# Continuous Correlated Beta Processes

**Robby Goetschalckx**
**University of Dundee**
**Dundee, UK**
**robby.goetschalckx@gmail.com**

**Pascal Poupart, Jesse Hoey**
**University of Waterloo**
**Waterloo, Canada**
**[ ppoupart ; jhoey] @cs.uwaterloo.ca**

## Abstract

In this paper we consider a (possibly continuous) space of Bernoulli experiments. We assume that the Bernoulli distributions are correlated. All evidence data comes in the form of successful or failed experiments at different points. Current state-of-the-art methods for expressing a distribution over a continuum of Bernoulli distributions use logistic Gaussian processes or Gaussian copula processes. However, both of these require computationally expensive matrix operations (cubic in the general case). We introduce a more intuitive approach, directly correlating beta distributions by sharing evidence between them according to a kernel function, an approach which has linear time complexity. The approach can easily be extended to multiple outcomes, giving a continuous correlated Dirichlet process, and can be used for both classification and learning the actual probabilities of the Bernoulli distributions. We show results for a number of data sets, as well as a case-study where a mixture of continuous beta processes is used as part of an automated stroke rehabilitation system.

## 1 Introduction

In this paper, we investigate a class of problems in which the objective is to model a correlated set of Bernoulli experiments (weighted coin flips). This class arises, for example, in a rehabilitation setting where a patient can succeed or fail at an exercise, and success will depend on some parameter, $x$ (e.g. the resistance or weight in the exercise). As $x$ changes, so does the probability of success, but not necessarily monotonically. For example, a person attempting to regain use of his/her arm after stroke is attempting to rehabilitate their nervous system as well as their muscles or bones. Neurological rehabilitation can show unusual patterns, as different pathways in the nervous system can lead to different abilities in the patient. Thus, while for any fixed parameter setting $x$, the probability distribution is a Bernoulli, we may be able to assume some degree of correlation between the Bernoulli variables at different $x$ values. This correlation can lead to increased robustness in the face of limited or noisy training data. In this paper, we show how a set of correlated beta processes can be used to efficiently model this space, and how this can lead to efficiency and robustness gains in simulated and real-world domains.

There are several methods which could be used to correlate the belief distributions of different experiments. A well-known and often used method is using *logistic Gaussian Processes* [Tokdar and Ghosh, 2007]. Here a Gaussian Process (GP) is used to learn a real-valued function over $S$ by keeping a multivariate normal distribution, and this function is then 'squashed' by using a sigmoid function to give values between 0 and 1. An elaborate explanation of this approach can be found in [Rasmussen and Williams, 2006]. One problem with this approach is that the posterior distribution after observing the outcome of an experiment is *not* a normal-distribution. The posterior is approximated by a Laplace approximation or by using Monte-Carlo sampling.

Another approach is by using a *copula* to map a GP to beta-distributions [Wilson and Ghahramani, 2010]. In this case, too, the posterior distribution has to be approximated by a Gaussian. A major drawback of using either of these GP methods is that they require a matrix inversion or the computation of a Cholesky decomposition of the covariance matrix of all the observed data points, which has cubic time complexity in general. Even with an incremental approach each incremental step still has quadratic time complexity. When the amount of data grows large, this is a serious disadvantage.

Instead of working with a latent GP, which is then translated to a distribution over probabilities, it would be more intuitive to work in the space of the probabilities straight away, keeping a beta-distribution for all the experiments and *assuming* these are correlated in such a way that experience can be easily shared according to a kernel function between the experiments. In this paper we suggest a method for easily sharing the data between the experiments and keeping a continuous, correlated multi-variate beta-distribution, or a *continuous correlated beta process*[1] (CCBP). Time complexity for prediction is, for this approach, linear in the amount of data.

The rest of this paper is structured as follows. In Section 3 the underlying theory and outline of the resulting algorithm are explained. In Section 4 we show various experimental re-

[1] It is unfortunate that there is already a graphical model with the name *beta process* [Ghahramani *et al.*, 2006], however, our name is chosen to make the correspondence with GPs clear.

sults, comparing the CCBP approach to the experiments of [Rasmussen and Williams, 2006] using logistic GPs. A specific case study showing how CCBPs could be used in an automated stroke rehabilitation system, where exercises can be chosen from a continuous space of difficulties is presented in Section 5. Finally, Section 6 presents the conclusions and discusses possible routes for future work.

## 2 Background

Consider a space $\mathcal{S}$ of Bernoulli experiments $S_x \in \mathcal{S}$ indexed by $x \in \mathcal{X}$ with an unknown probability $\theta(x) = \Pr(S_x = 1)$ of being successful. We denote successful experiments by 1 and unsuccessful experiments by 0. A conjugate prior for the distribution over $\theta(x)$ would be given by having a beta distribution for each possible experiment $S_x$:

$$\Pr(\theta(x)) = Beta(\alpha(x), \beta(x)) \propto \theta(x)^{\alpha(x)-1}(1-\theta_x)^{\beta(x)-1}$$

After observing the outcome $s$ of experiment $S_x$, we would update the distribution over $\theta(x)$ according to Bayes' rule to obtain the posterior:

$$\begin{aligned}\Pr(\theta(x)|S_x = s) &\propto \Pr(\theta(x)\Pr(S_x = s|\theta(x)) \\ &\propto \theta(x)^{\alpha(x)+\delta(s=1)-1}(1-\theta(x))^{\beta(x)+\delta(s=0)-1}\end{aligned}$$

Here $\delta(a)$ is a Kronecker delta that returns 1 when $a$ is true and 0 otherwise. Effectively, Bayes' rule increments the hyperparameter corresponding to the outcome by 1. To find accurate values for $\theta$ in the entire space $\mathcal{S}$, we would need to observe a good number of outcomes for all the experiments. However, if we know that the probabilities of different experiments are *correlated*, experience could be shared between experiments. For example, if the space $\mathcal{X}$ is a continuous set, we might know that $\theta(x)$ is a smooth, continuous function and have experiments close to each other according to some metric share their experience.

### 2.1 Beta Process

The Bernoulli-beta process [Ghahramani *et al.*, 2006] provides a framework to model a possibly infinite number of beta distributions over a space of Bernoulli experiments. However, the betas are assumed to be all independent and therefore there is no mechanism to share experience. In fact, it is particularly difficult to define a correlated beta process. The main issue is that the introduction of correlations does not seem to preserve conjugacy. There is a large literature on correlated multivariate beta distributions [Gupta and Wong, 1985; Olkin and Liu, 2003], however Bayesian updates are complicated and lead to approximative posteriors to keep the marginals of each Bernoulli in the class of beta distributions.

### 2.2 Logistic Gaussian Process (LGP)

Alternatively, one can leverage Gaussian Processes (GPs) to define a distribution over $\theta(x)$ since we can view $\theta$ as an unknown function over the space $\mathcal{X}$ that indexes the experiments. However, the range of $\theta$ is $[0, 1]$, while GPs assign non-zero probabilities to functions with a larger ranges. A common approach to restrict the range to $[0, 1]$ is to assume that the functions are "squashed" by a logistic sigmoid. In other words, it is possible to use a *logistic Gaussian process* that defines a GP over function outputs that may be anywhere in the reals but are re-mapped in $[0, 1]$ by the logistic sigmoid.

If $y \sim \mathcal{N}(\mu, \sigma)$ is a normal-distributed variable, the corresponding probability will be $\frac{1}{1+e^{-y}}$. After observing a successful outcome, the posterior for $y$ will become $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\frac{1}{1+e^{-y}}$, which is not a pdf of a Gaussian-distribution. It can, however, be approximated by a Gaussian-distribution by using a Laplace approximation (a second-order Taylor expansion of the logarithm of the posterior).

### 2.3 Gaussian Copula Process (GCP)

Another way of obtaining correlated beta-distributions from a Gaussian process is by the use of a copula [Nelsen, 2006], resulting in a so-called *copula process* [Wilson and Ghahramani, 2010]. A copula is a way of constructing a multi-variate distribution with arbitrary marginal distributions. This can be done by correlating the *cumulative* distributions through a copula function. For example, one could have a Gaussian process, and for one specific point (one Gaussian distribution) take a point generated according to this distribution, transform it through the cumulative distribution function to get a value between 0 and 1, and then transform this through the inverse of the cumulative distribution for a beta distribution. The result would be that a GP would be transformed into multi-variate correlated beta distribution. However, the posterior distribution after observing evidence would not be a proper multi-variate Gaussian distribution, but would have to be approximated by a Laplace approximation.

## 3 Continuous Correlated Beta Process

We propose a simple mechanism to share experience between a continuum of beta distributions. The idea is to use a kernel function $K(x, x')$ to indicate to what extent the experience for experiment $S_x$ should be shared with $S_{x'}$. Recall that for each outcome of an experiment $S_x$, the corresponding hyperparameter ($\alpha(x)$ or $\beta(x)$) is incremented by one. Intuitively, if $S_{x'}$ is correlated with $S_x$, it would make sense to use $S_x$'s outcomes to also increment the hyperparameters of $S_{x'}$, but perhaps by a fraction instead of 1. Hence, let $K : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be a kernel function that indicates the magnitude of the fractional updates for $S_{x'}$ based on the outcomes of $S_x$. In particular, $K(x, x') = 0$ when $S_x$ and $S_{x'}$ are independent, $K(x, x') = 1$ when $S_x$ and $S_{x'}$ are governed by the same Bernoulli distribution and $K(x, x') \in (0, 1)$ when $S_x$ and $S_{x'}$ are correlated. Suppose that we observe an outcome for $n$ different experiments (e.g. $S_{x_1} = s_1, S_{x_2} = s_2, \ldots, S_{x_n} = s_n$), then the posterior beta for the Bernoulli distribution of some experiment $S_x$ becomes:

$$\begin{aligned}&\Pr(\theta(x)|S_{x_1} = s_1, S_{x_2} = s_2, \ldots, S_{x_n} = s_n) \\ &\propto \quad \theta(x)^{\alpha(x)-1+\sum_{i=1}^{n}\delta(s_i=1)K(x_i,x)} \\ &\qquad \times(1-\theta(x))^{\beta(x)-1+\sum_{i=1}^{n}\delta(s_i=0)K(x_i,x)}\end{aligned}$$

This update can be done in time that is linear with respect to the amount of data. Note also that the approach generalizes immediately to multi-class problems if we replace betas by Dirichlets.

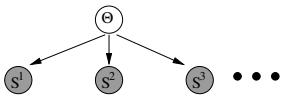


Figure 1: Naive Bayes model to learn a beta distribution.

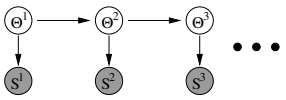


Figure 2: HMM model to learn a beta distribution.

### 3.1 Graphical Model

While the proposed approach is simple and efficient, it is not clear how it relates to Bayes' rule. In this section, we show how our update technique can viewed as inference in a graphical model that generalizes Bayes' rule.

**Simple Beta Distribution**

Consider a single experiment $S$ governed by a Bernoulli distribution with parameter $\theta = \Pr(S = 1)$. Suppose that the prior over $\theta$ is a beta distribution with hyperparameters $\alpha$ and $\beta$ and suppose that we observe $n$ samples $S^1, \ldots, S^n$ of the Bernoulli distribution.[2] We can represent the joint distribution over $\theta$ and $S^1, \ldots, S^n$ by the graphical model in Figure 1. Throughout the paper we use the convention that shaded nodes are observed while blank nodes are unobserved.

If the observations $S^1, \ldots, S^n$ are obtained sequentially, we can model the learning process by the HMM in Figure 2. The transition distribution is set to a Dirac distribution $\Pr(\theta^{t+1}|\theta^t) = \delta(\theta^{t+1} = \theta^t)$, which ensures that all the $\theta$'s are identical. As a result, the graphical models in Figures 1 and 2 are equivalent.

To derive the correlated beta process, it will be useful to model the hyperparameters $\alpha$ and $\beta$ as variables. Adding them to the graphical model yields the dynamic Bayesian network in Figure 3. Here, we use $A$ to denote the random variable corresponding to $\alpha$ and $B$ to denote the random variable corresponding to $\beta$. Since the prior over $\theta$ is a beta with known hyperparameters $\alpha$ and $\beta$, we set the prior over $A^1$ and $B^1$ to be the following Dirac distributions:

$$\Pr(A^1) = \delta(A^1 = \alpha)$$
$$\Pr(B^1) = \delta(B^1 = \beta)$$

After observing $S^t$, one of the hyperparameters is incremented, so we set the transition distributions for $A^{t+1}$ and $B^{t+t}$ as follows:

$$\Pr(A^{t+1}|A^t, S^t) = \begin{cases} \delta(A^{t+1} = A^t + 1) & \text{if } S^t = 1 \\ \delta(A^{t+1} = A^t) & \text{otherwise} \end{cases}$$

$$\Pr(B^{t+1}|B^t, S^t) = \begin{cases} \delta(B^{t+1} = B^t + 1) & \text{if } S^t = 0 \\ \delta(B^{t+1} = B^t) & \text{otherwise} \end{cases}$$

The graphical models in Figures 2 and 3 are equivalent in the sense that $\Pr(\theta^t|S^{1:t})$ is the same in both for all $t$'s.

[2]We use superscripts to index variables corresponding to different samples of the same experiment and subscripts to index variables corresponding to different experiments.

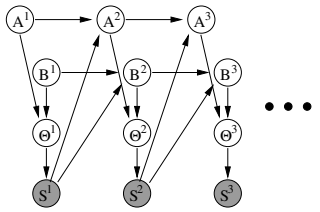


Figure 3: DBN model to learn a beta distribution.

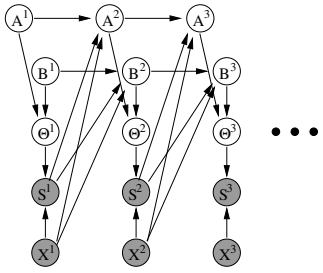


Figure 4: DBN model to learn multivariate beta distributions.

**Correlated Beta Process**

We are now ready to specify a simple bivariate beta distribution. Let $\theta^t = \langle \theta_1^t, \theta_2^t \rangle$ be a vector parameterizing two Bernoulli distributions for experiments $S_1$ and $S_2$. We first consider the case where the Bernoulli variables have independent beta distributions with hyperparameters $\alpha^t = \langle \alpha_1^t, \alpha_2^t \rangle$ and $\beta^t = \langle \beta_1^t, \beta_2^t \rangle$. Figure 4 describes an extended DBN where at every step $t$ one of the experiments is sampled. We denote by $S^t$ the observed outcome and by $X^t$ the index (1 or 2) of the experiment. The conditional distributions for $S^t$, $A^{t+1}$ and $B^{t+1}$ are adjusted as follows:

$$\Pr(S^t|\theta^t, X^t) = \theta^t_{X^t}$$

$$\Pr(A^{t+1}|A^t, S^t, X^t) = \begin{cases} \delta(A^{t+1} = A^t) & \text{if } S^t = 0 \\ \delta(\langle A_1^{t+1}, A_2^{t+1} \rangle = \langle A_1^t + 1, A_2^t \rangle) & \text{if } S^t = 1, X^t = 1 \\ \delta(\langle A_1^{t+1}, A_2^{t+1} \rangle = \langle A_1^t, A_2^t + 1 \rangle) & \text{if } S^t = 1, X^t = 2 \end{cases}$$

$$\Pr(B^{t+1}|B^t, S^t, X^t) = \begin{cases} \delta(B^{t+1} = B^t) & \text{if } S^t = 1 \\ \delta(\langle B_1^{t+1}, B_2^{t+1} \rangle = \langle B_1^t + 1, B_2^t \rangle) & \text{if } S^t = 0, X^t = 1 \\ \delta(\langle B_1^{t+1}, B_2^{t+1} \rangle = \langle B_1^t, B_2^t + 1 \rangle) & \text{if } S^t = 0, X^t = 2 \end{cases}$$

Now, suppose that the Bernoulli variables are correlated in such a way that an observation of one Bernoulli variable leads us to believe that the hyperparameters of the other Bernoulli variable should also be updated with a fractional increment $K$. Then we can simply revise the conditional distributions

for $A^{t+1}$ and $B^{t+1}$ as follows:

$$\Pr(A^{t+1}|A^t, S^t, X^t)$$
$$= \begin{cases} \delta(A^{t+1} = A^t) & \text{if } S^t = 0 \\ \delta(\langle A_1^{t+1}, A_2^{t+1}\rangle = \langle A_1^t + 1, A_2^t + K\rangle) & \text{if } S^t = 1, X^t = 1 \\ \delta(\langle A_1^{t+1}, A_2^{t+1}\rangle = \langle A_1^t + K, A_2^t + 1\rangle) & \text{if } S^t = 1, X^t = 2 \end{cases}$$
$$\Pr(B^{t+1}|B^t, S^t, X^t)$$
$$= \begin{cases} \delta(B^{t+1} = B^t) & \text{if } S^t = 1 \\ \delta(\langle B_1^{t+1}, B_2^{t+1}\rangle = \langle B_1^t + 1, B_2^t + K\rangle) & \text{if } S^t = 0, X^t = 1 \\ \delta(\langle B_1^{t+1}, B_2^{t+1}\rangle = \langle B_1^t + K, B_2^t + 1\rangle) & \text{if } S^t = 0, X^t = 2 \end{cases}$$

We can generalize the above construction to multivariate betas and in the limit to correlated beta processes. It is useful to think of $A$, $B$ and $\theta$ as functions of $X$ such that $\theta(X)$ is a Bernoulli variable indexed by $X$. Similarly, $A(X)$ and $B(X)$ are the hyperparameters of $\theta(X)$. Then we can generalize the conditional distributions for $A^{t+1}$ and $B^{t+1}$ by using a kernel update $K(X_t, x)$ that returns a value between 0 and 1 depending on how correlated $S_{X^t}$ and $S_x$ are.

$$\Pr(A^{t+1}|A^t, S^t, X^t)$$
$$= \begin{cases} \delta(A^{t+1}(x) = A^t(x)\ \forall x) & \text{if } S^t = 0 \\ \delta(A^{t+1}(x) = A^t(x) + K(X_t, x)\ \forall x) & \text{otherwise} \end{cases}$$
$$\Pr(B^{t+1}|B^t, S^t, X^t)$$
$$= \begin{cases} \delta(B^{t+1}(x) = B^t(x)\ \forall x) & \text{if } S^t = 1 \\ \delta(B^{t+1}(x) = B^t(x) + K(X_t, x)\ \forall x) & \text{otherwise} \end{cases}$$

### 3.2 Discussion

Although our technique to update betas can be viewed as inference in a graphical model, it does not correspond to Bayes' rule. While it would be desirable to use Bayes' rule, recall from Section 2 that all existing techniques that attempt to use Bayes' rule to update a joint distribution over correlated Bernoulli variables end up making an approximation to force the posterior into a convenient class of distributions and therefore do not use Bayes' rule either. This includes logistic Gaussian processes (LGPs) and Gaussian copula processes (GCPs). Our approach has the advantage that it is fairly easy to understand how the kernel function will induce sharing of experience among betas and computation is linear in the amount of data. In contrast, LGPs and GCPs need to invert the kernel function which takes cubic time with respect to the amount of data. Also, it is less obvious how the correlations encoded by the kernel function induce sharing of experience since a logistic sigmoid or copula is used to transform the output of the GPs. As a result, choosing a good kernel that will induce the right amount of experience sharing is more tricky with LGPs and GCPs.

The CCBP approach can be easily adapted to multiple outputs, giving a continuously correlated Dirichlet process. This will be demonstrated in a multi-class classification experiment in Section 4.2.

## 4 Experiments

### 4.1 One-dimensional Illustration

A first simple experiment uses a one-dimensional space $\mathcal{X} = [-2, 2]$, and actual Bernoulli probability given by $(1 + e^{-x^2} cos(10\frac{1-e^{-x}}{1+e^{-x}}))/2$. The kernel used for both LGP and CCBP is a squared exponential kernel $K(x_1, x_2) = e^{\frac{-(x_1-x_2)^2}{10}}$. Training data was obtained using a uniform distribution over $[-2, 2]$. Figure 5 shows the actual probabilities as well as the mean predictions as learned by the LGP and the CCBP after observing 1000 examples. The CCBP has less smooth behavior and might have benefitted from using a kernel with wider length scale.

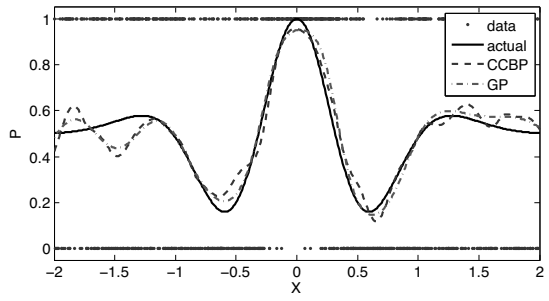


Figure 5: Estimated probability after 1000 observations.

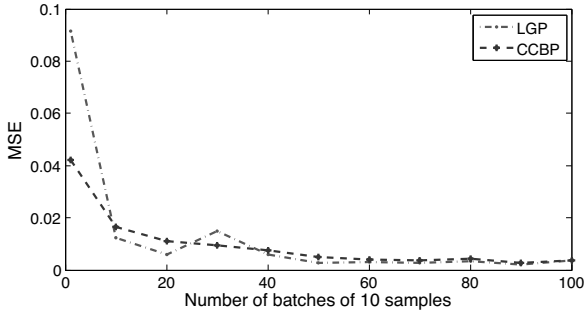


Figure 6: Mean squared error for varying size of training dataset.

In Figure 6 the mean squared error of the prediction of the probability is plotted for LGP and CCBP. The differences between the two approaches are small. However, the time needed to run the experiments was hugely different. This can be clearly seen in Figure 7, a log-log scale plot of the CPU time of CCBP and GP approach for varying amounts of data.

### 4.2 USPS Hand-written Digit dataset: Multi-class Classification

A second experiment shows the behaviour of CCBPs when used for multi-class classification. We use the well-known USPS dataset, with the modifications in training and test sets as explained in [Rasmussen and Williams, 2006]. The kernel used is the same as used in the description of multi-class classification at page 70 of this book, being an isometric squared exponential kernel, $K(x_1, x_2) = \sigma_f e^{\frac{||x_1-x_2||_2^2}{2l^2}}$. As was done with the LGPs, we varied the length-scale $l$ of the kernel. We did not vary the value of $\sigma_f$, as in CCBPs a value of $\sigma_f \neq 1$ would imply attributing too much or too little importance to

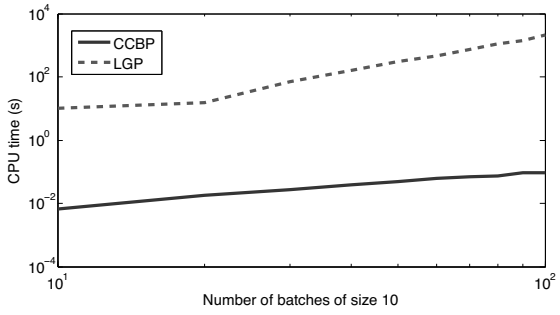


Figure 7: CPU time for varying size of training dataset. Note the log-log scale.

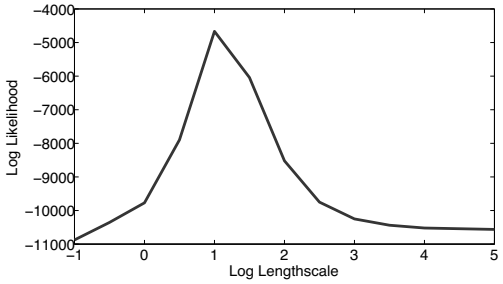


Figure 8: Log likelihood of training data for various length scales of the kernel.

empirical data, the correct value is $\sigma_f = 1$.

In Figure 8 the log likelihood of the training data is plotted for different values of the length scale. The highest likelihood occurs at $log(l) = 1.0$. In Figure 9 the misclassification rates on the test data set are shown. The best result occurs at $log(l) = 0.0$, where an error of $3\%$ was obtained - similar to the error reported for LGPs. At the point with highest likelihood, an error of $11\%$ is obtained. This shows that the use of a validation set (e.g. using cross-validation) to find good values for the kernel parameters is advisable. The best predictions of CCBPs are on par with those of LGPs, and cross-validation could be used to find good parameter settings.

## 5 A Case Study: Stroke Rehabilitation

Stroke is a leading cause of physical disability and death around the world [Heart and Stroke Foundation of Canada, 2011; American Heart Association, 2011]. Research has shown that post-stroke impairments can be reduced by repetitive and goal-directed rehabilitation, which improves motor function and cortical reorganization in stroke patients [Fasoli *et al.*, 2004]. The recovery process, however, is typically slow and labor-intensive, involving extensive interaction between a therapist and a patient. One of the main motivations for developing rehabilitation robotic devices is to automate repetitive interventions, which can alleviate strain on therapists.

The design of a successful robotic stroke rehabilitation device relies on an sufficient model of the rehabilitation schedule of the patient. This schedule describes the patient's abilities as a function of the parameters of the rehabilitation device, and describes the way in which these abilities change over time. Thus, the schedule can be used to plan a program (a series of exercises over along period of time) that will most effectively rehabilitate the patient.

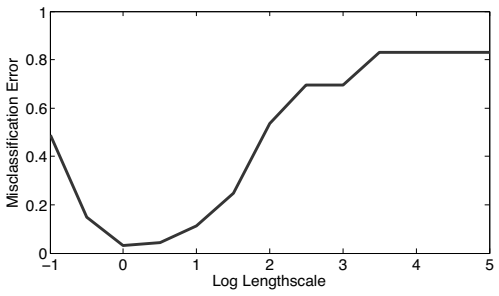


Figure 9: Misclassification rate for various length scales.

In this case study, we consider a device with a single parameter, e.g., the resistance of a haptic handle, where the rehabilitation exercise is to push the handle to a set target distance. For such a device, the rehabilitation schedule will include a function $P(success) \propto f(resistance)$, where $f(resistance)$ is a continuous function that relates the person's abilities (e.g. $success$, or whether they can reach the target) to the resistance level. One might expect this function to be monotonically decreasing with increasing resistance. However, stroke is a neurological condition, and this function may not always be strictly monotonic nor extremely smooth, since different pathways in the nervous system may be used for different stimuli (including the resistance level).

In this case study, we model the stroke rehabilitation schedule $P = f(x)$ using continuous correlated beta processes. We keep two such processes, one for when the person is fatigued, one for not fatigued. As the fatigue status is not directly observable, the belief state will be a mixture of the two CCBP models. Evidence is counted for both components, but only counts for a fraction (proportional to the current belief in the fatigue status). A few simulation results are shown. Figure 10 shows curves representing the priors for fatigue = no (blue) and fatigue = yes (red). The green lines represent the actual Bernoulli distributions the simulator uses, with the top one the distribution for non-fatigued. The horizontal axis represents the difficulty of the exercise. Note that, when fatigued, the success rate is estimated to be lower than when not fatigued, and success rate is believed, on average, to decrease with increasing difficulty. Figure 11 shows the posteriors after 5 sessions of 20 exercises, where the resistance level was chosen from a uniform distribution. At the start of a 20 exercise session, the belief in fatigue=yes is equal to 0, at each exercise there is a $10\%$ probability of becoming fatigued, once fatigued the person remains fatigued until the end of the session. Many different models for fatigue rate, improvement rate and priors can be modelled with this approach. Once integrated into a planning system, this kind of approach would

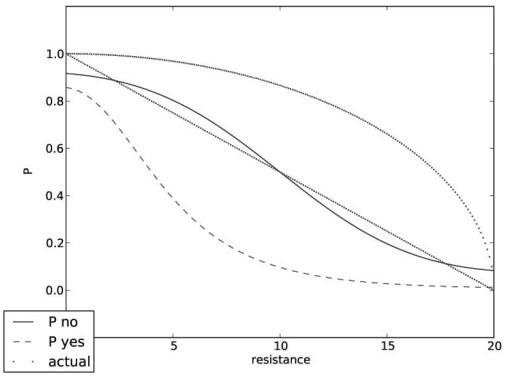


Figure 10: Means of priors for success rate for non-fatigued (blue) or fatigued (red) state. Green lines represent actual probabilities for the simulated person, the bottom line for fatigued.

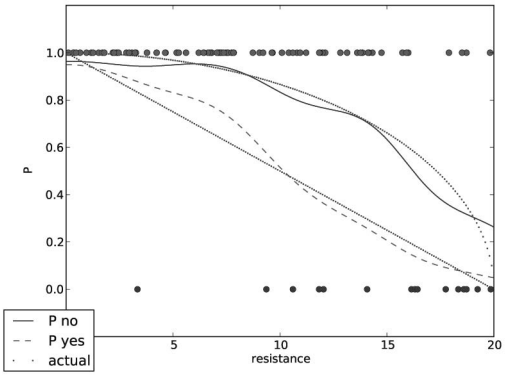


Figure 11: Means of posteriors for success rate for non-fatigued (blue) or fatigued (red) state after 100 exercises. Outcomes of the exercises are indicated by dots: red for successful, blue for unsuccessful.

require using the schedule model extensively, but time complexity of the predictions of the model are essential, and ideally should be done in a matter of seconds or less. This makes CCBPs a valid approach for use in such an automated system.

## 6 Conclusion

In this paper we introduced a novel approach for modelling correlated probabilities for a continuous space of Bernoulli experiments. In comparison to current state-of-the-art, we obtain results which are on par with logistic Gaussian Processes, but at a much better time complexity. GPs have cubic time complexity in the general case while continuous correlated beta processes have linear time complexity. We have illustrated how such continuous correlated beta processes can be viewed as inference in a graphical model. The approach is easily extended to Dirichlet distributions, allowing the learning of a correlated continuum of multinomial distributions.

In a number of experiments we have shown that the predictions made by CCBPs are as good as those of logistic GPs, but much faster. This increase in speed makes the approach valuable for a system such as an automated stroke rehabilitation device, where predictions have to be made fast so they can be used in a planning approach. There are several interesting directions for future work. For one, using the CCBP method in a POMDP planning system, such as the stroke rehabilitation case study, would be interesting. For this one would need an efficient way to do planning with continuous actions. This would also involve modeling multiple independent variables (e.g. resistance and target distance). Other possible routes for future work would exploit the fact that having beta distributions means that we have estimates of the variance of the estimate as well as the expected probability. This could be used to obtain *abstaining classifiers*, classifiers which can choose to refuse to classify an example if too uncertain, delegating the problematic example to another system, instead of giving a probably erroneous classification label. It would also allow for *active learning* where the system actively asks for the correct classification of examples in those parts of the space with highest uncertainty.

## 7 Acknowledgements

This publication has been supported by FONCICYT contract number 000000000095185. The content of this document reflects only the author's views. FONCICYT is not liable for any use that may be made of the contained information.

## References

[American Heart Association, 2011] American Heart Association. Stroke statistics, 2011.

[Fasoli *et al.*, 2004] S.E. Fasoli, H.I. Krebs, and N. Hogan. Robotic technology and stroke rehabilitation: Translating research into practice. *Topics in Stroke Rehabilitation*, 11(4):11–19, 2004.

[Ghahramani *et al.*, 2006] Z. Ghahramani, T.L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. In *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics*, 2006.

[Gupta and Wong, 1985] A. K. Gupta and C. F. Wong. On three and five parameter bivariate beta distributions. *Metrika*, 32:85–91, 1985.

[Heart and Stroke Foundation of Canada, 2011] Heart and Stroke Foundation of Canada. Stroke statistics, 2011.

[Nelsen, 2006] R.B. Nelsen. *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag, 2006.

[Olkin and Liu, 2003] Ingram Olkin and Ruixue Liu. A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407 – 412, 2003.

[Rasmussen and Williams, 2006] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes*. MIT Press, 2006.

[Tokdar and Ghosh, 2007] S.T. Tokdar and J.K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34 – 42, 2007.

[Wilson and Ghahramani, 2010] A.G. Wilson and Z. Ghahramani. Copula processes. In *Neural Information Processing Systems*, 2010.