Optimizing Dynamic Treatment Regimes via Volatile Contextual Gaussian Process Bandits

Ahmet Alparslan Celik [1] Cem Tekin [1]

Abstract

Management of chronic diseases such as diabetes mellitus requires adaptation of treatment regimes based on patient characteristics and response. There is no single treatment that fits all patients in all contexts; moreover, the set of admissible treatments usually varies over the course of the disease. In this paper, we address the problem of optimizing treatment regimes under time-varying constraints by using volatile contextual Gaussian process bandits. In particular, we propose a variant of GP-UCB with volatile arms, which takes into account the patient's context together with the set of admissible treatments when recommending new treatments. Our Bayesian approach is able to provide treatment recommendations to the patients along with confidence bounds which can be used for risk assessment. We use our algorithm to recommend bolus insulin doses for type 1 diabetes mellitus patients. Simulation studies show that our algorithm compares favorably with traditional blood glucose regulation methods.

1. Introduction

Treatment of chronic diseases requires long term commitment and adaptation to ever evolving conditions. As there is no one-size-fits-all approach, treatments must be adapted based on changing patient characteristics. Within this context, there has been a surge of interest in using machine learning techniques for identifying optimal personalized treatment regimes.

Since management of chronic diseases requires repeatedly making decisions, more data about the patient's response is accumulated over time. Moreover, data collected from the patient depends on the course of the treatment. Therefore, supervised learning methods—which require offline training data—are unable to produce accurate recommendations in the long run. As the patient characteristics evolve over time, treatment must be adjusted to maximize the benefit while minimizing the risks. This requires an intricate balance between exploration and exploitation. The best treatment under the current context must be identified with sequential experimentation while ensuring safety and efficiency at the same time.

In this work, we model optimization of dynamic treatment regimes as a volatile contextual Gaussian process (GP) bandit. The predictive power and non-parametric flexibility of GPs allow us to accurately model the relationship between treatment and response under different patient contexts. Moreover, sequential experimentation via upper confidence bounds constructed using the posterior mean and covariance functions of the GP allows us perform safe experimentation over a set of admissible treatments calculated based on the current context. Our framework allows flexibility in forming the set of admissible treatments, and enables safe experimentation among a set of treatments identified by clinical guidelines or baseline interpretable formula-based systems. In particular, we focus on using our framework for personalized bolus insulin dose recommendation for type 1 diabetes mellitus (T1DM) patients.

Modeling bolus insulin recommendation as a volatile contextual bandit. T1DM is a chronic autoimmune disease characterized by insulin deficiency due to pancreatic β cell loss. Lack of insulin regulation in diabetic patients can have serious adverse effects due to hypoglycemia and hyperglycemia, i.e., low and very high blood glucose (BG) levels, respectively, which might result in immediate hospitalization or even death, or long term damage to various organs at risk unless effectively treated. Therefore, T1DM patients must regulate their BG by regularly administering bolus insulin before the meals in order to avoid hyperglycemia and its adverse affects. This is a complicated process, since the optimal insulin dose depends on a variety of exogenous and endogenous contexts such as time of the day, pre-meal BG level, carbohydrate content of the meal, basal insulin levels, etc.

[1] Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey. Correspondence to: Ahmet Alparslan Celik <acelik@ee.bilkent.edu.tr>, Cem Tekin <cemtekin@ee.bilkent.edu.tr>.

Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 38^{th} International Conference on Machine Learning, 2021.

Our model treats each meal event as a decision epoch. For each context, it recommends a bolus insulin dose in order to achieve optimal BG regulation in the long run. In order to prevent short-term and long-term adverse effects due to possible hypoglycemia and hyperglycemia events, we employ safety constraints to identify the set of admissible treatments for each context. This induces volatility over the set of treatments, hence some treatments may not be explored at certain times. In particular, we consider safe experimentation around the dose suggested by a standard formula-based bolus insulin calculator in our simulations. We want to keep the BG as close as possible to the target value (e.g., 112.5 mg/dL) after each meal event by learning the best dose for each context. Obviously, this requires us to explore different doses. However, exploration needs to be safe by ensuring that recommended doses do not lead to hypoglycemia or hyperglycemia events, which translates to keeping the patients' postprandial BG between 70-180 mg/dL. In the end, we demonstrate that safe experimentation around a formula-based interpretable benchmark provides improved BG regulation.

Our contribution and comparison with related works. Our work builds on the formalism of contextual bandits and GP bandits. It allows us to recommend doses adapted to both patient and meal event characteristics. Prior works on contextual bandits such as (Lu et al., 2010) and (Langford & Zhang, 2007) mostly consider the problem within the context of online recommender systems. They propose non-Bayesian approaches to minimize the regret, which is a long-term performance metric. (Krause & Ong, 2011) proposes contextual GP bandits, and provides bounds on the regret that depend on information gain. GP bandit models have recently gained attention due to their trustable nature as the posterior covariance can fully capture the uncertainty in recommendations. Likewise, their performance can be theoretically proven via a rigorous regret analysis. These aspects are particularly important in healthcare applications, in which the prevention of unwanted consequences during the course of treatment is of the highest priority. We consider contextual GP bandits with dynamic (possibly context-dependent) arm availability and base our analysis on a tighter form of information gain, which is adapted to the sequence of observed contexts.

Our contributions are two fold. First, we adapt contextual Gaussian Process Upper Confidence Bound (CGP-UCB) algorithm given in (Krause & Ong, 2011) to the volatile arm setup, and call this adaptation VCGP-UCB. For any admissible sequence of treatments, we show that the regret of VCGP-UCB is $\tilde{O}(\sqrt{T\gamma_T^{vol}})$, where γ_T^{vol} represents the volatility-adapted maximum information gain. This term is always less than or equal to the unrestricted maximum information gain γ_T in (Krause & Ong, 2011). Therefore, for typical covariance functions such as squared-exponential or Matern, regret growth is sublinear in time. In terms of diabetes treatment, this implies that the average number of decision-epochs in which a suboptimal treatment is suggested converges to zero. To the best of our knowledge, this paper is the first to apply Bayesian bandit strategies for optimal BG control.

Another related strand of literature makes use of Markov decision processes (MDPs) for clinical decision-making (Bennett & Hauser, 2013). MDPs are especially suitable for problems in which past and present outcomes in the course of treatment are highly correlated. Within the context of diabetes management, using MDPs is more suitable for insulin-pump therapy with continuous glucose monitoring (CGM), where insulin is administered continuously over time. On the other hand, in this paper, we focus on diabetes treatment with multiple daily injections (e.g., 3-4 times a day). Since bolus insulin is rapid acting, its effect mostly wears out before the next meal. Moreover, our context includes fasting BG, which efficiently captures residual effect of the previous insulin dose. Thus, conditioned on the context, outcome of each dose can be regarded as i.i.d., and hence, contextual bandits are a better fit for our problem. Indeed, bandit algorithms can be applied in other dynamic treatment assignment and drug dosage problems, where the effect of treatment significantly decays before the next decision-epoch.

Second half of our contribution is focused on automatic BG control in diabetes mellitus patients. In the literature, there are mainly three categories of models for BG regulation: standard formula-based bolus insulin calculators (Schmidt & Nørgaard, 2014), closed-loop control systems (aka artificial pancreas) (Cobelli et al., 2011) and models based on reinforcement learning (Tejedor et al., 2020). The block diagram of our model, which is a closed-loop system that uses VCGP-UCB as the controller, is given in Figure 1.

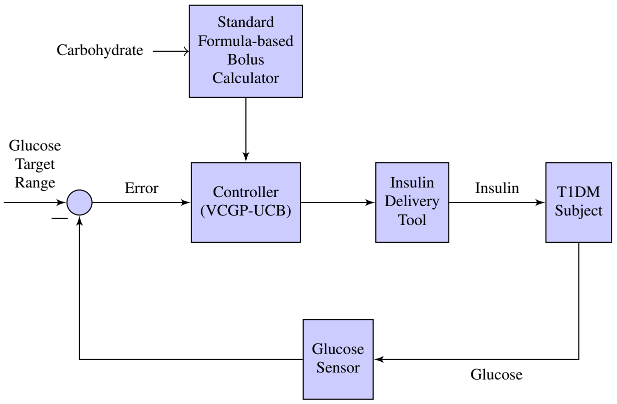

Figure 1. Our system model. We use volatile CGP-UCB to perform safe experimentation around the recommendation produced by a standard formula-based bolus calculator.

Table 1. Comparison with the related bandit models.

Properties	This work	GP-UCB (Srinivas et al., 2012)	CGP-UCB (Krause & Ong, 2011)	SAFE-OPT (Sui et al., 2015)	STAGE-OPT (Sui et al., 2018)	Safe-LUCB (Amani et al., 2019)	SGP-UCB (Amani et al., 2020)
Contextual	Yes	No	Yes	No	No	No	No
GP prior	Yes	Yes	Yes	Yes	Yes	No	Yes
Arm volatility	Exogenous†	No	No	Endogenous‡	Endogenous‡	Endogenous‡	Endogenous‡
Arm set	Finite	Can be infinite	Can be infinite	Finite	Can be infinite	Can be infinite	Finite
Context set	Can be infinite	-	Can be infinite	-	-	-	-
Information gain term in the regret bound	Volatility-adapted maximum	Maximum	Maximum	-	Maximum	-	Maximum

†Volatility induced by an exogenous process (e.g., context arrivals). ‡Volatility induced by safety constraints.

Table 2. Comparison with the diabetes treatment models.

Properties	This work	Standard bolus calculator (Schmidt & Nørgaard, 2014)	Adaptive bolus calculator (Herrero et al., 2017)	Bio-inspired artificial pancreas (Reddy et al., 2014)	Actor-Critic (Daskalaki et al., 2013)
Model	GP bandits	Formula-based	CBR & R2R†	Bio-modeling via DEs‡	Actor-Critic RL
Adaptiveness (to patient response)	High	Limited	High	Limited	High
Theoretical performance (regret) guarantees	Yes	No	No	No	No
Safe exploration	Yes	No	No	No	No

†CBR: Case-Based Reasoning, R2R: Run-To-Run control. ‡DE: Differential Equation.

Standard formula-based bolus insulin calculators are widely used to calculate pre-meal bolus insulin dose thanks to their simplicity and interpretability (Walsh et al., 2011). An example is

$$\text{Bolus insulin} = \left(\frac{\text{CHO}}{\text{ICR}} + \frac{\text{G}_M - \text{G}_T}{\text{CF}} \right)^+ \tag{1}$$

where $(a)^+ = \max\{a, 0\}$, CHO (g) the estimated amount of carbohydrate intake, ICR (g/U) is the insulin-to-carbohydrate-ratio, G_M (mg/dL) is the measured meal time BG, G_T (mg/dL) is the target BG and CF (mg/dL/U) is the insulin correction factor (Schmidt & Nørgaard, 2014). Recently, more sophisticated versions of the standard formula-based bolus calculator are developed such as advanced (Herrero et al., 2014) and adaptive (Herrero et al., 2017) calculators. While standard bolus insulin calculators serve as a transparent and interpretable baseline, they lack sophistication and ignore other contextual variables which might have an effect on post-meal BG levels (Schmidt & Nørgaard, 2014). Thus, precise regulation of BG control requires carefully adjusting bolus doses around a safe baseline.

A different line of research exists under the name *artificial pancreas*. It refers to the closed-loop control of BG in diabetes, which is usually achieved by subcutaneous CGM and insulin infusion. Within this context, control-theoretic approaches such as proportional-integral-derivative (PID) (Sherr et al., 2013) and model predictive control (MPC) (Kovatchev et al., 2013) have been widely investigated. In recent years, reinforcement learning (RL) based control models are also developed (Tejedor et al., 2020). Luckett et al. (2020) consider V-learning for estimating dynamic treatment regimes using a parametric class of policies, where exploration can be achieved using ϵ-greedy style randomization. They use V-learning to estimate an optimal treatment policy for T1DM patients. However, most of these methods are designed for use with insulin-pump therapy and CGM—tools which may not be readily available for most diabetic patients. In contrast, our algorithm provides recommendations tailored for traditional BG regulation, which can be performed by using finger-stick BG meters and multiple daily bolus insulin injections, and does not require access to expensive medical equipment. Compared to our work, most of other RL-based solutions are not interpretable due to their black-box nature, and they do not come with rigorous performance analysis.

We provide a detailed comparison of our work with related works in bandits and BG control for diabetes in Tables 1 and 2.

2. Problem Formulation

Let S denote the set of all treatments (finite), Z denote the set of all patient contexts (can be infinite) and T denote the number of iterations. For $t \in [T] := \{1, \ldots, T\}$, let z_t be the patient context and $S_t \subseteq S$ be the set of admissible treatments in round t. Our setup is general in the sense that it allows S_t to be a function of z_t. Since only a subset of treatments are available in each round, this setup corresponds to a bandit problem with volatile arms. We consider the problem of sequentially optimizing effectiveness of the given treatments, which is characterized by an unknown reward function $f : S \times Z \to \mathbb{R}$. Within the context of bolus insulin recommendation, reward measures closeness of postprandial BG to target BG (see Section 5 for details).

At the beginning of reach round t, the learner receives a patient context $z_t \in Z$, chooses a treatment $s_t \in S_t$, and then, observes a noisy reward $y_t = f(s_t, z_t) + \epsilon_t$, where ϵ_t denotes the zero mean Gaussian noise with σ^2 variance, independent across the rounds. The objective is to maximize the cumulative reward $\sum_{t=1}^{T} f(s_t, z_t)$ without knowing f beforehand. The optimal treatment in round t is denoted by $s_t^* = \operatorname{argmax}_{s \in S_t} f(s, z_t)$. Suboptimality of the treatment in round t is given as $r_t = f(s_t^*, z_t) - f(s_t, z_t)$, which is also called the instantaneous regret. The cumulative regret is the sum of instantaneous regrets, i.e., $R_T = \sum_{t=1}^{T} r_t$. It is well known that maximizing the cumulative reward is equivalent to minimizing the cumulative regret.

Let $X = S \times Z$ denote the set of treatment-context pairs. We assume that f is sampled from a known $GP(\mu, k)$ which is fully characterized by its mean function $\mu : X \to \mathbb{R}$, $\mu(x) = \mathbb{E}[f(x)]$ and covariance (kernel) function $k : X \times X \to \mathbb{R}, k(x, x') = \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))]$. We assume that $\mu \equiv 0$ and that the variance is bounded, i.e., $k(x, x) \leqslant 1$ for all $x \in X$. Given an observation history $\mathcal{A}_t = \{x_1, \ldots, x_t\}$, where $x_t = (s_t, z_t)$, the posterior mean and variance of the GP at point x can be calculated as follows (Rasmussen, 2003):

$$\begin{aligned}
\mu_t(x) &= \boldsymbol{k}_t(x)^T (\boldsymbol{K}_t + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_t, \\
k_t(x, x') &= k(x, x') - \boldsymbol{k}_t(x)^T (\boldsymbol{K}_t + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_t(x'), \\
\sigma_t^2(x) &= k_t(x, x),
\end{aligned}$$

where $\boldsymbol{k}_t(x) = [k(x_1, x), \ldots, k(x_t, x)]^T$, $\boldsymbol{K}_t$ is the kernel matrix $[k(x, x')]_{x, x' \in \mathcal{A}_t}$ and $\boldsymbol{y}_t = [y_1, \ldots, y_t]^T$.

3. Volatile CGP-UCB Algorithm

As our learning algorithm, we use contextual Gaussian Process (CGP) algorithm in (Krause & Ong, 2011) adapted to the volatile setting. Its pseudocode is given in Algorithm 1. Basically, at round t, VCGP-UCB selects treatment

$$s_t = \operatorname*{argmax}_{s \in S_t} \mu_{t-1}(s, z_t) + \beta_t^{1/2} \sigma_{t-1}(s, z_t),$$

where β_t is a non-decreasing function of t (which will be specified later). The selected treatment is the one with the highest upper confidence bound (UCB) index among all admissible treatments in round t. Within the context of bolus insulin dosage for diabetes treatment S_t is set by defining a safe experimentation region around the treatment suggested by a formula-based bolus calculator.

Postprandial BG predictions based on GPs: Treatment recommendations can be accompanied by postprandial BG predictions, by using another GP as a surrogate model for postprandial BG surface $g : S \times Z \to \mathbb{R}$. In this case, each noisy postprandial BG measurement is represented by $g_t = g(s_t, z_t) + \tilde{\eta}_t$, where $\tilde{\eta}_t$ represents the zero mean Gaussian measurement noise with known variance. The following remark explains how probabilities of hyperglycemia and hypoglycemia events can be computed using the GP posterior for g.

Remark 1. *Given a treatment-context pair (s_t, z_t) and BG target range (g_{low}, g_{high}), probability of hyperglycemia and hypoglycemia events, $P(g(s_t, z_t) < g_{low})$ and $P(g(s_t, z_t) > g_{high})$ respectively, can be computed from posterior distribution with mean $\tilde{\mu}_{t-1}(s_t, z_t)$ and variance $\tilde{\sigma}_{t-1}^2(s_t, z_t)$ at round t where $g(s_t, z_t) \sim \mathcal{N}(\tilde{\mu}_{t-1}(s_t, z_t), \tilde{\sigma}_{t-1}^2(s_t, z_t))$. Suppose that $X \sim \mathcal{N}(0, 1)$ and F_X denotes cumulative distribution function of random variable X. Thus, $P(g(s_t, z_t) < g_{low}) = F_X(\frac{g_{low} - \tilde{\mu}_{t-1}(s_t, z_t)}{\tilde{\sigma}_{t-1}(s_t, z_t)})$ and $P(g(s_t, z_t) > g_{high}) = 1 - F_X(\frac{g_{high} - \tilde{\mu}_{t-1}(s_t, z_t)}{\tilde{\sigma}_{t-1}(s_t, z_t)})$.*

Algorithm 1 VCGP-UCB algorithm

```
Input: Input space $X$; GP prior $\mu_0 = 0, \sigma_0, k$
for $t$ = 1 to $T$ do
    Observe context $z_t$, and the set of admissible treatments $S_t$
    Choose $s_t = \operatorname{argmax}_{s \in S_t} \mu_{t-1}(s, z_t) + \sqrt{\beta_t} \sigma_{t-1}(s, z_t)$
    Observe $y_t = f(s_t, z_t) + \epsilon_t$
    Update GP posterior to obtain $\mu_t$ and $\sigma_t$
end for
```

4. Theoretical Analysis

4.1. Main Regret Bound

Consider a fixed sequence of patient contexts $\boldsymbol{z}_T = [z_1, \ldots, z_T]$. Let $X_t = (z_t, S_t)$ and Let $\boldsymbol{X}_T = X_1 \times \cdots \times X_T$ represent the Cartesian product of all admissible treatment-context pairs up to round T. For a given sequence of treatment-context pairs A, let $\boldsymbol{y}_A$ denote the $|A|$-dimensional vector whose ith element corresponds to the reward observed for the ith treatment-context pair in A. The quantity governing our regret bounds is the volatility-adapted maximum information gain after T rounds (given $\boldsymbol{z}_T$), which is defined as

$$\gamma_T^{vol} = \max_{A \in \boldsymbol{X}_T} I(\boldsymbol{y_A}; \boldsymbol{f_A}),$$

where $\boldsymbol{f_A} = [f(x)]_{x \in A}$ and $I(\boldsymbol{y_A}; \boldsymbol{f_A})$ is the mutual information between f and reward observations at points in A.

First, we state our main theorem. Its proof is built on the proofs by (Srinivas et al., 2012) and (Krause & Ong, 2011).

Theorem 1. *Fix $\delta \in (0, 1)$. When volatile CGP-UCB is run with $\beta_t = 2log(|S|t^2\pi^2/6\delta)$, for an arbitrary fixed* *sequence of contexts $\boldsymbol{z}_T$, we have*

$$Pr\{R_T \leqslant \sqrt{C_1 T \beta_T \gamma_T^{vol}} \quad \forall T \geqslant 1\} \geqslant 1 - \delta \ ,$$

where $C_1 = 8/\log(1 + \sigma^{-2})$.

The above theorem provides an information-type regret bound for volatile CGP-UCB. The maximum information gain γ_T^{vol} depends on both the context sequence and the set of admissible treatments. This contrasts with the information gain terms defined in (Srinivas et al., 2012) and (Krause & Ong, 2011), which for our setting can be written as

$$\gamma_T = \max_{A \in \tilde{\boldsymbol{X}}_T} I(\boldsymbol{y_A}; \boldsymbol{f_A}) \ ,$$

where $\tilde{\boldsymbol{X}}_T = \times_{t=1}^T X$ is the Cartesian product of T copies of $X = S \times Z$. It is obvious that $\gamma_T^{vol} \leqslant \gamma_T$, and in practice γ_T^{vol} might be much smaller than γ_T. The reason for this is that we are constrained in each round to pick a treatment from the set of admissible treatments S_t instead of S, and we incur regret only when the selected treatment differs from the best available treatment of that round. Therefore, sublinear regret bounds in (Srinivas et al., 2012) and (Krause & Ong, 2011) obtained for squared exponential and Matern kernels also hold in our setting.

As a result, Theorem 1 shows that when such kernels are used, time-averaged regret given as R_T/T converges to zero with high probability. In the context of BG regulation, this implies that our algorithm's insulin recommendations converge to the optimal insulin recommendations over time.

4.2. Proof of Theorem 1

Lemma 1. *(Lemma 5.1 in (Srinivas et al., 2012)) Fix the sequence of contexts $\boldsymbol{z}_T$. Let $\delta \in (0, 1)$ and $\beta_t = 2log(|S|\pi_t\delta)$, where $\sum_{t\geqslant 1} \pi_t^{-1} = 1$, $\pi_t > 0$. Then, the following event holds with probability at least $1 - \delta$.*

$$|f(s, z_t) - \mu_{t-1}(s, z_t)| \leqslant \beta_t^{1/2} \sigma_{t-1}(s, z_t) \quad \forall s \in S, \forall t \geqslant 1 \ .$$

Lemma 2. *Fix $t \geqslant 1$. If $|f(s, z_t) - \mu_{t-1}(s, z_t)| \leqslant \beta_t^{1/2}\sigma_{t-1}(s, z_t)$ for all $s \in S_t$, then the instantaneous regret r_t is bounded by $2\beta_t^{1/2}\sigma_{t-1}(x_t)$.*

Proof. Let $s_t^* \in \mathrm{argmax}_{s \in S_t} f(s, z_t)$ be an optimal action. By definition of s_t : $\mu_{t-1}(s_t, z_t) + \beta_t^{1/2}\sigma_{t-1}(s_t, z_t) \geqslant \mu_{t-1}(s_t^*, z_t) + \beta_t^{1/2}\sigma_{t-1}(s_t^*, z_t) \geqslant f(s_t^*, z_t)$, where the last inequality is due to Lemma 1. Therefore, $r_t = f(s_t^*, z_t) - f(s_t, z_t) \leqslant \beta_t^{1/2}\sigma_{t-1}(s_t, z_t) + \mu_{t-1}(s_t, z_t) - f(s_t, z_t) \leqslant 2\beta_t^{1/2}\sigma_{t-1}(s_t, z_t)$. □

Lemma 3 (Lemma 5.3 in (Srinivas et al., 2012)). *Given $\boldsymbol{z}_T$, the information gain for the points selected can be expressed in terms of the predictive variances. If $\boldsymbol{f_T} = (f(x_t)) \in \mathbb{R}^T$:*

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) = \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2}\sigma_{t-1}^2(x_t)).$$

Lemma 4. *Fix the sequence of contexts $\boldsymbol{z}_T$. Pick $\delta \in (0, 1)$ and let β_t be defined as in Lemma 1.1. Then, the following holds with probability at least $1 - \delta$:*

$$R_T \leqslant \sqrt{T\beta_T C_1 I(\boldsymbol{y}_T; \boldsymbol{f}_T)} \leqslant \sqrt{TC_1\beta_T\gamma_T^{vol}} \quad \forall T \geqslant 1 \ ,$$

where $C_1 = 8/\log(1 + \sigma^{-2})$.

Proof. The proof is similar to the proof of Lemma 5.4 in (Srinivas et al., 2012). By Lemmas 1.1 and 1.2, we have that

$$Pr\{r_t^2 \leqslant 4\beta_t\sigma_{t-1}^2(x_t) \quad \forall t \geqslant 1\} \geqslant 1 - \delta \ .$$

Also note that

(i) β_t is non-decreasing in t, and thus, $\beta_T \geqslant \beta_t$ for $\forall T \geqslant t$.

(ii) Note that $\sigma^{-2}\sigma_{t-1}^2(x_t) \leqslant \sigma^{-2}k(x_t, x_t) \leqslant \sigma^{-2}$, where $k(x_t, x_t) \leqslant 1$ due to assumption of bounded variance. Let $C_2 = \sigma^{-2}/\log(1 + \sigma^{-2})$. Note that $C_2 \geqslant 1$ and $\frac{s^2}{\log(1+s^2)} \leqslant C_2$ for $s^2 \in [0, \sigma^{-2}]$.

(iii) By cannons i, ii and Lemma 3 we bound r_t.

$$\begin{aligned} \sum_{t=1}^T r_t^2 &\leqslant \sum_{t=1}^T 4\beta_T\sigma^2 C_2 \log(1 + \sigma^{-2}\sigma_{t-1}^2(x_t)) \\ &= 8\beta_T\sigma^2 C_2 \frac{1}{2} \sum_{t \geqslant 1} \log(1 + \sigma^{-2}\sigma_{t-1}^2(x_t)) \\ &\leqslant \beta_T C_1 \gamma_T^{vol} \ . \end{aligned}$$

Finally, by cannons i, ii and iii, we bound R_T. Result follows from Cauchy-Schwarz inequality.

$$\begin{aligned} R_T^2 = (\sum_{t=1}^T r_t)^2 &\leqslant T \sum_{t=1}^T r_t^2 \\ &\leqslant T\beta_T C_1 \gamma_T^{vol} \ . \end{aligned}$$

Thus, $R_T \leqslant \sqrt{T\beta_T C_1 \gamma_T^{vol}}$. □

5. Experiments

We perform *in silico* evaluation with UVa/PADOVA T1DM 2008 Simulator (Kovatchev et al., 2009), (Xie, 2018). The glucose-insulin model was created to substitute certain preclinical trials and approved by U.S. FDA in 2008 as a reliable framework for *in silico* trials and for closed-loop hormone controller design, testing, and validation. We use all 10 *in-silico* adult patients included in the simulator in

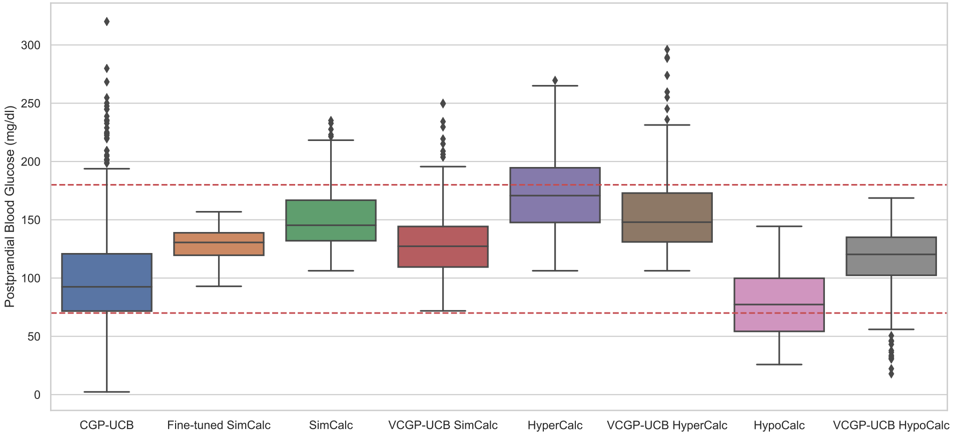

Figure 2. Box plot of postprandial BG distributions of different methods.

our experiments. We use (GPy, 2012) to implement VCGP-UCB. We compare our model against CGP-UCB (Krause & Ong, 2011) with non-volatile arms and formula-based bolus calculators (Walsh et al., 2011) with different amount of miscalibration. We demonstrate that our algorithm can successfully personalize well-accepted formula-based calculator results with safe exploration. We highlight robustness by introducing various miscalibrations to the standard formula-based calculator.

As our evaluation metrics, we use glycemic outcome measures that are widely accepted by the diabetes management community to evaluate glycemic control (Maahs et al., 2016) and glycemic risk measures (Kovatchev et al., 2000). Glycemic outcome metrics are mean BG, percentage time in BG target range [70,180] mg/dL (%inT), percentage time below target (i.e. hypoglycemia) (%<T) and percentage time above target (i.e. hyperglycemia) (%>T). Glycemic risk indices are low blood glycemic index (i.e. risk of hypoglycemia) (LBGI), high blood glycemic index (i.e. risk of hyperglycemia) (HBGI) and risk index (RI=LBGI+HBGI). LBGI is used to group subjects regarding to their long-term risk for hypoglycemia. The risk categories are minimal, low, moderate and high risk, with LBGI of below 1.1, 1.1-2.5, 2.5-5.0, and above 5.0, respectively (Kovatchev et al., 2003).

The BG measurement scale (20-600 mg/dL) is asymmetric. Hypoglycemia range (below 70 mg/dL) is much narrower than the hyperglycemia range (above 180 mg/dL). BG values are mapped to risk space where the minimum value of 0 is achieved at BG value of 112.5 mg/dL while its maximum value of 100 is achieved at 20 mg/dL and 600 mg/dL. We get risk value of 7.7 at the BG values of 70 and 180 mg/dL. Given a set of postprandial BG measurements, HBGI and LBGI are defined as the average hyperglycemia and hypoglycemia risk scores, respectively, where RI denotes overall risk score and is equal to the sum of HBGI and LBGI. The lower the risk values of LBGI and HBGI gets, the less the risk of hypoglycemia and hyperglycemia becomes since the postprandial BG gets closer to 112.5 mg/dL (Kovatchev et al., 2003).

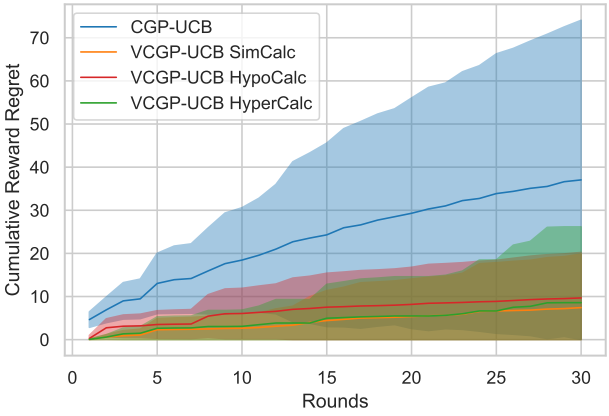

Figure 3. Cumulative reward regrets of volatile CGP-UCB algorithms. Error bars represent ± one standard deviation.

5.1. Setup

The context variables for our algorithms are the amount of carbohydrate intake, fasting BG level, time between meal and insulin intake, and time between meal and postprandial BG measurement. We generate 30 different meal events with different fasting BG values. Carbohydrate intake and fasting BG values are uniformly sampled from ranges 20-80 g and 70-180 mg/dL, respectively. Time between meal and insulin intake is set to 0. Time between meal and postprandial BG measurement is set to 150 minutes. Bolus insulin dose ranges from 0 to 30 units with 0.1 increments. Min-max normalization is used for all contexts and arms by mapping to unit interval.

Table 3. Glycemic results averaged over glycemic outcome of all ten in-silico adult patients. For postprandial BG, mean and ± one standard deviation are reported.

Method	BG (mg/dL)	%inT	%<T	%>T	RI	LBGI	HBGI
CGP-UCB	104.38±50.90	0.65	0.23	0.11	9.42	7.08	2.34
Fine-tuned SimCalc	128.85±10.55	1.00	0.00	0.00	0.37	0.03	0.34
SimCalc	152.05±18.00	0.83	0.00	0.17	1.30	0.00	1.30
VCGP-UCB SimCalc	129.77±23.32	0.94	0.00	0.06	1.36	0.24	1.12
HyperCalc	174.60±30.33	0.60	0.00	0.40	2.54	0.00	2.54
VCGP-UCB HyperCalc	155.94±25.05	0.79	0.00	0.21	1.55	0.00	1.55
HypoCalc	78.13±18.06	0.58	0.42	0.00	4.38	4.33	0.05
VCGP-UCB HypoCalc	115.40±25.12	0.92	0.08	0.00	3.59	3.26	0.34
VCGP-UCB SimCalc (± 40% exploration margin)	137.57±17.89	0.91	0.00	0.09	1.01	0.07	0.94
VCGP-UCB SimCalc (± 80% exploration margin)	123.97±29.57	0.88	0.06	0.06	1.79	0.51	1.28
VCGP-UCB SimCalc ($\tilde{\epsilon}_5$*)	130.73±24.56	0.93	0.00	0.07	1.35	0.27	1.07
VCGP-UCB SimCalc ($\tilde{\epsilon}_{10}$*)	130.40±25.70	0.89	0.02	0.09	1.50	0.36	1.14
VCGP-UCB SimCalc ($\tilde{\epsilon}_{20}$*)	131.22±32.28	0.87	0.04	0.09	2.21	1.07	1.14

*$\tilde{\epsilon}_s$ denotes zero mean Gaussian noise with s^2 variance added on top of postprandial BG value returned by the simulator.

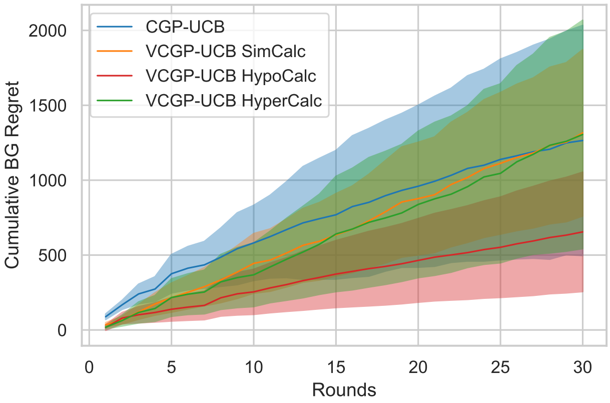

Figure 4. Cumulative BG regrets of volatile CGP-UCB algorithms. Error bars represent ± one standard deviation.

We use a composite ARD kernel function defined over joint context-arm set for GP-based algorithms. The kernel consists of additive combination of Matern 5/2 and linear covariance functions. Matern 5/2 length scales are set as 0.5 for carbohydrate intake, fasting BG level and bolus insulin, and 5 for others. Variances are set as 1 for all kernels. Noise variance is set to 1. β_t is set to 4.

The algorithms have the BG target of 112.5 mg/dL, which is regarded as the clinical center of the BG scale, i.e. the BG value associated with zero risk index (Kovatchev et al., 2003). In accordance with this, in order to evaluate performance of bolus recommendations given contexts, we define the following loss function based on resulting postprandial BG, similar to (Keramati et al., 2020)

$$loss(\tilde{g}) = \begin{cases} (\tilde{g}-6)^2/5 & \tilde{g} < 6 \\ (\tilde{g}-6)^2/10 & \tilde{g} \geqslant 6 \end{cases} \tag{2}$$

where $\tilde{g} = g/18.75$ is such that $\tilde{g} = 6$ corresponds to $g = 112.5$ mg/dL postprandial BG. Note that the BG measurement scale is asymmetric since the hypoglycemia range (below 70 mg/dL) is numerically much narrower than the hyperglycemia range (above 180 mg/dL). Moreover, hypoglycemia is considered to be more risky than hyperglycemia. This justifies choosing a higher loss for hypoglycemia. In each round t, learning algorithms observe the context z_t, choose a bolus dose s_t, observe the postprandial BG measurement g_t, and receive reward $y_t = -loss(\tilde{g}_t)$.

During the experiments, arm volatility is introduced through safety constraints for the insulin dose. These constraints are defined by taking formula-based bolus calculators (see Eq. 1) as baseline models. Our first calculator uses ICR and CF values of each patient that comes with the simulator, which are not very well-tuned. We name this calculator as SimCalc. We also provide results for a manually tuned version of SimCalc named as Fine-tuned SimCalc. For this, we scale down ICR and CF by a fixed amount in order to prevent hypoglycemia or hyperglycemia as much as possible. In addition, we test against two types of miscalibrated formula-based calculators that cause either a certain amount of hypoglycemia or hyperglycemia by scaling down and up ICR and CF values by the same fixed constant, respectively. We name these calculators as HypoCalc and HyperCalc. For each of the calculators SimCalc, HypoCalc and HyperCalc, we perform personalization through safe-exploration with margins of ± 60% of corresponding formula-based calculator recommendation by using VCGP-UCB. We report the corresponding results under the names VCGP-UCB SimCalc, VCGP-UCB HypoCalc and VCGP-UCB HyperCalc.

We train separate VCGP-UCB models for each patient. GPs are initialized with 2 samples per patient with carbohydrate intake, fasting BG levels 30 g, 70 g and 100 mg/dL, 150 mg/dL respectively. Bolus doses for these events are given using SimCalc.

5.2. Results

For each method of treatment, we report postprandial BG distribution of all patients in Figure 2, and the average glycemic metrics of all patients in Table 3. The results indicate that VCGP-UCB sufficiently compensates for miscalibrations in calculators as it shifts the mean BG towards the target value even when the baseline calculators are functioning inefficiently. VCGP-UCB fine-tunes glycemic control by exploring around the doses recommended by calculators. Also, it is safer than CGP-UCB as it has lower LBGI and HBGI values. This results from restricting dose exploration around baseline calculators.

Regret of VCGP-UCB with respect to the best available treatment over time given in Section 2 is shown in Figure 3. This regret is computed by averaging over all patients. VCGP-UCB achieves smaller regret compared to CGP-UCB since its exploration is restricted to set of admissible doses around the formula-based calculator.

In addition, we define the regret with respect to BG target as

$$R_T^g = \sum_{t=1}^{T} |g_{target} - \mathbb{E}[g_t|s_t, z_t]|,$$

where $g_{target} = 112.5$ mg/dL. Figure 4 shows how R_T^g averaged over all patients evolves over time. R_T^g increases faster for VCGP-UCB with SimCalc and HyperCalc because the safe exploration margins are not wide enough to include optimal insulin dose for some patients and contexts. Note that these calculators are more skewed towards the hyperglycemia region.

In Table 3, we also investigate the affect of BG measurement noise on performance of VCGP-UCB when used with SimCalc. Results show that BG noise up to 20 standard deviation is well tolerated by VCGP-UCB. In addition, we also study how the change of safety margin affects the performance of VCGP-UCB when used with SimCalc. Expanding the margin from 60% to 80% results in a slight increase in risk indices.

5.3. Limitations and Future Research Directions

While experimental results demonstrate the effectiveness of the proposed approach in BG regulation, the current model and algorithm have certain limitations. Next, we discuss these and potential remedies.

Safety. Safety of VCGP-UCB depends on the *reliability* of baselines around which exploration is performed. Exploration helps improving performance when clinically accepted baselines are inaccurate or miscalibrated up to a certain degree. However, utmost care should be taken when choosing the baselines, as faulty baselines can result in unsafe recommendations.

Model mismatch. In real-world, BG measurements are recorded by BG sensors. Device specific measurement noise can disturb the accuracy of the GP-based model. In particular, noise on the BG measurement will propagate through Eq. 2, resulting in non-Gaussian noise on rewards. This contamination will create model mismatch and can reduce the effectiveness of VCGP-UCB especially when BG measurement noise has high variance. Nevertheless, for most of the commercially available BG meters (fingerstick testing) measurement errors are small enough to meet International Organization for Standardization (ISO) criteria (Bergenstal, 2008).

There are two possible solutions to mitigate model mismatch. One can use Warped GPs (Snelson et al., 2004; Lázaro-Gredilla, 2012) to find a nonlinear transformation of the reward data, which can be accurately modeled using GPs. Alternatively, one can assume that BG measurement error is Gaussian and use a GP to model BG surface. This GP can be used to construct confidence intervals of BG values given treatment and context. Optimistic reward of each admissible treatment can be found by minimizing $loss(\tilde{g})$ over the confidence interval of BG values. Then, the treatment that minimizes the optimistic loss can be recommended.

Other sources of errors that require further investigation include errors made by the patient in reporting the correct value of carbohydrate intake (context) and insulin dosage (arm).

6. CONCLUSION

We adapted Contextual Gaussian Process Upper Confidence Bound algorithm to the volatile bandit setup, and proposed volatile CGP-UCB. We showed how volatile CGP-UCB can be used to optimize treatment regimes under time-varying constraints. Volatile CGP-UCB achieves $\tilde{O}(\sqrt{T\gamma_T^{vol}})$ regret, and enables safe exploration around a formula-based treatment strategy. This demonstrates the applicability of bandit algorithms in fine-tuning treatment decisions around interpretable baseline treatment strategies employed in clinical practice. We used our algorithm as a closed-loop system for BG regulation in type 1 diabetes mellitus patients. Simulation results show that our algorithm has the potential to improve BG regulation compared to formula-based methods.

Acknowledgements

This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 215E342 and in part by the BAGEP Award of the Science Academy.

References

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pp. 9256–9266, 2019.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Regret bound for safe Gaussian process bandit optimization. In *Learning for Dynamics and Control*, pp. 158–159. PMLR, 2020.

Bastani, M. Model-free intelligent diabetes management using machine learning. 2014.

Bennett, C. C. and Hauser, K. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial Intelligence in Medicine*, 57(1):9–19, 2013.

Bergenstal, R. M. Evaluating the accuracy of modern glucose meters. *Insulin*, 3(1):5–14, 2008.

Cobelli, C., Renard, E., and Kovatchev, B. Artificial pancreas: past, present, future. *Diabetes*, 60(11):2672–2682, 2011.

Daskalaki, E., Diem, P., and Mougiakakou, S. G. An actor–critic based controller for glucose regulation in type 1 diabetes. *Computer Methods and Programs in Biomedicine*, 109(2):116–125, 2013.

GPy. GPy: A Gaussian process framework in python. `http://github.com/SheffieldML/GPy`, 2012.

Herrero, P., Pesl, P., Reddy, M., Oliver, N., Georgiou, P., and Toumazou, C. Advanced insulin bolus advisor based on run-to-run control and case-based reasoning. *IEEE Journal of Biomedical and Health Informatics*, 19(3): 1087–1096, 2014.

Herrero, P., Bondia, J., Adewuyi, O., Pesl, P., El-Sharkawy, M., Reddy, M., Toumazou, C., Oliver, N., and Georgiou, P. Enhancing automatic closed-loop glucose control in type 1 diabetes with an adaptive meal bolus calculator–in silico evaluation under intra-day variability. *Computer Methods and Programs in Biomedicine*, 146:125–131, 2017.

Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. Being optimistic to be conservative: Quickly learning a CVaR policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4436–4443, 2020.

Kovatchev, B. P., Straume, M., Cox, D. J., and Farhy, L. S. Risk analysis of blood glucose data: Az quantitative approach to optimizing the control of insulin dependent diabetes. *Computational and Mathematical Methods in Medicine*, 3(1):1–10, 2000.

Kovatchev, B. P., Cox, D. J., Kumar, A., Gonder-Frederick, L., and Clarke, W. L. Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data. *Diabetes Technology & Therapeutics*, 5(5):817–828, 2003.

Kovatchev, B. P., Breton, M., Dalla Man, C., and Cobelli, C. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes, 2009.

Kovatchev, B. P., Renard, E., Cobelli, C., Zisser, H. C., Keith-Hynes, P., Anderson, S. M., Brown, S. A., Chernavvsky, D. R., Breton, M. D., Farret, A., et al. Feasibility of outpatient fully integrated closed-loop control: first studies of wearable artificial pancreas. *Diabetes Care*, 36 (7):1851–1858, 2013.

Krause, A. and Ong, C. S. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pp. 2447–2455, 2011.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems*, 20(1):96–1, 2007.

Lázaro-Gredilla, M. Bayesian warped Gaussian processes. *Advances in Neural Information Processing Systems*, 25: 1619–1627, 2012.

Lu, T., Pál, D., and Pál, M. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, pp. 485–492. JMLR Workshop and Conference Proceedings, 2010.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020.

Maahs, D. M., Buckingham, B. A., Castle, J. R., Cinar, A., Damiano, E. R., Dassau, E., DeVries, J. H., Doyle, F. J., Griffen, S. C., Haidar, A., et al. Outcome measures for artificial pancreas clinical trials: a consensus report. *Diabetes Care*, 39(7):1175–1179, 2016.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Reddy, M., Herrero, P., El Sharkawy, M., Pesl, P., Jugnee, N., Thomson, H., Pavitt, D., Toumazou, C., Johnston, D., Georgiou, P., et al. Feasibility study of a bio-inspired artificial pancreas in adults with type 1 diabetes. *Diabetes Technology & Therapeutics*, 16(9):550–557, 2014.

Schmidt, S. and Nørgaard, K. Bolus calculators. *Journal of Diabetes Science and Technology*, 8(5):1035–1041, 2014.

Sherr, J. L., Cengiz, E., Palerm, C. C., Clark, B., Kurtz, N., Roy, A., Carria, L., Cantwell, M., Tamborlane, W. V., and Weinzimer, S. A. Reduced hypoglycemia and increased time in target using closed-loop insulin delivery during nights with or without antecedent afternoon exercise in type 1 diabetes. *Diabetes Care*, 36(10):2909–2914, 2013.

Snelson, E., Rasmussen, C. E., and Ghahramani, Z. Warped Gaussian processes. *Advances in Neural Information Processing Systems*, 16:337–344, 2004.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

Sui, Y., Gotovos, A., Burdick, J., and Krause, A. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning*, pp. 997–1005, 2015.

Sui, Y., Burdick, J., Yue, Y., et al. Stagewise safe bayesian optimization with Gaussian processes. In *International Conference on Machine Learning*, pp. 4781–4789, 2018.

Tejedor, M., Woldaregay, A. Z., and Godtliebsen, F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artificial Intelligence in Medicine*, pp. 101836, 2020.

Walsh, J., Roberts, R., and Bailey, T. Guidelines for optimal bolus calculator settings in adults. *Journal of Diabetes Science and Technology*, 5(1):129–135, 2011.

Xie, J. Simglucose v0.2.1. https://github.com/jxx123/simglucose, 2018.