Exploiting Concavity Information in Contextual Bandit Optimization

Kevin Li[1] **Eric Laber**[1]

[1]Department of Statistical Science., Duke University, Durham, North Carolina, USA

Abstract

The contextual bandit models sequential decision-making problems in which rewards depend on both the chosen action and observed context. In many domains such as medicine, business, and engineering, prior knowledge provides structural information about the reward function that can be exploited to improve optimization efficiency. This paper studies settings where, for each observable context, the conditional mean reward is known to be concave with respect to the action variable. To leverage this structure, we develop a contextual bandit algorithm that conditions a Bayesian Gaussian Process posterior on concavity constraints. We propose a novel reward model that combines a concavity-preserving regression spline basis with a constrained GP posterior, yielding a tractable shape-constrained estimator. Building on this model, we construct a UCB algorithm and establish new posterior concentration inequalities for the constrained posterior, leading to regret guarantees that are never worse than applying standard GP-UCB without concavity information and can be strictly tighter when concavity is informative. Experiments on benchmark problems and a Warfarin dosing test application demonstrate substantial reductions in cumulative regret relative to state-of-the-art baselines.

1 INTRODUCTION

The contextual bandit framework is used for sequential optimization in settings where the reward (utility) of any decision depends on auxiliary context variables [Chu et al., 2011, Bouneffouf et al., 2020, Silva et al., 2022]. Any efficient sequential optimization algorithm makes use of prior knowledge to impose structure on the considered class of reward models. In the case of contextual bandits, such structure often comes in the form of assumptions on the conditional mean reward given context and action. Popular assumptions include linearity [Chu et al., 2011], samples from a Gaussian Process (GP) [Srinivas et al., 2009, Krause and Ong, 2011], smoothness of the function defined by bounds on RKHS norm [Valko et al., 2013, Chowdhury and Gopalan, 2017], or effective dimensionality as derived from the tangent kernel of a deep network [Zhou et al., 2020, Zhang et al., 2021, Kassraie and Krause, 2022].

Within medicine and business applications, the reward function is often known to be concave with respect to the action variable at each context value. In biology and pharmacology, this recurring pattern in the reward function is widely known as a U-shaped, bi-phasic dose-response curve [Calabrese and Baldwin, 2001]. At low doses, an agent often provides beneficial function for a biological system. However, at a certain dosage level, benefits become saturated and increased dosage causes harmful effects. This phenomenon results in concave dose-response curves that have been studied extensively for decision making in cancer treatment [Portier and Ye, 1998, Reynolds, 2010], environmental management/public health [Portier and Ye, 1998, Davis and Svendsgaard, 1990], psychological/behavioral intervention [Baldi and Bucherelli, 2005], and dose optimization in personalized medicine [Chen et al., 2016]. In business and economics, concave utility functions are foundational structural assumptions for decision making and modeling. A well-known example occurs in bidding and pricing systems in commerce, where increasing price increases the per-unit profit but decreases overall sales volume [Meyer and Pratt, 1968, Varian, 1982, Carroll and Kimball, 1996, Boehm and Pandalai-Nayar, 2022].

This paper shows how incorporating this concavity information into the GP contextual bandit algorithm can improve bandit optimization performance. GP contextual bandits have a key advantage of being effective, yet simple to implement. Therefore they have been successfully applied in a wide variety of applications such as experimental design in

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3762–3784.

biology [Durand et al., 2018], database routing optimization [Cereda et al., 2021], and industrial design [Zhou, 2025].

Independent of bandit optimization, significant research has shown the advantages of incorporating convexity/concavity constraints in function estimation [Hannah and Dunson, 2013, Wang and Berger, 2016, Mazumder et al., 2019, Aubin-Frankowski and Szabó, 2020]. However, it is challenging to design bandit algorithms with provable sub-linear regret bounds from these bespoke models.

This paper introduces Concave Spline Gaussian Process Upper Confidence Bound (CSGP-UCB), a contextual bandit algorithm that leverages known concavity of the mean reward in the action (for each fixed context) to accelerate sequential optimization while retaining sub-linear cumulative regret guarantees. By pairing a concavity-preserving regression-spline GP with an exact translation of concavity into finite linear inequality constraints on spline coefficients, the approach induces a truncated (shape-conditioned) GP posterior that remains computationally and theoretically tractable. A central technical contribution is a new set of concentration inequalities for linear functionals of this truncated posterior, enabling UCB confidence bounds that are never wider than those from unconstrained GP-UCB and can be substantially tighter when concavity is informative. These results yield (to the author's knowledge) the first sub-linear regret guarantees for a GP contextual bandit that explicitly conditions on concavity information under Bayesian assumptions. Empirically, CSGP-UCB achieves markedly lower regret than a broad set of state-of-the-art baselines across benchmark functions, simulations, and a Warfarin dosing test application, highlighting the value of concavity constraints in the moderate-sample regimes that motivate many real applications.

2 BACKGROUND

2.1 CONCAVE CONTEXTUAL BANDIT

We consider a sequential decision making problem consisting of T rounds. Our goal is to maximize the expected cumulative reward. In each round, the decision maker receives a d-dimensional context $\mathbf{x}_t \in \mathcal{X}$ and must choose $a_t \in \mathcal{A} \subseteq [0,1]$. After choosing the action a_t, the system produces a noisy reward

$$y_t = f(a_t, \mathbf{x}_t) + \epsilon_t, \tag{1}$$

where $\epsilon_t \sim_{\text{iid}} N(0, \sigma^2)$. Thus, $f(a_t, \mathbf{x}_t)$ represents the mean reward given action a_t and context $\mathbf{x}_t$. We further assume that $f(a, \mathbf{x})$ is twice continuously differentiable at every $a \in \mathcal{A}$ for each $\mathbf{x} \in \mathcal{X}$. At each time t, the decision maker also receives concavity information such that for each $t' \leq t$ the second partial with respect to a satisfies $\frac{\partial^2 f(a, \mathbf{x}_{t'})}{\partial a^2} \leq 0$ for all $a \in [0,1]$, i.e., the map $a \mapsto f(a, \mathbf{x}_{t'})$ is concave with respect to a at the current and all previously observed contexts. This assumption is natural when the decision maker only wants to condition on concavity conditions for contexts $\mathbf{x}$ that are feasible for the system and the feasible region is not known *a priori*.

The action selected at time t can be evaluated by the regret of not choosing the optimal action given the context at time t: $r_t \triangleq \sup_{a \in \mathcal{A}} f(a, \mathbf{x}_t) - f(a_t, \mathbf{x}_t)$. After the T rounds, the cumulative regret is $R_T \triangleq \sum_{t=1}^{T} r_t$. Our goal is to construct an algorithm that effectively makes use of concavity information to improve performance in small samples while also ensuring sub-linear regret as T grows large.

2.2 GAUSSIAN PROCESSES

A Gaussian Process $h(\mathbf{x})$ is a stochastic process for which its evaluation on any finite subset of inputs follows a Multivariate Normal distribution [Rasmussen and Williams, 2006]. A Gaussian process is fully defined by its mean function $\mu(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. A convenient property of Gaussian Processes is that upon observing some realizations of a Gaussian Process $\mathbf{h} = \{h(\mathbf{x}_1) \dots h(\mathbf{x}_n)\}^\intercal$, the posterior distribution of unobserved samples $\mathbf{h}^* = \{h(\mathbf{x}_1^*), \dots, h(\mathbf{x}_m^*)\}^\intercal$ follows a Multivariate Normal distribution with mean $\boldsymbol{\mu}^*$ and covariance $\mathbf{V}^*$ defined by

$$\begin{aligned} \boldsymbol{\mu}^* &= \boldsymbol{\mu}_{h^*} + \mathbf{K}_{m,n}\mathbf{K}_{n,n}^{-1}(\mathbf{h} - \boldsymbol{\mu}_h), \\ \mathbf{V}^* &= \mathbf{K}_{m,m} - \mathbf{K}_{m,n}\mathbf{K}_{n,n}^{-1}\mathbf{K}_{n,m}, \end{aligned} \tag{2}$$

where: $\mathbf{K}_{n,n}$ is the covariance matrix of the observed realizations $\mathbf{h}$ such that (i,j)th entry is $k(\mathbf{x}_i, \mathbf{x}_j)$; $\mathbf{K}_{m,m}$ is the covariance matrix of the unobserved realizations; $\mathbf{K}_{m,n}$ is the cross covariance matrix such that the (i,j)th entry is $k(\mathbf{x}_i^*, \mathbf{x}_j)$; and $\boldsymbol{\mu}_h, \boldsymbol{\mu}_{h^*}$ are the mean function evaluated on the observed and unobserved samples respectively.

2.3 CONCAVE SPLINE REGRESSION

For each context, the mean reward is a concave function of the action. Thus, it is useful to have a mechanism for modeling a one-dimensional function, say $g(a)$, that is expressive but also easily constrained to be concave. In this section, we discuss modeling such a function using regression splines; we omit dependence on the context and show how to make the overall mean function smooth across contexts in Section 4.1. We choose regression splines due to their representative power. Given enough knots, our chosen constrained spline model can uniformly approximate continuous concave functions defined on a closed interval [De Boor and De Boor, 1978, Meyer, 2008].

For regression splines of polynomial order k, with l knots we can derive the $l + k$ piece-wise polynomial M-Spline basis functions [Ramsay, 1988] $M_j(a)$, $j = 1, \dots, l + k$

such that $M_j(a) \geq 0$. We model $g''(a)$ using these M-Spline basis functions

$$g''(a) = \sum_{j=1}^{l+k} M_j(a)\beta_j.$$

A crucial property of the M-Spline basis is that for $k \leq 4$, $g''(a) \leq 0$ if and only if $\beta_j \leq 0$ for $j = 1, \ldots, l+k$ [Ramsay, 1988]. Integrating each M-Spline basis function $M_j(a)$ twice yields the C-Spline Basis functions introduced by Meyer [2008]: $C_j(a) = \int_0^a \int_0^u M_j(s)dsdu$. We consider a model for $g(a)$ that uses these C-Spline basis functions along with the constant and identity functions:

$$g(a) = \sum_{j=1}^{l+k} C_j(a)\beta_j + \beta_{l+k+1}a + \beta_{l+k+2}. \quad (3)$$

From the properties of $g''(a)$, it follows that $g(a)$ is concave if and only if $\beta_j \leq 0$ for $j = 1, \ldots, l+k$. Thus the resulting constrained spline function can represent all concave piecewise polynomials of order $k+2$.

3 RELATED WORK

Our work is most closely related to the literature of Gaussian Process Bandits which were originally introduced by Srinivas et al. [2009] and extended to the contextual case in Krause and Ong [2011]. In a similar vein, Valko et al. [2013] developed a Kernelized UCB algorithm where they bound regret using the effective dimension of the kernel. Chowdhury and Gopalan [2017] derived improved bounds for context-free GP/Kernelized UCB bandits and the first frequentist regret bounds for GP Thompson sampling.

Previous GP-bandit literature has incorporated complex structural assumptions into the prior of the reward function. The strategy has been to directly condition on the event that a sample from a standard GP satisfies the assumption. Kandasamy et al. [2019] develops an algorithm conditioned on GP-bandits with a multi-fidelity structure. Scarlett [2018] conditions on constraints to derive Bayesian optimization cumulative regret bounds. The methods presented in these papers do not translate well to our goal. In their methodology, conditioning on the structural constraints inflates the regret bounds relative to the unconditioned case. In addition, it is not clear how to practically condition a standard GP covariance function on infinite dimensional concavity constraints. Both outcomes are unsatisfactory for our goal of using concavity information to achieve more efficient optimization.

Independently of the bandit literature, the problem of shape-constrained function estimation in the form of monotone/convex GP regression has been extensively studied. One strategy has been to approximate the assumption by only conditioning on the constraints within a discrete grid of the input space [Da Veiga and Marrel, 2012, Wang and Berger, 2016]. These methods quickly become impractical, as conditioning on higher derivatives is notoriously numerically unstable [He and Chien, 2018] and the grid size must grow exponentially with dimension. Maatouk and Bay [2017] propose a more similar approach to ours. They tile the covariate space with a tensor grid basis of local basis functions with independent Gaussian priors on the coefficients. Unfortunately, their tensor grid grows exponentially with context dimension and does not scale to contexts with $d > 2$. Our paper extends their approach for the contextual bandit setting by allowing the coefficients for each basis function to vary smoothly with the context variables and showing that this leads to sub-linear cumulative regret in the bandit optimization context.

4 CONCAVE SPLINE GP CONTEXTUAL BANDIT

4.1 REWARD MODEL

Conditional on each context, we assume that the expected reward function takes the form of a regression spline as described in equation 3 with the generalization that the spline regression coefficients vary as smooth functions of $\mathbf{x}$. We define the CSGP model for the reward as:

$$\begin{aligned} f(a, \mathbf{x}) &= \sum_{j=1}^{J-2} C_j(a)\boldsymbol{\beta}_j(\mathbf{x}) + \boldsymbol{\beta}_{J-1}(\mathbf{x})a + \boldsymbol{\beta}_J(\mathbf{x}) \quad (4) \\ &:= \boldsymbol{\phi}(a)^T \boldsymbol{\beta}(\mathbf{x}), \end{aligned}$$

where each $C_j(a)$ is the C-Spline basis function described in Section 2.3, $\boldsymbol{\beta}(\mathbf{x}) := \{\boldsymbol{\beta}_1(\mathbf{x}), \ldots, \boldsymbol{\beta}_J(\mathbf{x})\}^T$ and $\boldsymbol{\phi}(a) := \{C_1(a), C_2(a), \ldots, a, 1\}^T$. We place Gaussian Processes priors on $\boldsymbol{\beta}_j(\mathbf{x})$ with mean function $\boldsymbol{\mu}_j$ and covariance function $k_j(\mathbf{x}, \mathbf{x}')$. These priors induce a Gaussian Process on $f(a, \mathbf{x})$.

4.2 INCORPORATING CONCAVITY INFORMATION

To incorporate concavity information and complete the CSGP formulation, we assume that at each time step t we observe the context $\mathbf{x}_t$ and the information that $f(a, \mathbf{x}_t)$ is concave in a at all $\mathbf{x}_1 \ldots \mathbf{x}_t$. This form of concavity constraint is natural when the decision maker only wants to condition on concavity conditions for contexts $\mathbf{x}$ that are feasible for the system and the feasible region is not known exactly *a priori*. Under our choice of spline basis (see Section 2.3) this information is exactly equivalent to conditioning on $\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq 0$ for $j = 1 \ldots J-2, t' \leq t$. An important property of our reward and observation models is that after conditioning on such information, the posterior for

the finite realizations of $\boldsymbol{\beta}(\mathbf{x})$ is a truncated Gaussian. Let $\mathbb{R}_e = \mathbb{R} \cup \{\infty\}$ denote the (positively) extended reals. Given $\boldsymbol{\nu} \in \mathbb{R}_e^p$, $\boldsymbol{\omega} \in \mathbb{R}^p$, and $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$, we let $N_{\boldsymbol{\nu}}(\boldsymbol{\omega}, \boldsymbol{\Omega})$ denote a multivariate normal distribution with mean $\boldsymbol{\omega}$ and covariance $\boldsymbol{\Omega}$ that has been truncated above by $\boldsymbol{\nu}$. The following result is proved in Appendix 8.

Lemma 1. *Suppose that we observe* $\mathbf{y}_{t-1} = (y_1, \ldots, y_{t-1})^T$. *Define* $\boldsymbol{\beta}_t := \{\boldsymbol{\beta}(\mathbf{x}_1), \ldots, \boldsymbol{\beta}(\mathbf{x}_t)\}^T$ *to be the concatenated vector of coefficients from every context up to time* t. *Let* $\mathcal{C}_{t'}$ *denote the event that* $\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq 0$, *for* $j = 1, \ldots, J-2$. *Then the time* t *posterior* $p(\boldsymbol{\beta}_t | \mathbf{y}_{t-1}, \cap_{t'=1}^t \mathcal{C}_{t'})$ *is the upper truncated Multivariate Normal distribution* $N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, *where* $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ *are the posterior quantities obtained from the standard GP time* t *posterior without truncation, and* $\boldsymbol{\nu}_j = 0, j \leq J-2$ *and* $\boldsymbol{\nu}_j = \infty$ *otherwise.*

Proving regret bounds under our model requires quantifying the posterior uncertainty surrounding the expected reward. The mean reward function can be expressed as $f(a, \mathbf{x}_t) = \mathbf{c}_t(a)^T \boldsymbol{\beta}_t$, where $\mathbf{c}_t(a) \in \mathbb{R}^{Jt}$ is defined so that the last J elements of $\mathbf{c}_t(a)$ equal to $\boldsymbol{\phi}(a)$ while its remaining entries are zero. Therefore, our time t CSGP posterior for $f(a, \mathbf{x}_t)$ is a Linear Combination of a Truncated Multivariate Normal (LCTMVN). We refer to this random variable as LCTMVN throughout the paper.

Without conditioning on concavity information the posterior distribution of $f(a, \mathbf{x}_t)$ is Gaussian with variance:

$$\sigma^2_{\mathbf{c}_t}(a, \mathbf{x}_t) = \mathbf{c}_t(a)^T \boldsymbol{\Sigma}_t \mathbf{c}_t(a) = \tag{5}$$
$$k_f\{(a, \mathbf{x}_t), (a, \mathbf{x}_t)\} - \mathbf{k}_{t,t-1}(\mathbf{K}_{t-1,t-1} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_{t,t-1},$$

where $\mathbf{k}_{t,t-1} \in \mathbb{R}^{1 \times t-1}$ and the ith element is $k_f\{(a, \mathbf{x}_t), (a_i, \mathbf{x}_i)\}$ and $\mathbf{K}_{t-1,t-1}$ is the prior covariance matrix of the $\{f(a_{t'}, \mathbf{x}_{t'})\}_{t'=1}^t$.

The challenge for regret analysis is that the concavity conditioned LCTMVN posterior has no canonical form or theoretically amenable quantile function. To derive regret guarantees, we will instead rely on novel concentration results.

4.3 LCTMVN CONCENTRATION PROPERTIES

Effective UCB bandit algorithms require confidence bounds that are narrow as possible while maintaining sufficiently high coverage probability of the true reward function. In this section, we derive concentration inequalities for concavity-conditioned LCTMVN variables that produce a composite confidence bound with the same coverage guarantees as the standard un-truncated Gaussian case, but that is never wider—and is often strictly narrower in practice. To our knowledge, these concentration results are new and may be of independent interest for the Bayesian constrained regression community. Proofs can be found in Appendix 9.

We first prove coverage guarantees for confidence bounds centered around the LCTMVN posterior mean. The degree of concentration depends on Bregman divergences that are sensitive to the geometry of the truncation.

Proposition 1. *Assume the concavity conditioned posterior of our coefficient vector* $\boldsymbol{\beta}$ *at time* t *is* $\boldsymbol{\beta}_t \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$. *Let* $\boldsymbol{\mu}_t^* = \mathbb{E}(\boldsymbol{\beta}_t | \mathbf{y}_{t-1}, \cap_{t'=1}^t \mathcal{C}_{t'})$. *Then given* $b \geq 0$:

$$Pr\left\{|\mathbf{c}_t(a)^T \boldsymbol{\beta}_t - \mathbf{c}_t(a)^T \boldsymbol{\mu}_t^*| \geq b\right\} \leq \exp\left(-\frac{b^2}{2\sigma^2_{\mathbf{c}_t}(a, \mathbf{x}_t)}\right) \times P_t(a, b)$$

Where

$$P_t(a,b) = \left\{ \exp\left[-D^{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t}_{G, \mathbf{c}_t(a)}\left(\boldsymbol{\mu}_t + \frac{b}{\sigma^2_{\mathbf{c}_t}(a, \mathbf{x}_t)} \Sigma_t \mathbf{c}_t(a), \boldsymbol{\mu}_t\right)\right] + \exp\left[-D^{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t}_{G, \mathbf{c}_t}\left(\boldsymbol{\mu}_t - \frac{b}{\sigma^2_{\mathbf{c}_t}(a, \mathbf{x}_t)} \Sigma_t \mathbf{c}_t(a), \boldsymbol{\mu}_t\right)\right] \right\}$$

$D^{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t}_{G, \mathbf{c}}(\cdot, \cdot)$ *represents the Bregman divergence with respect to the convex function* $G(\boldsymbol{\mu}) = -\log\left\{\int_{\boldsymbol{\beta} \preccurlyeq \nu} p_{\boldsymbol{\mu}, \Sigma_t}(\boldsymbol{\beta}) d\boldsymbol{\beta}\right\}$. *Here* $p_{\boldsymbol{\mu}, \Sigma_t}(\boldsymbol{\mu})$ *is the standard multivariate normal density with mean* $\boldsymbol{\mu}$ *and covariance* Σ_t.

The resulting inequality differs from the standard Gaussian concentration inequality (without concavity information/truncation) in the GP bandit literature by the factor $P^t_{\mathbf{c}_t}(a, b)$ [Srinivas et al., 2009]. Interpreting the Bregman divergences, we can see that $P^t_{\mathbf{c}_t}(a, b) < 1$ when the truncation causes curvature for $G(\boldsymbol{\mu})$ in the direction $\pm \Sigma_t \mathbf{c}_t(a)$. As $G(\boldsymbol{\mu})$ represents the probability of truncation region as a function of the mean parameters, $P^t_{\mathbf{c}_t}(a, b)$ represents the severity of the truncation effect on the LCTMVN variable. When the data $\mathbf{y}_t$ is not sufficient to concentrate the posterior in the truncation region, the LCTMVN bounds will have higher coverage probability than the standard Gaussian bounds for the same width. In Lemma 2, we show that $P^t_{\mathbf{c}_t}(a, b) < 1$ implies the existence of a mean-centered confidence bound with width parameter $b' < b$ that achieves the same coverage probability as the standard Gaussian confidence bound with width parameter b.

It is possible that $P^t_{\mathbf{c}_t}(a, b) > 1$, in which case the inequality becomes looser than the standard (unconstrained) Gaussian inequality. As shown in Appendix 12, this is unavoidable: LCTMVN variables can exhibit weaker mean-centered concentration than a Gaussian. To overcome this issue, our algorithm will also use confidence bounds centered around the posterior median that match the coverage guarantees achieved by the standard Gaussian case. We use tools from optimal transport [Caffarelli, 2000, Kim and Milman, 2012] and Gaussian isoperimetric inequalities [Ledoux, 2006] to derive the following result.

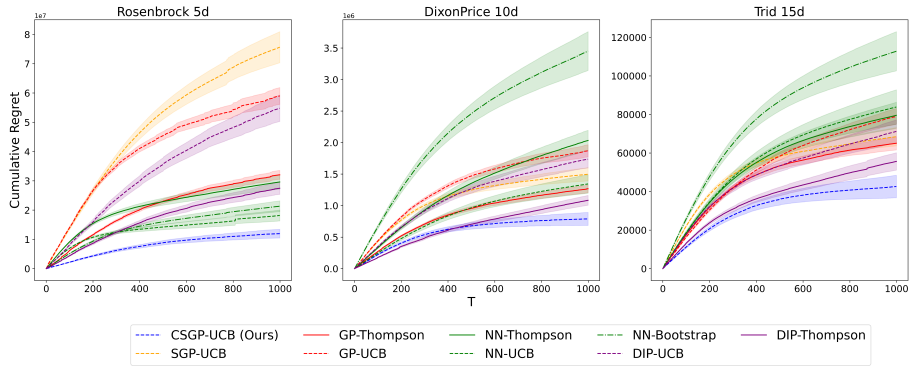

Figure 1: Mean cumulative regret for contextual bandit methods on benchmark functions across T, with $\pm$ 1 standard error bars.

Proposition 2. *Suppose* $\boldsymbol{\beta}_t \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$. *Let* $m_t(a, \mathbf{x}_t)$ *denote the median of* $\mathbf{c}^T(a)\boldsymbol{\beta}_t$. *Then*

$$Pr\left\{|\mathbf{c}(a)^T\boldsymbol{\beta}_t - m_t(a,\mathbf{x}_t)| \geq b\right\} \leq \exp\left(\frac{-b^2}{2\sigma^2_{\mathbf{c}_t}(a,\mathbf{x}_t)}\right)$$

Combining the two concentration inequalities allows us to create a LCTMVN confidence bound that achieves the same concentration as the Gaussian case is at least as narrow. The following lemma designs this confidence bound with corresponding coverage guarantee:

Lemma 2. *Suppose* $\boldsymbol{\beta}_t \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$ *and Let* χ_t *be a predefined constant. Define*

$$\kappa_t(a,\mathbf{x}_t) = \begin{cases} \mathbf{c}_t(a)^T\boldsymbol{\mu}^*_t, & \text{if } P_t\left[a, \chi_t\sigma_{\mathbf{c}_t}(a,\mathbf{x}_t)\right] < 1, \\ m_t(a,\mathbf{x}_t), & \text{else.} \end{cases}.$$

$$\alpha_t(a,\mathbf{x}_t) = \begin{cases} \chi^*_t(a,\mathbf{x}_t), & \text{if } P_t\left[a, \chi_t\sigma_{\mathbf{c}_t}(a,\mathbf{x}_t)\right] < 1, \\ \chi_t, & \text{else.} \end{cases}.$$

$$\chi^*_t(a,\mathbf{x}_t) := \inf_s \left\{s : \exp\left(-\frac{s}{2}\right) P_t\left[a, s\sigma_{\mathbf{c}_t}(a,\mathbf{x}_t)\right] \leq \exp\left(-\frac{\chi_t}{2}\right)\right\}.$$

$\chi^*_t(a,\mathbf{x}_t)$ *exists and we have the following concentration bound:*

$$Pr\left\{|f(a,\mathbf{x}_t) - \kappa(a,\mathbf{x}_t)| \geq \alpha_t(a,\mathbf{x}_t)\sigma_{\mathbf{c}_t}(a,\mathbf{x}_t)\right\} \leq \exp\left(-\frac{\chi_t}{2}\right).$$

In practice, we numerically compute χ^*_t which we discuss in Appendix 13. We investigate the empirical behavior of $\alpha_t(a, x_t)$ in section 5.3.

4.4 UCB ALGORITHM

We now define our CSGP-UCB algorithm. Recall that $\sigma^2_{\mathbf{c}_t}(a,\mathbf{x}_t)$ is the variance of the unconstrained CSGP posterior defined in equation 5. Let $\kappa(a,\mathbf{x}_t), \alpha(a,\mathbf{x}_t)$ be defined as in Lemma 2 using $\boldsymbol{\mu}_t, \Sigma_t$ from the time t concavity conditioned CSGP posterior. Given a predefined sequence χ_t, our CSGP-UCB algorithm selects action a_t as

$$a_t = \underset{a\in\mathcal{A}}{\text{argmax}}\, \kappa_t(a,\mathbf{x}_t) + \alpha_t(a,\mathbf{x}_t)^{\frac{1}{2}}\sigma_{\mathbf{c}_t}(a,\mathbf{x}_t)$$

Our UCB algorithm chooses the narrower of the mean and median centered confidence bounds for each $a \in \mathcal{A}$. If $P_t(a,\chi_t) < 1$, then we will use the mean bound as it has a smaller width. Otherwise the median centered bound will have smaller width. The sequence χ_t is chosen as a gradually increasing sequence to ensure that the bound covers the true value of f with high probability. We then select the action a that has the highest bound.

Computing the bound requires computing the LCTMVN variable means ($\boldsymbol{\mu}^*$), medians ($m(a,\mathbf{x}_t)$), log orthant probabilities ($G(\boldsymbol{\mu})$), and optimal confidence parameter χ^*_t. We discuss computation of these quantities in the Appendix 13.

4.5 REGRET ANALYSIS

We present the regret analysis for our UCB algorithm under the Bayesian assumption that the expected reward function $f(a,\mathbf{x})$ is a sample from CSGP prior described in Section 4.1. This assumption is commonly made and analyzed in the bandit literature [Srinivas et al., 2009, Krause and Ong, 2011, Chen et al., 2012, Wang et al., 2015, Bogunovic et al., 2016, Kathuria et al., 2016, Kassraie and Krause, 2022]. Proofs are contained in the Appendix.

To our knowledge, we provide the first sub-linear regret guarantees under the Bayesian assumption and the scenario that the bandit-algorithm receives concavity information. Any GP-posterior that is not conditioned on this concavity information will be mis-specified. The authors are not aware of algorithms that achieve sub-linear regret under the Bayesian assumption and mis-specified posteriors.

Our regret bound uses the maximum information gain concept first exploited for GP Bandit problems by Srinivas

et al. [2009]. We represent the maximal information gain by γ_T and prove in Appendix 10 that for our CSGP kernel, $\gamma_T = \mathcal{O}\left\{J \log(T)^{d+1}\right\}$.[1]

To highlight the source of concavity advantage, we note how the confidence parameter α_t relates to the time-t instantaneous regret r_t. The following lemma, which follows directly from Lemma 4.1 of Krause and Ong [2011], provides the following high probability bound on r_t for GP-UCB algorithms in discrete action spaces.

Lemma 3. *For $t \geq 1$, with probability $\geq 1-\delta$, $r_t \leq 2\alpha_t(a_t, \mathbf{x}_t)^{\frac{1}{2}}\sqrt{\gamma_T}$*

Thus, lowering the confidence parameter directly lowers the bound on r_t. Under standard GP-UCB, the confidence parameter does not depend on posterior quantities or the action/context. Thus $\alpha_t(a, \mathbf{x}_t) = \chi_t$. Concavity information allows $\alpha_t(a, \mathbf{x}_t) \ll \chi_t$ when the truncation is severe. Intuitively, having an tighter confidence parameter prevents over-exploration of uncertain regions. This translates to the following cumulative regret bound.

Proposition 3. *Let the expected reward function f be a sample from a known CSGP prior with noise model $N(0, \sigma^2)$. Also suppose that at each time step t the bandit algorithm receives noisy reward y_t as well as concavity information that $\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq 0$ for $j = 1 \ldots J-2, t' \leq t$. Pick $\delta \in (0,1)$ and suppose one of the following assumptions holds:*

- *$\mathcal{A}$ is finite with $\chi_t = 2\log(|\mathcal{A}|t^2\pi^2/6\delta)$*
- *$\mathcal{A} \subset [0,1]$ is compact and convex and the CSGP kernel $k_f(\cdot,\cdot)$ satisfies the following bound on the GP sample paths.[2] For some constants η, ζ*

$$\Pr\{\sup_{\mathbf{v}\in\mathcal{A}\times\mathcal{X}} |\partial f/\partial \mathbf{v}_j| > L\} \leq \eta \exp\left\{-(L/\zeta)^2\right\},$$

for $\forall j \in \{1, \ldots, d+1\}$. And we choose

$$\alpha_t = 2\log(2\pi^2 t^2/3\delta) + 2\log\left\{t^2\zeta\sqrt{\log\left(4\eta/\delta\right)}\right\}.$$

Define $\alpha_T^* = \max\{\alpha_t(a_t, \mathbf{x}_t)\}_{t=1}^T$ so that $\alpha_T^* \leq \chi_T$. Running CSGP-UCB obtains a regret bound $\mathcal{O}^*(\sqrt{T\gamma_T\alpha_T})$ with high probability. Precisely let $C_1 = \frac{8}{\log(1+\sigma^{-2})}$ then:

$$\Pr\left\{R_t \leq \sqrt{C_1 T \alpha_T^* \gamma_T} + 2, \forall T \geq 1\right\} \geq 1-\delta,$$

Our choices for χ_t are the default confidence parameters used by GP-UCB bandits [Srinivas et al., 2009, Krause and Ong, 2011]. As $\alpha_T^* \leq \chi_T$, our cumulative regret bounds under concavity conditions are never worse than the bound obtained for the case without concavity information. For moderate time horizons T, concavity is often highly informative and induces substantial posterior concentration, so $\alpha_t < \chi_t$ and $\alpha_T^* < \chi_T$. This directly shrinks the confidence bounds and translates into substantially lower cumulative regret. We explore the behavior of α_T^* in Figure 3 and Section 5.3. As $T \to \infty$, the data $\mathbf{y}_t$ increasingly identifies the function's concavity and the posterior mass concentrates within the feasible (truncated) region. The benefits of concavity information fades, and the regret rate is expected to recover the standard Gaussian-case asymptotic regret. This behavior is consistent with the literature on convex/constrained regression which has found that the asymptotic learning rates of constrained methods match the asymptotic rates of their unconstrained counterparts [Geyer, 1994, Tsybakov, 2008, Kur et al., 2020]. We also note that an important line of future work is to determine whether a sequence of α_t can be developed such that $\alpha_T^* < \chi_T$.

4.6 ROBUSTNESS TO MILD VIOLATIONS OF CONCAVITY.

The exact CSGP posterior conditions on the curvature coefficients satisfying $\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq 0$ for $j = 1, \ldots, J-2$, which enforces action-wise concavity. When domain knowledge suggests that concavity may hold only approximately, this hard constraint can be relaxed to

$$\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq c, \qquad j = 1, \ldots, J-2,$$

for a small scalar $c > 0$. The posterior remains an upper-truncated Gaussian, so the posterior computation and confidence-bound construction are unchanged after replacing the truncation threshold 0 by c. The choice $c = 0$ recovers the exact concavity model, while $c > 0$ permits controlled violations of concavity in the C-spline coefficient scale. Appendix 14 provides additional implementation details and empirical results under non-concave perturbations, showing that a modest positive value of c can reduce performance degradation when the reward is only mildly misspecified.

5 NUMERICAL EXPERIMENTS

We provide an empirical validation of CSGP-UCB and examine the evolution of the confidence parameter α_T^*. In section 5.1 we evaluate our method on widely used optimization benchmarks provided by Surjanovic and Bingham, Balandat et al. [2020]. In section 5.2 we study the behavior of CSGP-UCB as we vary the dimensions/smoothness of the underlying reward function. Section 5.3 examines the evolution of CSGP confidence parameter α_T^*. Finally in section 5.4, we test our method on a Warfarin dosing function application provided by Chen et al. [2016].

We benchmark our algorithms against state of the art non-parametric bandit baselines based on GPs and Neural Networks. We compare to GP-UCB/TS [Krause and Ong, 2011,

[1] See Appendix 10 for more discussion

[2] The condition on $k_f(\cdot,\cdot)$ is satisfied with a spline order 6 and kernels such as the common Gaussian kernel and the Matern kernel with $\nu > 2$.

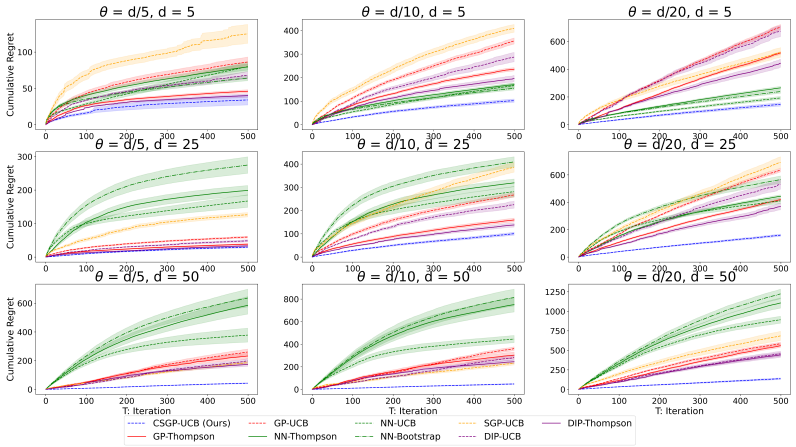

Figure 2: Cumulative regret for numerical simulation study. The proposed methods, CSGP-Thompson and CSGP-UCB, obtain markedly lower regret than competing methods across a range of problem dimensions and scalings.

Russo and Van Roy, 2014], neural contextual bandits (NN-UCB/NN-TS/NN-Bootstrap) [Kassraie and Krause, 2022, Zhang et al., 2021, Riquelme et al., 2018], our CSGP model without conditioning on concavity (SGP-UCB), and GP-UCB/TS where the second derivative of the GP posterior is conditioned to be negative on a grid as proposed by Da Veiga and Marrel [2012], Wang and Berger [2016] (DIP-UCB/TS). Details for the hyper-parameters and implementations of the methods can be found in Appendix 15. As Zhang et al. [2021], Kassraie and Krause [2022] provide algorithms only for a finite action space, we test the bandits with $\mathcal{A}$ defined as a uniform grid of size 100 on $[0, 1]$.

5.1 BENCHMARK FUNCTIONS

We test CSGP-UCB on popular benchmark functions widely used in the optimization literature that also satisfy our concavity assumptions [Surjanovic and Bingham, Balandat et al., 2020]. We consider the Rosenbrock ($d = 5$), Dixon-Price ($d = 10$), and Trid ($d = 15$). To adapt these benchmarks to the contextual bandit application, we sample context vectors uniformly from the domains of $d - 1$ variables and select actions on the held out dimension. For more details of implementation please see the Appendix 15.2.

Figure 1 shows that CSGP-UCB significantly outperforms other methods. SGP-UCB performs similarly to its GP-UCB counterpart which indicates that the performance of CSGP-UCB can be attributed to conditioning on the concavity information and not the novel spline model. We also see that the approximate shape constraints of DIP-UCB/Thompson outperforms the standard GP methods for smaller T. However, we also see that the DIP bandit converges to optimal actions more slowly than the standard GP methods. This behavior can be attributed to modeling inaccuracy caused by the known numerical instability caused by conditioning on higher derivative information [He and Chien, 2018]. We suspect that this is why Da Veiga and Marrel [2012], Wang and Berger [2016] do not implement/demonstrate second order constraints in their works.

5.2 NUMERICAL SIMULATIONS

This section examines the performance of CSGP-UCB across problem difficulties by varying the context dimension and underlying smoothness of the reward function. We simulate functions of the form:

$$f(a, \mathbf{x}) = \sum_{s=1}^{10} h_s(a)\eta_s(\mathbf{x})^2 + \gamma(x),$$

where $h_s(a)$ is sampled from a Gaussian Process prior conditioned on $h_s''(a) \leq 0$ on a very fine grid in $[0, 1]$. This conditioning produces functions $h_s(a)$ that are effectively concave in the interval [Wang and Berger, 2016]. We sample $\eta_s(\mathbf{x})$ and $\gamma(\mathbf{x})$ from independent Gaussian Processes equipped with isotropic Gaussian kernels. The concavity of $f(a, \mathbf{x})$ with respect to a follows as it is a sum of functions concave in a. We sample the contexts uniformly such that $\mathbf{x} \in [0, 1]^d$. A crucial property of our simulation set up is that the CSGP model is mis-specified for this simulation; i.e., $f(a, \mathbf{x})$ does not take the form of piecewise polynomials for any $\mathbf{x}_t$. We sample the noisy rewards from $y = f(a, \mathbf{x}) + \epsilon$ where $\epsilon \sim N(0, 0.1)$. We initialize each bandit algorithm with $n = 25$ observations and run each algorithm for $T = 500$ iterations.

We vary simulation scenarios to control the difficulty of function estimation and optimization. We vary context dimension and the length-scales of the kernels governing the GPs for $\eta_s(\mathbf{x}), \gamma(\mathbf{x})$. Smaller length-scales for GPs lead to reward function samples that vary more quickly and are

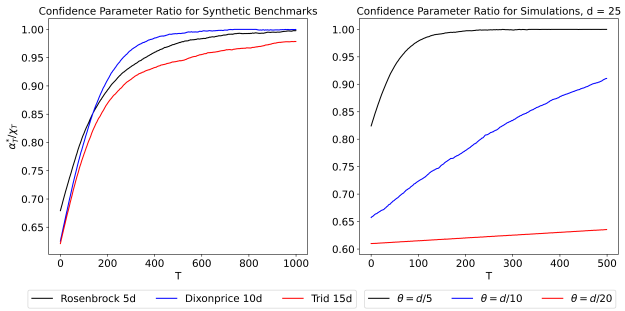

Figure 3: Mean ratio between maximum confidence parameter selected by CSGP-UCB up to time T α_T^*, and default GP-UCB time T confidence parameter χ_T.

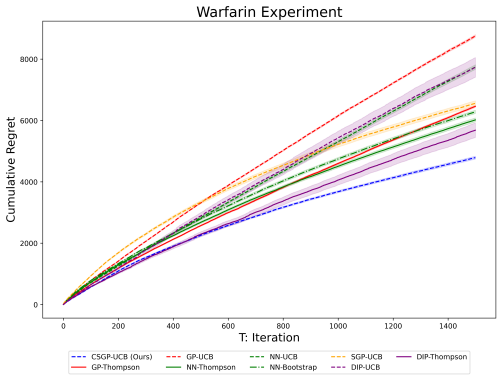

Figure 4: Cumulative regret for the Warfarin test function experiment with $\pm$ 1 SE bars.

thus more difficult to learn. We examine the performance of the various algorithms for $d = 5, 25, 50$ with length-scales $\theta = d/5, d/10, d/25, d/50$. Scaling the length-scale by dimension ensures the problem does not become intractable without massive sample sizes.

Figure 2 shows the mean cumulative regret of the various bandit algorithms across $n = 25$ replications for each experimental setting along with 1 SE bars. CSGP-UCB increasingly outperforms competing benchmarks as the difficulty of the optimization task grows with respect to both length-scale and dimension. SGP-UCB frequently under-performs GP-UCB which indicates that the superior performance of CSGP-UCB is due to the concavity information and not due to the structure of the model/simulation. CSGP-UCB outperforms the approximately concavity conditioned DIP-UCB/Thompson, especially as length-scales become small. We suspect this is due to the fact that DIP-UCB/Thompson requires an increasingly fine mesh of concavity conditioned inducing points to be effective depending on the difficulty of the problem. However, the optimal grid is unknown before running the algorithm and may be unfeasible due to numerical instability issues [He and Chien, 2018].

5.3 BEHAVIOR OF α_T^*

This section studies the behavior of the ratio α_T^*/χ_T, which is the factor by which conditioning on concavity tightens the CSGP-UCB confidence bound (and hence its regret bound) relative to standard GP-UCB with no concavity information. Smaller values indicate that concavity imposes strong truncation on the posterior, leading to a substantial reduction in uncertainty and improved function estimation. Values near one indicate a weak truncation effect, so concavity information provides little additional benefit.

Figure 3 plots the evolution of α_T^*/χ_T, averaged over 25 replications for the benchmarks and numerical simulations with $d = 25$. In both the benchmark and simulated experiments, α_T^*/χ_T is well below 1 during the exploration phase when the algorithm is still moving toward the optimum. We see from the regret experiments that this leads to significant reduction in cumulative regret for finite T. After exploration behavior reduces, the ratio approaches 1, suggesting that concavity becomes less informative. We see that for the most difficult simulation setting, $\theta = d/20$, $\alpha_T^*/\chi_T \ll 1$ throughout the experiment which reflects how concavity conditioning significantly contributes to CSGP-UCB's superior performance relative to all the baseline comparisons. Overall, these results indicate that concavity constraints can materially accelerate learning in the moderate-T regime that is often most relevant in medical and business settings.

5.4 WARFARIN DOSING TEST FUNCTION

In this section, we examine a test function used by Chen et al. [2016] to study optimal strategies for personalized dosing of Warfarin, an anti-blood clotting medication. Dose optimization is a natural application of concavity assumptions. Calabrese and Baldwin [2001] review the prevalence of the concavity condition in other medical applications. Patient response monotonically improves as dosage approaches the optimal amount from below, and will monotonically deteriorate afterwards due to excessive blood thinning [Consortium, 2009]. Outcomes also depend on patient context variables such as weight, age, and other patient biometrics. The test function's reward measures the difference between the observed value of a patient's biometric values and the optimal value. The function has 7 context dimensions representing patient biometrics/demographics and the action variable is the dosage. Discussion of the application can be found in Appendix 15.3.

Figure 4 displays the mean cumulative regret after running the algorithms for $T = 1500$ rounds across 25 replications along with 1 SE bars. We see that CSGP-UCB performs best, followed by DIP-Thompson. Again, we see that CSGP outperforms the unconstrained SGP model.

6 DISCUSSION

We proposed a contextual bandit algorithm for settings in which the mean reward is concave in the action given context information. We encoded concavity information into a novel spline-based GP model via truncation of the posterior distribution. Novel concentration properties of the resulting truncated posterior were used to construct the first UCB-algorithm with regret guarantees under Bayesian assumptions and concavity information. In simulation experiments and an illustrative biomedical application our algorithm significantly outperformed state-of-the-art non-linear bandit algorithms when the concavity conditions hold.

There are interesting directions for future work. It is an open question whether one can obtain regret guarantees under the Frequentist assumptions that the function belongs to some RKHS rather than being a draw from a GP prior. Such a bound would allow theoretical examination of the advantage of conditioning on concavity information. It would also be interesting to scale our algorithm using methods such as sparse GPs [Titsias, 2009, Hensman et al., 2013, Vakili et al., 2021].

Acknowledgements

We thank Amazon for funding this research. We also thank Piotr Suder, Joseph Lawson, and Yen-Chun Liu from the Reinforcement Learning Group at Amazon for some extremely helpful discussions.

References

Pierre-Cyril Aubin-Frankowski and Zoltán Szabó. Hard shape-constrained kernel machines. *Advances in Neural Information Processing Systems*, 33:384–395, 2020.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020. URL `http://arxiv.org/abs/1910.06403`.

Elisabetta Baldi and Corrado Bucherelli. The inverted "u-shaped" dose-effect relationships in learning and memory: modulation of arousal and consolidation. *Nonlinearity in biology, toxicology, medicine*, 3(1):nonlin–003, 2005.

Christoph E Boehm and Nitya Pandalai-Nayar. Convex supply curves. *American Economic Review*, 112(12): 3941–3969, 2022.

Ilija Bogunovic, Jonathan Scarlett, and Volkan Cevher. Time-varying gaussian process bandit optimization. In *Artificial Intelligence and Statistics*, pages 314–323. PMLR, 2016.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

Luis A Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563, 2000.

Edward J Calabrese and Linda A Baldwin. U-shaped dose-responses in biology, toxicology, and public health. *Annual review of public health*, 22(1):15–33, 2001.

Christopher D Carroll and Miles S Kimball. On the concavity of the consumption function. *Econometrica: Journal of the Econometric Society*, pages 981–992, 1996.

Stefano Cereda, Stefano Valladares, Paolo Cremonesi, Stefano Doni, et al. Cgptuner: a contextual gaussian process bandit approach for the automatic tuning of it configurations under varying workload conditions. *Proceedings of the VLDB Endowment*, 14(8):1401–1413, 2021.

Bo Chen, Rui Castro, and Andreas Krause. Joint optimization and variable selection of high-dimensional gaussian processes. *arXiv preprint arXiv:1206.6396*, 2012.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516): 1509–1521, 2016.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8): 753–764, 2009.

Sébastien Da Veiga and Amandine Marrel. Gaussian process modeling with inequality constraints. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 21, pages 529–555, 2012.

J Michael Davis and David J Svendsgaard. U-shaped dose-response curves: Their occurrence and implications for risk assessment. *Journal of Toxicology and Environmental Health, Part A Current Issues*, 30(2):71–83, 1990.

Carl De Boor and Carl De Boor. *A practical guide to splines*, volume 27. springer New York, 1978.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

Max Fathi, Nathael Gozlan, and Maxime Prod'homme. A proof of the caffarelli contraction theorem via entropic regularization. *Calculus of Variations and Partial Differential Equations*, 59(3):96, 2020.

Alan Genz and Giang Trinh. Numerical computation of multivariate normal probabilities using bivariate conditioning. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, pages 289–302. Springer, 2016.

Charles J Geyer. On the asymptotics of constrained m-estimation. *The Annals of statistics*, pages 1993–2010, 1994.

Lauren A. Hannah and David B. Dunson. Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research*, 14(102):3261–3294, 2013. URL `http://jmlr.org/papers/v14/hannah13a.html`.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

Xu He and Peter Chien. On the instability issue of gradient-enhanced gaussian process emulators for computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):627–644, 2018.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Raymond Kan and Cesare Robotti. On moments of folded and truncated multivariate normal distributions. *Journal of Computational and Graphical Statistics*, 26(4):930–934, 2017.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabas Poczos. Multi-fidelity gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, 66:151–196, 2019.

Parnian Kassraie and Andreas Krause. Neural contextual bandits without regret. In *International Conference on Artificial Intelligence and Statistics*, pages 240–278. PMLR, 2022.

Tarun Kathuria, Amit Deshpande, and Pushmeet Kohli. Batched gaussian process bandit optimization via determinantal point processes. *Advances in neural information processing systems*, 29, 2016.

Young-Heon Kim and Emanuel Milman. A generalization of caffarelli's contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354(3):827–862, 2012.

Andreas Krause and Cheng Ong. Contextual gaussian process bandit optimization. *Advances in neural information processing systems*, 24, 2011.

Gil Kur, Fuchang Gao, Adityanand Guntuboyina, and Bodhisattva Sen. Convex regression in multidimensions: Suboptimality of least squares estimators. *arXiv preprint arXiv:2006.02044*, 2020.

Michel Ledoux. Isoperimetry and gaussian analysis. In *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXIV—1994*, pages 165–294. Springer, 2006.

Hassan Maatouk and Xavier Bay. Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582, 2017.

BG Manjunath and Stefan Wilhelm. Moments calculation for the doubly truncated multivariate normal density. *Journal of Behavioral Data Science*, 1(1):17–33, 2021.

Rahul Mazumder, Arkopal Choudhury, Garud Iyengar, and Bodhisattva Sen. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525):318–331, 2019.

Mary C Meyer. Inference using shape-restricted regression splines. 2008.

Richard F Meyer and John W Pratt. The consistent assessment and fairing of preference functions. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):270–278, 1968.

Christopher J Portier and Frank Ye. U-shaped dose-response curves for carcinogens. *Human & Experimental Toxicology*, 17(12):705–707, 1998.

András Prékopa. Logarithmic concave measures with applications to stochastic programming. *ACTA SCIENTIARUM MATHEMATICARUM*, 1971.

James O Ramsay. Monotone regression splines in action. *Statistical science*, pages 425–441, 1988.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Andrew R Reynolds. Potential relevance of bell-shaped and u-shaped dose-responses for the therapeutic targeting of angiogenesis in cancer. *Dose-response*, 8(3):dose–response, 2010.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyYe6k-CW.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Jonathan Scarlett. Tight regret bounds for bayesian optimization in one dimension. In *International Conference on Machine Learning*, pages 4500–4508. PMLR, 2018.

Nícollas Silva, Heitor Werneck, Thiago Silva, Adriano CM Pereira, and Leonardo Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved February 15, 2026, from http://www.sfu.ca/~ssurjano.

Michalis K Titsias. Variational model selection for sparse gaussian process regression. *Report, University of Manchester, UK*, 2009.

Giang Trinh and Alan Genz. Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5):989–996, 2015.

Alexandre B Tsybakov. Nonparametric estimators. In *Introduction to Nonparametric Estimation*, pages 1–76. Springer, 2008.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

Michal Valko, Nathaniel Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

Hal R Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

Xiaojing Wang and James O Berger. Estimating shape constrained functions using gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25, 2016.

Zi Wang, Bolei Zhou, and Stefanie Jegelka. Optimization as estimation with gaussian processes in bandit settings. *arXiv preprint arXiv:1510.06423*, 2015.

Kaiwen Wu and Jacob R Gardner. A fast, robust elliptical slice sampling method for truncated multivariate normal distributions. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tkAtoZkcUnm.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

Tianshi Zhou. A gaussian process approach for contextual bandit-based battery replacement. In *ITM Web of Conferences*, volume 73, page 01020. EDP Sciences, 2025.

Exploiting Concavity Information in Gaussian Process Contextual Bandit Optimization (Supplementary Material)

Kevin Li[1] **Eric Laber**[1]

[1]Department of Statistical Science., Duke University, Durham, North Carolina, USA

7 PROOF OF γ_T BOUND

Proof. Recall that the covariance function on the GP induced on our function $f(a, \mathbf{x})$ takes form:

$$k_f\left((a, \mathbf{x}), (a', \mathbf{x}')\right) = \sum_{j=1}^{J} \boldsymbol{\phi}_j(a) k_j(\mathbf{x}, \mathbf{x}') \boldsymbol{\phi}_j(a').$$

This is an additive composition of product kernels. The product kernels are product of a Gaussian kernel defined on contexts with a linear kernel defined on actions. We can use the following identity from Krause and Ong [2011] to derive the maximum information gain γ_T for the composition of kernels k_S, k_Z defined on S, Z respectively:

$$\gamma(T, k_S \otimes k_Z, S \times Z) \leq d\gamma(T, k_S, S) + d\log(T)$$
$$\gamma(T, k_S \oplus k_Z, S \times Z) \leq \gamma(T, k_S, S) + \gamma(T, k_Z, Z) + 2\log(T),$$

Where Z has at most dimension d. The identities hold with S, Z reversed. Srinivas et al. [2009] show that for the d dimensional Gaussian $\gamma_T = \mathcal{O}(\log(T)^d)$ and the d dimensional Linear kernel is $\gamma_T = \mathcal{O}(d\log(T))$. Putting these results together gives us the Lemma and that for a Gaussian Kernel kernel on k_j we have that for k_f:

$$\gamma(T, k_f, \times\mathcal{A} \times \mathcal{X}) = \mathcal{O}(J\log(T)^{d+1}),$$

where $\mathcal{X} \subset \mathcal{R}^d$. □

8 PROOF OF POSTERIOR FORM FOR $\boldsymbol{\beta}_t$

Lemma 1. *Suppose that we observe $\mathbf{y}_{t-1} = (y_1, \ldots, y_{t-1})^T$. Define $\boldsymbol{\beta}_t := \{\boldsymbol{\beta}(\mathbf{x}_1), \ldots, \boldsymbol{\beta}(\mathbf{x}_t)\}^T$ to be the concatenated vector of coefficients from every context up to time t. Let $\mathcal{C}_{t'}$ denote the event that $\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq 0$, for $j = 1, \ldots, J-2$. Then the time t posterior $p\left(\boldsymbol{\beta}_t | \mathbf{y}_{t-1}, \cap_{t'=1}^{t} \mathcal{C}_{t'}\right)$ is the upper truncated Multivariate Normal distribution $N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ are the posterior quantities obtained from the standard GP time t posterior without truncation, and $\boldsymbol{\nu}_j = 0, j \leq J-2$ and $\boldsymbol{\nu}_j = \infty$ otherwise.*

Proof. We can write the posterior as

$$\begin{aligned} p(\boldsymbol{\beta}_t | \mathbf{y}_{t-1}, \boldsymbol{\beta}_t \leq \nu) &= \frac{p(\mathbf{y}_{t-1} | \boldsymbol{\beta}_t) p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_t \leq \nu)}{C_1} \\ &= \frac{p(\mathbf{y}_{t-1} | \boldsymbol{\beta}_t) p(\boldsymbol{\beta}_t) \mathbb{1}(\boldsymbol{\beta}_t \leq \nu)}{C_1 C_2} \\ &\propto p(\mathbf{y}_{t-1} | \boldsymbol{\beta}_t) p(\boldsymbol{\beta}_t) \mathbb{1}(\boldsymbol{\beta}_t \leq \nu). \end{aligned}$$

Where $C_1 = \int_{-\infty}^{\boldsymbol{\nu}} p(\mathbf{y}_{t-1}|\boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_t \leq \nu)d\boldsymbol{\beta}_t$ and $C_2 = \int_{-\infty}^{\boldsymbol{\nu}} p(\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t$. Given our observational model $p(\mathbf{y}_{t-1}|\boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t)$ is proportional to Multivariate Normal distribution (Both likelihood and prior are Multivariate Densities), giving us that the posterior in question is a truncated Multivariate Normal. □

9 PROOFS FOR LCTMVN CONCENTRATION PROPERTIES

9.1 KEY RESULTS FROM THE LITERATURE

We state relevant results from the literature.

Proposition 4 (Gaussian isoperimetric inequality [Ledoux, 2006])**.** *Let γ_d be the standard Gaussian measure on $\mathbb{R}^d$. For a Borel set $A \subseteq \mathbb{R}^d$ and $r \geq 0$ define $A^{(r)}$ to be the Minkowski sum of A and the d-dimensional ball with radius r. If $A \subseteq \mathbb{R}^d$ is Borel and H is any half-space with $\gamma_d(A) \geq \gamma_d(H)$, then for all $r \geq 0$ we have:*

$$\gamma_d(A^{(r)}) \geq \gamma_d(H^{(r)}).$$

Equivalently, if Φ is the standard 1d standard normal CDF and $\gamma_d(A) \geq \Phi(a)$ then

$$\gamma_d(A^{(r)}) \geq \Phi(a + r).$$

In particular,

$$\Phi^{-1}(\gamma_d(A^r)) \geq \Phi^{-1}(\gamma_d(A)) + r$$

Proposition 5 (Generalized Caffarelli Contraction Theorem Kim and Milman [2012], Fathi et al. [2020])**.** *Let $\mu = \exp\{-Q(\mathbf{x})\}d\mathbf{x}$ and $\nu = \exp\{-Q(\mathbf{x}) - V(\mathbf{x})\}dx$ denote two borel probability measures on Euclidean space $(\mathbb{R}^d, |\cdot|)$, where Q denotes a quadratic function, i.e.:*

$$Q(\mathbf{x}) = \langle A\mathbf{x}, \mathbf{x}\rangle + \langle \mathbf{b}, \mathbf{x}\rangle + c$$

with A positive-definite, and V is a convex function that can possibly take value $+\infty$. Then the Brenier optimal-transport map $T = T_{opt}$ pushing forward μ onto ν is a contraction:

$$\forall \mathbf{x}, \mathbf{y} \in R^d \quad ||T(\mathbf{x}) - T(\mathbf{y})|| \leq ||\mathbf{x} - \mathbf{y}||.$$

9.2 PROOF OF TRUNCATION-INSENSITIVE CONCENTRATION THEOREM AROUND MEDIAN

Lemma 4. *Define $\mathbf{z} \sim N(0, \mathbf{I}_d)$. Define convex region K and random variable $\mathbf{z}'$ such that $\mathbf{z}' \stackrel{d}{=} \mathbf{z}|\mathbf{z} \in K$. There exists a contraction map T such that $T(\mathbf{z}) \stackrel{d}{=} \mathbf{z}'$*

Proof: The conditional density of $\mathbf{z}|\mathbf{z} \in K$ is proportional to:

$$\begin{aligned} p(\mathbf{z}|\mathbf{z} \in K) &\propto \exp(-\frac{1}{2}\mathbf{z}^2)\mathbb{1}(\mathbf{z} \in K) \\ &= \exp\left\{-\left(\frac{1}{2}\mathbf{z}^2 + V_K(\mathbf{z})\right)\right\}, \end{aligned}$$

where we define $V_K(\mathbf{z})$ as

$$V_K(\mathbf{z}) = \begin{cases} 0, & \text{if } \mathbf{z} \in K, \\ \infty, & \text{else.} \end{cases}$$

Note that $V_K(\boldsymbol{\beta})$ is a convex function as K is convex. Let $Q(\mathbf{z}) \propto -\frac{1}{2}\mathbf{z}^2$ and $V(\mathbf{z}) = V_K(\mathbf{z})$ be defined above. Then we can apply Proposition 5 to prove the existence of a contraction map T that pushes forward the base Gaussian measure to the Gaussian measure conditioned on the truncation K. In other words, there exists T such that $T(\mathbf{z}) \stackrel{d}{=} z'$. □

Lemma 5. *Let $f(\cdot)$ be a L-Lipshitz function and $\mathbf{z} \sim \gamma_d$ where γ_d is a standard Gaussian measure in $\mathbb{R}^d$. Let $A = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq m\}$ where m is the median of $f(\mathbf{z})$. Fix $t \geq 0$ and $r = t/L$, then we have*

$$A^{(r)} \subseteq \{\mathbf{z} : f(\mathbf{z}) \leq m + t\}$$

Proof: By definition of $A^{(r)}$, for any $\mathbf{z} \in A^{(r)}$ there exists $\mathbf{a} \in A$ such that $||\mathbf{z} - \mathbf{a}|| \leq r$. Because $f(\cdot)$ is L-Lipshitz:

$$f(\mathbf{z}) - f(\mathbf{a}) \leq L||\mathbf{z} - \mathbf{a}|| \implies f(\mathbf{z}) \leq f(\mathbf{a}) + L||\mathbf{z} - \mathbf{a}|| \leq m + Lr = m + t,$$

which implies the inclusion. □

Proposition 2. *Suppose $\boldsymbol{\beta}_t \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$. Let $m_t(a, \mathbf{x}_t)$ denote the median of $\mathbf{c}^T(a)\boldsymbol{\beta}_t$. Then*

$$Pr\left\{|\mathbf{c}(a)^T\boldsymbol{\beta}_t - m_t(a, \mathbf{x}_t)| \geq b\right\} \leq \exp\left(\frac{-b^2}{2\sigma^2_{\mathbf{c}_t}(a, \mathbf{x}_t)}\right)$$

Proof: Define $A = \Sigma^{\frac{1}{2}}$ and let $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_d)$. We can write $\boldsymbol{\beta} \stackrel{d}{=} \mu + A\mathbf{z}'$, where $\mathbf{z}' \stackrel{d}{=} \mathbf{z} \mid \mathbf{z} \leq A^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu})$

Define the truncation set $K = \{\mathbf{z} \in \mathbb{R}^d : (\boldsymbol{\mu} + A\mathbf{z})_i \leq \boldsymbol{\nu}_i, i = 1, \cdots, d\}$. K is a convex regions because it is a affine transformation of the original half-space truncation region. By Lemma 4 there exists a transport map T such that $T(\mathbf{z}) \stackrel{d}{=} \mathbf{z}'$.

Now define function $h(\mathbf{z}) = \mathbf{c}^T\{\boldsymbol{\mu} + AT(\mathbf{z})\}$. We can show that for constants $\mathbf{z}_1, \mathbf{z_2} \in \mathbb{R}^d$ that

$$||h(\mathbf{z}_1) - h(\mathbf{z}_2)|| \leq ||A^T\mathbf{c}||\,||T(\mathbf{z}_1) - T(\mathbf{z}_2)|| \leq \sqrt{\mathbf{c}^T\Sigma\mathbf{c}}\,||\mathbf{z}_1 - \mathbf{z}_2||,$$

Where we use that $T(\cdot)$ is a contraction, $||A^T\mathbf{c}|| = \sqrt{\mathbf{c}^T AA^T\mathbf{c}}$ and $A = \boldsymbol{\Sigma}^{\frac{1}{2}}$. This shows that $h(\mathbf{z})$ has Lipshitz property $L = \sqrt{\mathbf{c}^T\Sigma\mathbf{c}}$.

Define the set $A := \{\mathbf{z} \in \mathbb{R}^d : h(\mathbf{z}) \leq m\}$. By definition $\gamma_d(A) \geq \frac{1}{2}$. Fix $t \geq 0$ and set $r = \frac{b}{\sqrt{\mathbf{c}^T\Sigma\mathbf{c}}}$. By using lemma 9.2 and taking complements, we have

$$\{\mathbf{z} \in \mathbb{R}^d : h(\mathbf{z}) > m + t\} \subseteq (A^{(r)})^C \implies \Pr\{h(\mathbf{z}) > m + b\} \leq 1 - \gamma_d(A^{(r)})$$

We can now apply the Isoperimetric inequality from Proposition 4 to $\gamma_d(A^{(r)})$:

$$\gamma_d(A) \geq \Phi(0) \implies \gamma_d(A^{(r)}) \geq \Phi(0 + r)$$

Thus we arrive at the one tail inequality

$$\Pr\{h(\mathbf{z}) > b + m\} = \Pr\{\boldsymbol{\beta} - m > t\} \leq 1 - \Phi\left\{\frac{b}{\sqrt{\mathbf{c}^T\Sigma\mathbf{c}}}\right\}$$

To bound left tail, $\Pr\{\boldsymbol{\beta} \leq m - c\}$ note that $-m$ is the median of $-\boldsymbol{\beta}$, and that $-h(\mathbf{z})$ has the same Lipchitz constant as $h(\mathbf{z})$. Therefore we can apply the same argument above and obtain,

$$\Pr\{\boldsymbol{\beta} \leq m - c\} = \Pr\{-h(\boldsymbol{\beta}) \geq -m + b\} \leq 1 - \Phi\left\{\frac{b}{\sqrt{\mathbf{c}^T\Sigma\mathbf{c}}}\right\}.$$

The union bound gives us the final two tailed bound. □

9.3 PROOFS OF TRUNCATION SENSITIVE CONCENTRATION AROUND MEAN

Lemma 6. *Let C be a convex region and $p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\beta})$ be the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance Σ. Then $q(\boldsymbol{\mu}) = \int_C p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\beta})d\boldsymbol{\beta}$ is log concave in $\boldsymbol{\mu}$. Equivalently $-log\{q(\boldsymbol{\mu})\}$ is convex.*

Proof: We can rewrite the multivariate normal density as

$$p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\beta}) = p_{\mathbf{0},\Sigma}(\boldsymbol{\beta} - \boldsymbol{\mu}),$$

which shows that it is log-concave with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. By Prekopa's theorem, integration over a convex region preserves log concavity [Prékopa, 1971]. Therefore $q(\boldsymbol{\mu})$ is log-concave.

Lemma 7. *Suppose $\boldsymbol{\beta} \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}, \Sigma)$ and let $p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\beta})$ be the standard multivariate normal density with mean $\boldsymbol{\mu}$ and covariance Σ. Define*

$$q(\boldsymbol{\mu}) = \int_K p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

Where $K = \{\boldsymbol{\beta} \in \mathbb{R}^d : \boldsymbol{\beta}_i \leq \boldsymbol{\nu}_i, i, \cdots, d\}$. Let $\boldsymbol{\mu}^* = \mathbb{E}(\boldsymbol{\beta})$, then:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} + \Sigma \nabla g(\boldsymbol{\mu}),$$

where $g(\boldsymbol{\mu}) = \log\{q(\boldsymbol{\mu})\}$.

Proof: We can differentiate under the integral of $q(\boldsymbol{\mu})$ as the density is gaussian:

$$\nabla_{\boldsymbol{\mu}} q(\boldsymbol{\mu}) = \int_K \nabla_{\boldsymbol{\mu}} p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\beta}) d\boldsymbol{\beta} = \int_K \Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}) p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\beta}) d\boldsymbol{\beta} = \Sigma^{-1} \left[\mathbb{E}\{\mathbf{z}\mathbb{1}_K(\boldsymbol{\beta})\} - \boldsymbol{\mu} q(\boldsymbol{\mu})\right].$$

From this we can divide both of the expression by $q(\boldsymbol{\mu})$ to obtain:

$$\frac{\nabla q(\boldsymbol{\mu})}{q(\boldsymbol{\mu})} = \Sigma^{-1}(\boldsymbol{\mu}^* - \boldsymbol{\mu}).$$

Noting that $\frac{\nabla q(\boldsymbol{\mu})}{q(\boldsymbol{\mu})} = \nabla g(\boldsymbol{\mu})$ we arrive at the result:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} + \Sigma \nabla g(\boldsymbol{\mu}).$$

□

Lemma 8. *Suppose $\boldsymbol{\beta} \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}, \Sigma)$ and define $\boldsymbol{\mu}^* = \mathbb{E}(\boldsymbol{\beta})$. Let $\mathbf{c}$ be a vector of constants in $\mathbb{R}^d$ and define $u = \mathbf{c}^T(\boldsymbol{\beta} - \boldsymbol{\mu}^*)$. Then the moment generating function of u, $\psi(k)$, is:*

$$\psi(k) := \mathbb{E}\{\exp(ku)\} = \exp\left(\frac{1}{2}k^2\sigma^2\right) \exp\left\{g(\boldsymbol{\mu} + k\Sigma\mathbf{c}) - g(\boldsymbol{\mu}) - k\mathbf{c}^T(\boldsymbol{\mu}^* - \boldsymbol{\mu})\right\},$$

where $\sigma^2 = \mathbf{c}^t \Sigma \mathbf{c}$ and $g(\boldsymbol{\mu})$ is defined as in Lemma 7

Proof:

$$\mathbb{E}\{\exp(ku)\} = \exp(-k\mathbf{c}^T\boldsymbol{\mu}^*) \frac{\mathbb{E}\left\{\exp(k\mathbf{c}^T\boldsymbol{\beta})\mathbb{1}(\boldsymbol{\beta} \preccurlyeq \boldsymbol{\nu})\right\}}{q(\boldsymbol{\mu})}.$$

Completing the square we can derive using standard gaussian identities that:

$$\exp(k\mathbf{c}^T\boldsymbol{\beta}) p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\beta}) = \exp\left(k\mathbf{c}^T\boldsymbol{\mu} + \frac{1}{2}k^2\sigma^2\right) p_{\boldsymbol{\mu}+k\Sigma\mathbf{c},\Sigma}(\boldsymbol{\beta}).$$

Therefore

$$\mathbb{E}\left\{\exp(k\mathbf{c}^T\boldsymbol{\beta})\mathbb{1}(\boldsymbol{\beta} \preccurlyeq \boldsymbol{\nu})\right\} = \exp\left(k\mathbf{c}^T\boldsymbol{\mu} + \frac{1}{2}k^2\sigma^2\right) q(\boldsymbol{\mu} + k\Sigma\mathbf{c})$$

Plugging this into back into the original expression we get our result

$$\mathbb{E}\{\exp(ku)\} = \exp\left(\frac{1}{2}k^2\sigma^2\right) \exp\left\{g(\boldsymbol{\mu} + k\Sigma\mathbf{c}) - g(\boldsymbol{\mu}) - k\mathbf{c}^T(\boldsymbol{\mu}^* - \boldsymbol{\mu})\right\}$$

□

Proposition 1. *Assume the concavity conditioned posterior of our coefficient vector $\boldsymbol{\beta}$ at time t is $\boldsymbol{\beta}_t \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$. Let $\boldsymbol{\mu}_t^* = \mathbb{E}(\boldsymbol{\beta}_t | \mathbf{y}_{t-1}, \cap_{t'=1}^t \mathcal{C}_{t'})$. Then given $b \geq 0$:*

$$Pr\left\{|\mathbf{c}_t(a)^T\boldsymbol{\beta}_t - \mathbf{c}_t(a)^T\boldsymbol{\mu}_t^*| \geq b\right\} \leq \exp\left(-\frac{b^2}{2\sigma^2_{\mathbf{c}_t}(a, \mathbf{x}_t)}\right) \times P_t(a, b)$$

Where

$$P_t(a,b) = \left\{ \exp\left[-D_{G,\mathbf{c}_t(a)}^{\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t}\left(\boldsymbol{\mu}_t + \frac{b}{\sigma^2_{\mathbf{c}_t}(a,\mathbf{x}_t)}\Sigma_t \mathbf{c}_t(a), \boldsymbol{\mu}_t \right) \right] \right.$$
$$\left. + \exp\left[-D_{G,\mathbf{c}_t}^{\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t}\left(\boldsymbol{\mu}_t - \frac{b}{\sigma^2_{\mathbf{c}_t}(a,\mathbf{x}_t)}\Sigma_t \mathbf{c}_t(a), \boldsymbol{\mu}_t \right) \right] \right\}$$

$D_{G,\mathbf{c}}^{\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t}(\cdot,\cdot)$ *represents the Bregman divergence with respect to the convex function* $G(\boldsymbol{\mu}) = -\log\left\{ \int_{\boldsymbol{\beta} \preccurlyeq \nu} p_{\boldsymbol{\mu},\Sigma_t}(\boldsymbol{\beta}) d\boldsymbol{\beta} \right\}$. *Here* $p_{\boldsymbol{\mu},\Sigma_t}(\boldsymbol{\mu})$ *is the standard multivariate normal density with mean* $\boldsymbol{\mu}$ *and covariance* Σ_t.

Proof: We drop the subscripts for clarity. We first examine the one sided bound, $\Pr\left\{ \mathbf{c}^T(\boldsymbol{\beta} - \boldsymbol{\mu}^* \geq b \right\}$. We apply the Chernoff bound using the moment generating function derived in Lemma 8 with the optimal $k = \frac{b}{\sigma^2}$:

$$\Pr\left\{ \mathbf{c}^T(\boldsymbol{\beta} - \boldsymbol{\mu}^*) \geq b \right\} \leq \inf_k \exp(-kb) \exp\left(\frac{1}{2} k^2 \sigma^2 \right) \exp\left\{ g(\boldsymbol{\mu} + k\Sigma\mathbf{c}) - g(\boldsymbol{\mu}) - k\mathbf{c}^T(\boldsymbol{\mu}^* - \boldsymbol{\mu}) \right\}$$
$$= \exp\left(-\frac{b^2}{2\sigma^2} \right) \frac{q\left(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c} \right)}{q(\boldsymbol{\mu})} \exp\left(-\frac{b}{\sigma^2} d \right)$$

We can obtain the other tail by applying the Chernoff bound to $\mathbf{c}^T(\boldsymbol{\beta} - \boldsymbol{\mu}^*)$ and taking the union bound with the above bound. This gives the first line of the result.

To relate the inequality to Bregman divergences recall from Lemma 7 that $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \Sigma \nabla g(\boldsymbol{\mu})$. Using this we can write:

$$\mathbf{c}^T(\boldsymbol{\mu}^* - \boldsymbol{\mu}) = \langle \nabla g(\boldsymbol{\mu}), \Sigma\mathbf{c} \rangle$$

So that we can rewrite factors in terms of G

$$-\log\left\{ \frac{q\left(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c} \right)}{q(\boldsymbol{\mu})} \exp\left(-\frac{b}{\sigma^2} d \right) \right\} = G(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c}) - G(\boldsymbol{\mu}) - \langle \nabla G(\boldsymbol{\mu}), \frac{b}{\sigma^2}\Sigma\mathbf{c} \rangle.$$

Notice that by Lemma 6, $G(\theta)$ is strictly convex and this value is greater than zero. Therefore it is a valid Bregman divergence and:

$$G\left(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c} \right) - G(\boldsymbol{\mu}) - \langle \nabla G(\boldsymbol{\mu}), \frac{b}{\sigma^2}\Sigma\mathbf{c} \rangle = D_{G,\mathbf{c}}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c}, \boldsymbol{\mu} \right).$$

The left tail follows the exact same argument. Plugging this form into the bound provides the result. □

Lemma 9. *Define function*

$$h(b) := \exp\left[-D_{G,\mathbf{c}}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c}, \boldsymbol{\mu} \right) \right] + \exp\left[-D_{G,\mathbf{c}}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\boldsymbol{\mu} - \frac{b}{\sigma^2}\Sigma\mathbf{c}, \boldsymbol{\mu} \right) \right]$$

$h(b)$ *is decreasing for* $b \geq 0$.

Proof: Define $\mathbf{v} = \frac{1}{\sigma^2}\Sigma\mathbf{c}$ and write:

$$k(b) = D_{G,\mathbf{c}}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\boldsymbol{\mu} + \frac{b}{\sigma^2}\Sigma\mathbf{c}, \boldsymbol{\mu} \right) = G(\boldsymbol{\mu} + b\mathbf{v}) - G(\boldsymbol{\mu}) - b\langle \nabla G(\boldsymbol{\mu}), \mathbf{v} \rangle$$

Differentiating k with respect to b we obtain:

$$k'(b) = \langle \nabla G(\boldsymbol{\mu} + b\mathbf{v}) - \nabla G(\boldsymbol{\mu}), \mathbf{v} \rangle.$$

Given the strict-convexity of G, we know that its gradients are monotone. Therefore:

$$\langle \nabla G(\boldsymbol{\mu} + b\mathbf{v}) - \nabla G(\boldsymbol{\mu}), b\mathbf{v} \rangle \geq 0 \implies k'(b) \geq 0.$$

Therefore we see that $k(b)$ is increasing. The exact same argument applies to $D_{G,\mathbf{c}}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\boldsymbol{\mu} - \frac{b}{\sigma^2}\Sigma\mathbf{c}, \boldsymbol{\mu} \right)$, which shows that it is increasing. Therefore $h(b)$ is decreasing. □

9.4 COMBINED BOUND CONCENTRATION PROOF

Lemma 10. *If $P_t(a, \chi_t) < 1$ and $\chi_t \geq 1$ then $\chi_t^* < \chi_t$ exists.*

Proof: From Lemma 9 we have that $P_t(a, x)$ is monotone decreasing with respect to s. Therefore $\exp\left(-\frac{s}{2}\right) P_t(a, s\sigma_{\mathbf{c}_t})$ is monotone decreasing in s. We also have by assumption:

$$\exp\left(-\frac{\chi_t}{2}\right) P_t(a, \chi_t \sigma_{\mathbf{c}_t}(a, \mathbf{x}_t)) < \exp\left(-\frac{\chi_t}{2}\right)$$

All functions are monotone decreasing and continuous. In addition $\exp\left(-\frac{0}{2}\right) P_t(a, 0) = 1 \geq \exp\left(-\frac{\chi_t}{2}\right)$. Therefore by the intermediate value theorem, there exists $s < \chi_t$ such that the inequality is fulfilled. □

Lemma 2. *Suppose $\boldsymbol{\beta}_t \sim N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$ and Let χ_t be a predefined constant. Define*

$$\kappa_t(a, \mathbf{x}_t) = \begin{cases} \mathbf{c}_t(a)^T \boldsymbol{\mu}_t^*, & \textit{if } P_t\left[a, \chi_t \sigma_{\mathbf{c}_t}(a, \mathbf{x}_t)\right] < 1, \\ m_t(a, \mathbf{x}_t), & \textit{else}. \end{cases}.$$

$$\alpha_t(a, \mathbf{x}_t) = \begin{cases} \chi_t^*(a, \mathbf{x}_t), & \textit{if } P_t\left[a, \chi_t \sigma_{\mathbf{c}_t}(a, \mathbf{x}_t)\right] < 1, \\ \chi_t, & \textit{else}. \end{cases}.$$

$$\begin{aligned} &\chi_t^*(a, \mathbf{x}_t) \\ &:= \inf_s \left\{ s : \exp\left(-\frac{s}{2}\right) P_t\left[a, s\sigma_{\mathbf{c}_t}(a, \mathbf{x}_t)\right] \leq \exp\left(-\frac{\chi_t}{2}\right) \right\}. \end{aligned}$$

$\chi_t^*(a, \mathbf{x}_t)$ exists and we have the following concentration bound:

$$\begin{aligned} &Pr\left\{|f(a, \mathbf{x}_t) - \kappa(a, \mathbf{x}_t)| \geq \alpha_t(a, \mathbf{x}_t)\sigma_{\mathbf{c}_t}(a, \mathbf{x}_t)\right\} \\ &\leq \exp\left(-\frac{\chi_t}{2}\right). \end{aligned}$$

Proof: First note that at time t, upon observing information according to our assumptions, $f(a, \mathbf{x}_t) = \mathbf{c}_t(a)^T \boldsymbol{\beta}_t$ where $\boldsymbol{\beta}_t$ is distributed according to the upper truncated normal distribution $N_{\boldsymbol{\nu}}(\boldsymbol{\mu}_t, \Sigma_t)$.

First consider the case where $P_t(a, \chi_t) \geq 1$. Then the confidence region bound is centered around the median using χ_t. Applying Proposition 2 and noting from Srinivas et al. [2009] that $2\left\{1 - \Phi\left(\chi_t^{1/2}\right)\right\} \leq \exp\left(-\frac{\chi_t}{2}\right)$ the statement holds.

Now consider the case where $P_t(a, \mathbf{x}_t, \chi_t) < 1$. By construction and application of Proposition 1 we can write:

$$\Pr\left\{|f(a, \mathbf{x}_t) - \mu_t^*(a, \mathbf{x}_t)| \geq \alpha_t(a, \mathbf{x}_t)\sigma_{\mathbf{c}_t}\right\} \leq \exp\left(-\frac{\chi_t^*}{2}\right) P_t(a, \chi_t^* \sigma_{\mathbf{c})t}(a, \mathbf{x}_t)) \leq \exp\left(-\frac{\chi_t}{2}\right).$$

□

10 MAXIMUM INFORMATION GAIN

Following Srinivas et al. [2009], we will use information theoretical quantities to bound the regret of our algorithm. Therefore we will be interested in the mutual information $\mathrm{I}(\mathbf{y}_A, f)$ between the set of observations $\mathbf{y}_A$ and the underlying function f. $\mathrm{I}(\mathbf{y}_A, f)$ represents the reduction in uncertainty surrounding the mean reward function f upon observing rewards $\mathbf{y}_A$. Our regret bounds will take the form $\mathcal{O}^*(\sqrt{T\gamma_T})$ where γ_T is the *maximal information gain* defined as

$$\gamma_T := \max_{A \subset \mathcal{A}, |A| = T} \mathrm{I}(\mathbf{y}_A, f).$$

The maximal information gain, γ_T, for our model can be bounded using the results of Krause and Ong [2011]. The GP induced on $f(a, \mathbf{x})$ by our model has a covariance function composed of J additive components. Each additive component is the product of a generic d-dimensional covariance function (such as Gaussian or Matérn) and a 1-dimensional linear kernel derived from the C-Spline features. Krause and Ong [2011] show how to bound γ_T for additive and product compositions. Thus, we are able to obtain the following asymptotic bounds for the CSGP model.

Lemma 11. *Let the prior for the expected reward function $f(a, \mathbf{x})$ be the CSGP model defined in Equation 4 such that the maximal information gain of each $k_j(\cdot, \cdot)$ at time T is bounded by γ_T^*. Then the maximal information gain for the CSGP covariance function satisfies*

$$\gamma_T \leq J\gamma_T^* + 4J\log(T).$$

Vakili et al. [2021] provides bounds for γ_T^* for common kernels. Throughout this paper we select $k_j(\cdot, \cdot)$ to be Gaussian kernels so that the maximum information gain for CSGP satisfies $\gamma_T = \mathcal{O}\left\{J\log(T)^{d+1}\right\}$.

10.1 PROOF OF LEMMA 11

Proof. Recall that the covariance function on the GP induced on our function $f(a, \mathbf{x})$ takes form:

$$k_f\left((a, \mathbf{x}), (a', \mathbf{x}')\right) = \sum_{j=1}^{J} \boldsymbol{\phi}_j(a) k_j(\mathbf{x}, \mathbf{x}') \boldsymbol{\phi}_j(a').$$

This is an additive composition of product kernels. The product kernels are product of a Gaussian kernel defined on contexts with a linear kernel defined on actions. We can use the following identity from Krause and Ong [2011] to derive the maximum information gain γ_T for the composition of kernels k_S, k_Z defined on S, Z respectively:

$$\gamma(T, k_S \otimes k_Z, S \times Z) \leq d\gamma(T, k_S, S) + d\log(T)$$
$$\gamma(T, k_S \oplus k_Z, S \times Z) \leq \gamma(T, k_S, S) + \gamma(T, k_Z, Z) + 2\log(T),$$

Where Z has at most dimension d. The identities hold with S, Z reversed. Srinivas et al. [2009] show that for the d dimensional Gaussian $\gamma_T = \mathcal{O}(\log(T)^d)$ and the d dimensional Linear kernel is $\gamma_T = \mathcal{O}(d\log(T))$. Putting these results together gives us the Lemma and that for a Gaussian kernel on k_j we have that for k_f:

$$\gamma(T, k_f, \times\mathcal{A} \times \mathcal{X}) = \mathcal{O}(J\log(T)^{d+1}),$$

where $\mathcal{X} \subset \mathcal{R}^d$.

□

11 REGRET PROOFS

Our proof for the UCB regret follows the ideas presented by Srinivas et al. [2009], Krause and Ong [2011]. First we prove that given our choice of χ_t the our upper confidence bound holds with high probability over the finite decision set over all t and $a \in \mathcal{A}$:

Lemma 12. *Pick $\delta \in (0, 1)$. Set $\chi_t = 2\log(|\mathcal{A}|t^2\pi^2/6\delta)$. Then*

$$Pr\left\{\forall t, \forall a \in \mathcal{A}, |f(a, \mathbf{x}_t) - \kappa_t(a, \mathbf{x}_t)| \leq \sigma_{\mathbf{c}_t}(a_t, \mathbf{x}_t)\alpha_t(a, \mathbf{x}_t)^{\frac{1}{2}}\right\} \geq 1 - \delta.$$

Proof. We apply the Union bound over a in the bound obtained in Lemma 2 so that

$$\Pr\left\{\forall a \in \mathcal{A}, |f(a, \mathbf{x}_t) - \kappa_t(a, \mathbf{x}_t)| \leq \sigma_{\mathbf{c}_t}(a_t, \mathbf{x}_t)\alpha_t(a, \mathbf{x}_t)^{\frac{1}{2}}\right\} = \Pr\left\{\forall a \in \mathcal{A}, |\mathbf{c}_t(a)^T\boldsymbol{\beta}_t - \kappa_t(a, \mathbf{x}_t)| \leq \sigma_{\mathbf{c}_t}(a_t, \mathbf{x}_t)\alpha_t(a, \mathbf{x}_t)^{\frac{1}{2}}\right\}$$
$$\geq 1 - |\mathcal{A}|\exp\left(-\frac{1}{2}\chi_t\right)$$

Where the inequality follows from the concentration inequality proved in Lemma 2. Taking a Union bound with respect to t, choosing χ_t such that $|\mathcal{A}|\exp\left(-\frac{\chi_t}{2}\right) = \frac{6\delta}{\pi^2 t^2}$, then the statement holds. □

We also prove the following concentration results that will used for the continuous case where there are increasingly fine discretizations $\mathcal{A}_t$.

We first state the following lemma that adapts lemma 5.5 from [Srinivas et al., 2009] for our purposes:

Lemma 13. *Pick $\delta \in (0,1)$ and $\alpha_t = 2\log(\pi_t/\delta)$ where $\sum_{t\geq 1}\pi_t = 1, \pi_t > 0$ then*

$$|f(a_t,\mathbf{x}_t) - \kappa(a_t,\mathbf{x}_t)| \leq \sigma_{\mathbf{c}_t}(a_t,\mathbf{x}_t)\alpha_t^{\frac{1}{2}}(a_t,\mathbf{x}_t) \quad \forall t \geq 1,$$

holds with probability $1-\delta$.

Proof. Fix $t \geq 1$ and $a \in \mathcal{A}, \mathbf{x} \in \mathcal{X}$. Lemma 2 already showed that conditioning on the action/contexts, rewards, and concavity observed before taking action time t :

$$\Pr\left\{|f(a,\mathbf{x}_t) - \kappa_t(a,\mathbf{x}_t)| > \sigma_{\mathbf{c}_t}(a,\mathbf{x}_t)\alpha_t(a,\mathbf{x}_t)^{\frac{1}{2}}\right\} \leq \exp\left(\frac{-\chi_t}{2}\right)$$

We have that $\exp\left(-\frac{\chi_t}{2}\right) = \frac{\delta}{\pi_t}$ so we can take the union bound for $t \in \mathbb{N}$ to prove the statement. □

Lemma 3. *For $t \geq 1$, with probability $\geq 1-\delta$, $r_t \leq 2\alpha_t(a_t,\mathbf{x}_t)^{\frac{1}{2}}\sqrt{\gamma_T}$*

Proof Given the inequality proven in lemma 12, we can directly apply lemma 4.1 from Krause and Ong [2011] which gives

$$r_t \leq \alpha_t(a,\mathbf{x}_t)^{\frac{1}{2}}\sigma_{\mathbf{c}_t}(a,\mathbf{x}_t).$$

An argument from Russo and Van Roy [2014] shows that that $\sum_{t=1}^{T}\sigma^2_{\mathbf{c}_t}(a,\mathbf{x}_t) \leq \gamma_T$. The argument follows. □

Proposition 3. *Let the expected reward function f be a sample from a known CSGP prior with noise model $N(0,\sigma^2)$. Also suppose that at each time step t the bandit algorithm receives noisy reward y_t as well as concavity information that $\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \leq 0$ for $j = 1\ldots J-2, t' \leq t$. Pick $\delta \in (0,1)$ and suppose one of the following assumptions holds:*

- *$\mathcal{A}$ is finite with $\chi_t = 2\log(|\mathcal{A}|t^2\pi^2/6\delta)$*
- *$\mathcal{A} \subset [0,1]$ is compact and convex and the CSGP kernel $k_f(\cdot,\cdot)$ satisfies the following bound on the GP sample paths.*[1] *For some constants η, ζ*

$$\Pr\{\sup_{\mathbf{v}\in\mathcal{A}\times\mathcal{X}}|\partial f/\partial \mathbf{v}_j| > L\} \leq \eta\exp\left\{-(L/\zeta)^2\right\},$$

for $\forall j \in \{1,\ldots,d+1\}$. And we choose

$$\alpha_t = 2\log(2\pi^2t^2/3\delta) + 2\log\left\{t^2\zeta\sqrt{\log(4\eta/\delta)}\right\}.$$

Define $\alpha_T^* = \max\{\alpha_t(a_t,\mathbf{x}_t)\}_{t=1}^T$ so that $\alpha_T^* \leq \chi_T$. Running CSGP-UCB obtains a regret bound $\mathcal{O}^*(\sqrt{T\gamma_T\alpha_T})$ with high probability. Precisely let $C_1 = \frac{8}{\log(1+\sigma^{-2})}$ then:

$$\Pr\left\{R_t \leq \sqrt{C_1T\alpha_T^*\gamma_T} + 2, \forall T \geq 1\right\} \geq 1-\delta,$$

Proof. First we study the discrete case. Conditional on our high probability bound results proved in Lemmas 12, 13, the arguments follow Krause and Ong [2011] up to Theorem 5. Given the arguments up to that point we have with probability $\geq 1-\delta$ that

$$r_t^2 \leq 4\alpha_t\sigma_t^2(a,\mathbf{x}_t)$$

However, in our case, the confidence parameter α_t is no longer strictly increasing. Therefore α_T is not necessarily the largest value. So instead we take $\alpha_T^* = \max\{\alpha_t\}_{t=1}^T$:

$$r_t^2 \leq 4\alpha_t^*\sigma_t^2(a,\mathbf{x}_t)$$

and the theorem follows from the same arguments as in Krause and Ong [2011]. Note that $\alpha_T^* \leq \chi_T$ by construction.

The continuous case follows Krause and Ong [2011] similarly conditional on Lemma 13. We replace the upper bound confidence parameter $\alpha_T(a,\mathbf{x}_t)$ with α_T^*, the maximum over all $\alpha_t(a,\mathbf{x}_t)$ up to time T.

□

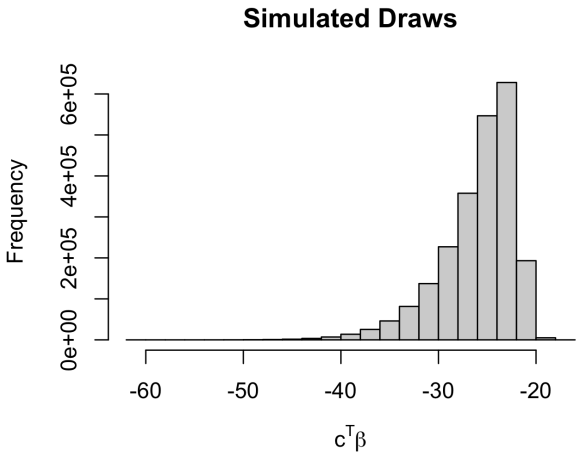

Figure 5: Example draws of the LCTMVN variable. The distribution is very skewed.

12 EXAMPLE OF LCTMVN TAILS HEAVIER THAN ORIGINAL GAUSSIAN

For exposition, we provide an easily understandable and reproducible LCTMVN produced from a truncated bivariate normal distribution. Take $\mathbf{c} = (1, 10)^T$ and define the truncation parameter to be $\boldsymbol{\nu} = (-2, 10)$. Define a bivariate normal distribution with unit variance and correlation parameter $\rho = .995$. Simulated draws of the the LCTMVN variable of the given MVN distribution with truncation parameter $\boldsymbol{\nu}$ are shown in Figure 5. we see that the distribution has a heavy right tail. In fact one can calculate analytically/numerically that this distribution has weaker concentration properties than a standard gaussian. Let $\boldsymbol{\beta}'$ be a draw from the un-truncated MVN distribution defined above. Then we have

$$\Pr\left\{\mathbf{c}^T\boldsymbol{\beta} < \ E(\mathbf{c}^T\boldsymbol{\beta}) + 1\right\} \approx .501$$
$$\Pr\left\{\mathbf{c}^T\boldsymbol{\beta} < \ E(\mathbf{c}^T\boldsymbol{\beta'}) + 1\right\} \approx .536$$

Which shows that the LCTMVN has exceeds standard Gaussian concentration for this width parameter. In addition, this probability for the LCTMVN violates the standard gaussian bandit bound $\frac{1}{2}\exp(-1/\sigma_c^2) = .51$. Although, the correlation here is extreme and the confidence tight, we have found empirically that this violation of gaussian concentration can get much worse in common cases for higher dimensions.

13 UCB COMPUTATION

Computing our bound requires computing the LCTMVN variable means ($\boldsymbol{\mu}^*$), medians ($m(a, \mathbf{x}_t)$), orthant probabilities ($\Pr\{\boldsymbol{\beta}_t \preccurlyeq \boldsymbol{\nu}\}$), and optimal confidence parameter χ_t^*. Luckily there exists a vast literature for truncated multivariate normal computation. The mean can be calculated with high accuracy using identities from Kan and Robotti [2017], Manjunath and Wilhelm [2021]. Orthant probabilities can be computed efficiently using algorithms and integral approximations provided by Trinh and Genz [2015], Genz and Trinh [2016]. We calculate posterior medians via efficient monte-carlo methods for generating truncated Multivariate Normals [Wu and Gardner, 2024]. We make use of these algorithms through their implementations in the BoTorch library [Balandat et al., 2020]. We calculate $\chi_t^*(a, \mathbf{x}_t)$ efficiently using the bisection method/binary search that exploits the monotonicity of $P_t(a, b)$ w.r.t to b (see Lemma 9 in the Supplementary material). We note that the numerical and monte-carlo methods introduce approximation error. However, we find that these methods perform well in numerical experiments. We leave regret analysis on approximation error to future work.

[1]The condition on $k_f(\cdot, \cdot)$ is satisfied with a spline order 6 and kernels such as the common Gaussian kernel and the Matern kernel with $\nu > 2$.

Table 1: Final cumulative regret after $T = 500$ rounds under perturbations that violate concavity. Columns denote the perturbation scale $\rho = \mathrm{sd}(g_\rho)/\,\mathrm{sd}(f_0)$. Parentheses give ± 1 standard error. Lower is better.

CSGP-UCB with relaxed coefficient upper bound				
	Perturbation scale ρ			
Coefficient upper bound c	0	0.05	0.10	0.25
0	101 ($\pm$7)	124 ($\pm$8)	181 ($\pm$9)	539 ($\pm$42)
0.1	104 ($\pm$7)	120 ($\pm$8)	167 ($\pm$8)	479 ($\pm$40)
0.2	111 ($\pm$8)	115 ($\pm$8)	143 ($\pm$8)	424 ($\pm$37)

Unconstrained GP and spline baselines				
	Perturbation scale ρ			
Method	0	0.05	0.10	0.25
GP-Thompson	160 ($\pm$8)	169 ($\pm$8)	178 ($\pm$8)	209 ($\pm$9)
GP-UCB	268 ($\pm$10)	271 ($\pm$10)	283 ($\pm$10)	301 ($\pm$11)
SGP-UCB	387 ($\pm$21)	400 ($\pm$21)	422 ($\pm$21)	456 ($\pm$22)

14 ROBUSTNESS TO VIOLATIONS OF CONCAVITY

The experiments in the main text focus on settings where the expected reward is concave in the action for every context. We also examine the behavior of CSGP-UCB when this assumption is mildly violated. Let $f_0(a, \mathbf{x})$ denote one reward function generated from the simulation design in Section 5.2 with $d = 25$ and kernel length-scale $\theta = d/10$. We construct perturbed rewards of the form

$$f_\rho(a, \mathbf{x}) = f_0(a, \mathbf{x}) + g_\rho(a, \mathbf{x}),$$

where g_ρ is sampled from a zero-mean Gaussian process with a Gaussian kernel. The kernel for g_ρ is anisotropic, with context-space length-scale $\theta_x = 4$ and action-space length-scale $\theta_a = 0.75$. The smaller action-space length-scale induces substantial variation along the action dimension, so f_ρ need not be concave in a. We vary the prior standard deviation of g_ρ as a proportion

$$\rho = \frac{\mathrm{sd}(g_\rho)}{\mathrm{sd}(f_0)}$$

of the prior standard deviation of f_0 over the input region, with $\rho \in \{0, 0.05, 0.10, 0.25\}$.

We also evaluate a simple robustness modification to the CSGP posterior. Exact concavity corresponds to conditioning on

$$\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \le 0, \qquad j = 1, \ldots, J-2.$$

However, the posterior construction and concentration arguments only require an upper-truncated Gaussian posterior. Thus, when mild violations of concavity are plausible, we can replace the hard concavity constraint by the relaxed constraint

$$\boldsymbol{\beta}_j(\mathbf{x}_{t'}) \le c, \qquad j = 1, \ldots, J-2,$$

for some scalar coefficient upper bound $c \ge 0$. The choice $c = 0$ recovers the exact concavity constraint, while $c > 0$ permits bounded positive curvature coefficients in the C-spline representation. The CSGP-UCB algorithm is otherwise unchanged; only the truncation vector $\boldsymbol{\nu}$ is modified so that the entries corresponding to the curvature coefficients are set to c rather than 0.

Table 1 reports final cumulative regret after $T = 500$ bandit rounds. Parentheses denote ± 1 standard error. We omit DIP-UCB/Thompson from this table because the misspecification exacerbated its previously observed numerical stability issues, leading to substantially worse performance.

The results indicate that CSGP-UCB is robust to minor violations of the concavity condition. For perturbations up to $\rho = 0.10$, CSGP-UCB matches or outperforms the strongest benchmark, GP-Thompson. Relaxing the coefficient upper bound improves robustness under misspecification: for example, increasing c from 0 to 0.2 reduces regret from 181 to 143 when $\rho = 0.10$, and from 539 to 424 when $\rho = 0.25$. However, for perturbations as large as $\rho = 0.25$, CSGP-UCB performs poorly relative to GP-Thompson. This is expected, since such a large perturbation removes much of the concave structure of the reward function. In this regime, one should not expect a method designed to exploit concavity to be advantageous.

15 EXPERIMENTAL DETAILS

15.1 HYPER-PARAMETER SELECTION AND IMPLEMENTATION

For all algorithms we set the noise parameter to the truth across all experiments. For the NN-UCB algorithm we follow [Kassraie and Krause, 2022] as set the learning rate for SGD at $\eta = .01$. Note that the use of the SGD optimizer is crucial for the theoretical properties of the algorithm. For NN Thompson [Zhang et al., 2021] we do a grid search on $\nu \in \{1e-1, 1e-2, 1e-3, 1e-4, 1e-5\}$ and report the best performance. For both NN-UCB and NN-Thompson, we follow both papers and use a 1-Layer neural network. We set the number of nodes to the layer 512. For NN-Bootstrap as in [Riquelme et al., 2018, Zhou et al., 2020, Zhang et al., 2021], we simultaneously train 5 neural networks with a probability of data point inclusion $p = .8$. We train all neural networks for 500 iterations. For CSGP-UCB, use 1000 samples from the elliptical slice sampler of Wu and Gardner [2024] to sample monte-carlo samples to calculate medians.

For the Gaussian process based methods, we learn the kernel hyper- parameters by maximizing the marginal likelihood on the observed data [Rasmussen and Williams, 2006]. All GP based methods use Gaussian kernels. For neural network based methods, we train/update the appropriate network's parameters at every round. For all UCB algorithms, we follow Krause and Ong [2011], Srinivas et al. [2009] and choose the confidence parameter α_t to satisfy the bounds with probability $\delta = 0.1$. To compute χ_t^* we use a bisection algorithm with a tolerance of .025.

Due to computational considerations we also make simplifications in the implementation of the algorithms. Following Zhou et al. [2020], Zhang et al. [2021], Kassraie and Krause [2022] for the neural network based methods, we only calculate the posterior variance using the diagonal of the empirical tangent kernel matrix.

For DIP-UCB/Thompson we condition the second derivative to be negative on a uniform grid of 7 points for the current context $\mathbf{x}_t$ and the 5 nearest observed contexts. We experimented conditioning on a finer grid, but this led to poorer performance due to numerical stability issues.

For CSGP-UCB, we use a posterior that is only conditioned on the concavity information of the current round, $\mathcal{C}_t$ and the 5 nearest observed contexts to the current $\mathbf{x}_t$. This simplification guarantees that the posterior functions obey the concave constraint for the current context. Empirical results below demonstrate that this works well in practice. For all experiments, CSGP-UCB uses a cubic C-splines model with 5 uniform knots on $[0, 1]$. Cubic splines are the widely used default due to their empirically observed balance between function complexity and smoothness of the resulting curve [Hastie et al., 2009]. We also use 1000 samples to compute the posterior median. To compute χ_t^* we use a bisection algorithm with a tolerance of .025.

15.2 SYNTHETIC BENCHMARKS

We initialize all algorithms with $n = 25$ training points before starting contextual bandit optimization. We use the Rosenbrock and Dixon-price function as implemented in the BoTorch library [Balandat et al., 2020]. For the Trid function we follow the specification presented in Surjanovic and Bingham and we sample from the hypercube $[-d, d]^d$. For all the functions, we choose the action from the first variable and sample the contexts from the last $d - 1$ variables.

15.3 WARFARIN DOSING DETAILS

The primary biometric measured for patients on Warfarin is the International Normalized Ratio (INR), a measure of how quickly the blood clots. Within the safe Warfarin dosage regime, the INR has a positive, affine relationship with Warfarin dose Consortium [2009]. In addition patient outcomes are best when their post-dose INR matches the optimal INR given their demographics. Therefore a reasnoble reward function is the absolute difference between the post-dose INR with the patient's optimal INR. As the post-dose INR is an affine function of the dosage, the reward function is concave with respect to the action for every context.

We summarize the test function used by Chen et al. [2016] as follows:

$$
\begin{aligned}
&f(a, \mathbf{x}) = 8 + 4\cos(2\pi \mathbf{x}_2) - 2\mathbf{x}_3 - 8\mathbf{x}_5^3 - 15|g(\mathbf{x}) - 2a|^2 \\
&g(\mathbf{x}) = 0.6\mathbb{1}(-0.5 \leq \mathbf{x}_1 \leq 0.5) + 1.2\mathbb{1}(\mathbf{x}_1 > 0.5) \\
&+ 1.2\mathbb{1}(\mathbf{x}_1 < -0.5) + \mathbf{x}_4^2 + 0.5\log(|\mathbf{x}_7| + 1) - 0.6,
\end{aligned}
$$

where $\mathbf{x} \in [-1, 1]^7$, $a \in [0, 1]$. We define $\mathcal{A}$ as a uniform grid in $[0, 1]$ with $|\mathcal{A}| = 100$. We initialize the data with $n = 25$ data points with contexts and actions drawn uniformly from the domain. We sample the noisy rewards from $y = f(a, \mathbf{x}) + \epsilon$ where $\epsilon \sim N(0, 0.1)$ as in Chen et al. [2016].