \bar{P}_i + \alpha \left(\bar{P}_i + \frac{1}{2} \Delta P_{i,\max} \Delta P_{i,\max}^T \right) \\ &\quad - \bar{P}_i B K_{i+1} - K_{i+1}^T B^T \bar{P}_i + \gamma \Delta P_{i,\max} \Delta P_{i,\max}^T + \frac{1}{\gamma} K_{i+1}^T B^T B K_{i+1}] x_t. \end{aligned} \tag{40}$$

Using the Lyapunov theory, the system will be quadratic stable with the convergent rate α if $\dot{V}_i(x(t)) \leq -\alpha V_i(x(t))$. This condition is satisfied if

$$x_t^T [-Q - K_i^T R K_i + \bar{P}_i B K_i + K_i^T B^T \bar{P}_i + \alpha \left(\bar{P}_i + \frac{1}{2} \Delta P_{i,\max} \Delta P_{i,\max}^T \right) - \bar{P}_i B K_{i+1} - K_{i+1}^T B^T \bar{P}_i + \gamma \Delta P_{i,\max} \Delta P_{i,\max}^T + \frac{1}{\gamma} K_{i+1}^T B^T B K_{i+1}] x_t \leq 0.$$

The above condition can be written in the matrix form, as shown in Theorem 1. □

By using Theorem 1, it can be seen that with the proposed improved policy, the closed-loop system will be asymptotically stable. It is also noted that Theorem 1 is also applicable for unknown nonlinear systems if they can be approximated by a linear state-space equation (Equation (1)) and if their nonlinearity is within the uncertainty bound ΔP calculated from Δw in Equation (16).

2.1.4. Robust Reinforcement Learning Algorithm

The robust RL algorithm for controlling partially unknown dynamically systems includes the following steps:

Initialization

(Step $i = 0$)

- Select an initial policy $u(t) = -K_0 x(t)$.

Estimation of the Value Function

(Step $i = 1, 2, \dots$)

- Apply the control action $u(t)$ based on the current policy $u(t) = -K_i x(t)$.
- At time $t + T$, collect and compute the dataset (X, Y) , which are defined in Equations (11) and (13).
- Update vector w by using the batch least-square method (Equation (10)).

Control Policy Update

- Transform vector w into the kernel matrix P using the Kronecker transformation.
- Update the policy by solving the LMI in Theorem 1.

Figure 1 shows the simplified diagram of the above algorithm. It is noted that the estimation of the value function is an on-policy learning since it updates $V^\pi(x(t))$ using the V-value of the next state and the current policy's action.

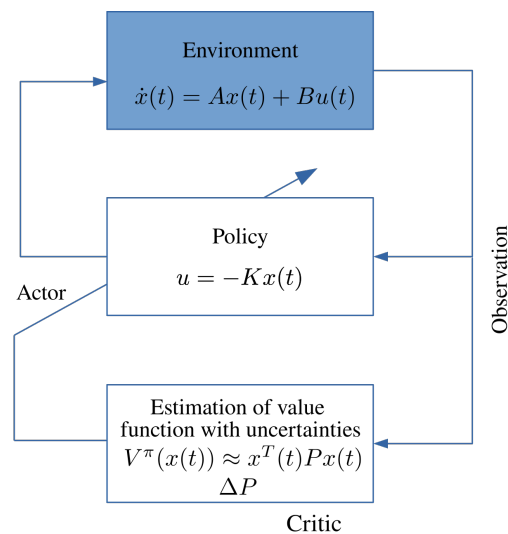


Figure 1. Data-driven robust reinforcement learning diagram.

2.1.5. Simulation Setup

A simulation study of the proposed robust RL controller was conducted on a glucose kinetics model, which can be described by [20–23]:

$$\frac{dD_1(t)}{dt} = A_G D(t) - \frac{D_1(t)}{\tau_D}, \tag{41}$$

$$\frac{dD_2(t)}{dt} = \frac{D_1(t)}{\tau_D} - \frac{D_2(t)}{\tau_D}, \tag{42}$$

$$\frac{dg(t)}{dt} = -p_1 g(t) - \chi(t)g(t) + \frac{D_2(t)}{\tau_D} + w(t) \tag{43}$$

and

$$\frac{d\chi(t)}{dt} = -p_2 \chi(t) + p_3 V(i(t) - i_b(t)). \tag{44}$$

In this model, parameter and variable descriptions can be found in Table 1 and Table 2, respectively. The values of the parameters are selected based on [20,21]. Variable $w(t)$ in Equation (43) is the process noise. The measured blood glucose value is affected by a random noise $v(t)$:

$$\hat{g}(t) = g(t) + v(t). \tag{45}$$

The inputs of the model are the amount of carbohydrate intake D and the insulin concentration i . The value of $i(t) - i_b(t)$ must be non-negative:

$$i(t) - i_b(t) \geq 0. \tag{46}$$

Table 1. Glucose kinetics model parameters.

Parameter	Description	Unit
p_1	Glucose effectiveness	min^{-1}
p_2	Insulin sensitivity	min^{-1}
p_3	Insulin rate of clearance	min^{-1}
A_G	Carbohydrate bioavailability	min^{-1}
τ_D	Glucose absorption constant	min
V	Plasma volume	mL
$i_b(t)$	Initial basal rate	$\mu\text{IU}/(\text{mL}\cdot\text{min})$

Table 2. Variables of the glucose kinetics model.

Variable	Description	Unit
D	Amount of carbohydrate intake	mmol/min
D_1	Glucose in compartment 1	mmol
D_2	Glucose in compartment 2	mmol
$g(t)$	Plasma glucose concentration	mmol/L
$\chi(t)$	Interstitial insulin activity	min^{-1}
$i(t)$	Plasma insulin concentration	$\mu\text{IU}/\text{mL}$

3. Results and Discussion

In order to evaluate the performance of the robust RL controller, we implemented the controller on the glucose kinetics model as described in the previous section under a daily scenario of patients with Type 1 diabetes. In order to make the scenario realistic, three different levels of uncertainties were used in the model. Uncertainties include process noise ($w(t)$) and measurement noise ($v(t)$). It is assumed that the noises are Gaussian distributions with standard deviations for each case as shown in Table 3.

Table 3. Standard deviations of process and measurement noises.

Uncertainty Case	Process Noise ($w(t)$)	Measurement Noise ($v(t)$)
1	0	0
2	0	0.002
3	0.1	0.1
4	0.1	1

3.1. Without Meal Intake

This part describes the simulation results during the fasting period (without meal intake). The purpose of the simulation is to compare the performances of the robust RL algorithm with the conventional optimal RL algorithm [24] in the nominal condition (uncertainty case 1). The initial blood glucose for both scenario was set at 290 mg/dL and the target blood glucose is 90 mg/dL. The initial policy at the beginning of the simulation was chosen as follows:

$$u(t) = -K_0x(t) = -0.27g(t) + 266.00\chi(t). \tag{47}$$

Figure 2 shows the comparison in blood glucose level between the robust RL and the optimal RL algorithm in the nominal condition. From the results, it can be seen that the robust RL successfully reduces the blood glucose level while the optimal RL becomes unstable when the blood glucose approaches the desired value. The instability of the optimal RL in this case can be explained by the nonlinearity of the system (due to the coupling term $\chi(t)g(t)$ in Equation (43)), the saturation of the insulin concentration (Equation (46)), and the lack of perturbed data when the blood glucose approaches the steady-state value. The insulin concentration during the simulation can be found in Figure 3. In this figure, the dotted blue line indicates the unstable insulin profile.

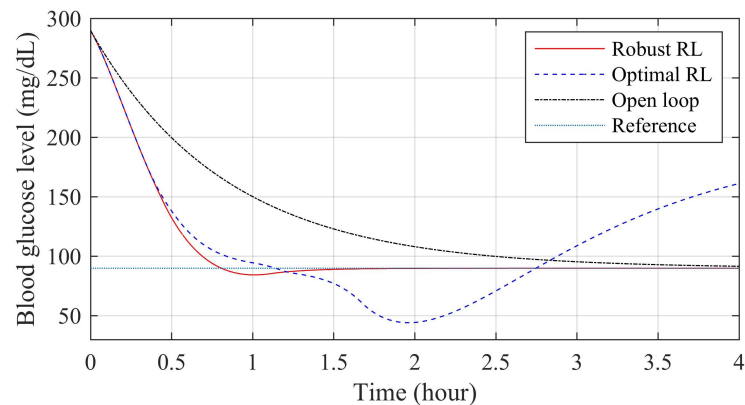


Figure 2. Comparison of blood glucose responses in nominal case without meal intake.

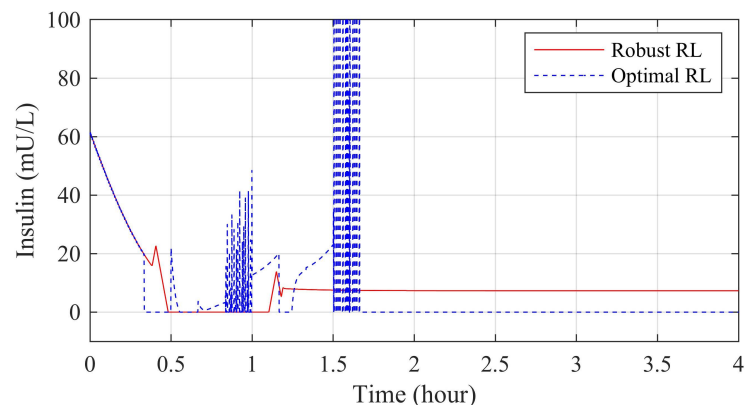


Figure 3. Comparison of insulin concentration in nominal case without meal intake.

Figure 4 shows the blood glucose responses from the robust RL in different uncertain conditions without meal intake. The results show similar and stable responses in all the uncertain conditions with settling time to the desired blood glucose level of approximately 45 min. The insulin concentration and the update of controller gains can be found in Figures 5 and 6.

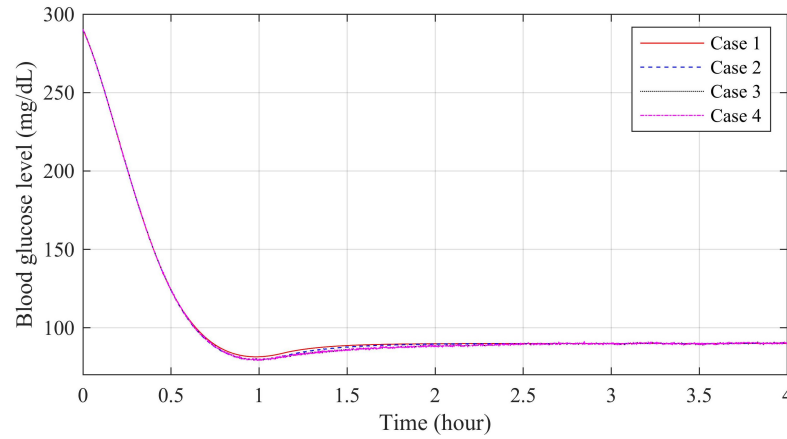


Figure 4. Comparison of blood glucose responses in uncertain cases without meal intake.

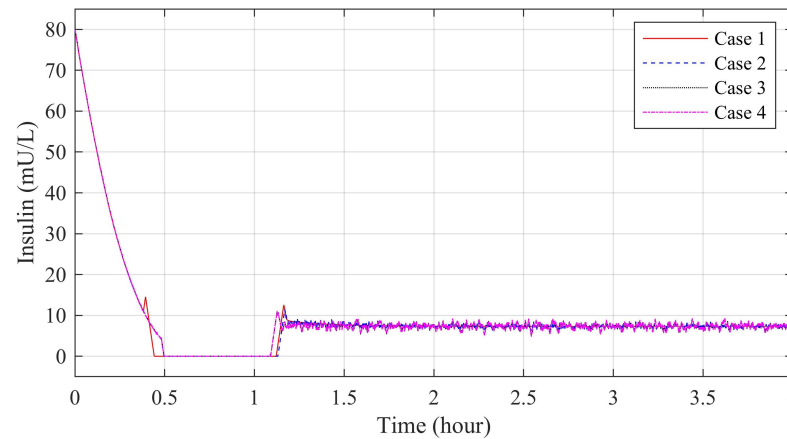


Figure 5. Insulin concentration in uncertain cases without meal intake.

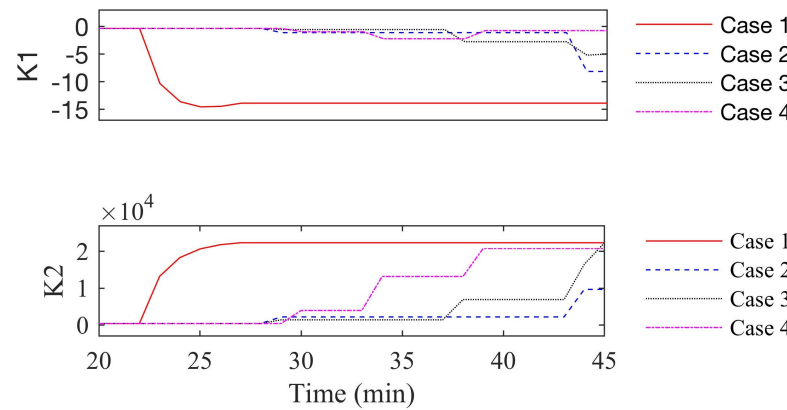


Figure 6. Update of controller gains during the learning process ($K1$ and $K2$ represent the first and second element of the controller gain vector K).

3.2. With Meal Intake

In this part, the performance of the robust RL controller was tested under conditions for which the system is subjected to meal intakes with the carbohydrate profile as shown in Figure 7.

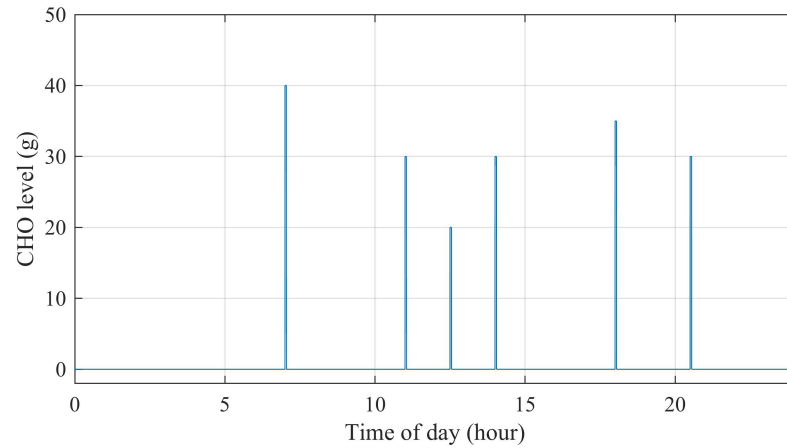


Figure 7. Carbohydrate intake per meal.

During the simulation period with meal intakes, blood glucose responses throughout the day of the robust RL control systems under four uncertain cases are shown in Figure 8. The insulin concentration during the process can also be found in Figure 9. The results show that the controller provides the most aggressive action under case 1 (no uncertainty) and the least aggressive action under case 4 (with highest level of measurement and process noises). This leads to the largest and smallest reduction of postprandial blood glucose in case 1 and case 4, respectively. Most importantly, the robust RL algorithm kept the system in stable condition and there is no hypoglycemia event during the simulation for all four cases under different level of uncertainties.

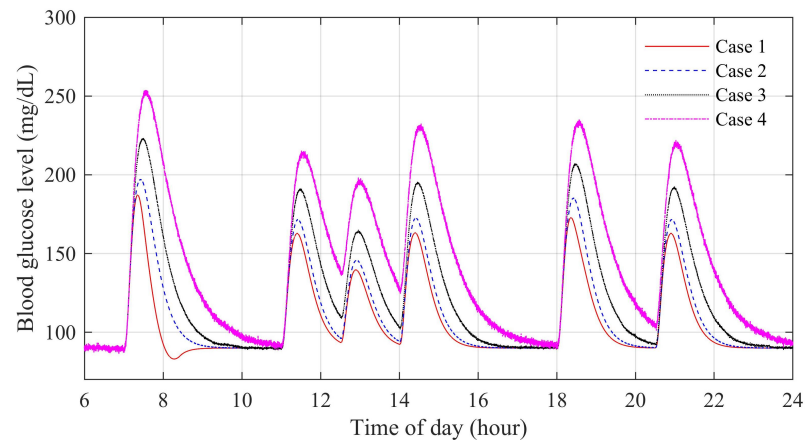


Figure 8. Blood glucose responses in simulation with meals.

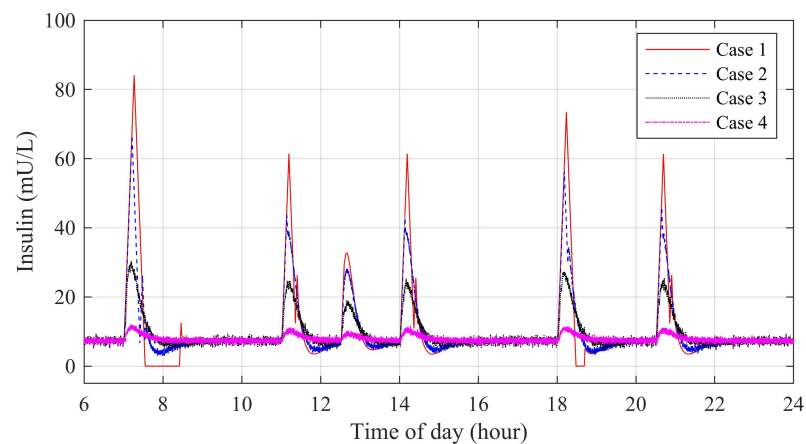


Figure 9. Insulin concentration in simulation with meals.

4. Conclusions

The paper proposes a robust RL algorithm for dynamical systems with uncertainties. The uncertainties can be approximated by the critic and represented in the value function. LMI techniques were used to improve the controller gain. The algorithm was simulated on a blood glucose model for patients with Type 1 diabetes. The objective of the simulation is to control and maintain a healthy blood glucose level. The comparison between the robust RL algorithm and the optimal RL algorithm shows a significant improvement in the robustness of the proposed algorithm. Simulation results show that the algorithm successfully regulated the blood glucose and kept the system stable under different levels of uncertainty.

Author Contributions: P.D.N. conceptualized ideas, developed algorithms, performed training and validation, numerical simulations, and led the writing process. M.T. contributed to the development of algorithms, provided critical feedback, analyzed results, and read and approved the final manuscript. F.G. acquired funding and resource, managed the project, and provided critical feedback leading to this publication. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tromsø Research Foundation under project “A smart controller for T1D using RL and SS representation” with grant/award number: A3327. The article processing charge was funded by a grant from the publication fund of UiT The Arctic University of Norway.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed or generated during the study is available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LMI Linear matrix inequalities
 RL Reinforcement learning

References

1. Lee, H.; Tomizuka, M. Robust adaptive control using a universal approximator for SISO nonlinear systems. *IEEE Trans. Fuzzy Syst.* **2000**, *8*, 95–106. [\[CrossRef\]](#)
2. Goyal, V.; Deolia, V.K.; Sharma, T.N. Robust sliding mode control for nonlinear discrete-time delayed systems based on neural network. *Intell. Control Autom.* **2015**, *06*, 75–83. [\[CrossRef\]](#)
3. Chadli, M.; Guerra, T.M. LMI solution for robust static output feedback control of discrete Takagi-Sugeno fuzzy models. *IEEE Trans. Fuzzy Syst.* **2012**, *20*, 1160–1165. [\[CrossRef\]](#)

4. Ngo, P.D.; Shin, Y.C. Modelling of unstructured uncertainties and robust controlling of nonlinear dynamic systems based on type-2 fuzzy basis function networks. *Eng. Appl. Artif. Intell.* **2016**, *53*, 74–85. [[CrossRef](#)]
5. Bothe, M.K.; Dickens, L.; Reichel, K.; Tellmann, A.; Ellger, B.; Westphal, M.; Faisal, A.A. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev. Med. Devices* **2013**, *10*, 661–673. [[CrossRef](#)]
6. De Paula, M.; Ávila, L.O.; Martínez, E.C. Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes. *Appl. Soft Comput. J.* **2015**, *35*, 310–332. [[CrossRef](#)]
7. Ouyang, Y.; He, W.; Li, X. Reinforcement learning control of a single-link flexible robotic manipulator. *IET Control Theory Appl.* **2017**, *11*, 1426–1433. [[CrossRef](#)]
8. Vrabie, D.; Vamvoudakis, K.G.; Lewis, F.L. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*, 1st ed.; Institution of Engineering and Technology: London, UK, 2012; Volume 81. [[CrossRef](#)]
9. Sutton, R.; Barto, A. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018; p. 129.
10. Yang, Y.; Guo, Z.; Xiong, H.; Ding, D.W.; Yin, Y.; Wunsch, D.C. Data-Driven Robust Control of Discrete-Time Uncertain Linear Systems via Off-Policy Reinforcement Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3735–3747. [[CrossRef](#)]
11. Karimi, A.; Kammer, C. A data-driven approach to robust control of multivariable systems by convex optimization. *Automatica* **2017**, *85*, 227–233. [[CrossRef](#)]
12. Takabe, S.; Imanishi, M.; Wadayama, T.; Hayakawa, R.; Hayashi, K. Trainable Projected Gradient Detector for Massive Overloaded MIMO Channels: Data-Driven Tuning Approach. *IEEE Access* **2019**, *7*, 93326–93338. [[CrossRef](#)]
13. Na, J.; Zhao, J.; Gao, G.; Li, Z. Output-Feedback Robust Control of Uncertain Systems via Online Data-Driven Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2650–2662. [[CrossRef](#)]
14. Makarem, S.; Delibas, B.; Koc, B. Data-Driven Tuning of PID Controlled Piezoelectric Ultrasonic Motor. *Actuators* **2021**, *10*, 148. [[CrossRef](#)]
15. Wang, Q.; Molenaar, P.; Harsh, S.; Freeman, K.; Xie, J.; Gold, C.; Rovine, M.; Ulbrecht, J. Personalized state-space modeling of glucose dynamics for type 1 diabetes using continuously monitored glucose, insulin dose, and meal intake: An extended Kalman filter approach. *J. Diabetes Sci. Technol.* **2014**, *8*, 331–345. [[CrossRef](#)] [[PubMed](#)]
16. Kothare, M.V.; Balakrishnan, V.; Morari, M. Robust constrained model predictive control using linear matrix inequalities. *Automatica* **1996**, *32*, 1361–1379. [[CrossRef](#)]
17. Fu, J.H.; Abed, E. Linear feedback stabilization of nonlinear systems. In Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, UK, 11–13 December 1991; pp. 58–63. [[CrossRef](#)]
18. Eker, S.A.; Nikolaou, M. Linear control of nonlinear systems: Interplay between nonlinearity and feedback. *AIChE J.* **2002**, *48*, 1957–1980. [[CrossRef](#)]
19. Khalil, H. *Nonlinear Systems*; Prentice Hall: Hoboken, NJ, USA, 2002; p. 218.
20. Bergman, R.N.; Ider, Y.Z.; Bowden, C.R.; Cobelli, C. Quantitative estimation of insulin sensitivity. *Am. J. Physiol. Endocrinol. Metab.* **1979**, *236*, E667. [[CrossRef](#)] [[PubMed](#)]
21. Hovorka, R.; Canonico, V.; Chassin, L.J.; Haueter, U.; Massi-Benedetti, M.; Orsini Federici, M.; Pieber, T.R.; Schaller, H.C.; Schaupp, L.; Vering, T.; et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiol. Meas.* **2004**, *25*, 905–920. [[CrossRef](#)] [[PubMed](#)]
22. Wilinska, M.E.; Chassin, L.J.; Schaller, H.C.; Schaupp, L.; Pieber, T.R.; Hovorka, R. Insulin kinetics in type-1 diabetes: Continuous and bolus delivery of rapid acting insulin. *IEEE Trans. Biomed. Eng.* **2005**, *52*, 3–12. [[CrossRef](#)] [[PubMed](#)]
23. Mösching, A. Reinforcement Learning Methods for Glucose Regulation in Type 1 Diabetes. Master’s Thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2016.
24. Ngo, P.D.; Wei, S.; Holubova, A.; Muzik, J.; Godtlielsen, F. Reinforcement-learning optimal control for type-1 diabetes. In Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, USA, 4–7 March 2018; pp. 333–336. [[CrossRef](#)]