

---

# Efficient learning of smooth probability functions from Bernoulli tests with guarantees

---

Paul Rolland<sup>1</sup> Ali Kavis<sup>1</sup> Alex Immer<sup>1</sup> Adish Singla<sup>2</sup> Volkan Cevher<sup>1</sup>

## Abstract

We study the fundamental problem of learning an unknown, smooth probability function via point-wise Bernoulli tests. We provide a scalable algorithm for efficiently solving this problem with rigorous guarantees. In particular, we prove the convergence rate of our posterior update rule to the true probability function in  $L_2$ -norm. Moreover, we allow the Bernoulli tests to depend on contextual features and provide a modified inference engine with provable guarantees for this novel setting. Numerical results show that the empirical convergence rates match the theory, and illustrate the superiority of our approach in handling contextual features over the state-of-the-art.

## 1. Introduction

One of the central challenges in machine learning relates to learning a continuous probability function from point-wise Bernoulli tests (Casella & Berger, 2002; Johnson & Wichern, 2002). Examples include, but are not limited to, clinical trials (DerSimonian & Laird, 1986), recommendation systems (McNee et al., 2003), sponsored search (Pandey & Olston, 2007), and binary classification. Due to the curse of dimensionality, we often require a large number of tests in order to obtain an accurate approximation of the target function. It is thus necessary to use a method that scalably constructs this approximation with the number of tests.

A widely used method for efficiently solving this problem is the Logistic Gaussian Process (LGP) algorithm (Tokdar & Ghosh, 2007). While this algorithm has no clear provable guarantees, it is shown to be very efficient in practice in approximating the target function. However, the time required for inferring the posterior distribution at some point

grows cubically with the number of tests, and can thus be inapplicable when the amount of data becomes large. There has been extensive work to resolve this cubic complexity associated with GP computations (Rasmussen, 2004). However, these methods require additional approximations on the posterior distribution, which impacts the efficiency and make the overall algorithm even more complicated, leading to further difficulties in establishing theoretical convergence guarantees.

Recently, Goetschalckx et al. (2011) tackled the issues encountered by LGP, and proposed a scalable inference engine based on Beta Processes called Continuous Correlated Beta Process (CCBP) for approximating the probability function. By scalable, we mean the algorithm complexity scales linearly with the number of tests. However, no theoretical analysis is provided, and the approximation error saturates as the number of tests becomes large (*cf.*, section 5.1). Hence, it is unclear whether provable convergence and scalability can be obtained simultaneously.

This paper bridges this gap by designing a simple and scalable method for efficiently approximating the probability functions with provable convergence. Our algorithm constructs a posterior distribution that allows inference in linear time (w.r.t. the number of tests) and converges in  $L_2$ -norm to the true probability function (uniformly over the feature space), see Theorem 1.

In addition, we also allow the Bernoulli tests to depend on contextual parameters influencing the success probabilities. To ensure convergence of the approximation, these features need to be taken into account in the inference engine. We thus provide the first algorithm that efficiently treats these contextual features while performing inference, and retain guarantees. As a motivation for this setting, we demonstrate how this algorithm can efficiently be used for treating bias in the data (Agarwal et al., 2018).

### 1.1. Basic model and the challenge

In its basic form, we seek to learn an unknown, smooth function  $\pi : \mathcal{X} \rightarrow [0, 1]$ ,  $\mathcal{X} \subset \mathbb{R}^d$  from point-wise Bernoulli tests, where  $d$  is the features space dimension. We model such tests as  $s_i \sim \text{Bernoulli}(\pi(x_i))$ , where  $\sim$  means *dis-*

---

<sup>1</sup>Ecole Polytechnique Fédérale de Lausanne, Switzerland <sup>2</sup>Max Planck Institute for Software Systems, Saarbrücken, Germany. Correspondence to: Paul Rolland <paul.rolland@epfl.ch>.

tributed as,  $x_i \in \mathcal{X}$ , and we model our knowledge of  $\pi$  at point  $x$  via a random variable  $\tilde{\pi}(x)$ .

Without additional assumptions, this problem is clearly hard, since experiments are performed only at points  $\{x_i\}_{i=1,\dots,t}$ , which constitute a negligible fraction of the space  $\mathcal{X}$ . In this paper, we will make the following assumption about the probability function:

**Assumption 1.** *The function  $\pi$  is  $L$ -Lipschitz continuous, i.e., there exists a constant  $L \in \mathbb{R}$  such that*

$$|\pi(x) - \pi(y)| \leq L\|x - y\| \quad (1)$$

$\forall x, y \in \mathcal{X}$  for some norm  $\|\cdot\|$  over  $\mathcal{X}$ .

In order to ensure convergence of  $\tilde{\pi}(x)$  to  $\pi(x)$  for all  $x \in \mathcal{X}$ , we must design a way of sharing experience among variables using this smoothness assumption.

Our work uses a prior for  $\pi$  based on the Beta distribution and designs a simple sharing scheme to provably ensure convergence of the posterior.

**Dynamic setting** In a more generic setting that we call “dynamic setting,” we assume that each Bernoulli test can be linearly influenced by some contextual features. Each experiment is then described by a quadruplet  $S_i = (x_i, s_i, A_i, B_i)$  and we study the following simple model for its probability of success:

$$Pr(s_i = 1) := A_i\pi(x_i) + B_i. \quad (2)$$

We have to restrict  $0 \leq B_i \leq 1$ ,  $0 \leq A_i + B_i \leq 1$  to ensure that this quantity remains a probability given that  $\pi(x_i)$  lies in  $[0, 1]$ . We assume that we have knowledge of estimates for  $A_i$  and  $B_i$  in expectation.

Such contextual features naturally arise in real applications (Krause & Ong, 2011). For example, in the case of clinical trials (DerSimonian & Laird, 1986), the goal is to learn the patient’s probability of succeeding at an exercise with a given difficulty. A possible contextual feature can then be the state of fatigue of the patient, which can influence its success probability. Here, LGP algorithm could be used, but the contextual feature must be added as an additional parameter. We show that, if we know how this feature influences the Bernoulli tests, then we can achieve faster convergence.

## 1.2. Our contributions

We summarize our contributions as follows:

1. We provide the first theoretical guarantees for the problem of learning a smooth probability function over a compact space using Beta Processes.

2. We provide an **efficient** and **scalable** algorithm that is able to handle contextual parameters explicitly influencing the probability function.
3. We demonstrate the efficiency of our method on synthetic data, and observe the benefit of treating contextual features in the inference. We also present a real-world application of our model.

**Roadmap** We analyze the simple setting without contextual features (referred to as *static* setting). We start by designing a Bayesian update rule for point-wise inference, and then include experience sharing in order to ensure  $L_2$  convergence over the whole space with a provable rate. We then treat the dynamic setting in the same way, and finally demonstrate our theoretical findings via extensive simulations and on a case-study of clinical trials for rehabilitation (*cf.*, Section 5).

## 2. Related Work

**Correlated inference via GPs** The idea of sharing experience of experiments between points with similar target function value is inspired by what was done with Gaussian Processes (GPs) (Williams & Rasmussen, 1996). GPs essentially define a prior over real-valued functions defined on a continuous space, and use a kernel function that represents how experiments performed at different points in the space are correlated.

GP-based models are not directly applicable to our problem setting given that our function  $\pi$  represents probabilities in the range  $[0, 1]$ . For our problem setting, a popular approach is Logistic Gaussian Processes (LGP) (Tokdar & Ghosh, 2007)—it learns an intermediate GP over the space  $\mathcal{X}$  which is then squashed to the range  $[0, 1]$  via a logistic transformation. Experience sharing is then done by modeling the covariance between tests performed at different points through a predefined kernel. This allows constructing a covariance matrix between test points, which can be used to estimate the posterior distribution at any other sample point.

Gaussian Copula Processes (GCP) (Wilson & Ghahramani, 2010) is another GP-based approach that learns a GP and uses a copula to map it to Beta distributions over the space.

More recently, Ranganath and Blei (Ranganath & Blei, 2017) explored correlated random measures including correlated Beta-Bernoulli extension. However, GPs are still used in order to define these correlations.

There are at least two key limitations with these “indirect” approaches: First, the posterior distributions after observing a Bernoulli outcome is analytically intractable, and needs to be approximated, e.g. using Laplace approximation (Tokdar & Ghosh, 2007). Second, the time complexity of prediction

grows cubically  $\mathcal{O}(t^3)$  with respect to the number of samples  $t$ . There is extensive work to resolve this cubic complexity associated to GP computations (Rasmussen, 2004). However, these methods require additional approximations on the posterior distribution, which impacts the efficiency, and make the overall algorithm even more complicated, leading to further difficulties in establishing theoretical guarantees.

Methods based on GPs that take context variables into account have also been designed (Krause & Ong, 2011). However, they simply allow for the use of specific kernels for these variables and still require an increase in the feature space dimension. In this work, by directly modifying the inference process, we compute a posterior that takes into account contextual features without increasing the feature space dimension.

**Correlated Beta Processes** In contrast to GPs, it is very challenging to define correlated Beta distributions. The first work introducing a Beta process without using GPs is the one of Hjort (1990) (Hjort, 1990), but lacks the correlation aspect. Some other works studied multi-variate Beta distributions for simple settings considering only a few variables (Gupta & Wong, 1985; Olkin & Liu, 2003).

Goetschalckx et al. (2011) (Goetschalckx et al., 2011) proposed an approach named Continuous Correlated Beta Processes (CCBP) to deal with a continuous space of Beta distributions and to share experience between them via using a kernel. CCBP is shown to achieve results comparable to the state of the art approach based on LGP. Furthermore, it is shown that CCBP is much more time efficient—linear  $\mathcal{O}(t)$  runtime for CCBP in comparison to cubic  $\mathcal{O}(t^3)$  runtime of GP-based methods.

CCBP approach has been used in several real-world application settings, e.g., for learning patient’s fitness function in rehabilitation (Goetschalckx et al., 2011), learning the wandering behavior of people with dementia (Hoey et al., 2012), and in the application of analyzing genetic ancestry in admixed populations (Gompert, 2016).

However, the method presented in (Goetschalckx et al., 2011), by simply using a heuristic kernel, gives an approximation which does not converge to the target function as the number of samples increases. In order for the method to converge, this kernel must depend on the number of samples. In this paper, we provide an explicit kernel to use which ensures convergence of the approximated probability function to the target function with provable rate.

### 3. Inference for the static setting

We start by analyzing the static setting, in which no contextual features influence the Bernoulli tests. We first design a Bayesian update rule for point-wise inference, then we pro-

pose an experience sharing method and prove convergence guarantees.

#### 3.1. Uncorrelated case: a Bayesian approach

Suppose we do not use the smoothness assumption of  $\pi$ . Then a naive solution is to model each random variable  $\tilde{\pi}(x)$  by the conjugate prior of the Bernoulli distribution, which is the Beta distribution. Then, starting from a prior  $\tilde{\pi}(x) \sim \text{Beta}(\alpha(x), \beta(x)) \forall x \in \mathcal{X}$ , the Bayesian posterior  $\tilde{\pi}(x|\mathcal{S})$  conditioned on  $\mathcal{S} = \{(x_i, s_i)\}_{i=1, \dots, t}$  is:

$$\tilde{\pi}(x|\mathcal{S}) \sim \text{Beta} \left( \alpha(x) + \sum_{i=1}^t \delta_{s_i=1} \delta_{x_i=x}, \beta(x) + \sum_{i=1}^t \delta_{s_i=0} \delta_{x_i=x} \right)$$

where  $\delta_a$  is the Kronecker delta.

This particular update scheme does not take smoothness assumption of the function into account and any experiment  $S_i = (x_i, s_i)$  only influences the corresponding random variable  $\tilde{\pi}(x_i)$ . In particular, if no experiment is performed at  $x$ , then our belief of  $\pi(x)$  remains unchanged. It is thus necessary to make use of the smoothness assumption.

#### 3.2. Leveraging smoothness of $\pi$ via experience sharing

Goetschalckx et al. (2011) propose a mechanism of experience sharing among correlated variables. To this purpose, they introduce a kernel  $K : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$  where  $K(x, x_i)$  indicates to what extent the experience for experiment at  $x_i$  should be shared with any other point  $x$ . Indeed, thanks to the Lipschitz continuity assumption (1), we expect close points to have similar probabilities. However, although the Beta distribution is the conjugate prior of the Bernoulli distribution, this conjugacy does not hold anymore when we use experience sharing. Instead of using the Bayesian posterior, we use the following update rule:

$$\tilde{\pi}(x|\mathcal{S}) \sim \text{Beta} \left( \alpha(x) + \sum_{i=1}^t \delta_{s_i=1} K(x, x_i), \beta(x) + \sum_{i=1}^t \delta_{s_i=0} K(x, x_i) \right). \quad (3)$$

With this update rule, the result of experiment  $S_i$  influences all variables  $\tilde{\pi}(x)$  for which  $K(x, x_i) > 0$ , and the magnitude of influence is proportional to  $K(x, x_i)$ . Note that this update rule is no more Bayesian. However, all existing methods, including LGP and GCP, also involve non-Bayesian updates.

In Goetschalckx et al. (2011), authors do not specify any particular choice of kernel function and the selection process is left as a heuristic. We show that proper selection of

**Algorithm 1** Smooth Beta Process (SBP)

**Input:** experiments points and observations  $\mathcal{S} = \{x_i, s_i\}_{i=1,\dots,t}$ , query point  $x \in \mathcal{X}$ , prior knowledge  $\tilde{\pi}(x) \sim \mathcal{B}(\alpha(x), \beta(x))$   
**Output:** Posterior distribution  $\tilde{\pi}(x|\mathcal{S})$   
 1. Set  $\Delta \propto \frac{1}{t^{\frac{1}{d+2}}}$   
 2. Compute the posterior as in (3) using kernel  $K(x, x') = \delta_{\|x-x'\| \leq \Delta}$

kernel is essential for convergence to the true underlying distributions at all points. In particular, to ensure convergence in  $L_2$  norm, this kernel must shrink as the number of observations increases (Algorithm 1). We can see that our algorithm, called Smooth Beta Process (SBP), allows for fast inference at any point  $x \in \mathcal{X}$ , since it simply requires to find the tests that are performed at most  $\Delta$  far from  $x$ , and compute the posterior distribution as in (3) depending on the number of successes and failures within these tests.

**Remark 1.** *The particular dependence of the kernel on the number of samples ensuring optimal convergence is not trivial, and can only be found via a theoretical analysis of the model.*

**Theorem 1.** *Let  $\pi : [0, 1]^d \rightarrow [0, 1]$  be  $L$ -Lipschitz continuous. Suppose we measure the results of experiments  $\mathcal{S} = \{(x_i, s_i)\}_{i=1,\dots,t}$  where  $s_i$  is a sample from a Bernoulli distribution with parameter  $\pi(x_i)$ . Experiment points  $\{x_i\}_{i=1,\dots,t}$  are assumed to be i.i.d. and uniformly distributed over the space. Then, starting with a uniform prior  $\alpha(x) = \beta(x) = 1 \forall x \in [0, 1]^d$ , the posterior  $\tilde{\pi}(x|\mathcal{S})$  obtained from SBP uniformly converges in  $L_2$ -norm to  $\pi(x)$ , i.e.*

$$\sup_{x \in [0,1]^d} \mathbb{E}_{\mathcal{S}} (\mathbb{E} ((\tilde{\pi}(x|\mathcal{S}) - \pi(x))^2)) = \mathcal{O} \left( L^{\frac{2d}{d+2}} t^{-\frac{2}{d+2}} \right), \quad (4)$$

where the outer expectation is performed over experiment points  $\{x_i\}_{i=1,\dots,t}$  and their results  $\{s_i\}_{i=1,\dots,t}$ . SBP also computes point-wise posterior in time  $\mathcal{O}(t)$ .

**Remark 2.** *This theorem provides an upper bound for the  $L_2$  norm over any point of the space, and takes into account where the experiments are performed in the feature space. If all experiments are performed at the same point, then we recover the familiar square-root rate at that point (Ghosal, 1997), but we would not converge at points that are far away.*

**Remark 3.** *The constraint on the input space being  $[0, 1]^d$  can easily be extended to any compact space  $\mathcal{X} \subset \mathbb{R}^d$ . This would simply modify the convergence rate by a factor equal to the volume of  $\mathcal{X}$ .*

**Remark 4.** *The dependence of the convergence rate on the feature space dimension is due to the curse of dimensionality, and the fact that we provide convergence uniformly*

*over the whole space. However, this is not due to the particular algorithm we use, and we empirically show that LGP suffers the same dependence on the space dimension (see Section 5). Similar issues are also prevalent in GP optimization despite which great applications success has been obtained (Shahriari et al., 2016).*

In appendix ??, we show how this algorithm naturally applies to binary classification. Restricted to classification, our algorithm becomes similar to the fixed-radius nearest neighbour algorithm (Chen et al., 2018), but the current framework allows for error quantification and precise prior injection.

## 4. Inference for the dynamic setting

We analyze the dynamic setting where Bernoulli tests are influenced by contextual features as in (2).

### 4.1. Uncorrelated case: a Bayesian approach

As previously, we start by analyzing the uncorrelated case, i.e., how to update the distribution of  $\tilde{\pi}(x)$  conditioned on the outputs of experiments  $\mathcal{S} = \{(s_i, x_i, A_i, B_i)\}_{i=1,\dots,t}$  all performed at  $x$ .

Since experiments are not samples from Bernoulli variables with parameter  $\pi(x)$ , Bayesian update is not straightforward but can be achieved using sums of Beta distributions, as shown in Theorem 2.

**Theorem 2.** *Suppose  $\tilde{\pi}(x) \sim \mathcal{B}(\alpha, \beta)$  and we observe the result of a sample  $s \sim \text{Bernoulli}(A\pi(x) + B)$ . Then the Bayesian posterior for  $\tilde{\pi}(x)$  conditioned on this observation is given by*

$$\tilde{\pi}(x|s) \sim C_0 \mathcal{B}(\alpha + 1, \beta) + C_1 \mathcal{B}(\alpha, \beta + 1), \quad (5)$$

where in the case of success ( $s = 1$ ), we have

$$C_0 = \frac{B\beta}{B\beta + (A+B)\alpha}, \quad C_1 = \frac{(A+B)\alpha}{B\beta + (A+B)\alpha}$$

and in the case of failure ( $s = 0$ ), we have

$$C_0 = \frac{(1-B)\beta}{(1-B)\beta + (1-A-B)\alpha},$$

$$C_1 = \frac{(1-A-B)\alpha}{(1-B)\beta + (1-A-B)\alpha}.$$

In (5), we mean that the density function of the posterior random variable  $\tilde{\pi}(x|s)$  is the weighted sum of the two density functions given on the right-hand side.

Then, by using this result recursively on a set of experiments  $\mathcal{S} = \{(s_i, x_i, A_i, B_i)\}_{i=1,\dots,t}$ , we can obtain a general update rule.

**Corollary 1.** Suppose  $\tilde{\pi}(x) \sim \mathcal{B}(\alpha, \beta)$  and we observe the outputs of experiments  $\mathcal{S} = \{(s_i, x, A_i, B_i)\}_{i=1, \dots, t}$  where  $s_i$ 's are sampled from Bernoulli random variables, each with parameter  $A_i\pi(x) + B_i$ . Then the Bayesian posterior  $\tilde{\pi}(x|\mathcal{S})$  is given by

$$\tilde{\pi}(x|\mathcal{S}) \sim \sum_{i=0}^t C_i^t \mathcal{B}(\alpha + i, \beta + t - i) \quad (6)$$

where  $C_i^t$ 's can be computed via an iterative procedure starting from  $C_0^0 = 1$  and  $\forall n = 0, \dots, t$ :

$$C_i^{n+1} = \frac{1}{E_s^n} (B_i C_i^n (\beta + n - i) + (A_i + B_i) C_{i-1}^n (\alpha + i - 1))$$

if  $s_n = 1$ ; and

$$C_i^{n+1} = \frac{1}{E_f^n} ((1 - B_i) C_i^n (\beta + n - i) + (1 - A_i - B_i) C_{i-1}^n (\alpha + i - 1))$$

if  $s_n = 0$ .  $E_s^n$  and  $E_f^n$  are normalization factors that ensure  $\sum_{i=0}^n C_i^n = 1 \forall n$ . For simplicity of notation, we use  $C_{-1}^n = C_{n+1}^n = 0 \forall n$ .

This gives us a way of updating, in a fully Bayesian manner, the distribution of  $\tilde{\pi}(x)$  conditioned on observations of experiments performed at  $x$ . It involves a linear combination of Beta distributions, with coefficients depending on successes and contextual features.

**Certainty invariance assumption** For simplicity, we will assume that if an event is certain (i.e., occurs with probability 1), then context variables cannot lower this probability (i.e.,  $\forall x_i$  if  $\pi(x_i)$ , then  $A_i\pi(x_i) + B_i = 1$ ). This is trivially equivalent to the constraint  $A_i + B_i = 1 \forall i$ . If, on the contrary, there is an impossibility invariance, i.e., context variables cannot make an impossible event possible, we can make a change of variable  $f \leftrightarrow 1 - f$  (i.e., invert the meanings of ‘‘success’’ and ‘‘failure’’) in order to satisfy the certainty invariance assumption.

In Corollary 1, we see that the time complexity required for computing  $C_i^t \forall i = 0, \dots, t$  is  $\mathcal{O}(t^2)$ . However, in the particular case where  $A_i = A$ ,  $B_i = B \forall i$  and  $A + B = 1$ , the update rule of Corollary 1 becomes much simpler, i.e., computable in time  $\mathcal{O}(t)$ :

**Corollary 2.** Suppose  $\tilde{\pi}(x) \sim \mathcal{B}(\alpha, \beta)$  and we observe the outputs of experiments  $\mathcal{S} = \{(s_i, x, 1 - B, B)\}_{i=1, \dots, t}$  where  $s_i \sim \text{Bernoulli}((1 - B)\pi(x) + B)$ . Then the Bayesian posterior  $\tilde{\pi}(x|\mathcal{S})$  conditioned on these observations is given by

$$\tilde{\pi}(x|\mathcal{S}) \sim \sum_{i=0}^S C_i^t \mathcal{B}(\alpha + i, \beta + t - i) \quad (7)$$

**Algorithm 2** Inference engine for the simplified dynamic setting: Constant  $A, B$

**Input:** experiments descriptions  $\mathcal{S} = \{(x_i, s_i, A, B)\}_{i=1, \dots, t}$ , point of interest  $x \in \mathcal{X}$ , prior knowledge  $\tilde{\pi}(x) \sim \mathcal{B}(\alpha(x), \beta(x))$

**Output:** Posterior distribution  $\tilde{\pi}(x|\mathcal{S})$

1. Set  $\Delta \propto \frac{1}{t^{\frac{1}{d+2}}}$
2. Build the set of neighboring experiments  $\mathcal{S}_x = \{(x_i, s_i, A, B) : \|x - x_i\| \leq \Delta\}$
3. Compute the posterior as in Corollary 1 (or 2 if  $A + B = 1$ ) using the results of experiments  $\mathcal{S}_x$  as if performed at  $x$ .

where  $S = \sum_{i=1}^t s_i$  is the total number of successes and

$$C_i^t \propto \binom{S}{i} (\alpha - 1 + i)! (\beta + t - 1 - i)! B^{S-i} \quad (8)$$

$\forall i = 0, \dots, S$ . We can compute all  $C_i^t$ 's in time  $\mathcal{O}(t)$  via the relation  $C_{i+1}^t = \frac{(S-i)(\alpha+i)}{B(i+1)(\beta+t-1-i)} C_i^t$ .

## 4.2. Leveraging smoothness of $\pi$ via experience sharing: Simplified setting

We now introduce the use of correlations between samples in the update rule, in a similar way as done in the static setting. We first analyze a simplified setting where the contextual parameters are constant among all experiments, i.e.  $A_i = A$ ,  $B_i = B \forall i$ .

Due to the more complex form of the Bayesian update in the uncorrelated case, it turns out that the previous technique is not straightforward to apply. One way to introduce this idea of experience sharing would be to modify the Bayesian update rule of Theorem 2 as:

$$\tilde{\pi}(x|s_i) \sim C_0 \mathcal{B}(\alpha + K(x, x_i), \beta) + C_1 \mathcal{B}(\alpha, \beta + K(x, x_i))$$

where  $K$  is the same kernel as defined previously.

However, if  $K(x, x')$  is real valued and that we apply such an update for each experiment, then it turns out that the number of terms required for describing the posterior distribution grows exponentially with the number of observations. In order to ensure the tractability of the posterior, we can restrict the kernel to values in  $\{0, 1\}$ , e.g.  $K(x, x') = \delta_{\|x-x'\| \leq \Delta}$  for some kernel width  $\Delta$ . This means that, each time we make an observation at  $x_i$ , all random variables  $\tilde{\pi}(x)$  with  $\|x - x_i\| \leq \Delta$  are updated as if the same experiments had been performed at  $x$  (Algorithm 2).

We provide guarantees for convergence of the posterior distribution generated by Algorithm 2 under the certainty invariance assumption (Theorem 3). This constraint on  $A$  and  $B$  is equivalent to saying that the contextual parameters necessarily increases the success probability.

**Algorithm 3** Contextual Smooth Beta Process (CSBP)

**Input:** experiments descriptions  $\mathcal{S} = \{(x_i, s_i, A_i, B_i)\}_{i=1,\dots,t}$ , point of interest  $x \in \mathcal{X}$ , prior knowledge  $\tilde{\pi}(x) \sim \mathcal{B}(\alpha(x), \beta(x))$   
**Output:** Posterior distribution  $\tilde{\pi}(x|\mathcal{S})$   
 1. Set  $\Delta \propto \frac{1}{t^{\frac{1}{d+2}}}$   
 2. Build the set of neighboring experiments  $\mathcal{S}_x = \{(x_i, s_i, A_i, B_i) : \|x - x_i\| \leq \Delta\}$   
 3. Compute means  $B = \frac{1}{|\mathcal{S}_x|} \sum_{i: \|x-x_i\| \leq \Delta} B_i$ ,  $A = \frac{1}{|\mathcal{S}_x|} \sum_{i: \|x-x_i\| \leq \Delta} A_i$   
 4. Compute the posterior as in Corollary 1 (or 2 if  $A + B = 1$ ) using the results of experiments  $\mathcal{S}_x$  as if performed at  $x$ , and with constant parameters  $A, B$ .

**Theorem 3.** Let  $\pi : [0, 1]^d \rightarrow ]0, 1]$  be  $L$ -Lipschitz continuous. Suppose we observe the results of experiments  $\mathcal{S} = \{(x_i, s_i, 1 - B, B)\}_{i=1,\dots,t}$  where  $s_i \sim \text{Bernoulli}((1 - B)\pi(x) + B)$ . Experiment points  $\{x_i\}_{i=1,\dots,t}$  are assumed to be i.i.d. uniformly distributed over the space. Then, starting with a uniform prior  $\alpha(x) = \beta(x) = 1 \forall x \in [0, 1]^d$ , the posterior  $\tilde{\pi}(x|\mathcal{S})$  obtained from Algorithm 2 uniformly converges in  $L_2$ -norm to  $\pi(x)$ , i.e.,

$$\sup_{x \in [0, 1]^d} \mathbb{E}_{\mathcal{S}} (\mathbb{E} ((\tilde{\pi}(x|\mathcal{S}) - \pi(x))^2)) = \mathcal{O} \left( L^{\frac{2d}{d+2}} ((1 - B)t)^{-\frac{2}{d+2}} \right).$$

Moreover, Algorithm 2 computes the point-wise posterior in time  $\mathcal{O}(t)$ .

**Remark 5.** We observe that adding the contextual parameter  $B$  does not modify the convergence rate compared to the static case. By using other algorithms such as LGP, parameter  $B$  should be added to the feature space, increasing its dimension by 1, which impacts the convergence rate as we demonstrate in the sequel.

### 4.3. Leveraging smoothness of $\pi$ via experience sharing: General setting

We finally analyze the general setting where contextual parameters are noisy and may vary among the experiments. Instead of using each  $A_i, B_i$  in the update rule of the posterior, we can perform this update as if all experiments were performed with the same coefficients  $A$  and  $B$ , which are the means of the coefficients  $A_i$  and  $B_i$  respectively. This approximation simplifies the analysis and does not influence the error rate.

Algorithm 3, called Contextual Smooth Beta Process (CSBP), is general and can be applied to any contextual parameters  $A_i, B_i$  with no constraints. We provide guarantees for convergence of the posterior distribution generated

by CSBP under the certainty invariance assumption (Theorem 4).

**Theorem 4.** Let  $\pi : [0, 1]^d \rightarrow ]0, 1]$  be  $L$ -Lipschitz continuous. Suppose we observe the results of experiments  $\mathcal{S} = \{(x_i, s_i, 1 - B_i, B_i)\}_{i=1,\dots,t}$  where  $s_i \sim \text{Bernoulli}((1 - B_i + \epsilon_i)\pi(x_i) + B_i + \epsilon_i)$ , i.e., contextual features are noisy. We assume  $\epsilon_i$ 's are independent random variables with zero mean and variance  $\sigma^2$ . The points  $\{x_i\}_{i=1,\dots,t}$  are assumed to be i.i.d. uniformly distributed over the space. Then, starting with a uniform prior  $\alpha(x) = \beta(x) = 1 \forall x \in [0, 1]^d$ , the posterior  $\tilde{\pi}(x|\mathcal{S})$  obtained from Algorithm 3 uniformly converges in  $L_2$ -norm to  $\pi(x)$ , i.e.,

$$\sup_{x \in [0, 1]^d} \mathbb{E}_{\mathcal{S}} (\mathbb{E} ((\tilde{\pi}(x|\mathcal{S}) - \pi(x))^2)) = \mathcal{O} \left( c(B, \sigma^2) L^{\frac{2d}{d+2}} t^{-\frac{2}{d+2}} \right),$$

where  $c(B, \sigma^2)$  is a constant depending on  $\{B_i\}_{i=1,\dots,t}$  and the noise  $\sigma^2$ . Moreover, CSBP computes the posterior in time  $\mathcal{O}(t)$ .

## 5. Numerical experiments

We devise a set of experiments to demonstrate the capabilities of our inference engine and validate the theoretical bounds for static and dynamic settings. We start with synthetic experiments in 1D and 2D, and finally reproduce the case study used in (Goetschalckx et al., 2011) to show the efficiency of our dynamic algorithm.

### 5.1. Synthetic examples

We construct a function  $\pi : \mathcal{X} \rightarrow [0, 1]$ , uniformly select points  $\{x_i\}_{i=1,\dots,t}$ , and sample  $s_i \sim \text{Bernoulli}(\pi(x_i))$ ,  $i = 1, \dots, t$ . From these data, SBP constructs the posterior distributions  $\tilde{\pi}(x|\mathcal{S}) \forall x \in \mathcal{X}$ . This experiment is performed both in 1D setting using a feature space  $\mathcal{X} = [0, 1]$ , and in 2D with  $\mathcal{X} = [0, 1]^2$ . We also apply LGP and CCBP (with fixed square exponential kernel) to this problem for comparison. Explicit forms of the chosen functions are presented in the Appendix.

For the dynamic setting, contextual parameters  $\{B_i\}_{i=1,\dots,t}$  are sampled independently and uniformly from  $[0, 1]$ , and the tests are then performed by sampling  $s_i \sim \text{Bernoulli}((1 - B_i)\pi(x_i) + B_i)$ ,  $i = 1, \dots, t$ . The posterior is constructed using CSBP. We also applied LGP to this dynamic setting by including the parameter  $B$  as an additional feature. In order to evaluate  $\pi$ , LGP returns the approximated distribution associated with  $B = 0$ .

For the static setting (1D and 2D), Figures 1 (left) and 2 (left) show the  $L_2$  errors of the posterior distributions averaged across all  $x \in \mathcal{X}$ , and over 20 runs, as functions of the number of samples  $t$ . We can observe the convergence upper

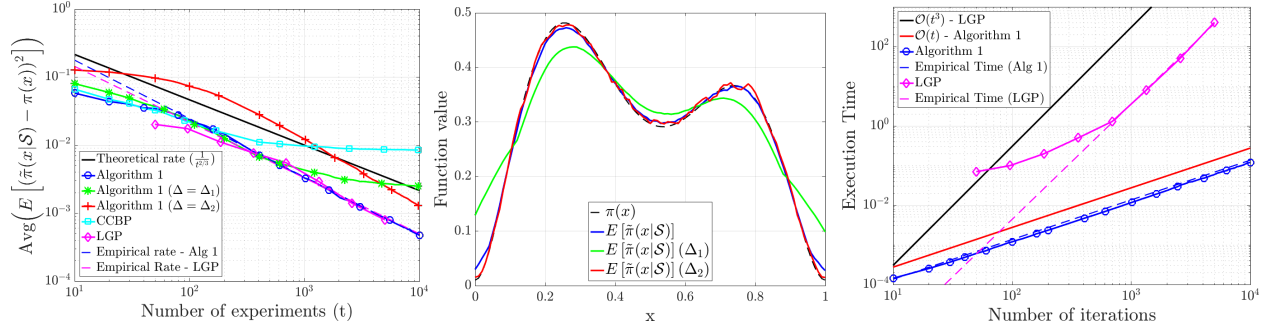


Figure 1. Left:  $L_2$  error for 1D static setting. Middle: Mean posterior estimates  $\mathbb{E}[\tilde{\pi}(x|S)]$  generated by Algorithm 1 for different kernel widths. Right: Running time for Algorithm 1 and LGP

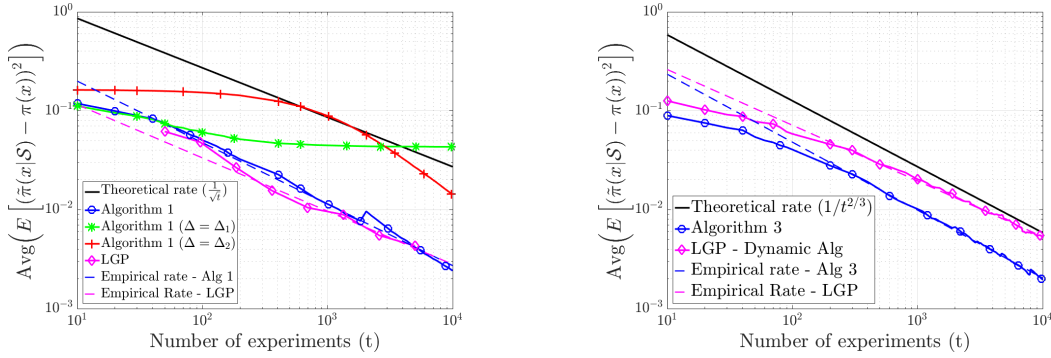


Figure 2.  $L_2$  error of posterior  $\mathbb{E}[\tilde{\pi}(x|S)]$  for 2D static and 1D dynamic settings, averaged over all points  $x \in \mathcal{X}$  versus number of samples.

bounds  $\mathcal{O}(1/t^{2/3})$  in 1D, and  $\mathcal{O}(1/\sqrt{t})$  in 2D as predicted by Theorem 1.

We observe that LGP and our method perform similarly, as pointed out in (Goetschalckx et al., 2011). However, running LGP takes significantly more time than our method since its time complexity is  $\mathcal{O}(t^3)$  compared to  $\mathcal{O}(t)$  for our algorithm, as demonstrated numerically in Figure 1 (right). We observe that CCBP saturates after some time since the kernel is independent of the number of samples.

Additionally, Figure 1 (left) demonstrates two sets of error curves for variations of Algorithm 1. To argue about optimality of kernel width specification, we run SBP with fixed kernel widths  $\Delta_1 = 50^{-\frac{1}{d+2}}$  and  $\Delta_2 = 500000^{-\frac{1}{d+2}}$ . When  $\Delta \ll t^{-\frac{1}{d+2}}$ , the  $L_2$  error initially decays at a slow rate and error remains larger than the optimal setting (green curves). On the contrary if we fix the kernel width  $\Delta \gg t^{-\frac{1}{d+2}}$ , the error saturates at early iterations (blue curves).

Figure 1 (middle) shows how the built posterior distribution approximates the true synthetic probability function by plotting the posterior mean over the space  $\mathcal{X}$  for the different kernel widths. We observe that using a wide kernel ( $\Delta_1$ ) leads to a posterior which is too smooth, due to experience oversharing. On the other hand, using a narrow kernel ( $\Delta_2$ )

leads to a highly non-smooth posterior, due to insufficient sharing.

Figure 2 (right) similarly shows the  $L_2$  error of the posterior distributions for the dynamic setting, also averaged over all  $x \in \mathcal{X}$ , and over 20 runs. We again observe the convergence upper bounds  $\mathcal{O}(1/\sqrt{t})$  in 2D as predicted by Theorem 4. We observe that our algorithm performs much better than LGP since it applies on a lower dimensional space.

## 5.2. Application to biased data: a case study from Goetschalckx et al. (2011)

Handling biased data is currently one of the major problems in machine learning. In this section, we investigate how CSBP can treat bias by the mean of contextual features.

In line with Goetschalckx et al. (2011), we conduct a case study with synthetic stroke rehabilitation data. The goal is to determine the probability that a patient succeeds in an exercise based on its difficulty. However, the patient can in some cases be fatigued, which influences the success probability and thus introduces a bias in the experiments.

Let  $f(x)$  denote the success probability function for exercise with difficulty  $x \in [0, 1]$ , when the patient is not fatigued. We assume that the patient has a certain level of fatigue

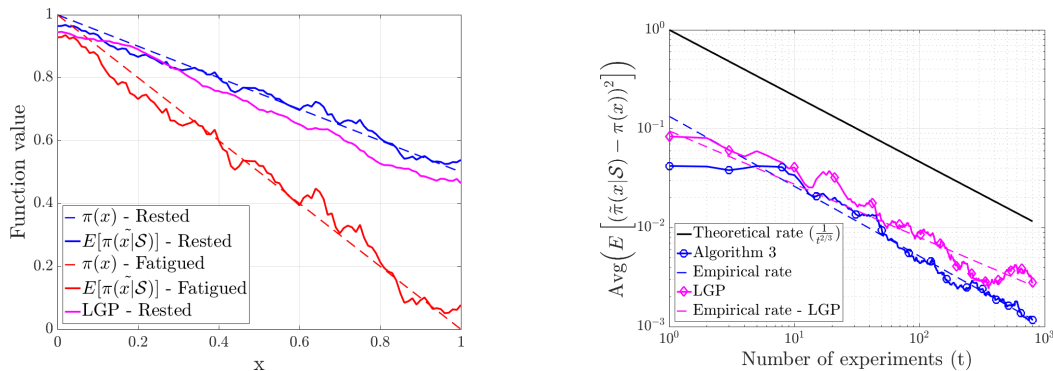


Figure 3. Left: Mean posteriors  $\mathbb{E}[\tilde{\pi}(x|\mathcal{S})]$  for target functions representing rested (blue) and highest fatigue (red) states. Right:  $L_2$  error for rested state averaged over all points versus sample size  $t$ .

$\alpha_f \in [0.5, 1]$ , in which case his success probability function becomes  $\alpha_f f$ .

Note that the impossibility invariance assumption holds in this case since being fatigue cannot make possible a task which was already impossible. As mentioned previously, we can then simply make a change of variable in order to satisfy the certainty invariance assumption, and safely apply the dynamic algorithm.

Alternatively, by treating the level of fatigue as a new dimension in the feature space, LGP can be applied to the rehabilitation case study.

We assume that the difficulty of the exercise influences (in an unknown way) the success probability as  $f(x) = 1 - x$ ,  $x \in [0, 1]$ . We construct a synthetic dataset by uniformly sampling exercise difficulties  $x$  and fatigue levels  $\alpha_f$ , and then sampling the success from  $f_{\alpha_f}(x)$ .

Figure 3 shows the reconstructed success probability distributions when the patient is either not fatigued (rested state) or in the final fatigued state ( $\alpha_f = 0.5$ ), as well as the  $L_2$  error of the posterior for the rested state. Since LGP operates on a higher dimensional space, we observe that the  $L_2$  error decays slower and the approximation of the target function for the rested state is worse than CSBP.

## 6. Conclusions

In this paper, we build an inference engine for learning smooth probability functions from a set of Bernoulli experiments, which may be influenced by contextual features. We design an efficient and scalable algorithm for computing a posterior converging to the target function with provable rate, and demonstrate its efficiency on synthetic and real-world problems. These characteristics together with the simplicity of SBP make it a competitive tool compared to LGP, which has been shown to be an important tool in many real-world applications. We thus expect practitioners to

apply this method to such problems.

**Discussion and future work** The current analysis can only model a particular type of contextual influence, which modifies the success probability as  $A_i \pi(x_i) + B_i$ . It turns out that Theorem 2 can be generalized to any polynomial transformation of the success probability (i.e.  $\sum_{j=0}^p a_i^{(j)} \pi(x_i)^j$ ), allowing for a wider class of contextual influences.

Moreover, the theoretical framework we provide seems to be applicable to a large class of problems, such as risk tracking, Bandit setting, active learning, etc. Extending this model to such applications would also be an interesting research direction.

## 7. Acknowledgement

This work was supported by the Swiss National Science Foundation (SNSF) under grant number 407540\_167319.

## References

- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.
- Audibert, J.-Y., Tsybakov, A. B., et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2): 608–633, 2007.
- Casella, G. and Berger, R. L. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.
- Chen, G. H., Shah, D., et al. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018.
- DerSimonian, R. and Laird, N. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.



- Ghosal, S. A review of consistency and convergence of posterior distribution. In *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1997.
- Goetschalckx, R., Poupart, P., and Hoey, J. Continuous correlated beta processes. In *IJCAI*, 2011.
- Gompert, Z. A continuous correlated beta process model for genetic ancestry in admixed populations. *PLoS one*, 11(3):e0151047, 2016.
- Gupta, A. K. and Wong, C. On three and five parameter bivariate beta distributions. *Metrika*, 32(1):85–91, 1985.
- Hjort, N. L. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pp. 1259–1294, 1990.
- Hoey, J., Yang, X., Grzes, M., Navarro, R., and Favela, J. Modeling and learning for lacasa, the location and context-aware safety assistant. In *NIPS 2012 Workshop on Machine Learning Approaches to Mobile Context Lake Tahoe, NV*, 2012.
- Johnson, R. A. and Wichern, D. *Multivariate analysis*. Wiley Online Library, 2002.
- Knapik, B. T., van der Vaart, A. W., van Zanten, J. H., et al. Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011.
- Krause, A. and Ong, C. S. Contextual gaussian process bandit optimization. In *NIPS*, pp. 2447–2455, 2011.
- Krichevsky, R. and Trofimov, V. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.
- McNee, S. M., Lam, S. K., Konstan, J. A., and Riedl, J. Interfaces for eliciting new user preferences in recommender systems. In *International Conference on User Modeling*, pp. 178–187. Springer, 2003.
- Olkin, I. and Liu, R. A bivariate beta distribution. *IEEE Transactions on Information Theory*, 62(4):407–412, 2003.
- Pandey, S. and Olston, C. Handling advertisements of unknown quality in search advertising. In *NIPS*, pp. 1065–1072, 2007.
- Ranganath, R. and Blei, D. M. Correlated random measures. *Journal of the American Statistical Association*, pp. 1–14, 2017.
- Rasmussen, C. E. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pp. 63–71. Springer, 2004.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- Tokdar, S. T. and Ghosh, J. K. Posterior consistency of logistic gaussian process priors in density estimation. *Journal of statistical planning and inference*, 137(1):34–42, 2007.
- van der Vaart, A. W., van Zanten, J. H., et al. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.
- Williams, C. K. and Rasmussen, C. E. Gaussian processes for regression. In *NIPS*, pp. 514–520, 1996.
- Wilson, A. G. and Ghahramani, Z. Copula processes. In *NIPS*, pp. 2460–2468, 2010.