2026 European Control Conference (ECC)
July 7-10, 2026. Reykjavík, Iceland

Safe Deep Reinforcement Learning Control of Type 1 Diabetes

F. BALDISSERI*,[1], G. LOPS[2], M. M. H. ATANASIOUS[1], D. MENEGATTI[1], V. BECCHETTI[1], F. DELLI PRISCOLI[1], S. MASCOLO[2], V. A. RACANELLI[2], and A. WRONA[1]

***Abstract*— Achieving safe and autonomous glycemic regulation for Type-1 Diabetes care is an urgent challenge. Although Deep Reinforcement Learning (DRL) emerged as a promising paradigm, practical deployment is hindered by the risk of uncontrolled hyperglycemia or hypoglycemia. This work adapts two safe DRL approaches in the context of automated insulin delivery. The first consists of a Lagrangian constrained Markov decision process that solves a primal–dual scheme with adaptive multipliers, thereby delivering constraint satisfaction in expectation; the second adopts a Barrier–Lyapunov Actor–Critic framework that embeds discrete-time control-barrier conditions and Lyapunov decrease into the learning updates, ensuring stepwise feasibility and promoting stability by design. Simulations under randomized meal timing and size, benchmarked against a standard clinical practice protocol and an unconstrained DRL baseline, indicate improved time-in-range with reduced hypoglycemic events.**

I. INTRODUCTION AND RELATED WORKS

A. Problem Statement

Type-1 Diabetes (T1D) is a chronic condition that requires lifelong insulin therapy to regulate blood glucose levels. Improper dosing can lead to acute complications that can cause severe damage to organs and systems [1]. Currently, glycemic regulation in T1D patients often relies on manual intervention, guided by frequent glucose monitoring and heuristic or rule-based insulin administration protocols. This process is not only burdensome for patients but also susceptible to human error and suboptimal decision-making [2]. With the emergence of continuous glucose monitoring systems and insulin pumps, there is a growing opportunity to close the loop and implement fully autonomous glucose regulation strategies [3]. However, safety remains a paramount concern in this context. A controller that fails to respect physiological constraints can endanger the patient's health. Therefore, there is an urgent need for intelligent, data-driven algorithms that can autonomously manage insulin delivery while strictly ensuring safety during the deployment of insulin therapy.

B. Related Works and Proposed Innovations

Traditional control strategies such as Proportional-Integral-Derivative (PID) controllers have been widely used due to their simplicity and ease of implementation, showing reasonable performance under known patient conditions but limited adaptability to inter- and intra-patient variability [4]. Fuzzy logic controllers have been proposed to handle the nonlinear and uncertain dynamics of glucose-insulin interaction. These methods incorporate expert knowledge to make heuristic decisions about insulin dosing, as demonstrated in [5], offering improved robustness but limited scalability due to rule-based complexity. Another popular technique, Model Predictive Control (MPC), incorporates system constraints and prediction of future glucose trajectories. Studies such as [6] and [7] have shown that MPC-based artificial pancreas systems can provide safe and effective glucose regulation, especially when combined with physiological models of glucose-insulin dynamics. However, MPC's performance often relies on the accuracy of the underlying model, and tuning remains challenging for highly variable conditions.

More recently, Reinforcement Learning (RL) has gained attention as a promising alternative due to its data-driven nature and ability to learn optimal insulin delivery strategies directly from interaction with the system, without requiring precise physiological models [8]. Unlike traditional approaches, RL can adapt to patient-specific dynamics over time, offering a personalized and potentially more robust solution to automated glucose control, as demonstrated with in-silico trials in recent research [9]. Academic literature spans off-policy actor–critic (e.g., DDPG [10] or TD3 [11]), on-policy methods (e.g., PPO, Maskable PPO [12]), and modular/controller-assisted designs [13]. Recent work reports improvements through architectural choices and training heuristics (e.g., prioritized replay, delay-aware formulations) as well as modular pipelines that lessen reliance on meal announcement [14]. While results vary by scenario, several studies report gains over PID/MPC baselines, including better time-in-range (TIR) and fewer hypoglycemic events [15].

However, the main limitation of prior approaches is that they do not provide formal safety guarantees to enforce constraints at runtime. In contrast, this work exploits two separate safe-RL approaches. First, a Lagrangian Constrained Markov Decision Process (CMDP) formulation encodes clinically meaningful constraints and is solved via a primal–dual procedure with adaptive multipliers, thereby eliminating heuristic penalty tuning and delivering constraint satisfaction in expectation. Secondly, a Barrier–Lyapunov Actor–Critic method incorporates control-barrier conditions and Lyapunov functions decrease directly into the learning objective and updates, ensuring stepwise feasibility and promoting stability by construction. Under realistic variability in meal timing

This work has been partially funded from the European Union's Horizon Europe research and innovation programme under the project SHIELD, grant agreement No. 101156751.

[1]Department of Computer, Control and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185, Rome, Lazio, Italy.

[2]Department of Electrical and Information Engineering, Polytechnic University of Bari, Italy.

*Corresponding author. E-mail: baldisseri@diag.uniroma1.it

and size, both approaches achieve higher TIR with no observed hypoglycemic events, outperforming a standard-of-care protocol and an unconstrained deep-RL baseline. Thus, this work provides insights to improve glycemic regulation performance without compromising safety.

The remainder of the paper is organized as follows: Section II-A recalls general notions on MDPs, CMDPs, and RL, while Section II-B and Section II-C focus on the proposed RL algorithms; Section III describes the instantiation of the considered control problem through the two proposed algorithms; IV reports simulation results benchmarking the proposed control methodologies with a clinical practice baseline and a conventional DRL controller; lastly, Section V draws conclusions and outlines directions for future work.

II. Safe Reinforcement Learning Control Algorithms

A. Preliminaries

Reinforcement Learning (RL) concerns training decision-making agents to select actions in a dynamical environment using observed data and reward [16]. The environment is modeled as a Markov decision process (MDP), a framework for sequential decision making under uncertainty that satisfies the Markov property, i.e., future evolution is conditionally independent of the past given the present state.

An MDP consists of a tuple $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, where: $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $p(s' \mid s, a)$ is the transition probability that the system moves from state s to state s' upon applying action a, $r(s, a, s')$ is the one-step reward received after transitioning from s to s' with action a, $\gamma \in [0, 1)$ is the discount factor modulating preference for immediate ($\gamma \approx 0$) versus delayed ($\gamma \approx 1$) rewards. The objective is to learn an optimal policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected discounted return, characterized by the state–action value function $Q^\pi(s, a)$:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s,\ a_0 = a \right]. \tag{1}$$

While standard RL maximizes expected reward without explicit safety guarantees, many control tasks—including insulin dosing—require constraint satisfaction. When constraints are taken into account in the derivation of an optimal policy, one relies upon the Constrained MDP (CMDP) formulation which differs from the standard one by the introduction of the cost function $C : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$. The goal remains to derive an optimal safe policy π by maximizing the expected reward subject to a set of safety constraints.

Moreover, in order to enable effective policy learning in large-scale, intricate environments, one may rely upon the Deep RL (DRL) paradigm [17], that is to augment RL methodologies with deep neural function approximators to represent complex mappings over high-dimensional state, action, and reward spaces.

B. Lagrangian Constrained Markov Decision Process

Given a constrained Markov decision process (CMDP) with cost function $C : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$, initial distribution μ, and discount $\gamma \in [0, 1)$. For a stationary policy π_θ, define the discounted performance and a general (discounted or mean-value) constraint statistic as:

$$J_R(\pi_\theta) \triangleq \mathbb{E}_{\pi_\theta,\, s_0 \sim \mu}\Big[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t) \Big], \tag{2}$$

$$J_C(\pi_\theta) \triangleq \mathbb{E}_{\pi_\theta,\, s_0 \sim \mu}\big[C(s_0, s_1, \ldots) \big], \tag{3}$$

and seek:

$$\max_{\pi_\theta}\ J_R(\pi_\theta) \tag{4}$$

$$\text{s.t.} \qquad J_C(\pi_\theta) \leq d, \tag{5}$$

where d is a vector of scalars representing the safety boundaries. The solution of the corresponding optimization problem is performed via the introduction of a Lagrange multiplier $\lambda \geq 0$ yields the saddle problem:

$$\min_{\lambda \geq 0}\ \max_{\theta}\ \mathcal{L}(\theta, \lambda)\ \triangleq\ J_R(\pi_\theta) - \lambda\big(J_C(\pi_\theta) - d\big), \tag{6}$$

which is solved by stochastic primal–dual recursions with projection [18]:

$$\theta_{k+1} = \Gamma_\Theta\Big(\theta_k + \eta_\theta\, \widehat{\nabla}_\theta \mathcal{L}(\theta_k, \lambda_k)\Big), \tag{7}$$

$$\lambda_{k+1} = \Big[\lambda_k + \eta_\lambda\big(\widehat{J}_C(\pi_{\theta_k}) - d\big)\Big]_+. \tag{8}$$

Here, $\widehat{\nabla}_\theta \mathcal{L}(\theta_k, \lambda_k)$ denotes a stochastic estimate of the gradient of the Lagrangian with respect to the policy parameters, Γ_Θ denotes projection onto a compact feasible parameter set Θ, and $[\cdot]_+$ denotes projection onto the nonnegative reals $[0, \infty)$. The step sizes satisfy $\sum_k \eta_\theta = \sum_k \eta_\lambda = \infty$, $\sum_k (\eta_\theta^2 + \eta_\lambda^2) < \infty$. In actor–critic form, the actor updates the policy parameters to improve control actions, while the critic estimates the corresponding penalized value function. As a result, the timescale separation $\eta_\lambda / \eta_\theta \to 0$ allows the actor to run faster than the dual variable. The iterates converge to a local saddle point that is feasible for (5) [18].

To enable actor–critic learning with explicit constraints, the critic is trained on a penalized reward while the dual update still uses the original constraint statistic. For notational convenience, let $c(s, a)$ denote the instantaneous stage-wise constraint cost, i.e., the one-step quantity whose cumulative expectation defines $J_C(\pi_\theta)$ (thus c is the per-step counterpart of the cumulative constraint functional C introduced above). The penalized reward and value are then defined as:

$$\hat{r}_\lambda(s, a) \triangleq r(s, a) - \lambda\, c(s, a), \tag{9}$$

$$V_\lambda^\pi(s) \triangleq \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t \hat{r}_\lambda(s_t, a_t) \,\Big|\, s_0 = s \Big], \tag{10}$$

and estimate V_λ^π (or Q_λ^π) by TD learning using $\hat{r}_\lambda$. The policy update ascends on the penalized objective induced by (10), whereas the dual ascent (8) uses an episodic (Monte-Carlo) estimate of the original constraint $J_C(\pi_\theta)$, ensuring feasibility while benefiting from low-variance critic updates.

For a deterministic actor $a = \pi_\lambda(s;\theta)$ with critic $Q_\lambda(s,a;\omega)$, a practical instantiation is:

$$\omega \leftarrow \omega - \eta_v \nabla_\omega \Big(Q_\lambda(s,a;\omega) - \big[\hat{r}_\lambda(s,a) + \tag{11}$$

$$+ \gamma\, Q_\lambda(s', \pi_\lambda(s';\theta);\omega^-)\big] \Big)^2, \tag{12}$$

$$\theta \leftarrow \theta + \eta_\theta\, \mathbb{E}_s \Big[\nabla_\theta \pi_\lambda(s;\theta)\, \nabla_a Q_\lambda(s,a;\omega)\big|_{a=\pi_\lambda(s;\theta)} \Big], \tag{13}$$

$$\lambda \leftarrow \big[\lambda + \eta_\lambda\, (\widehat{J}_C(\pi_\theta) - d)\big]_+. \tag{14}$$

Target-network parameters ω^- and standard stabilizers (e.g., projection/clipping for λ) are used in practice. Under the aforementioned stepsize conditions and regularity assumptions, the coupled three-timescale process (critic on the fastest timescale, actor on the intermediate, dual on the slowest) converges almost surely to a feasible stationary point of (6). This method interprets each clinical requirement (e.g., hypoglycaemia avoidance, IOB compliance) as soft constraints, whose violation increases the dual penalty. Over successive episodes, the adaptive multiplier raises the cost of unsafe behaviour, steering the policy toward feasible regions while maintaining exploration. This mechanism adaptively increases the penalty associated with unsafe behaviour, thereby reducing the need for manually tuned safety penalties in the reward design.

C. Barrier–Lyapunov Actor–Critic

Consider a discrete-time, control-affine Markov system driven by a stochastic policy:

$$\begin{aligned} x_{t+1} &= f(x_t) + g(x_t)\, u_t, \\ x_t \in \mathcal{X} \subset \mathbb{R}^n, &\qquad u_t \in \mathcal{U} \subset \mathbb{R}^m. \end{aligned} \tag{15}$$

A policy $\pi_\theta(u|x)$ is trained with an entropy-regularized actor–critic, namely the Soft Actor Critic (SAC) [19], while enforcing stepwise safety and stability via control barrier functions (CBFs) and a control-Lyapunov function surrogate through an augmented Lagrangian [20], denoted in the following as BLAC.

Let $h_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, k$, define safe sets $\mathcal{C}_i \doteq \{x : h_i(x) \geq 0\}$ and $\mathcal{C} \doteq \cap_i \mathcal{C}_i$. Discrete-time CBF safety with shrinkage $\eta \in [0,1]$ requires, step by step:

$$h_i(x_{t+1}) \;-\; h_i(x_t) \;\geq\; -\eta\, h_i(x_t), \qquad \forall t,\ \forall i. \tag{16}$$

If (16) holds along the trajectory for all constraints, the safe set $\mathcal{C}$ is forward invariant. Stability is shaped by the discounted value of a nonnegative cost c_π under policy π:

$$L_\pi(x_t) \;=\; \mathbb{E}_{\tau\sim\pi}\left[\sum_{k=0}^{\infty} \gamma^k\, C_\pi(x_{t+k}) \right], \tag{17}$$

$$L_\pi(x_{t+1}) \;\leq\; L_\pi(x_t) - \beta\, L_\pi(x_t), \tag{18}$$

where $0 < \gamma_c < 1$, $\beta > 0$, and L_π is approximated by a Lyapunov critic L_ν. During learning, analytic enforcement of (16)–(18) is replaced by data-based residuals on replay batches. With transitions (x_t, u_t, x_{t+1}) and a one-step predictor $\hat{x}_{t+1} \doteq f(x_t) + g(x_t)u_t$, violations are mapped to nonnegative ReLU residuals:

$$\phi_i(x_t,u_t) \;\doteq\; \mathrm{ReLU}\Big(h_i(x_t) \;-\; h_i(\hat{x}_{t+1}) \;-\; \eta\, h_i(x_t) \Big), \tag{19}$$

$$\psi(x_t,u_t) \;\doteq\; \mathrm{ReLU}\Big(L_\nu(\hat{x}_{t+1}) \;-\; L_\nu(x_t) \;+\; \beta\, L_\nu(x_t) \Big), \tag{20}$$

with $i = 1, \ldots, k$. Batch means $\overline{\phi}_i$ and $\overline{\psi}$ are driven toward zero by the actor update.

The actor minimizes the SAC objective plus an augmented-Lagrangian penalty. Writing $\tilde{u} = \tanh\big(\mu_\theta(x) + \sigma_\theta(x) \odot \xi\big)$ for the stochastic action used inside the loss:

$$\begin{aligned} J_{\mathrm{actor}}(\theta) = \; & \mathbb{E}_{x,\xi}\Big[\alpha\, \log \pi_\theta(\tilde{u}|x) \;-\; \min_{j=1,2} Q_{\phi_j}(x,\tilde{u}) \Big] + \\ & + \sum_{i=1}^{k} \Big(\lambda_i\, \overline{\phi}_i \;+\; \tfrac{\rho_i}{2}\, \overline{\phi}_i^{\,2} \Big) \;+\; \zeta\, \overline{\psi} \;+\; \tfrac{\rho_{\mathrm{clf}}}{2}\, \overline{\psi}^{\,2}. \end{aligned} \tag{21}$$

Dual variables ascend and quadratic weights tighten geometrically:

$$\lambda_i \leftarrow \big[\, \lambda_i \;+\; \ell\, \overline{\phi}_i \,\big]_+, \qquad \zeta \leftarrow \big[\, \zeta \;+\; \ell\, \overline{\psi} \,\big]_+, \tag{22}$$

$$\rho_i \leftarrow C_\rho\, \rho_i, \qquad \rho_{\mathrm{clf}} \leftarrow C_\rho\, \rho_{\mathrm{clf}}. \tag{23}$$

Twin critics, the temperature α, and L_ν are trained with standard SAC MSE losses and Polyak-averaged targets.

If $\overline{\phi}_i \to 0$ and $\overline{\psi} \to 0$, then safety is enforced with margin:

$$\mathbb{E}[h_i(x_{t+1}) \;-\; (1-\eta)\, h_i(x_t)] \;\geq\; -L_{h_i}\, \varepsilon \;-\; o(1), \tag{24}$$

and stability is enforced in mean cost with bias:

$$\mathbb{E}[L_\nu(x_{t+1})] \;\leq\; (1-\beta)\, \mathbb{E}[L_\nu(x_t)] \;+\; L_L\, \varepsilon + \delta_L, \tag{25}$$

hence:

$$\mathbb{E}[L_\nu(x_t)] \;\leq\; (1-\beta)^t\, \mathbb{E}[L_\nu(x_0)] \;+\; \frac{L_L\, \varepsilon + \delta_L}{\beta}. \tag{26}$$

Moreover, stationarity of the actor for the penalized problem is achieved with geometric growth of $\rho_i, \rho_{\mathrm{clf}}$ and projected ascent for λ_i, ζ, any limit point of the joint updates is a first-order stationary point of the constrained actor problem.

Compared with the CMDP formulation, BLAC operates at a finer temporal granularity: instead of penalizing long-horizon constraint violations, it enforces feasibility at each transition by construction. This per-step conditioning is expected to translate into more stable insulin modulation, particularly around critical thresholds

III. Control Framework Implementation

A. Lagrangian CMDP Implementation

The CMDP is instantiated with a state vector $s_t \in \mathcal{S}$ that aggregates clinically relevant variables: current glucose (G_t), short-horizon trend ($\mathrm{d}\mathrm{G}_t/\mathrm{d}t$), insulin on board (IOB). The control action $a_t \in [0,1]$ is affinely mapped to a nominal per-minute bolus $u_t \in [0, \bar{u}]$, with $\bar{u}$ the rate ceiling. The instantaneous performance signal r_t is a shaped reward: a banded term favors the 90–140 $[\mathrm{mg/dL}]$ range with a small center bonus near 120 $[\mathrm{mg/dL}]$; linear penalties increase

above $140\,[\mathrm{mg/dL}]$ (stronger beyond $180\,[\mathrm{mg/dL}]$) and below $90\,[\mathrm{mg/dL}]$ (much stronger below $70\,[\mathrm{mg/dL}]$); a gentle slope penalty discourages dosing into falling glucose.

Safety constraints are imposed as bounds on expected discounted costs under a stationary policy π:

$$J_{C_i}(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t\, C_i(s_t)\Big] \le d_i,$$

with tolerances set to $d_{\mathrm{hypo}} = 0.05$, $d_{\mathrm{hyper}} = 0.20$, and $d_{\mathrm{iob}} = 0.10$). Hypoglycaemia avoidance and adherence to IOB constraints are enforced as stronger constraints, whereas time-above-range (hyperglycemia) is permitted for at most 20% of the time horizon to accommodate postprandial excursions. The instantiated costs are:

$$\begin{aligned} C_{\mathrm{hypo}}(s_t) &= \mathbf{1}\{\mathrm{G}_t < 90\},\\ C_{\mathrm{hyper}}(s_t) &= \mathbf{1}\{\mathrm{G}_t > 140\},\\ C_{\mathrm{iob}}(s_t) &= \big(\mathrm{IOB}_t - L_{\mathrm{IOB}}(t)\big)_+, \end{aligned} \tag{27}$$

where the time-varying IOB limit is $L_{\mathrm{IOB}}(t) = 6\,[\mathrm{U}]$ during the day (06:00–22:00) and $L_{\mathrm{IOB}}(t) = 3\,[\mathrm{U}]$ at night (22:00–06:00).

Training maximizes the discounted Lagrangian return with nonnegative multipliers λ:

$$\begin{aligned} \mathcal{L}(\pi,\lambda) = \mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t\Big(r_t - \lambda_{\mathrm{hypo}}\big(C_{\mathrm{hypo}}(s_t) - d_{\mathrm{hypo}}\big) -\\ + \lambda_{\mathrm{hyper}}\big(C_{\mathrm{hyper}}(s_t) - d_{\mathrm{hyper}}\big) -\\ + \lambda_{\mathrm{iob}}\big(C_{\mathrm{iob}}(s_t) - d_{\mathrm{iob}}\big)\Big)\Bigg]. \end{aligned} \tag{28}$$

A deterministic actor–critic TD3 algorithm optimizes the primal policy using the augmented reward inside the bracket, while the dual variables are updated at episode boundaries via projected ascent:

$$\lambda_i^{k+1} = \Pi_{[0,\lambda_{\max}]}\big(\lambda_i^k + \eta\,(\bar{C}_i^{\,k} - d_i)\big)\,, \tag{29}$$

where $\bar{C}_i^{\,k}$ denotes the empirical mean cost for episode k, $\eta > 0$ is the dual step size, and Π projects onto $[0, \lambda_{\max}]$. This implementation realizes the CMDP program end-to-end: primal learning drives performance under the Lagrangian objective, dual updates steer long-horizon constraint satisfaction, and the runtime projector ensures per-step safety.

B. BLAC Implementation

The state and action follow the earlier formulation. Barrier and Lyapunov specifications are encoded by the discrete-time control-barrier functions (CBFs):

$$h_1(s) = \mathrm{G} - b_{\min}, \quad h_2(s) = b_{\max} - \mathrm{G}, \tag{30}$$

with $b_{\min} = 90\,[\mathrm{mg/dL}]$, $b_{\max} = 140\,[\mathrm{mg/dL}]$. Stepwise feasibility is promoted by the discrete-time CBF conditions $h_j(s_{t+1}) \ge (1-\eta)\,h_j(s_t)$, $j = 1, 2$. A Lyapunov function $L(s) \ge 0$ encourages stability through $L(s_{t+1}) \le (1-\beta)\,L(s_t)$. The Ruan linear model of a diabetic patient is used as the one-step prediction model. The details about such model can be found in [21]. Critics and Lyapunov learning use temporal-difference targets on the realized next state s_{t+1}:

$$\begin{aligned} y_t = r_t + \gamma(1-d_t)\,\Big(Q_{\min}(s_{t+1},a_{t+1}) -\\ + \alpha\log\pi(a_{t+1}\mid s_{t+1})\Big), \end{aligned} \tag{31}$$

$$y_t^L = c_t + \gamma(1-d_t)\,L(s_{t+1}). \tag{32}$$

where $d_t \in \{0,1\}$ is the terminal flag equal to 1 only at the last step of the episode, α is the SAC entropy temperature, initialized at 0.2 and adapted online toward the target entropy -2.0, $Q_{\min}(s,a) = \min\{Q_1(s,a),\, Q_2(s,a)\}$ is the minimum of twin critics, and $c_t = |\mathrm{G}_t - 120|\ [\mathrm{mg/dL}]$ is a tracking cost.

The actor is optimized with a SAC objective augmented by BLAC penalties:

$$\begin{aligned} J_{\mathrm{actor}} = \mathbb{E}\big[\alpha\,\log\pi(a_t\mid s_t) - Q_{\min}(s_t,a_t)\big] +\\ + \sum_{j=1}^{2}\left(\lambda_j\,\delta_j^{\mathrm{cbf}} + \tfrac{\rho_j^{\mathrm{cbf}}}{2}\,(\delta_j^{\mathrm{cbf}})^2\right) +\\ + \zeta\,\delta^{\mathrm{clf}} + \tfrac{\rho^{\mathrm{clf}}}{2}\,(\delta^{\mathrm{clf}})^2, \end{aligned} \tag{33}$$

where $\lambda_j, \zeta \ge 0$ are Lagrange multipliers and $\rho_j^{\mathrm{cbf}}, \rho^{\mathrm{clf}} > 0$ are quadratic penalty weights. In implementation, the multipliers are initialized to zero and updated by projected ascent on batch-averaged residuals, while the penalties follow a geometric growth schedule:

$$\lambda_j \leftarrow \big[\lambda_j + \eta_\lambda\,\overline{\delta_j^{\mathrm{cbf}}}\big]_+, \qquad \zeta \leftarrow \big[\zeta + \eta_\lambda\,\overline{\delta^{\mathrm{clf}}}\big]_+, \tag{34}$$

$$\rho_j^{\mathrm{cbf}} \leftarrow c_\rho\,\rho_j^{\mathrm{cbf}}, \qquad \rho^{\mathrm{clf}} \leftarrow c_\rho\,\rho^{\mathrm{clf}}, \tag{35}$$

with $j \in \{1,2\}$, stepsize $\eta_\lambda = 1.5\times10^{-3}$, and growth factor $c_\rho = 1.08$. Initial values are $\lambda_j^{(0)} = 0$ $(j = 1,2)$, $\zeta^{(0)} = 0$, and $\rho_j^{\mathrm{cbf}}(0) = \rho^{\mathrm{clf}}(0) = 1.0$.

IV. Simulations and Results

The proposed control methodologies are validated in silico on the Simglucose simulator – namely the Python version of the Uva/Padova (2018) simulator of a T1D patient. The meal scenarios were randomly generated as in [22], in order to model realistic fluctuations in meal patterns, encompassing noise in meal times (± 60 [min]) and quantities (± 20 [g] of carbohydrates). The Yale Insulin Infusion Protocol [23] is used as a benchmark that reflects standard clinical practice; it consists of a rule-based protocol that adjusts insulin infusion according to tabulated rules that incorporate the current glucose level and its recent trajectory under the prevailing rate. Moreover, the standard TD3 algorithm is used to train a baseline DRL method which does not encompass safety. Unlike the prior L-CMDP [18]/BLAC [20] designs, the policy is enhanced with a risk-aware exploration scheme that is formally shielded by the barrier constraints. Specifically, action noise is projected onto the CBF-feasible cone, ensuring first-order satisfaction of $h_i(x_{t+1}) \ge (1-\eta)\,h_i(x_t)$ during exploration. Each training episode lasts three full simulated days, i.e., 4320 one–minute steps. An MLP with two hidden layers of sizes 256 and 256 with ReLU activations

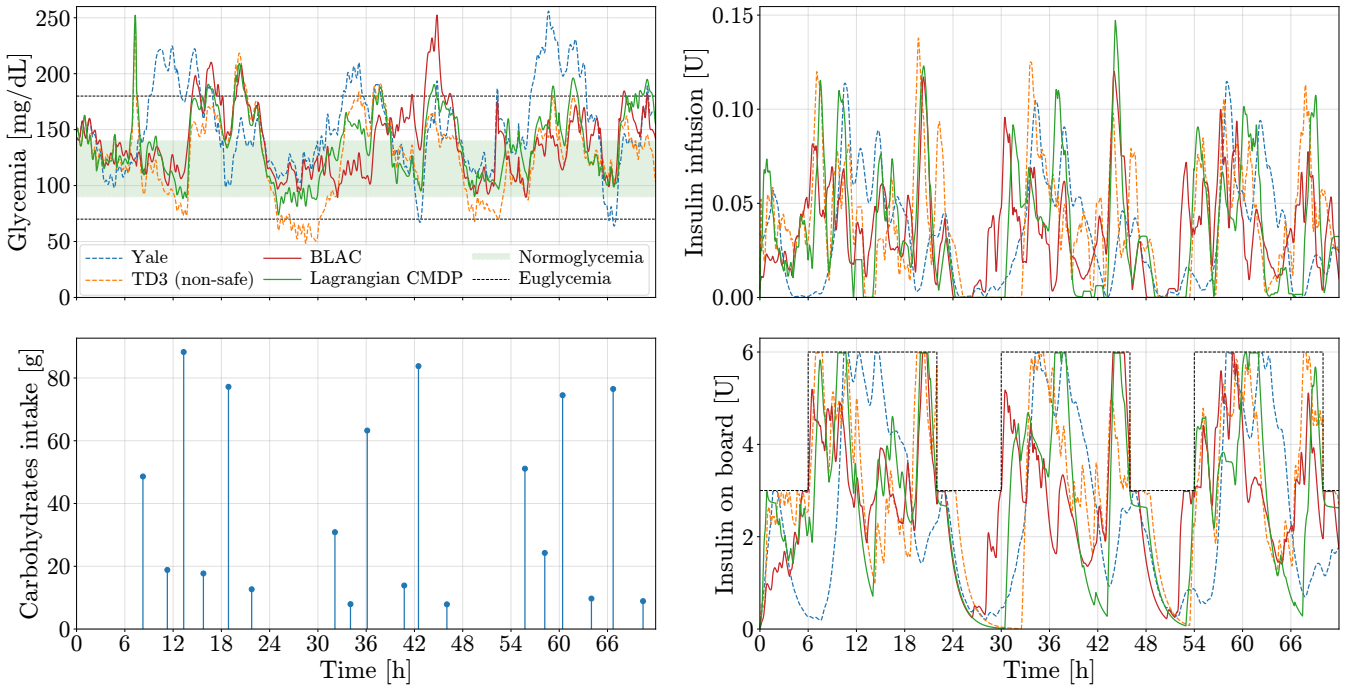

Fig. 1. Closed-loop evaluation over 3 days. Top left: glycemia trajectories with the $90-140$ mg/dL normoglycemia reference and the $70-180$ mg/dL euglycemia band; top right: insulin infusion rate; bottom left: meal carbohydrate intake; bottom right: insulin on board (IOB). The Lagrange Constrained-MDP (L-CMDP) and Barrier-Lyapunov Actor-Critic (BLAC) proposed controllers are compared against the Yale standard clinical practice baseline and the standard TD3 baseline. Day/night IOB constraints are indicated and respected by the safe controllers.

is used; Adam is used to train all networks at learning rate $3{\times}10^{-4}$;training uses batch size 256 and discounts $\gamma = 0.99$. Optimization runs for a fixed 100 episodes. The simulations were carried out on a workstation equipped with an Intel Core i9 9900k processor with 128 GB of RAM and a Nvidia GeForce RTX 3090 as GPU.

Figure 1 illustrates the closed-loop evaluation over 3 days of the 4 control methods for patient *Adult#001* of the Simglucose simulator. Table I summarizes performance metrics that are relevant in the context of glycemic regulation. Note that two glycemic bands are introduced: the reference normoglycemia band $90-140$ [mg/dL] and the wider euglycemia band $70-180$ [mg/dL]. Overall, both safe RL methods improve time-in-range (TIR) without inducing hypoglycemia, and BLAC attains its highest value. Time-below-range (TBR) is essentially eliminated by the safe controllers, whereas TD3 exhibits 6.41% and the baseline 1.09%,

TABLE I

COMPARISON OF CLOSED-LOOP PERFORMANCE METRICS FOR THE 4 CONTROL METHODS.

Metric	Yale	TD3	L-CMDP	BLAC
Mean G (mg/dL)	151.6	**125.9**	137.8	139.2
TIR 90–140 (%)	38.98	43.81	50.73	**54.68**
TIR 70–180 (%)	74.51	88.91	89.23	**89.81**
TAR >140 (%)	57.36	**37.90**	44.99	45.16
TAR >180 (%)	24.40	**4.68**	10.77	10.19
TBR <90 (%)	3.66	18.29	4.28	**0.16**
TBR <70 (%)	1.09	6.41	**0.00**	**0.00**

underscoring the practical value of explicit safety mechanisms for hypoglycemia avoidance. While TD3 achieves the lowest average glucose and best time-above-range (TAR), it does so at the cost of substantial hypoglycemia. These results show that embedding safety in DRL yields clinically desirable trade-offs: improved TIR, suppressed hypoglycemia, and competitive hyperglycemia control, without resorting to aggressive dosing patterns.

In the following, the closed-loop analysis is extended to the entire cohort of ten adult virtual patients of the Simglucose simulator. Similar outcomes to those illustrated in Fig. 1 are obtained for the remaining nine adult subjects and over the horizons extended up to 14 days. Figure 2 summarizes the inter-patient performance of the four controllers in terms of both time-in-range (TIR, $90-140$ [mg/dL]) and time-below-range (TBR <90 [mg/dL]). Cohort mean $\pm$ standard deviation for the TIR are: Yale baseline $27.5 \pm 8.0\%$, TD3 baseline $42.7 \pm 8.4\%$, L-CMDP $39.0 \pm 11.7\%$, and BLAC $40.1 \pm 11.8\%$. For the TBR, the corresponding values are: Yale baseline $12.4 \pm 10.9\%$, TD3 baseline $15.6 \pm 10.9\%$, L-CMDP $7.5 \pm 6.7\%$, and BLAC $3.6 \pm 0.9\%$. Although the unconstrained TD3 achieves the highest mean TIR, this improvement is accompanied by the largest hypoglycemic exposure. Conversely, both safe controllers substantially suppress hypoglycemia—most notably BLAC, whose TBR remains below 4% across the cohort—while preserving competitive TIR values. Moreover, the BLAC controller achieves reduced inter-patient variability: the standard deviation of the TBR across the ten adult subjects decreases from 9.6% for

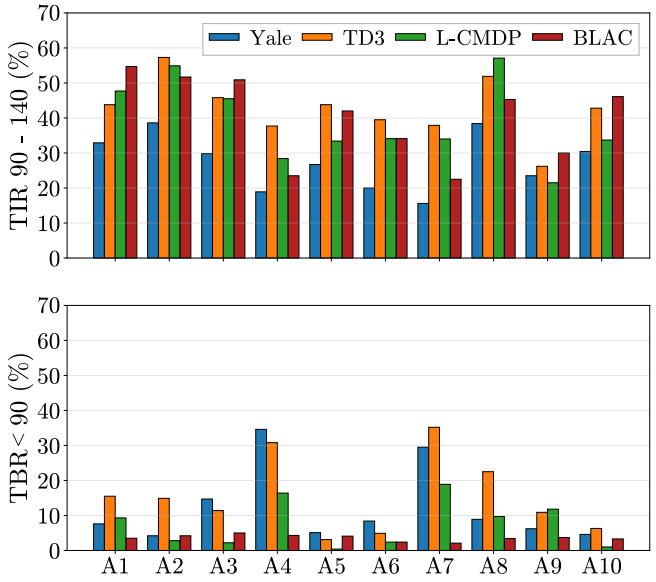

Fig. 2. Time-in-range (TIR, $90 - 140[\mathrm{mg/dL}]$) and time-below-range (TBR $< 90[\mathrm{mg/dL}]$) over the 10 adult virtual patients of the Simglucose simulator. Each group of bars corresponds to a different patient (i.e., Patient *Adult #001* is denoted as A1), comparing the Yale clinical-practice baseline, the standard TD3 baseline, the proposed Lagrangian CMDP controller (L-CMDP), and the Barrier–Lyapunov Actor–Critic (BLAC).

the Yale baseline, 9.9% for TD3, and 5.9% for the L-CMDP, to only 0.9% for BLAC. Thus, the BLAC controller enhances policy generalization across heterogeneous virtual subjects, an essential feature for prospective clinical deployment.

V. Conclusions and Future Works

This work investigated the application of two safe deep-RL strategies in the context of automated insulin delivery: a Lagrangian CMDP with adaptive multipliers encoding hypoglycaemia avoidance, and a Barrier–Lyapunov Actor–Critic scheme embedding discrete-time control-barrier and Lyapunov conditions to ensure stepwise feasibility. In-silico evaluations with randomized meals showed higher time-in-range and reduced observed hypoglycemic events relative to the Yale protocol and an unconstrained TD3 baseline. These results indicate that DRL can improve glycemic regulation without compromising safety. Future research will address also the safety during learning; specifically, methods that minimize training-phase constraint violations will be developed, thereby enabling not only safe deployment but also clinically acceptable training of DRL controllers in real settings.

References

[1] M. J. Haller, M. A. Atkinson, and D. Schatz, "Type 1 diabetes mellitus: etiology, presentation, and management," *Pediatric Clinics*, vol. 52, no. 6, pp. 1553–1578, 2005.

[2] J. McKnight, S. Wild, M. Lamb, M. Cooper, T. Jones, E. Davis, S. Hofer, M. Fritsch, E. Schober, J. Svensson, *et al.*, "Glycaemic control of type 1 diabetes in clinical practice early in the 21st century: an international comparison," *Diabetic Medicine*, vol. 32, no. 8, pp. 1036–1050, 2015.

[3] L. P. Plotnick, L. M. Clark, F. L. Brancati, and T. Erlinger, "Safety and effectiveness of insulin pump therapy in children and adolescents with type 1 diabetes," *Clinical Diabetology*, vol. 4, no. 3, pp. 245–252, 2003.

[4] B. W. Bequette, "A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas," *Diabetes technology & therapeutics*, vol. 7, no. 1, pp. 28–47, 2005.

[5] K. Turksoy, L. Quinn, E. Littlejohn, and A. Cinar, "Multivariable adaptive identification and control for artificial pancreas systems," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 883–891, 2013.

[6] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering, *et al.*, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiological measurement*, vol. 25, no. 4, p. 905, 2004.

[7] E. Dassau, H. Zisser, R. A. Harvey, M. W. Percival, B. Grosman, W. Bevier, E. Atlas, S. Miller, R. Nimri, L. Jovanovič, *et al.*, "Clinical evaluation of a personalized artificial pancreas," *Diabetes care*, vol. 36, no. 4, pp. 801–809, 2013.

[8] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, "Deep reinforcement learning for closed-loop blood glucose control," in *Machine Learning for Healthcare Conference*, pp. 508–536, PMLR, 2020.

[9] F. Baldisseri, D. Menegatti, and A. Wrona, "Deep deterministic policy gradient control of type 1 diabetes," in *2024 European Control Conference (ECC)*, pp. 868–873, IEEE, 2024.

[10] G. Lops, V. A. Racanelli, G. Manfredi, L. De Cicco, and S. Mascolo, "A deep deterministic policy gradient control algorithm for automatic insulin delivery," *IFAC-PapersOnLine*, vol. 59, no. 2, pp. 13–18, 2025. 1st IFAC Workshop on Engineering Diabetes Technologies EDT 2025.

[11] F. Baldisseri, M. M. Atanasious, V. Becchetti, A. Di Paola, G. Lops, D. Menegatti, A. Wrona, S. Mascolo, and F. D. Priscoli, "A quantitative comparison of deep reinforcement learning algorithms for type 1 diabetes control," in *2025 11th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1, pp. 1073–1078, IEEE, 2025.

[12] G. Lops, G. Manfredi, V. A. Racanelli, L. De Cicco, and S. Mascolo, "A safety aware deep reinforcement learning technique for automated insulin delivery," in *2025 33rd Mediterranean Conference on Control and Automation (MED)*, pp. 120–125, 2025.

[13] M. M. Atanasious, V. Becchetti, F. Baldisseri, D. Menegatti, and A. Wrona, "Deep reinforcement learning control of type-1 diabetes with cross-patient generalization," in *2024 32nd Mediterranean Conference on Control and Automation (MED)*, pp. 221–226, IEEE, 2024.

[14] S. Wang and W. Gu, "An improved strategy for blood glucose control using multi-step deep reinforcement learning," in *Proceedings of the 2024 16th International Conference on Bioinformatics and Biomedical Technology*, pp. 196–203, 2024.

[15] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1223–1232, 2020.

[16] R. S. Sutton, A. G. Barto, *et al.*, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[18] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv preprint arXiv:1805.11074*, 2018.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, pp. 1861–1870, Pmlr, 2018.

[20] L. Zhao, K. Gatsis, and A. Papachristodoulou, "Stable and safe reinforcement learning via a barrier-lyapunov actor-critic approach," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 1320–1325, IEEE, 2023.

[21] P. Abuin, P. S. Rivadeneira, A. Ferramosca, and A. H. González, "Artificial pancreas under stable pulsatile mpc: Improving the closed-loop performance," *Journal of Process Control*, vol. 92, pp. 246–260, 2020.

[22] H. Emerson, M. Guy, and R. McConville, "Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes," *Journal of Biomedical Informatics*, vol. 142, p. 104376, 2023.

[23] M. R. Marvin, S. E. Inzucchi, and B. J. Besterman, "Computerization of the yale insulin infusion protocol and potential insights into causes of hypoglycemia with intravenous insulin," *Diabetes technology & therapeutics*, vol. 15, no. 3, pp. 246–252, 2013.