A Proof of Theorem 1 and Corollary 2

To prove Theorem 1 and Corollary 2, we use the following particular solution of QCQP problems.

Lemma 1. *Suppose that $A \in \mathbb{R}^{n\times n}$ is a positive definite matrix, a vector $b \in \mathbb{R}^n$, a scalar c, and positive constants k_{x}, k_{y}. The solution of this problem*

$$\begin{aligned} \textbf{minimize} \quad & k_{\mathrm{x}} x^{\mathrm{T}} A x + k_{\mathrm{y}} |y| \\ \textbf{subject to} \quad & y \geq x^{\mathrm{T}} A x - 2b^{\mathrm{T}} x + c. \end{aligned}$$

is given by

$$\begin{aligned} x^* &= \left(1 - \sqrt{\mathrm{C}(1-\tilde{c}; \tilde{k}_{\mathrm{x}}^2, 1)}\right) A^{-1} b, \\ y^* &= \mathrm{R}\left(c - (1-\tilde{k}_x^2) b^{\mathrm{T}} A^{-1} b\right), \end{aligned}$$

where

$$\tilde{c} = \frac{c}{b^{\mathrm{T}} A^{-1} b}, \quad \tilde{k}_{\mathrm{x}} = \frac{k_{\mathrm{x}}}{k_{\mathrm{x}} + k_{\mathrm{y}}}.$$

Proof: See Appendix B. □

To keep the notation short in the following proofs, we define

$$\beta \triangleq \nabla V.$$

Proof of Theorem 2: For the Hamilton-Jacobi function $\mathrm{HJ}(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$ of the modified dynamics $(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$, the map $f_{\mathbf{m}}$ and each row of the map $G_{\mathbf{m}}$ don't depend on the orthogonal vector of β. Hence, $f_{\mathbf{m}}$ and $G_{\mathbf{m}}$ are written as

$$f_{\mathbf{m}} = f_{\mathbf{n}} - \beta p, \quad G_{\mathbf{m}} = G_{\mathbf{n}} - \beta q^{\mathrm{T}},$$

where p is scalar and $q \in \mathbb{R}^m$. $h_{\mathbf{m}}$ depends only on the norm in $\mathrm{HJ}(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$, i. e.,

$$h_{\mathbf{m}} = (1-r)hn,$$

where r is scalar.

The optimal function (5a) is rewritten as the p, q, and r, i.e.,

$$\begin{aligned} &\frac{k_1}{\|\beta\|}\|f_{\mathbf{m}} - f_{\mathbf{n}}\| + \frac{k_2}{2\gamma^2}\|G_{\mathbf{m}} - G_{\mathbf{n}}\|^2 + \frac{k_2}{2\|\beta\|^2}\|h_{\mathbf{m}} - h_{\mathbf{n}}\| \\ &= k_1|p| + k_2\frac{\|\beta\|^2}{2\gamma^2}\|q\|^2 + k_2\frac{\|h_{\mathbf{n}}\|}{2\|\beta\|^2} r^2. \end{aligned}$$

Also, the condition of constraint (5b) is rewritten as

$$\begin{aligned} &\mathrm{HJ}_\beta(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}}) \\ &= \beta^{\mathrm{T}} f_{\mathbf{m}} + \frac{1}{2\gamma^2}\|G_{\mathbf{m}}^{\mathrm{T}}\beta\|^2 + \frac{1}{2}\|h_{\mathbf{m}}\|^2 \\ &= \beta^{\mathrm{T}}(f_{\mathbf{n}} - \beta p) + \frac{1}{2\gamma^2}\|(G_{\mathbf{n}} - \beta q^{\mathrm{T}})^{\mathrm{T}}\beta\|^2 + \frac{1}{2}\Big\|(1-r)h_{\mathbf{n}}\|^2 \\ &= -\|\beta\|^2 p + \frac{1}{2\gamma^2}\|\beta\|^4\|q\|^2 + \frac{1}{2}\|\hat{h}\|^2 r^2 - \frac{1}{\gamma^2}\|\beta\|^2\beta^{\mathrm{T}} G_{\mathbf{n}} q - \|h_{\mathbf{n}}\|^2 r + \mathrm{HJ}_\beta(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}}) \\ &\leq 0. \end{aligned}$$

If $x \triangleq [q^{\mathrm{T}}, r]^{\mathrm{T}}$, $y \triangleq p$, and

$$A \triangleq \begin{bmatrix} \frac{\|\beta\|^2}{2\gamma^2} I_m & 0 \\ 0 & \frac{\|h_{\mathbf{n}}\|^2}{2\|\beta\|^2} \end{bmatrix}, \quad b \triangleq \begin{bmatrix} \frac{1}{2\gamma^2} G_{\mathbf{n}}^{\mathrm{T}}\beta \\ \frac{\|h_{\mathbf{n}}\|^2}{2\|\beta\|^2} \end{bmatrix} \quad c \triangleq \frac{1}{\|\beta\|^2}\Big(\mathrm{HJ}_\beta(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}})\Big),$$

then this optimal problem (5) of this theorem becomes a problem used in Lemma 1. The optimal point of p, q and r is given by

$$
\begin{aligned}
p^* &= \frac{1}{\|\beta\|^2}\mathrm{R}\left(V_{\hat{f}} + \hat{k}^2 V_{G_{\mathbf{n}},h_{\mathbf{n}}}\right),\\
q^* &= \frac{1}{\|\beta\|^2}\left(1-\sqrt{\mathrm{C}\Big(-\tfrac{V_{f_{\mathbf{n}}}}{V_{G_{\mathbf{n}},h_{\mathbf{n}}}};\tilde{k}^2,1\Big)}\right)G_{\mathbf{n}}^{\mathrm{T}}\beta,\\
r^* &= 1-\sqrt{\mathrm{C}\Big(-\tfrac{V_{f_{\mathbf{n}}}}{V_{G_{\mathbf{n}},h_{\mathbf{n}}}};\tilde{k}^2,1\Big)}.
\end{aligned}
$$

Therefore, Theorem 1 is derived.

Proof of Theorem 2:

Optimal maps $f_{\mathbf{m}}$ and $G_{\mathbf{m}}$ are also written as

$$f_{\mathbf{m}} = f_{\mathbf{n}} - \beta p, \quad G_{\mathbf{m}} = G_{\mathbf{n}} - \beta q^{\mathrm{T}}.$$

Hence, the objective function (7a) is rewritten as p and q function, i.e.,

$$\frac{k_1}{\|\beta\|}\|f_{\mathbf{m}} - f_{\mathbf{n}}\| + \frac{k_2}{2\gamma^2}\|G_{\mathbf{m}} - G_{\mathbf{n}}\|^2 = k_1|p| + k_2\frac{\|\beta\|^2}{2\gamma^2}\|q\|^2.$$

Also the condition of constraint (7b) is written

$$
\begin{aligned}
&\mathrm{HJ}_\beta(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{n}})\\
&= \beta^{\mathrm{T}} f_{\mathbf{m}} + \frac{1}{2\gamma^2}\|G_{\mathbf{m}}^{\mathrm{T}}\beta\|^2 + \frac{1}{2}\|h_{\mathbf{m}}\|^2\\
&= \beta^{\mathrm{T}}(f_{\mathbf{n}} - \beta p) + \frac{1}{2\gamma^2}\|(G_{\mathbf{n}} - \beta q^{\mathrm{T}})^{\mathrm{T}}\beta\|^2 + \frac{1}{2}\|h_{\mathbf{n}}\|^2\\
&= -\|\beta\|^2 p + \frac{1}{2\gamma^2}\|\beta\|^4\|q\|^2 - \frac{1}{\gamma^2}\|\beta\|^2\beta^{\mathrm{T}}G_{\mathbf{n}}q + \mathrm{HJ}_\beta(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}})\\
&\le 0.
\end{aligned}
$$

If $x \triangleq q$, $y \triangleq p$, and

$$A \triangleq \frac{\|\beta\|^2}{2\gamma^2}I_m, \quad b \triangleq \frac{1}{2\gamma^2}G_{\mathbf{m}}^{\mathrm{T}}\beta, \quad c \triangleq \frac{1}{\|\beta\|^2}\Big(\mathrm{HJ}_\beta(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}})\Big),$$

then this optimal problem (7) becomes the problem used in Lemma 1. The optimal points of p and q are given by

$$
\begin{aligned}
p^* &= \frac{1}{\|\beta\|^2}\mathrm{R}\left(V_{f_{\mathbf{n}},h_{\mathbf{n}}} + \hat{k}_2^2(V_{G_{\mathbf{n}}})\right),\\
q^* &= \frac{1}{\|\beta\|^2}\left(1-\sqrt{\mathrm{C}\Big(-\tfrac{V_{f_{\mathbf{n}},h_{\mathbf{n}}}}{V_{G_{\mathbf{n}}}};\tilde{k}^2,1\Big)}\right)G_{\mathbf{n}}^{\mathrm{T}}\beta,
\end{aligned}
$$

Therefore, the solution of problem (7) becomes Eqs. (8).

B Proof of the particular solution of QCQP

This section presents the proof of the following QCQP problem.

$$
\begin{aligned}
\textbf{minimize} \quad & k_{\mathrm{x}} x^{\mathrm{T}}Ax + k_{\mathrm{y}}|y| && \text{(10a)}\\
\textbf{subject to} \quad & y \ge x^{\mathrm{T}}Ax - 2b^{\mathrm{T}}x + c, && \text{(10b)}
\end{aligned}
$$

where A is a positive definite matrix. We classify this problem according to the parameter A, b and c.

First, the solution of this optimal problem is switched depending on the positive or negative value of c.

Lemma 2. *If $c \leq 0$ then the solution of Eq. (10) is $(x, y) = (0, 0)$. If $c > 0$, the solution x^* of Eq. (10) equals the solution of the following problem:*

$$\textbf{minimize} \quad k_x x^{\mathrm{T}} A x + k_y |x^{\mathrm{T}} A x - 2b^{\mathrm{T}} x + c|. \tag{11}$$

Furthermore, the solution y^* is given by

$$y^* = k_x x^{*\mathrm{T}} A x^* - 2b^{\mathrm{T}} x^* + c. \tag{12}$$

Proof: The objective function (10a) is strictly convex and the minimum point is $(x, y) = (0, 0)$. If $c \leq 0$, the origin $(x, y) = (0, 0)$ satisfies the condition of constraint (10b). Therefore, the solution of Eq. (10) is $(x, y) = (0, 0)$.

If $c > 0$, $(x, y) = (0, 0)$ does not satisfies the constraint condition (10b), and the optimal solution belong the boundary of the region satisfying (10b). Therefore, the solution point (x^*, y^*) satisfies Eq. (12). $\square$

Also, the solution of the new optimal problem (11) is switched by the ratio between c and $b^{\mathrm{T}} A^{-1} b$.

Lemma 3. *If $\tilde{c} > 1 - \tilde{k}_x^2$ then the solution x^* of the optimal problem (11) is given by*

$$x^* = (1 - \tilde{k}_x) A^{-1} b, \tag{13}$$

where

$$\tilde{c} = \frac{c}{b^{\mathrm{T}} A^{-1} b}, \quad \tilde{k}_x = \frac{k_x}{k_x + k_y}.$$

If $\tilde{c} \leq 1 - \tilde{k}_x^2$, the solution x^* equals the solution of the following QCQP problem, such that

$$\textbf{minimize} \quad x^{\mathrm{T}} A x \tag{14a}$$

$$\textbf{subject to} \quad x^{\mathrm{T}} A x - 2b^{\mathrm{T}} x + c = 0. \tag{14b}$$

Proof: The optimal problem (11) is split on the sign of $x^{*\mathrm{T}} A x^* - 2b^{\mathrm{T}} x^* + c$.

Case $x^{*\mathrm{T}} A x^* - 2b^{\mathrm{T}} x^* + c < 0$:

The optimal problem of this case is written as

$$\textbf{minimize} \quad (k_x - k_y) x^{\mathrm{T}} A x + 2k_y b^{\mathrm{T}} x - k_y c.$$

If $k_x - k_y \leq 0$, there is no optimal point. Otherwise $k_x - k_y > 0$, the optimal point is written as

$$x^* = -\frac{k_y}{k_x - k_y} A^{-1} b.$$

The optimal point does not satisfies the condition $x^{*\mathrm{T}} A x^* - 2b^{\mathrm{T}} x^* + c < 0$, because

$$\begin{aligned} & x^{*\mathrm{T}} A x^* - 2b^{\mathrm{T}} x^* + c \\ &= \frac{k_y^2}{(k_x - k_y)^2} b^{\mathrm{T}} A^{-1} b + \frac{2k_y}{k_x - k_y} b^{\mathrm{T}} A^{-1} b + c > 0, \end{aligned}$$

where A^{-1} is a positive definite matrix and $c > 0$.

Case $x^{*\mathrm{T}} A x^* - 2b^{\mathrm{T}} x^* + c > 0$:

The problem (11) is written as

$$\textbf{minimize} \quad (k_x + k_y) x^{\mathrm{T}} A x - 2k_y b^{\mathrm{T}} x + k_y c.$$

Hence, this optimal point is written as

$$\begin{aligned} x^* &= \frac{k_y}{k_x + k_y} A^{-1} b \\ &= (1 - \tilde{k}_x) A^{-1} b, \end{aligned}$$

where $\sqrt{A}$ is a square root of the positive definite matrix A. The condition $x^{*\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c > 0$ is rewritten by the previous optimal solution, such that

$$\begin{aligned}
&x^{*\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c\\
&= (1-\tilde{k}_x)^2 b^{\mathrm{T}}A^{-1}b - 2(1-\tilde{k}_x)b^{\mathrm{T}}A^{-1}b + c\\
&= -\left(1-\tilde{k}_x^2\right) b^{\mathrm{T}}A^{-1}b + c > 0\\
&\Leftrightarrow \tilde{c} > 1-\tilde{k}_x^2.
\end{aligned}$$

Case $x^{\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c = 0$: The problem (11) is written as the QCQP problem (14). □

The solution of the simple QCQP problem (14) is easily derived using the method of Lagrange multiplier.

Lemma 4. *The solution of the simple QCQP problem (14) is given by*

$$x^* = \left(1-\sqrt{1-\tilde{c}}\right)A^{-1}b.$$

Proof: Supposing a Lagrange multiplier $\lambda > 0$, the Lagrange function is written as

$$L(x,\lambda) = x^{\mathrm{T}}Ax + \lambda(x^{\mathrm{T}}Ax - 2b^{\mathrm{T}}x + c).$$

The KKT condition is given by

$$\frac{\partial L(x^*,\lambda)}{\partial x} = 2(1+\lambda)Ax^* - 2\lambda b = 0. \tag{15}$$

$$\frac{\partial L(x^*,\lambda)}{\partial \lambda} = x^{*\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c = 0, \tag{16}$$

Eq. (15) is written as

$$x^* = \frac{\lambda}{1+\lambda}A^{-1}b,$$

and the Eq. (16) is given by

$$\begin{aligned}
&x^{*\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c = 0\\
\Leftrightarrow&\frac{\lambda^2}{(1+\lambda)^2}b^{\mathrm{T}}A^{-1}b - \frac{2\lambda}{1+\lambda}b^{\mathrm{T}}A^{-1}b + c = 0\\
\Leftrightarrow&-\frac{\lambda^2+2\lambda}{(1+\lambda)^2}b^{\mathrm{T}}A^{-1}b + c = 0\\
\Leftrightarrow&-(\lambda^2+2\lambda)b^{\mathrm{T}}A^{-1}b + (1+\lambda)^2 c = 0\\
\Leftrightarrow&(c-b^{\mathrm{T}}A^{-1}b)\lambda^2 + 2(c-b^{\mathrm{T}}A^{-1}b)\lambda + c = 0\\
\Leftrightarrow&\lambda^2 + 2\lambda + \frac{\tilde{c}}{(\tilde{c}-1)} = 0\\
\Leftrightarrow&\lambda = -1 \pm \sqrt{1-\frac{\tilde{c}}{\tilde{c}-1}}\\
\Leftrightarrow&\lambda = -1 \pm \sqrt{\frac{1}{1-\tilde{c}}}.
\end{aligned}$$

As $\lambda > 0$ and $\tilde{c} > 0$, the Lagrange multiplier is written as

$$\lambda = -1 + \sqrt{\frac{1}{1-\tilde{c}}}.$$

Therefore, the optimal point x^* is given by

$$\begin{aligned}
x^* &= \frac{\lambda}{1+\lambda}A^{-1}b\\
&= \frac{-1+\sqrt{\frac{1}{1-\tilde{c}}}}{\sqrt{\frac{1}{1-\tilde{c}}}}A^{-1}b\\
&= (1-\sqrt{1-\tilde{c}})A^{-1}b.
\end{aligned}$$

$\square$

Finally, we summarize three Lemmas 2-4 and solve the QCQP problem (10).

Lemma 1. *The solution of the problem (10) is given by*

$$x^* = \left(1 - \sqrt{\mathrm{C}\left(1 - \tilde{c}; \tilde{k}_x^2, 1\right)}\right) A^{-1}b, \tag{17a}$$

$$y^* = \mathrm{R}\left(c - (1 - \tilde{k}_x^2)b^{\mathrm{T}}A^{-1}b\right). \tag{17b}$$

Proof: Lemmas 2-4 split the solution x^* as three cases:

$$\begin{aligned} x^* &:= \begin{cases} (1 - \tilde{k}_x)A^{-1}b & 1 - \tilde{k}_x^2 < \tilde{c} \\ \left(1 - \sqrt{1 - \tilde{c}}\right) A^{-1}b & 0 < \tilde{c} \leq 1 - \tilde{k}_x^2 \\ 0 & c \leq 0 \end{cases} \\ &:= \begin{cases} \left(1 - \sqrt{\tilde{k}_x^2}\right) A^{-1}b & 1 - \tilde{c} < \tilde{k}_x^2 \\ \left(1 - \sqrt{1 - \tilde{c}}\right) A^{-1}b & \tilde{k}_x^2 \leq 1 - \tilde{c} < 1 \\ \left(1 - \sqrt{1}\right) A^{-1}b & 1 \leq 1 - \tilde{c} \end{cases} \\ &= \left(1 - \sqrt{\mathrm{C}\left(1 - \tilde{c}; \tilde{k}_x^2, 1\right)}\right) A^{-1}b. \end{aligned}$$

Furthermore, the solution of y^* is written as,

$$\begin{aligned} y^* &:= \begin{cases} x^{*\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c & 1 - \tilde{k}_x^2 < \tilde{c} \\ x^{*\mathrm{T}}Ax^* - 2b^{\mathrm{T}}x^* + c & 0 < \tilde{c} \leq 1 - \tilde{k}_x^2 \\ 0 & c \leq 0 \end{cases} \\ &= \begin{cases} c - (1 - \tilde{k}_x^2)b^{\mathrm{T}}A^{-1}b & 1 - \tilde{k}_x^2 < \tilde{c} \\ 0 & 0 < \tilde{c} \leq 1 - \tilde{k}_x^2 \\ 0 & \tilde{c} \leq 0 \end{cases} \\ &= \mathrm{R}(c - (1 - \tilde{k}_x^2)b^{\mathrm{T}}A^{-1}b). \end{aligned}$$

Therefore, the solution become Eq. (17). $\square$

C Overall schematic of the learning process

Algorithm 1 shows the overall schematic of the learning process. The first line defines the modified dynamics $(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$ from the nominal dynamics $(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}})$, defined by the neural network, where ϕ is a set of parameters of the nominal dynamics. The 2-7 line represents a training loop, where the gradient-based optimization methods can be used by using the forward and backward calculation. Note that an ODE solver is used for forward calculation, and Algorithm 2 shows the forward calculation when the Euler method is used. For simplicity, mini-batch computation omitted in this schematic.

D Neural network architecture and hyper parameters

This section details how to determine the neural network architecture. The architecture and hyper parameters of the neural networks were basically determined by Bayesian optimizing using the validation dataset.

Table 1 shows the search space of Bayesian optimization. The first three parameters: learning rate, weight decay, and batch size are parameters for training the neural networks. Also, an optimizer is selected from AdamW, Adam, and RMSProp. The structure of neural network is determined from

Algorithm 1 Training process

Input: x_0: initial state, u: input signal,y: output signal, $(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}})$: nominal dynamics, V: a designed function

define modified functions $(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$ from $(f_{\mathbf{n}}, G_{\mathbf{n}}, h_{\mathbf{n}})$ and V
for 1 to #iterations **do**
  $\hat{y} \leftarrow$ ODE with $(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$ from x_0, u (**Algorithm 2**)
  forward computation of Loss function (9) from y
  ∇_ϕLoss $\leftarrow$ backward computation with Loss
  $\phi \leftarrow$ Optimizer(ϕ, ∇_ϕLoss)
end for

Algorithm 2 Forward computation for dynamics Eq. (1)

Input: x_0: initial state, u: input signal, $(f_{\mathbf{m}}, G_{\mathbf{m}}, h_{\mathbf{m}})$: dynamics
Output: $\hat{y}$: output signal

for t $\leftarrow 0$ to T **do**
  $x_{t+1} \leftarrow x_t + \Delta t(f_{\mathbf{m}}(x_t) + G_{\mathbf{m}}(x_t)u_t)$
  $\hat{y}_t \leftarrow h_{\mathbf{m}}(x_t)$
end for
return $\hat{y}$

the number of intermediate layers and dimensions for each layer. One layer in our setting consists of a fully connected layer with a ReLU activation. The last three rows represent parameters related to our proposed methods. ϵ is a parameter of the loss function L_{HJ}. Initial scale parameter is multiplied with the output of $f_{\mathbf{n}}$ to prevent the value of $f_{\mathbf{n}}(x)$ from becoming large in the initial stages of learning. When $f_{\mathbf{n}}(x)$, which determine the behavior of the internal system, outputs a large value, it diverges due to time evolution, and the learning of the entire system may not progress. Therefore, it is empirically preferable to start with a small value for $f_{\mathbf{n}}(x)$ at the initial stage of learning. When the flag of 'stop gradient for projection' is false, backward computation related to the second term of modification of $f_{\mathbf{m}}$ and $G_{\mathbf{m}}$ is disabled. Note that modification related to $f_{\mathbf{m}}$ and $G_{\mathbf{m}}$ consists of two terms (see Theorem 1). Setting this parameter to false resulted in better performance in our all experiments.

We ran 300 trials using the Bayesian optimization for the bistable model benchmark and the glucose insulin benchmark with the above settings, and the hyper parameters obtained are shown in Table 2, where the number of dimensions for each hidden layer is shown in tuple from the order closest to the input layer.

Hyperparameters of comparative methods were determined by grid search using the validation dataset. ARX and PWARX have an order parameter n of the autoregressive model, and this parameter is searched in the range of $1-5$. The number of iterations was set to 10000 so that the optimization of PWARX converges sufficiently. MOESP and ORT have an internal dimension n ($1 \leq n \leq 20$) and the number of subsequences used for estimation k ($2n < k \leq 20$).

Table 1: The search space of Bayesian optimization

parameter name	range	type
learning rate	10^{-5} – 10^{-3}	log scale
weight decay	10^{-10} – 10^{-6}	log scale
batch size	10 – 100	integer
optimizer	{ AdamW, Adam, RMSProp}	categorical
#layer for $f_{\mathbf{n}}$	0 – 3	integer
#layer for $G_{\mathbf{n}}$	0 – 3	integer
#layer for $h_{\mathbf{n}}$	0 – 3	integer
#dim. for a hidden layer of $f_{\mathbf{n}}$	$8 - 32$	integer
#dim. for a hidden layer of $G_{\mathbf{n}}$	$8 - 32$	integer
#dim. for a hidden layer of $h_{\mathbf{n}}$	$8 - 32$	integer
ϵ	0 – 1.0	log scale
Initial scale parameter for $f_{\mathbf{n}}$	10^{-5} – 0.1	log scale
Stop gradient for projection	true, false	boolean

Table 2: Selected parameters for each benchmark

parameter name	bistable	glucose insulin
learning rate	3.01×10^{-4}	3.28×10^{-4}
weight decay	4.76×10^{-9}	2.28×10^{-9}
batch size	100	100
optimizer	RMSProp	RMSProp
#layer for $f_{\mathbf{n}}$	3	1
#layer for $G_{\mathbf{n}}$	1	2
#layer for $h_{\mathbf{n}}$	3	2
#dim. for a hidden layer of $f_{\mathbf{n}}$	(17,10,22)	(8)
#dim. for a hidden layer of $G_{\mathbf{n}}$	(34)	(27,29)
#dim. for a hidden layer of $h_{\mathbf{n}}$	(10,62,58)	(35,18)
ϵ	0.63	0.75
Initial scale parameter for $f_{\mathbf{n}}$	9.64×10^{-2}	8.94×10^{-2}
Stop gradient for projection	false	false

E Glucose-Insulin system

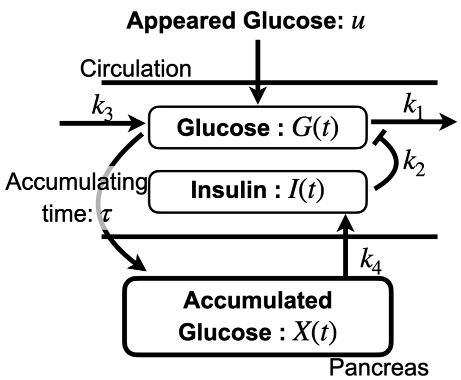
Glucose concentration in the blood is modeled as a time-delay system regulated by insulin concentration (See Fig. 7 (A)) [32]. Suppose that G I, and X are the glucose, insulin, and accumulated glucose plasma concentration ([mg/100ml],[μUI/ml],and [min mg/100ml], respectively) and u is the amount of ingested glucose per minute [min^{-1} mg/100ml]. The dynamics of each concentration is given by

$$\begin{aligned}
\dot{G}(t) &= -k_1 G(t) - k_2 G(t) I(t) + g_0 + u(t), \\
\dot{I}(t) &= -k_3 I(t) + \frac{k_4}{\tau} X(t), \\
\dot{X}(t) &= G(t) - G(t-\tau), \\
y(t) &= [G(t), I(t)]^{\mathrm{T}},
\end{aligned}$$

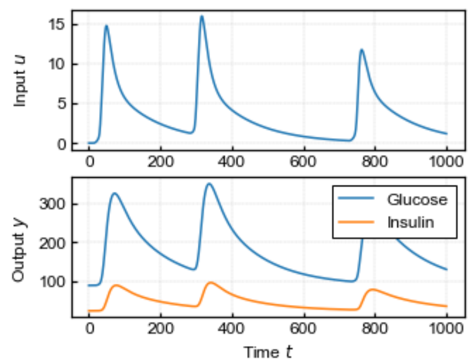
where k_1 is a spontaneous glucose disappearance rate, k_2 is an insulin-dependent glucose disappearance rate, g_0 is a constant increase in plasma glucose concentration, k_3 is an insulin disappearance rate, k_4 is an insulin release rate per the average glucose concentration within the last τ minute.

Table 3: Model parameters.

Parameter	Value	Unit
k_1	3.35×10^{-2}	$\frac{1}{\text{min}}$
k_2	5.22×10^{-5}	$\frac{1}{\text{min}(\mu\text{UI/ml})}$
k_3	1.055	$\frac{1}{\text{min}}$
k_4	0.293	$\frac{(\mu\text{UI/ml})}{\text{min(mg/100ml)}}$
g_0	3.13	$\frac{\text{(mg/100ml)}}{\text{min}}$
τ	6	min



(A)



(B)

Figure 7: (A) Overview and (B) input and output behavior of the glucose insulin system.

This system has a unique asymptotically stable equilibrium point $(G, I, X) \equiv (G^*, I^*, X^*)$ on the nonlinear plain such that

$$G^* = \frac{-k_1k_3 + \sqrt{(k_1k_3)^2 + 4k_2k_3k_4g_0}}{2k_2k_4},$$

$$I^* = \frac{-k_1k_3 + \sqrt{(k_1k_3)^2 + 4k_2k_3k_4g_0}}{2k_2k_3},$$

$$X^* = \frac{-k_1k_3 + \sqrt{(k_1k_3)^2 + 4k_2k_3k_4g_0}}{2k_2k_4}\tau.$$

Here we set the initial state of this model as $G(0) = G^*, I(0) = I^*, X(0) = X^*$ and set the model parameters as shown in the following Table 3. Furthermore, we adopt the output of the previous oral glucose absorption system [33] as u (See Fig. 7 (B)), and the glucose absorption amount u is normalized based on human blood volume per body weight (0.80[100 ml/kg]).