Stagewise Safe Bayesian Optimization with Gaussian Processes

Yanan Sui [1] **Vincent Zhuang** [1] **Joel W. Burdick** [1] **Yisong Yue** [1]

Abstract

Enforcing safety is a key aspect of many problems pertaining to sequential decision making under uncertainty, which require the decisions made at every step to be both informative of the optimal decision and also safe. For example, we value both efficacy and comfort in medical therapy, and efficiency and safety in robotic control. We consider this problem of optimizing an unknown utility function with absolute feedback or preference feedback subject to unknown safety constraints. We develop an efficient safe Bayesian optimization algorithm, StageOpt, that separates safe region expansion and utility function maximization into two distinct stages. Compared to existing approaches which interleave between expansion and optimization, we show that StageOpt is more efficient and naturally applicable to a broader class of problems. We provide theoretical guarantees for both the satisfaction of safety constraints as well as convergence to the optimal utility value. We evaluate StageOpt on both a variety of synthetic experiments, as well as in clinical practice. We demonstrate that StageOpt is more effective than existing safe optimization approaches, and is able to safely and effectively optimize spinal cord stimulation therapy in our clinical experiments.

1. Introduction

Bayesian optimization is a well-established approach for sequentially optimizing unknown utility functions. By leveraging regularity assumptions such as smoothness and continuity, such techniques offer efficient solutions for a wide range of high-dimensional problem settings such as experimental design and personalization in recommender systems. Many of these applications are also subject to a variety of safety constraints, so that actions cannot be freely chosen from the entire input space. For instance, in safe Bayesian optimization, any chosen action during optimization must be known to be "safe", regardless of the reward from the utility function. Typically, one is initially given a small region of the decision/action space that is known to be safe, and must iteratively expand the safe action region during optimization (Sui et al., 2015).

A motivating application of our work is a clinical setting, where physicians need to sequentially choose among a large set of therapies (Sui et al., 2017a). The effectiveness and safety of different therapies are initially unknown, and can only be determined through sequential tests starting from some initial set of well-studied therapies. A natural way to explore is to start from some therapies similar to these initial ones, since their efficacy and safety would not differ too greatly. By iteratively repeating this process, one can gradually explore the utility and safety landscapes in a safe fashion.

Our contributions. We propose a novel safe Bayesian optimization algorithm, StageOpt, to address the challenge of efficiently identifying the total safe region and optimizing the utility function within the safe region. In contrast to previous safe Bayesian optimization work (Sui et al., 2015; Berkenkamp et al., 2016a) which interleaves safe region expansion and optimization, StageOpt is a stagewise algorithm which first expands the safe region and then optimizes the utility function. StageOpt is well suited for settings in which the safety and utility functions are very different (e.g., temperature vs gripping force), i.e. lie on different scales or amplitudes. Furthermore, in settings in which the utility and safety functions are measured in different ways, it is natural to have a separate first stage dedicated to safe region expansion. For example, in clinical trials we may wish to spend the first stage only querying the patient about the comfort of the stimulus, as opposed to having to measure the utility and comfort simultaneously.

Conceptually, StageOpt models the safety function(s) and utility function as sampled functions from different Gaussian processes (GPs), and uses confidence bounds to assess the safety of unexplored decisions. We provide theoretical results for StageOpt under the assumptions that

[1]California Institute of Technology, Pasadena, CA, USA. Correspondence to: Yanan Sui <ysui@caltech.edu>, Vincent Zhuang <vzhuang@caltech.edu>, Joel W. Burdick <jwb@robotics.caltech.edu>, Yisong Yue <yyue@caltech.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

(1) the safety and utility functions have bounded norms in their Reproducing Kernel Hilbert Spaces (RKHS) associated with the GPs, and (2) the safety functions are Lipschitz-continuous, which is guaranteed by many common kernels. We guarantee (with high probability) the convergence of StageOpt to the safely reachable optimum decision. In addition to simulation experiments, we apply StageOpt to a clinical setting of optimizing spinal cord stimulation for patients with spinal cord injuries. Compared to expert physicians, we find that StageOpt explores a larger safe region and finds better stimulation strategy.

2. Related Work

Many Bayesian optimization methods often model the unknown underlying functions as Gaussian processes (GPs), which are smooth, flexible, nonparametric models (Rasmussen & Williams, 2006). GPs are widely used as a regularity assumption in many Bayesian optimization techniques, since they can easily encode prior knowledge and explicitly model variance.

The fundamental tradeoff between exploration and exploitation in sequential decision problems is commonly formalized as the multi-armed bandit problem (MAB), introduced by Robbins (1952). In MAB, each decision is associated with a stochastic reward with initially unknown distribution. The goal of a bandit algorithm is to maximize the cumulative reward. In a variant called "best-arm identification" (Audibert et al., 2010), one seeks to identify the decision with highest reward with minimal trials. It has been widely studied under a variety of different situations (cf., Bubeck & Cesa-Bianchi (2012) for an overview). Many efficient algorithms build on the methods of *upper confidence bounds* proposed in Auer (2002), and *Thompson sampling* proposed in Thompson (1933). Their key ideas are to use posterior distributions of rewards to implicitly negotiate the explore-exploit tradeoff by optimistic sampling. This idea naturally extends to bandit problems with complex (or even infinite) decision sets under certain regularity conditions of the reward function (Dani et al., 2008; Kleinberg et al., 2008; Bubeck et al., 2008).

In the kernelized setting, several algorithms with theoretical guarantees have been proposed. Srinivas et al. (2010) propose the GP-UCB algorithm, which uses confidence bounds to address bandit problems with a reward function modeled using a Gaussian process. Gotovos et al. (2013) studies active sampling for localizing level sets, finding where the objective crosses a specified threshold. Chowdhury & Gopalan (2017) extends the work of Srinivas et al. (2010) by proving tighter bounds as well as providing guarantees for a GP-based Thompson sampling algorithm. However, none of these algorithms are designed to work with safety constraints, and often violate them in practice (Sui et al., 2015). There are also algorithms without theoretical guarantees. Gelbart et al. (2014) studies a constrained Expected Improvement algorithm for Bayesian optimization with unknown constraints. Hernández-Lobato et al. (2016) considers a general framework for constrained Bayesian optimization using information-based search.

The problem of safe exploration has been considered in control and reinforcement learning (Hans et al., 2008; Gillula & Tomlin, 2011; Garcia & Fernandez, 2012; Turchetta et al., 2016). These methods typically consider the problem of safe exploration in MDPs. They ensure safety by restricting policies to be ergodic with high probability and able to recover from any state visited. The safe optimization problem has also been studied under the restriction of the bandit/optimization setting, where decisions do not cause state transitions. This leads to simpler algorithms (SafeOpt) with stronger guarantees (Sui et al., 2015; Berkenkamp et al., 2016a), and fits well to safe sampling problems and applications. There are other safe algorithms (Schreiter et al., 2015; Wu et al., 2016) under different active learning settings. Our work builds upon the SafeOpt approach, with stronger empirical performance and convergence rates on a broad class of safety functions.

3. Problem Statement

We consider a sequential decision problem in which we seek to optimize an unknown utility function $f : D \to \mathbb{R}$ from noisy evaluations at iteratively chosen sample points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \ldots \in D$. However, we further require that each of these sample points are "safe": that is, for each of n unknown safety functions $g_i : D \to \mathbb{R}$ at $g_i(\boldsymbol{x}_t)$ lies above some threshold $h_i \in \mathbb{R}$. We can formally write our optimization problem as follows:

$$\max_{\boldsymbol{x} \in D} f(\boldsymbol{x}) \quad \text{subject to} \quad g_i(\boldsymbol{x}) \geq h_i \text{ for } i = 1, \ldots, n \quad (1)$$

Regularity assumptions. In order to model the utility function and the safety functions, we use Gaussian processes (GPs), which are smooth yet flexible nonparametric models. Equivalently, we assume that f and all g_i have bounded norm in the associated Reproducing Kernel Hilbert Space (RKHS). A GP is fully specified by its mean function $\mu(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$; in this work, we assume WLOG GP priors to have zero mean (i.e. $\mu(\boldsymbol{x}) = 0$). We further assume that each safety function g_i is L_i-Lipschitz continuous with respect to some metric d on D. This assumption is quite mild, and is automatically satisfied by many commonly-used kernels (Srinivas et al., 2010; Sui et al., 2015).

Feedback models. We primarily consider noise-perturbed feedback, in which our observations are perturbed

by i.i.d. Gaussian noise, i.e., for samples at points $A_T = [\boldsymbol{x}_1 \dots \boldsymbol{x}_T]^T \subseteq D$, we have $\boldsymbol{y}_t = f(\boldsymbol{x}_t) + n_t$ where $n_t \sim N(0, \sigma^2)$. The posterior over f is then also Gaussian with mean $\mu_T(\boldsymbol{x})$, covariance $k_T(\boldsymbol{x}, \boldsymbol{x}')$ and variance $\sigma_T^2(\boldsymbol{x}, \boldsymbol{x}')$ that satisfy,

$$\begin{aligned}\mu_T(\boldsymbol{x}) &= \boldsymbol{k}_T(\boldsymbol{x})^T(\boldsymbol{K}_T + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}_T \\ k_T(\boldsymbol{x}, \boldsymbol{x}') &= k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_T(\boldsymbol{x})^T(\boldsymbol{K}_T + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_T(\boldsymbol{x}') \\ \sigma_T^2(\boldsymbol{x}) &= k_T(\boldsymbol{x}, \boldsymbol{x}),\end{aligned}$$

where $\boldsymbol{k}_T(\boldsymbol{x}) = [k(\boldsymbol{x}_1, \boldsymbol{x}) \dots k(\boldsymbol{x}_T, \boldsymbol{x})]^T$ and $\boldsymbol{K}_T$ is the positive definite kernel matrix $[k(\boldsymbol{x}, \boldsymbol{x}')]_{\boldsymbol{x}, \boldsymbol{x}' \in A_T}$.

We also consider the case in which only preference feedback is available for the utility function. This setting is often used to characterize real-world applications that elicit subjective human feedback. One way to formalize the online optimization problem is the dueling bandits problem (Yue et al., 2012; Sui et al., 2017b). In the basic dueling bandits formulation, given two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we stochastically receive binary 0/1 feedback according to a Bernoulli distribution with parameter $\phi(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2))$, where ϕ is a *link function* mapping $\mathbb{R} \times \mathbb{R}$ to $[0, 1]$. For example, a common link function is the logit function $\phi(x, y) = (1 + \exp(y - x))^{-1}$.

To our knowledge, there are no existing algorithms for the safe Bayesian dueling bandit setting. Although our proposed algorithm is amenable to the full dueling bandits setting (as discussed later), to compare against existing algorithms, we consider the restricted dueling problem in which at timestep t one receives preference feedback between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$. The pseudocode for our proposed algorithm under this type of dueling feedback can be found in Appendix **??**.

Safe optimization. Using a uniform zero-mean prior (as is typical in many Bayesian optimization approaches) does not provide sufficient information to identify any point as safe with high probability. Therefore, we additionally assume that we are given an initial "seed" set of safe decision(s), which we denote as $S_0 \subset D$. Note that given an arbitrary seed set, it is not guaranteed that we will be able to discover the globally optimal decision $\boldsymbol{x}^* = \operatorname{argmax}_{\boldsymbol{x} \in D} f(\boldsymbol{x})$, e.g. if the safe region around $\boldsymbol{x}^*$ is topologically separate from that of S_0. Instead, we can formally define the optimization goal for a given seed via the *one-step reachability* operator:

$$R_\epsilon(S) := S \cup \bigcap_i \big\{\boldsymbol{x} \in D \;\big|\; \exists \boldsymbol{x}' \in S, \; g_i(\boldsymbol{x}') - \epsilon - L_i d(\boldsymbol{x}', \boldsymbol{x}) \geq h_i\big\},$$

which gives the set all of points that can be established as safe given evaluations of f on S with ϵ noise. Then, given some finite horizon T, we can define the subset of D reachable after T iterations from the initial safe seed set S_0 as the following:

$$R_\epsilon^T(S_0) := \underbrace{R_\epsilon(R_\epsilon \dots (R_\epsilon}_{T \text{ times}}(S_0)) \dots).$$

Thus, our optimization goal is $\operatorname{argmax}_{\boldsymbol{x} \in R_\epsilon^T(S_0)} f(\boldsymbol{x})$.

4. Algorithm

We now introduce our proposed algorithm, StageOpt, for the safe exploration for optimization problem.

Overview. We start with a high-level description of StageOpt. StageOpt separates the safe optimization problem into two stages: an exploration phase in which the safe region is iteratively expanded, followed by an optimization phase in which Bayesian optimization is performed within the safe region. We assume that our algorithm runs for a fixed T time steps, and that the first safe expansion region has horizon $T_0 < T$ with the optimization phase being $T_1 = T - T_0$ time steps long.

StageOpt models the utility function and the safety functions via Gaussian processes, and leverages their uncertainty in order to safely explore and optimize. In particular, at each iteration t, StageOpt uses the confidence intervals

$$Q_t^i(\boldsymbol{x}) := \left[\mu_{t-1}^i(\boldsymbol{x}) \pm \beta_t \sigma_{t-1}^i(\boldsymbol{x})\right], \tag{2}$$

where β_t is a scalar whose choice will be discussed later. We use superscripts to denote the confidence intervals for the respective safety functions, and we use the superscript f for the utility function. In order to guarantee both safety and progress in safe region expansion, instead of using Q_t^i directly, StageOpt uses the confidence intervals C_t^i defined as $C_t^i(\boldsymbol{x}) := C_{t-1}^i(\boldsymbol{x}) \cap Q_t^i(\boldsymbol{x})$, $C_0^i(\boldsymbol{x}) = [h_i, \infty]$ so that C_t^i are sequentially contained in C_{t-1}^i for all t. We also define the upper and lower bounds of C_t^i to be u_t^i and ℓ_t^i respectively, as well as the width as $w_t^i = u_t^i - \ell_t^i$.

We defined the optimization goal with respect to a tolerance parameter ϵ, which can employed as a stopping condition for the expansion stage. Namely, if the expansion stage stops at T_0 under the condition $\max_{\boldsymbol{x} \in G_t} w_t(\boldsymbol{x}) \leq \epsilon$, then the ϵ-Reachable safe region $R_\epsilon^{T_0}(S_0)$ is guaranteed to be expanded. Similarly, we have a tolerance parameter ζ (in Algorithm 1) to control utility function optimization with time horizon T_1.

Stage One: Safe region expansion. StageOpt expands the safe region in the same way as that of SafeOpt (Sui et al., 2015; Berkenkamp et al., 2016a). An increasing sequence of safe subsets $S_t \subseteq D$ is computed based on the confidence intervals of the GP posterior:

$$S_t = \bigcap_i \bigcup_{\boldsymbol{x} \in S_{t-1}} \big\{\boldsymbol{x}' \in D \;\big|\; \ell_t^i(\boldsymbol{x}) - L_i d(\boldsymbol{x}, \boldsymbol{x}') \geq h_i\big\}.$$

At each iteration, StageOpt computes a set of *expander* points G_t (that is, points within the current safe region that are likely to expand the safe region) and picks the expander with the highest predictive uncertainty.

In order to define the set G_t, we first define the function:

$$e_t(\boldsymbol{x}) := \left| \bigcap_i \left\{ \boldsymbol{x}' \in D \setminus S_t \;\middle|\; u_t(\boldsymbol{x}) - L_i d(\boldsymbol{x}, \boldsymbol{x}') \geq h_i \right\} \right|,$$

which (optimistically) quantifies the potential enlargement of the current safe set after we sample a new decision $\boldsymbol{x}$. Then, G_t is simply given by:

$$G_t = \{\boldsymbol{x} \in S_t : e_t(\boldsymbol{x}) > 0\}.$$

Finally, at each iteration StageOpt selects x_t to be $x_t = \operatorname{argmax}_{\boldsymbol{x} \in G_t} w_n(\boldsymbol{x}, i)$.

Stage Two: Utility optimization. Once the safe region is established, StageOpt can use any standard online optimization approach to optimize the utility function within the expanded safe region. For concreteness, we present here the GP-UCB algorithm (Srinivas et al., 2010). For completeness, we present a version of StageOpt based on preference-based utility optimization in Appendix **??**. Our theoretical analysis is also predicated on using GP-UCB, since it offers finite-time regret bounds. Formally, at each iteration in this phase, we select the arm x_t as the following:

$$x_t = \operatorname*{argmax}_{\boldsymbol{x} \in S_t} \mu^f_{t-1}(\boldsymbol{x}) + \beta_t \sigma^f_{t-1}(\boldsymbol{x}) \tag{3}$$

Note that it possible (though typically unlikely) for the safe region to further expand during this phase.

Comparison between SafeOpt and StageOpt. Although StageOpt is similar to SafeOpt in that it constructs confidence intervals and defines the safe region in the same way, there are distinct differences in how these algorithms work. We illustrate the behavior of SafeOpt and StageOpt starting from a common safe seed in Figure 1. Initially, both algorithms select the same points since they use the same definition of safe expansion. However, StageOpt selects noticeably better optimization points than SafeOpt due its UCB criterion. We leave a more detailed discussion of this behavior for Section 6.

We also re-emphasize that since StageOpt separates the safe optimization problem into safe expansion and utility optimization phases, it is much more amenable to a variety of related settings than SafeOpt. For example, as discussed in detail in the appendix, dueling feedback can easily be incorporated into StageOpt: in the dueling setting, one can simply replace GP-UCB in the utility optimization stage with any kernelized dueling-bandit algorithm, such as KernelSelfSparring (Sui et al., 2017b).

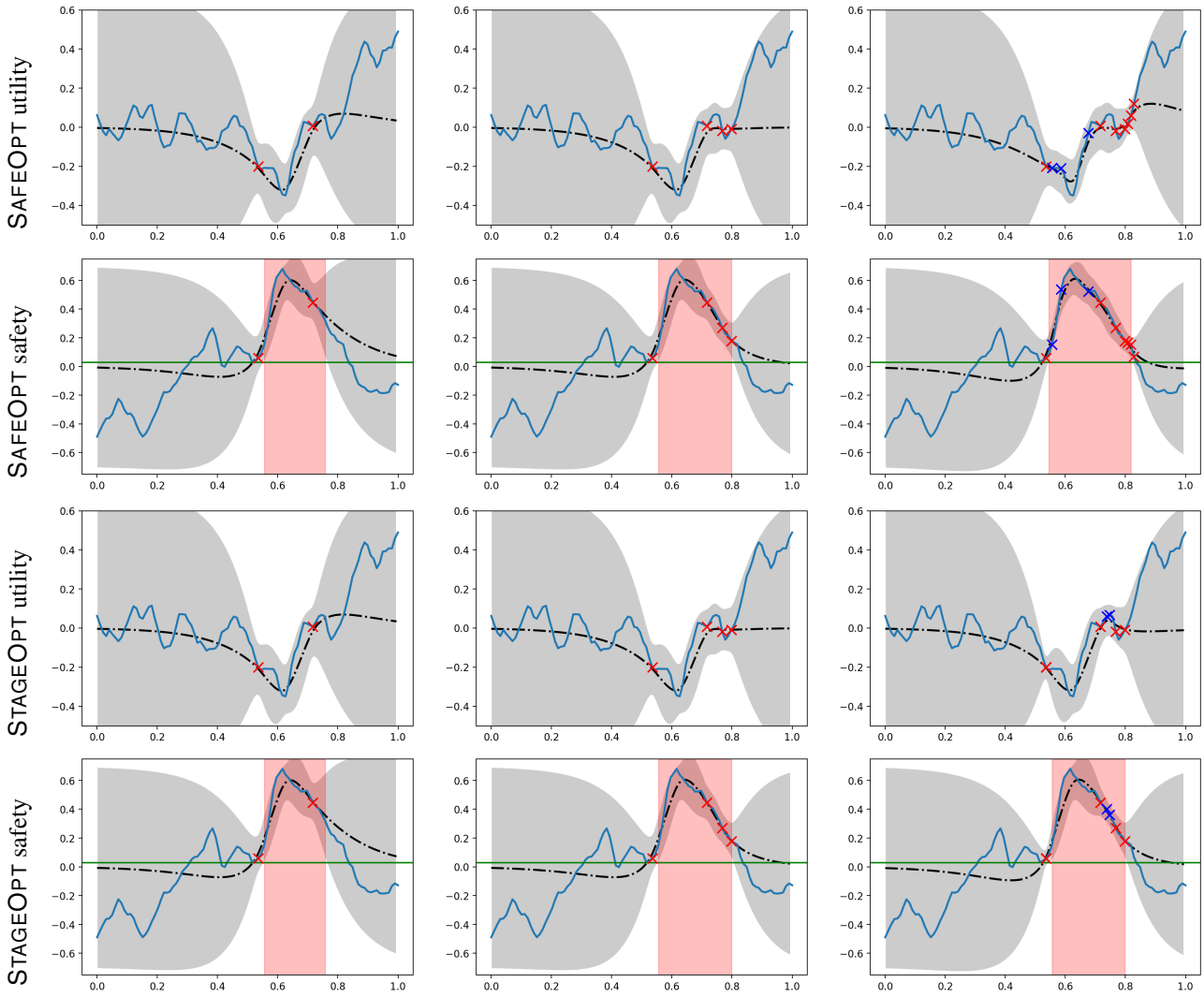
Algorithm 1 StageOpt

Input: sample set D, $i \in \{1, \ldots, n\}$,
GP prior for utility function f,
GP priors for safety functions g_i,
Lipschitz constants L_i for g_i,
safe seed set S_0,
safety threshold h_i,
accuracies ϵ (for expansion), ζ (for optimization).
$C^i_0(\boldsymbol{x}) \leftarrow [h_i, \infty)$, for all $\boldsymbol{x} \in S_0$
$C^i_0(\boldsymbol{x}) \leftarrow \mathbb{R}$, for all $\boldsymbol{x} \in D \setminus S_0$
$Q^i_0(\boldsymbol{x}) \leftarrow \mathbb{R}$, for all $\boldsymbol{x} \in D$
$C^f_0(\boldsymbol{x}) \leftarrow \mathbb{R}$, for all $\boldsymbol{x} \in D$
$Q^f_0(\boldsymbol{x}) \leftarrow \mathbb{R}$, for all $\boldsymbol{x} \in D$
for $t = 1, \ldots T_0$ **do**
  $C^i_t(\boldsymbol{x}) \leftarrow C^i_{t-1}(\boldsymbol{x}) \cap Q^i_{t-1}(\boldsymbol{x})$
  $C^f_t(\boldsymbol{x}) \leftarrow C^f_{t-1}(\boldsymbol{x}) \cap Q^f_{t-1}(\boldsymbol{x})$
  $S_t \leftarrow \bigcap_i \bigcup_{\boldsymbol{x} \in S_{t-1}} \left\{ \boldsymbol{x}' \in D \;\middle|\; \ell^i_t(\boldsymbol{x}) - L_i d(\boldsymbol{x}, \boldsymbol{x}') \geq h_i \right\}$
  $G_t \leftarrow \left\{ \boldsymbol{x} \in S_t \;\middle|\; e_t(\boldsymbol{x}) > 0 \right\}$
  if $\forall i, \epsilon^i_t < \epsilon$ **then**
    $\boldsymbol{x}_t \leftarrow \operatorname{argmax}_{\boldsymbol{x} \in G_t, i \in \{1, \ldots, n\}} w^i_t(\boldsymbol{x})$
  else
    $\boldsymbol{x}_t \leftarrow \operatorname{argmax}_{\boldsymbol{x} \in S_t} \mu^f_{t-1}(\boldsymbol{x}) + \beta_t \sigma^f_{t-1}(\boldsymbol{x})$
  end if
  $y_{f,t} \leftarrow f(\boldsymbol{x}_t) + n_{f,t}$
  $y_{i,t} \leftarrow g_i(\boldsymbol{x}_t) + n_{i,t}$
  Compute $Q_{f,t}(\boldsymbol{x})$ and $Q_{i,t}(\boldsymbol{x})$, for all $\boldsymbol{x} \in S_t$
end for
for $t = T_0 + 1, \ldots, T$ **do**
  $C^f_t(\boldsymbol{x}) \leftarrow C^f_{t-1}(\boldsymbol{x}) \cap Q^f_{t-1}(\boldsymbol{x})$
  $\boldsymbol{x}_t \leftarrow \operatorname{argmax}_{\boldsymbol{x} \in S_t} \mu^f_{t-1}(\boldsymbol{x}) + \beta_t \sigma^f_{t-1}(\boldsymbol{x})$
  $y_{f,t} \leftarrow f(\boldsymbol{x}_t) + n_{f,t}$
  $y_{i,t} \leftarrow g_i(\boldsymbol{x}_t) + n_{i,t}$
  Compute $Q_{f,t}(\boldsymbol{x})$ and $Q_{i,t}(\boldsymbol{x})$, for all $\boldsymbol{x} \in S_t$
end for

5. Theoretical Results

In this section, we show the effectiveness of StageOpt by theoretically bounding its sample complexity for expansion and optimization. The two stages of StageOpt are the expansion of the safe region in search for the total safe region, and the optimization within the safe region.

The correctness of StageOpt relies on the fact that the classification of sets S_t and G_t is sound. While this requires that the confidence bounds C_t are conservative, using bounds that are too conservative will slow down the algorithm considerably. The tightness of the confidence bounds is controlled by parameter β_t in Equation 2. This problem of properly tuning confidence bounds using Gaussian processes in exploration–exploitation trade-off has been studied by Srinivas et al. (2010); Chowdhury & Gopalan (2017). These algorithms are designed for the stochastic multi-armed bandit problem on a kernelized input space without safety constraints. However, their choice of confidence bounds can be generalized to our setting for expansion and optimization. In particular, for our theoretical results to



(a) 2 iterations (b) 4 iterations (c) 10 iterations

Figure 1. Evolution of GPs in SAFEOPT and STAGEOPT for a fixed safe seed; dashed lines correspond to the mean and shaded areas to ± 2 standard deviations. The first and third rows depict the utility function, and the second and fourth rows depict a single safety function. The utility and safety functions were randomly sampled from a zero-mean GP with a Matern kernel, and are represented with solid blue lines. The safety threshold is shown as the green line, and safe regions are shown in red. The red markers correspond to safe expansions and blue markers to maximizations and optimizations. We see that STAGEOPT identifies actions with higher utility than SAFEOPT.

hold it suffices to choose:

$$\beta_t = B + \sigma\sqrt{2(\gamma_{t-1} + 1 + \log(1/\delta))}, \tag{4}$$

where B is a bound on the RKHS norm of f, δ is the allowed failure probability, observation noise is σ-sub-Gaussian, and γ_t quantifies the effective degrees of freedom associated with the kernel function. Concretely,

$$\gamma_t = \max_{|A| \leq t} I(f; \boldsymbol{y}_A)$$

is the maximal mutual information that can be obtained about the GP prior from t samples.

We present two main theorems for STAGEOPT. Theorem 1 ensures convergence to the reachable safe region in the safe expansion stage. Theorem 2 ensures convergence towards optimal utility value within the safe region in the utility optimization stage. Both results are finite time bounds.

Theorem 1. *Suppose safety functions g_i satisfies $\|g_i\|_k^2 \leq B$ and g_i further is L_i-Lipschitz-continuous. $i \in \{1, \ldots, n\}$. Also, suppose $S_0 \neq \varnothing$, and $g_i(\boldsymbol{x}) \geq h_i$, for all $\boldsymbol{x} \in S_0$. Fix any $\epsilon > 0$ and $\delta \in (0, 1)$. Suppose we run the safe region expansion stage of* STAGEOPT *with seed set S_0, with noise n_t to be σ-sub-Gaussian, and $\beta_t = B + \sigma\sqrt{2(\gamma_{t-1} + 1 + \log(1/\delta))}$ with safety function hyper-*

parameters. Let t^* be the smallest positive integer satisfying

$$\frac{t^*}{\beta_{t^*}^2 \gamma_{nt^*}} \geq \frac{C_1 \left(|\bar{R}_0(S_0)| + 1\right)}{\epsilon^2},$$

where $C_1 = 8/\log(1+\sigma^{-2})$. Then, the following jointly hold with probability at least $1-\delta$:

- *$\forall t \geq 1$ and $i \in \{1, \ldots, n\}$, $g_i(\boldsymbol{x}_t) \geq h_i$,*
- *$\forall t \geq t^*$, ϵ-Reachable safe region $R_\epsilon^{T_0}(S_0)$ is guaranteed to be expanded.*

The detailed proof of Theorem 1 is presented in Appendix **??**. In Theorem 1, we count t from the beginning of expansion stage. We choose $T_0 = t^*$ with T_0 the expansion time defined in Section 4. We show that with high probability, the expansion stage of StageOpt guarantees safety, and expands the initial safe region S_0 to an ϵ-reachable set after at most t^* iterations. The size of t^* depends on the largest size of safe region $\bar{R}_0(S_0)$, the accuracy parameters ϵ, ζ, the confidence parameter δ, the complexity of the function B and the parameterization of the GP via γ_t.

The proof is based on the following idea. Within a stage, wherein S_t does not expand, the uncertainty $w_t(\boldsymbol{x}_t)$ monotonically decreases due to construction of G_t. We prove that, the condition $\max_{\boldsymbol{x} \in G_t} w(\boldsymbol{x}) < \epsilon$ implies either of two possibilities: S_t will expand after the next evaluation, i.e., the reachable region will increase, and, hence, the next stage shall commence; or, we have already established all decisions within $\bar{R}_\epsilon(S_0)$ as safe, i.e., $S_t = \bar{R}_\epsilon(S_0)$. To establish the sample complexity we use a bound on how quickly $w_t(\boldsymbol{x}_t)$ decreases.

Theorem 2. *Suppose utility function f satisfies $\|f\|_k^2 \leq B$, $\delta \in (0,1)$, and noise n_t is σ-sub-Gaussian. $\beta_t = B + \sigma\sqrt{2(\gamma_{t-1} + 1 + \log(1/\delta))}$ with utility function hyperparameters. T_1 the time horizon for optimization stage. Fix any $\zeta > 0$. Suppose we run the optimization stage of StageOpt within the expansion stage safe region $R_\epsilon^{T_0}(S_0)$. Let Y be the smallest positive integer satisfying*

$$\frac{4\sqrt{2}}{\sqrt{Y}}\left(B\sqrt{\gamma_Y} + \sigma\sqrt{2\gamma_Y(\gamma_Y + 1 + \log(1/\delta))}\right) \leq \zeta$$

Then with probability at least $1-\delta$, StageOpt finds ζ-optimal utility value: $f(\hat{\boldsymbol{x}}^*) \geq f(\boldsymbol{x}^*) - \zeta$.

The proof of Theorem 2 is presented in Appendix **??**. We count t from the beginning of the optimization stage in Theorem 2. We choose $T_1 = Y$ with T_1 the time horizon of optimization stage. We prove the existence of an ϵ-optimal decision $\hat{\boldsymbol{x}}^*$ within the expansion stage safe region.

Discussion. StageOpt separates safe region expansion and utility function maximization into two distinct stages.

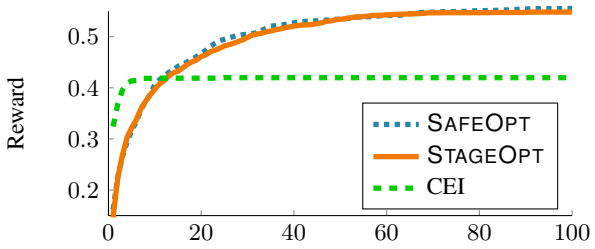


Figure 2. Comparison between SafeOpt, StageOpt, and constrained EI on synthetic data with one safety function. In this simple setting, both SafeOpt and StageOpt perform similarly. In order to achieve the same level of safety guarantees, constrained EI must be much more careful during exploration, and consequently fails to identify the optimal point.

Theorem 1 guarantees ϵ-optimal expansion in the first stage within time horizon T_0. Theorem 2 guarantees ζ-optimal utility value in the second stage within time horizon T_1. Compared to existing approaches which interleave between expansion and optimization, StageOpt does not require any similarity or comparability between safety and utility. In Section 6 we show empirically that StageOpt is more efficient and far more natural for some applications.

6. Experimental Results

We evaluated our algorithm on synthetic data as well as on a live clinical experiment on spinal cord therapy.

Modified StageOpt and SafeOpt. In real applications, it may be difficult to compute an accurate estimate of the Lipschitz constants, which may have an adverse effect on the definition of the safe region and its expansion dynamics. In these scenarios, one can use a modified version of SafeOpt that defines safe points using only the GPs (Berkenkamp et al., 2016b). This modification can be directly applied to StageOpt as well; for clarity, we state the details here. Under this alternative definition, a point is classified as safe if the lower confidence bound of each of its safety GPs lies above the respective threshold:

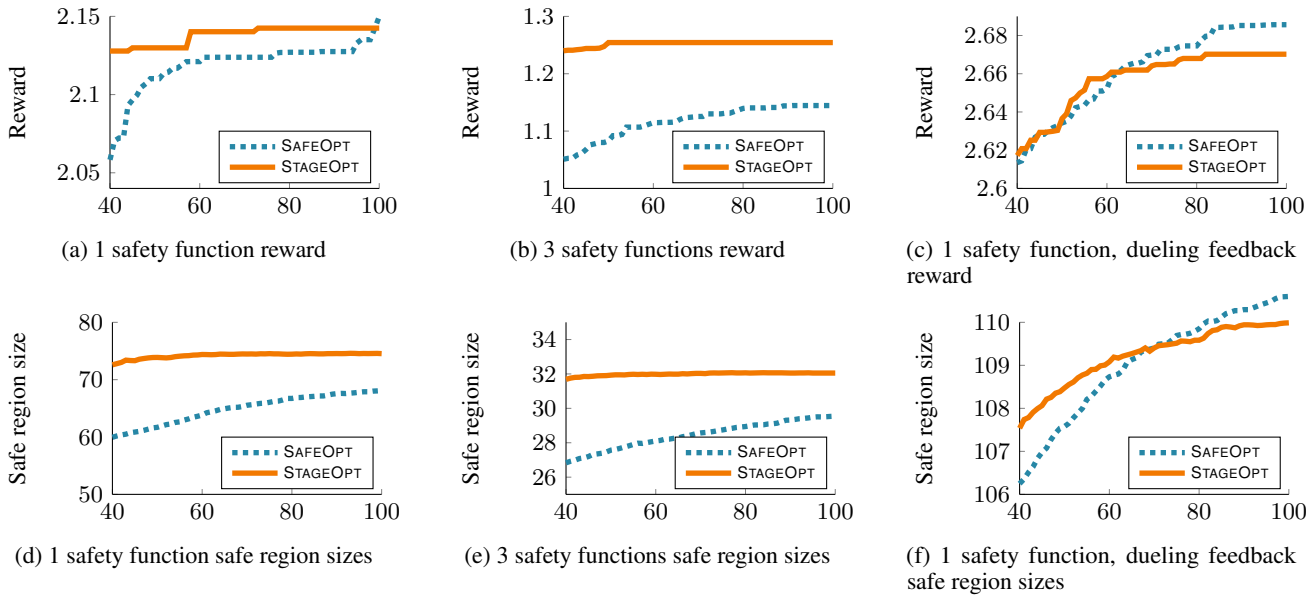
$$S_t = \bigcap_i \left\{\boldsymbol{x} \in D \;\middle|\; \ell_t^i(\boldsymbol{x}) \geq h_i\right\}.$$

A safe point is then an expander if an optimistic noiseless measurement of its upper confidence bound results in a non-safe point having all of its lower confidence bounds above the respective thresholds:

$$e_t(\boldsymbol{x}) := \left|\bigcap_i \left\{\boldsymbol{x}' \in D \setminus S_t \;\middle|\; \ell_{t,u_t(\boldsymbol{x})}(\boldsymbol{x}) \geq h_i\right\}\right|.$$

6.1. Synthetic Data

We evaluated on several synthetic settings with various types of safety constraints and feedback. In each setting, the utility



(a) 1 safety function reward

(b) 3 safety functions reward

(c) 1 safety function, dueling feedback reward

(d) 1 safety function safe region sizes

(e) 3 safety functions safe region sizes

(f) 1 safety function, dueling feedback safe region sizes

Figure 3. Results on three synthetic scenarios. The first row corresponds to the reward and the second row to the growth of the safe region sizes (higher is better for both). In both of these metrics, StageOpt performs at least as well as SafeOpt. For clarity, we omit the first 40 iterations for each setting since the algorithms similarly expand the safe region during that phase.

function was sampled from a zero-mean GP with Matérn kernel ($\nu = 1.2$) over the space $D = [0, 1]^2$ uniformly discretized into 25×25 points. We considered the following safety constraint settings: (i) One safety function g_1 sampled from a zero-mean GP with a Matérn kernel with $\nu = 1.2$. (ii) Three safety functions g_1, g_2, g_3, sampled from zero-mean GPs with length scales 0.2, 0.4 and 0.8.

We set the amplitudes of the safety functions to be 0.1 that of the utility function, and the safety threshold for each safety function g_i to be $\mu_i + \frac{1}{2}\sigma_i$. We define a point x to be a *safe seed* if it satisfies $g_i(x) > \mu_i + \sigma_i$.

We also considered several cases for feedback. For both safety settings, we examined the standard Gaussian noise-perturbed case, with $\sigma^2 = 0.0025$. We also ran experiments for the dueling feedback case and the first safety setting.

Algorithms. As discussed previously, SafeOpt is the only other known algorithm that has similar guarantees in our setting, and serves as the main competitor to StageOpt. In addition, we also compared against the constrained Expected Improvement (CEI) algorithm from Gelbart et al. (2014). Since CEI only guarantees stepwise safety as opposed to over the entire time horizon, we set the safety threshold to be δ/T with $\delta = 0.1$ in order to match our setting. Naturally, with such a stringent threshold, CEI is not very competitive compared to StageOpt and SafeOpt, as seen in Figure 2. In order to adequately distinguish between the latter two algorithms, we omit constrained EI results from all further figures.

Results. In each setting, we randomly sampled 30 combinations of utility and safety functions and ran StageOpt and SafeOpt for $T = 100$ iterations starting from each of 10 randomly sampled safe seeds. For StageOpt, we used a dynamic stopping criterion for the safe expansion phase (i.e. T_0) of when the safe region plateaus for 10 iterations, hard capped at 80 iterations. In these experiments, we primarily used GP-UCB in the utility optimization phase. We also tried two other common acquisition functions, Expected Improvement (EI) and Maximum Probability of Improvement (MPI). However, we observed similar behavior between all three acquisition functions, since our algorithm quickly identifies the reachable safe region in most scenarios.

In Figure 3, for each setting and algorithm, we plot both the growth of the size of the safe region as well as a notion of reward $r_t = \max_{1 \le i \le t} f(\boldsymbol{x}_i)$. Although there is some similarity between the performances of the algorithms, it is evident that StageOpt grows the safe region at least as fast as SafeOpt, while also reaching a optimal sample point more quickly.

6.2. Clinical Experiments

We finally applied StageOpt to safely optimize clinical spinal cord stimulation in order to help tetraplegic patients regain physical mobility. The goal is to find effective stimulation therapies for patients with severe spinal cord injuries without introducing undesirable side effects. For example, bad stimulations could have negative effects on the rehabilitation and are often painful. This application is easily

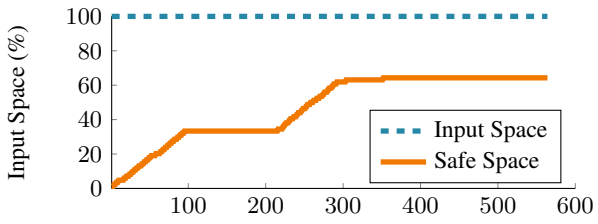


Figure 4. Expansion of the safe region for spinal cord injury therapy. The orange solid line represents the growth of safe region over time, and the blue dashed line the total size of the input space.

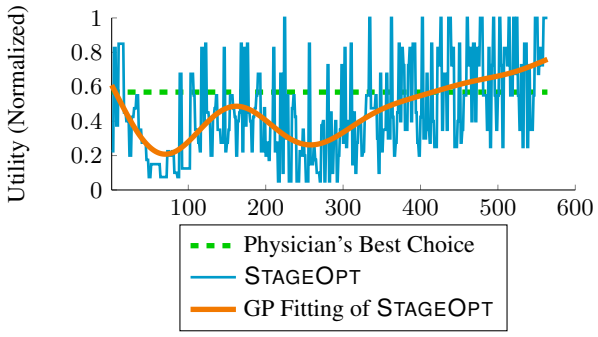


Figure 5. Utilities within the safe region (larger is better). The green dashed line denotes the physician's best choice. The thin blue line shows the utilities of StageOpt at each iteration, and the orange solid line is a GP curve fitting of these utilities.

framed under our problem setting; the chosen configurations must stay above a safety threshold.

A total of 564 therapeutic trials were done with a tetraplegic patient in gripping experiments over 10 weeks. In each trial, one stimulating pattern was generated by the 32-channel-electrode, and was fixed within each trial. For a fixed electrode configuration, the stimulation frequency and amplitude were modulated synergistically in order to find those best for effective gripping. A similar setup was studied in (Sui et al., 2017a). We optimized the electrode patterns with preference-based StageOpt (see Appendix **??**) and performed exhaustive search for stimulation frequency and amplitude over a narrow range.

Results. Figure 4 shows the reachable stimulating patterns by the algorithm under safety constraints. The physicians are confident that the total safe region has been reached between 300 and 400 iterations. In our experiments, StageOpt does not sample any unsafe stimulating patterns.

Figure 5 plots the utility measure of the stimulating pattern at each iteration. The orange solid line is a GP curve fitting of StageOpt (in thin blue). It clearly exceeds the physician's best choice (dotted green line) after around 400 iterations of online experiments. These results demonstrate the practicality of StageOpt to safely optimize in challenging settings, such as those involving live human subjects.

7. Conclusion & Discussion

In this paper, we study the problem of safe Bayesian optimization, which is well suited to any setting requiring safe online optimization such as medical therapies, safe recommender systems, and safe robotic control. We proposed a general framework, StageOpt, which is able to tackle non-comparable safety constraints and utility function. We provide strong theoretical guarantees for StageOpt with safety functions and utility function sampled from Gaussian processes. Specifically, we bound the sample complexity to achieve an ϵ-safe region and ζ-optimal utility value within the safe region. The whole sampling process is guaranteed to be safe with high probability.

We compared StageOpt with classical Bayesian optimization methods and state-of-the-art safe optimization algorithms. We evaluated multiple cases such as single safety function, multiple safety functions, real-valued utility, and dueling-feedback utility. Our extensive experiments on synthetic data show that StageOpt can achieve its theoretical guarantees on safety and optimality. Its performance on safe expansion is among the best and utility maximization outperforms the state-of-the-art.

This result also provides an efficient tool for online optimization in safety-critical applications. For instance, we applied StageOpt with dueling-feedback utility function on the gripping rehabilitation therapy for tetraplegic patients. Our live clinical experiments demonstrated good performance a real human experiment. The therapies proposed by StageOpt outperform the ones suggested by experienced physicians.

There are many interesting directions for future work. For instance, we assume a static environment that does not evolve in response to the actions taken. In our clinical application, this implies assuming that the patients' condition and response to stimuli do not improve over time. Moving forward, it would be interesting to incorporate dynamics into our setting, which would lead to the multi-criteria safe reinforcement learning setting (Moldovan & Abbeel, 2012; Turchetta et al., 2016; Wachi et al., 2018).

Another interesting direction is developing theoretically rigorous approaches outside of using Gaussian processes (GPs). Although highly flexible, GPs require a well-specified prior and kernel in order to be effective. While one could use uniformed priors to model most settings, such priors tend to lead to very slow convergence. One alternative is to automatically learn a good kernel (Wilson et al., 2016). Another approach is to assume a low-dimensional manifold within the high-dimensional uniformed kernel (Djolonga et al., 2013), which could also speed up learning.

Acknowledgements
This research was also supported in part by NSF Awards #1564330 & #1637598, JPL PDF IAMS100224, a Bloomberg Data Science Research Grant, and a gift from Northrop Grumman.

References

Audibert, J.-Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *Conference on Learning Theory (COLT)*, 2010.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research (JMLR)*, 2002.

Berkenkamp, F., Krause, A., and Schoellig, A. P. Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. Technical report, arXiv, February 2016a.

Berkenkamp, F., Schoellig, A. P., and Krause, A. Safe controller optimization for quadrotors with gaussian processes. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, pp. 491–496, May 2016b.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5:1–122, 2012.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 201–208, 2008.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pp. 844–853, 2017.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Djolonga, J., Krause, A., and Cevher, V. High-dimensional gaussian process bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1025–1033, 2013.

Garcia, J. and Fernandez, F. Safe exploration of state and action spaces in reinforcement learning. *Journal of Machine Learning Research*, 2012.

Gelbart, M. A., Snoek, J., and Adams, R. P. Bayesian optimization with unknown constraints. *Uncertainty in Artificial Intelligence (UAI)*, 2014.

Gillula, J. and Tomlin, C. Guaranteed safe online learning of a bounded system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.

Gotovos, A., Casati, N., Hitz, G., and Krause, A. Active learning for level set estimation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Hans, A., Schneegaß, D., Schäfer, A., and Udluft, S. Safe exploration for reinforcement learning. In *ESANN*, 2008.

Hernández-Lobato, J. M., Gelbart, M. A., Adams, R. P., Hoffman, M. W., and Ghahramani, Z. A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 2016.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing (STOC)*, pp. 681–690. Association for Computing Machinery, Inc., May 2008.

Moldovan, T. and Abbeel, P. Safe exploration in markov decision processes. In *International Conference on Machine Learning (ICML)*, 2012.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Schreiter, J., Nguyen-Tuong, D., Eberts, M., Bischoff, B., Markert, H., and Toussaint, M. Safe exploration for active learning with gaussian processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 133–149. Springer, 2015.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, 2010.

Sui, Y., Gotovos, A., Burdick, J. W., and Krause, A. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning (ICML)*, 2015.

Sui, Y., Yue, Y., and Burdick, J. W. Correlational dueling bandits with application to clinical treatment in large decision spaces. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017a.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling bandits with dependent arms. In *Uncertainty in Artificial Intelligence (UAI)*, 2017b.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Turchetta, M., Berkenkamp, F., and Krause, A. Safe exploration in finite markov decision processes with gaussian processes. In *Neural Information Processing Systems (NIPS)*, pp. 4305–4313, December 2016.

Wachi, A., Sui, Y., Yue, Y., and Ono, M. Safe exploration and optimization of constrained mdps using gaussian processes. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 370–378, 2016.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *International Conference on Machine Learning (ICML)*, pp. 1254–1262, 2016.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.