One-Shot Optimal Design for Gaussian Process Analysis of Randomized Experiments

Jelena Markovic-Voronov
Meta
jmarkovic@meta.com

Qing Feng
Meta
qingfeng@meta.com

Eytan Bakshy
Meta
ebakshy@meta.com

Abstract

Bayesian optimization provides a sample-efficient approach to optimize systems that are evaluated with randomized experiments, such as Internet experiments (A/B tests) and clinical trials. Such evaluations are often resource- and time-consuming in order to measure noisy and long-term outcomes. Thus, the initial randomized design, i.e., determining the number of test groups and their sample sizes, plays a critical role in building an accurate Gaussian Process (GP) model to optimize efficiently and decreasing experimentation cost. We develop a simulation-based method with meta-learned priors to decide the optimal design for the initial batch of GP-modeled randomized experiments. The meta-learning is performed on a large corpus of randomized experiments conducted at Meta, obtaining sensible GP priors for simulating across different designs. The one-shot optimal design policy is derived by training a machine learning model with simulation data to map experiment characteristics to an optimal design. Our evaluations show that our proposed optimal design significantly improves resource-efficiency while achieving a target GP model accuracy.

1 Introduction

Randomized experiments are the gold-standard method for Internet firms to evaluate the performance of changes to their service [Kohavi and Longbotham, 2017]. Running these experiments is often resource- and time-consuming since it requires using large sample sizes to detect the effects of high-variance outcomes and running the experiments over the long periods of time to observe long-term effects. Especially when experimenting with high-dimensional action spaces, such as continuous parameters in a ranking policy [Letham and Bakshy, 2019], the prototypical practice that runs a parallel set of treatments chosen adhocly and selects the best one based on observed outcomes is not sustainable.

Bayesian optimization (BayesOpt) provides a sample-efficient approach for optimizing black-box functions with expensive evaluations and is widely used to tune online systems that are evaluated with A/B tests [Letham et al., 2019, Feng et al., 2020, Liu et al., 2022]. The space of possible parameters of the underlying function or system to be optimized is the input space to BayesOpt, also referred to as the input parameter space. The underlying function's outcomes are only observed at several parameter values called *arms* or *test groups*. The observed arms together with the corresponding observed function outcomes represent the input and the output, respectively, of the Gaussian Process (GP) model used in BayesOpt. This model allows predicting the underlying hard-to-evaluate function values anywhere in the input parameter space along with the uncertainties. The BayesOpt framework then uses the acquisition function optimization to suggest the most promising parameter values at which to evaluate the underlying function next. The acquisition function uses the fitted GP model to balance exploration, choosing parameter values with high variance, and exploitation, choosing the parameter values with high predicted values for the underlying function.

6th Workshop on Meta-Learning at NeurIPS 2022, New Orleans.

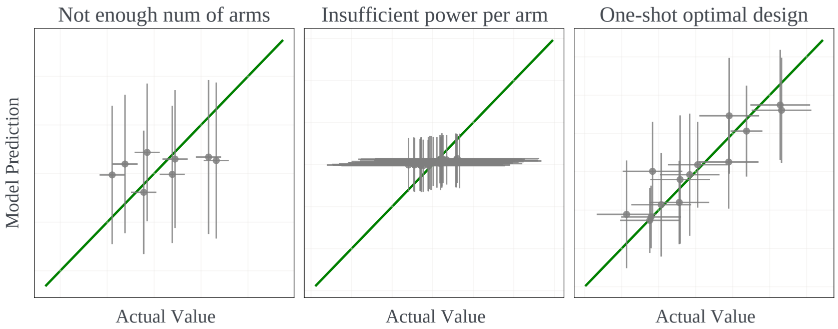

Figure 1: Leave-one-out cross validation predictions (mean, and 95% posterior predictive interval) for the GP fits to the results of a 3-d real-world experiment under three different initial designs: the first one on the left with not enough arms to cover the input parameter space but with large sample size per arm, the second one with more arms but less sample size per arm so much more noise per arm and the third one is somewhere in between. We see that the first two are sub-optimal as they lead to poor out-of-sample predictions. The last one allows us to predict the out-of-sample output well thus that design is preferable.

Considering the long evaluation time via randomized experiments, batch BayesOpt is used in practice to take full advantage of the experimental resources and test online multiple parameter values simultaneously. It starts with an initial batch with arms generated by space-filling designs, e.g., Latin hypercube, builds a Gaussian process (GP) model based on previous outcomes, and uses an acquisition function defined from this GP model to select a batch of arms to evaluate next [Jones et al., 1998, Snoek et al., 2012]. The process is then repeated until the optimal, or close to optimal, parameter value is achieved.

Although BayesOpt is sequential in nature, the experiment design of the initial random batch is critical for building an accurate GP model to generate good candidates sequentially and decreasing experimentation cost. Having enough unique arms is known to be important to model the underlying response surfaces, e.g., empirical rule $n = 10d$ such that the number of arms is 10 times the input dimension, which has been proposed and studied by Loeppky et al. [2009]. In addition, being able to isolate signal from noise is known to be important to the GP model accuracy and BayesOpt performance [Binois et al., 2019]. Thus, each arm requires a certain amount of samples (adequate statistical power) to detect effects for noisy outcomes. Figure 1 demonstrates the importance of the initial design to achieve a good GP model fit and optimization performance. This brings an important design question in the face of noisy evaluations: how to decide the number of unique arms and sample size per arm in order to balance exploring response surfaces with more arms while keeping an adequate signal-to-noise ratio with the experimentation resources, i.e. total sample size, in mind?

There are two main challenges in tackling this design question: 1) In practice, the prior knowledge of the underlying response surfaces, e.g., the variance of outcome and the distribution of GP hyperparameters, is typically unknown. Without such a priori information, one can hardly choose a design, as it largely depends on the effective dimensionality of the input parameter space and the smoothness of the response surfaces. Sequential design strategies that iterate between design and learning has been proposed and studied as a remedy [Chen and Zhou, 2017]. However, this does not solve the one-shot design that is desired in the batch BayesOpt, especially when evaluations take a long time. 2) Solving a full batched version of selecting locations of unique arms and the numbers of samples allocated for each arm is not computationally tractable as discussed in Binois et al. [2019].

In this paper, we propose a simulation-based method with meta-learned priors to decide on the optimal design for the initial batch of GP-modeled randomized experiments. The contributions of this work are: 1) We frame this one-shot design problem as a constrained optimization problem depending on the use case: a) obtaining the best GP model prediction accuracy given a fixed sample budget or b) minimizing the total sample size while keeping the GP model accuracy above a certain accuracy threshold. In both cases, we use a random uniform location for arms and use the optimization to decide on the number of arms and the sample size per arm. 2) We conduct a meta-analysis of a large corpus of historical randomized experiments at Meta to generate empirical GP priors. The meta-learned GP priors capture and transfer the information to facilitate the one-shot design for experiments with different characteristics, e.g., parameter dimensions and outcomes. 3) We devise a simulation-based approach to take the complexity of input-output relationship, i.e., the true underlying response surface and the noise, into account through sampling GPs from meta-learned priors. 4) We propose an easy-to-deploy one-shot design policy by building a machine learning model with the simulated datasets.

2 Methodology

2.1 One-Shot Optimal Design for Gaussian Processes

Let $\mathbf{Y} = [y_1, \dots, y_n]^T$ denote the outcomes of n equally sized arms $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ from a randomized experiment with total sample size N. We assume the observations are from the initial batch, implying the input arms $\mathbf{x}_i$ are chosen to be uniformly at random in the pre-specified input parameter space $\mathcal{X} \subset \mathbb{R}^d$. We assume the model $y_i = f(\mathbf{x}_i) + \epsilon_i$, $i = 1, \dots, n$, where the function $f : \mathcal{X} \to \mathbb{R}$ is the true response surface and $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$ is the observation noise in which σ_e^2 depends on the sample size per arm $\sigma_e^2 \propto \frac{n}{N}$. For simplicity and comparability, the outcomes are standardized to zero mean and unit variance; the observation noise is scaled accordingly. The input variables $\mathbf{X} = [\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T]$ are scaled to $[0, 1]^d$, and each arm $\mathbf{x}_i$ is a d-dimensional vector.

We assume $f \sim \mathcal{GP}(0, K)$ to be a zero-mean GP with the covariance matrix K, parameterized by lengthscales $\theta = (\theta_1, \cdots, \theta_d)$ and outputscale σ_f^2, represented as $K(\mathbf{x}_i, \mathbf{x}_j) = \text{Cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = \sigma_f^2 \cdot K_\theta(\mathbf{x}_i, \mathbf{x}_j)$. The covariance function K_θ is taken from the Matérn family with parameter $\gamma = 2.5$, as suggested in Rasmussen [2003]. The kernel parameters θ and σ_f^2 are estimated through optimizing the marginal log likelihood and σ_e^2 is assumed to be known.

In one-shot optimal design, the goal is to achieve good GP prediction accuracy, which is evaluated on out-of-sample mean squared error (MSE). The locations of arms are decided by a random uniform design. The decision parameters are thus the number of arms n and the total experiment sample size N. The problem can be formulated as a constrained optimization problem based on the use case: a) minimize out-of-sample MSE given a fixed sample size budget B, i.e., $\min_{n,N} \text{MSE}_f(n, N)$ s.t. $N \leq B$; or b) minimize sample size N to achieve a target out-of-sample MSE denoted as τ, i.e., $\min_{n,N} N$ s.t. $\text{MSE}_f(n, N) \leq \tau$.

2.2 Empirical Priors from Meta-learning

One main challenge of one-shot optimal design is that prior distributions of variance σ_f^2, noise σ_e^2, and distribution of kernel hyper-parameters θ are unknown. To optimize MSE that depends on these unknown priors, we perform a meta-analysis based on a large corpus of historical randomized experiments at Meta to obtain empirical priors. For each experiment, a GP with fixed (known) noise model is fitted on the outcomes from the initial batch of observations. A leave-one-out cross validation is performed to filter out estimates with poor fitted GP. This meta-analysis generates empirical prior distributions of kernel hyper-parameters and outcome noise across a large pool of experiments with different parameter dimensions, conducted in different product areas etc. Figure 2 visualizes the distribution of estimated lengthscales segmented based on input parameter dimension d. It can be seen that the effective dimension does not linearly increase with d: higher dimensional spaces tend to be sparse, i.e., with less important parameters impacting the outcome. Learning these function priors by meta-learning plays an important role in guiding efficient experiment designs.

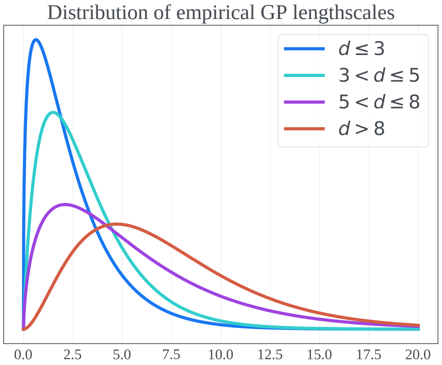

Figure 2: Empirical lengthscales distribution from meta-learning. Each density curve corresponds to the lengthscale distribution segmented based on the input parameter dimension d.

2.3 Surrogate Data Based on Meta-Learned Priors

Directly solving the one-shot optimal design is not computationally tractable. Therefore, we consider a simulation-based algorithm. For each design of n arms and N sample size, simulation is used to obtain an out-of-sample MSE for a given GP prior obtained from meta-analysis. Intuitively, this allows us to measure the GP model fit of realistic response surfaces learned from historical experiments using different experiment designs.

We simulate realistic "true" response surfaces by sampling GPs from the GP distributions obtained from the meta-learning. In each simulation run, given a (true) GP f from meta-learning, we evaluate a grid of designs (n, N). For each design, n training inputs $\mathbf{X}$ are generated by random uniform design and training outputs $f(\mathbf{X})$ are perturbed with noise randomly sampled from Gaussian distribution with variance being $\frac{n}{N}s^2$, in which s^2 is the observed outcome variance. We fit a GP on the simulated training data and compute the out-of-sample MSE, measuring the prediction accuracy of the fitted GP. We run this simulation multiple times for each design and compute the average of out-of-sample MSE. From the simulation, we measure the performance (out-of-sample MSE) of different designs for a variety of experiments with different characteristics. Algorithm 1 in the Appendix describes the process of generating the simulated data in more detail.

2.4 Optimal Design Policy

Based on the simulation data, we derive an optimal design policy by training a supervised machine learning algorithm to suggest (n, N) based on experiments' characteristics. In particular, we build a logistic regression model to predict the goodness of model fit (one if out-of-sample MSE is below the accuracy threshold τ and zero otherwise) using design (n, N) and additional experiment-specific characteristics as the features of the logistic regression.

In our case, we use the dimension d of the input parameter space of the experiment and the signal-to-noise ratio (SNR) s_f of the outcome metric that the experiment is trying to optimize as additional experiment characteristics. We derive the SNR metric estimate from historical randomized experiments tracking that particular metric. These two additional experiment characteristics d and s_f are both naturally important for predicting the goodness of the model fit given (n, N). For example, we expect the experiments with larger dimension d to require larger n for their model to be fit well as they will need more arms to cover the input parameter space. Similarly, we expect the experiments with larger s_f to allow for smaller N for the good model fit as less noisy outcomes require smaller sample size for the good model fit.

Other supervised machine learning models, e.g., a neural network, and additional experiment features, e.g., product area, can also be used. We decided to use the logistic regression for its interpretability. Logistic regression allows the predictions to follow the overall expected relationship between its input features and the predictions, e.g., output larger probabilities of a good model fit for smaller d, larger s_f and larger N.

For future experiments, the model can predict the performance of different designs and decide which one is optimal. Given the experiment characteristics d and s_f, the logistic regression predicts the probability of a good model fit for different designs across the whole (n, N) grid. Based on these predicted probabilities of a good model fit, one can pick the final design depending on their use case, i.e., depending on how they formulated the constrained optimization problem in Section 2.1. The winning design can be: a) the one maximizing the probability of a good model fit among all the designs with sample size below the given sample size budget B or b) the one having the smallest sample size among the designs with the predicted probability of a good model fit above a certain threshold.

3 Results and Discussions

Illustrative real-world experiments. Here we illustrate the simulation-based one-shot design with three real-world experiments with 2-dimensional parameter spaces. Figure 3 shows the out-of-sample MSE of different design choices for these experiments. The left plot shows the results of outcomes with lowest SNR, from which we see that increasing the number of arms is always harmful for such extremely noisy outcomes. Increasing statistical power per arm is preferable in order to isolate signal from noise. On the other hand, for outcomes with high SNR (right plot), having more arms is more

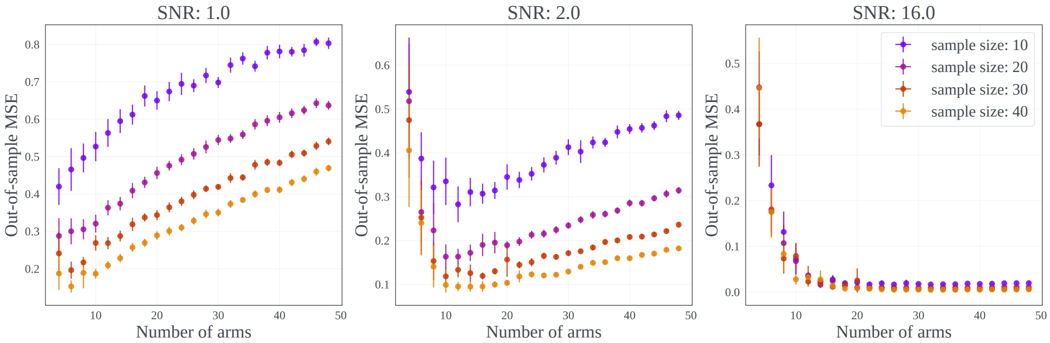

Figure 3: Out-of-sample MSE obtained from simulation for different designs i.e number of arms (n) and sample sizes (N, given here as a percentage of the population). The simulations are based on three 2-dimensional real-world randomized experiments. The signal-to-noise ratios of these three experiment outcome metrics vary from low (left) to high (right). Different sample sizes are indicated by different colors.

useful to achieve better GP model fit since each arm is adequately powered and the marginal gain from increasing the sample size decreases.

For outcomes with average SNR, the middle plot shows that the optimal design choice balances exploring new arms versus keeping enough power per arm. When there are a few arms, increasing the number of arms n improves out-of-sample MSE; and there is a sharp transition that once n passes a point, having more arms hurts the GP model fit.

Evaluation with a large corpus of randomized experiments. We conduct an evaluation of proposed one-shot optimal design on a large corpus of real-world randomized experiments. The evaluation is done in a train-test split fashion so that the optimal design policy is trained on training data and evaluated on a separate set of experiments. We compare our method against two different design strategies: 1) historical design used by experimenters; 2) the empirical rule $n = 10d$ recommended by Chapman et al. [1994] and N being historical (observed) sample size. The design is decided by finding a lowest N that achieves a target GP model accuracy and the performance is evaluated based on recommended sample size. The evaluation shows that our method is close to 3X more resource-efficient. See Table 2 in the Appendix for details.

4 Conclusion

This work develops a practical algorithm for the design of randomized experiments concurrently tuning multiple parameters via Bayesian optimization, leading to a more efficient use of experiment resources. A meta-analysis of a large pool of real randomized experiments allows us to learn the sensible prior distributions of the data, transferring information from historical to new experiments. This information is further boosted by simulating more datasets across a grid of designs. The final step of the method is a policy recommending an optimal design for brand new experiments, including the number of arms and their sizes based on the experiments' characteristics.

Future work will include experiment design based on other quantities of interest besides MSE and the design of high-dimensional experiments besides $d = 12$, which is the largest number of tuning parameters of any experiment in this work. Other possibilities include adding other experiment characteristics into the optimal design policy such as the specific information of the experiment parameters besides their total number.

References

Mickaël Binois, Jiangeng Huang, Robert B Gramacy, and Mike Ludkovski. Replication or exploration? sequential design for stochastic simulation experiments. *Technometrics*, 61(1):7–23, 2019.

William L Chapman, William J Welch, Kenneth P Bowman, Jerome Sacks, and John E Walsh. Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *Journal of Geophysical Research: Oceans*, 99(C1):919–935, 1994.

Xi Chen and Qiang Zhou. Sequential design strategies for mean response surface metamodeling via stochastic kriging with adaptive exploration and exploitation. *European Journal of Operational Research*, 262(2):575–585, 2017.

Qing Feng, Benjamin Letham, Hongzi Mao, and Eytan Bakshy. High-dimensional contextual policy search with unknown context rewards using Bayesian optimization. In *Advances in Neural Information Processing Systems 33*, NeurIPS, 2020.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

Benjamin Letham and Eytan Bakshy. Bayesian optimization for policy search via Online-Offline experimentation. April 2019.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, Eytan Bakshy, et al. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

Sulin Liu, Qing Feng, David Eriksson, Benjamin Letham, and Eytan Bakshy. Sparse bayesian optimization. *arXiv preprint arXiv:2203.01900*, 2022.

Jason L Loeppky, Jerome Sacks, and William J Welch. Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51(4):366–376, 2009.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25, 2012.

A Appendix

A.1 One-shot Optimal Design Algorithm Details

Starting with a list of real historical experiments, Algorithm 1 describes the steps for getting the surrogate experiment datasets. Function `Sample-GP` draws new training and test datasets from a given GP with the noise level computed based on the given SNR adjusted to correspond to the given sample size per arm. By drawing multiple datasets, function `Single-Design-MSE` computes the average out-of-sample MSE. Finally, `Surrogate-Data` fits a GP to each historical real experiment and uses it as the ground truth for simulating more experiment datasets from a grid of designs.

Algorithm 1 Simulation Algorithm

```
function SAMPLE-GP($f, s_f, n, N$)
    Sample input $\mathbf{X}_{n\times d}$ based on Sobol design and out-of-sample points $\mathbf{X}^\star{}_{m\times d} \overset{i.i.d.}{\sim} \mathbf{U}(0,1)$
    Sample ground truth GP $f([\mathbf{X}, \mathbf{X}^\star])$
    Set the noise level: $\sigma_\epsilon = s_f * N/n$
    Perturb input points with noise $\epsilon, \epsilon^\star \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$: $\mathbf{Y} = f(\mathbf{X}) + \epsilon$ and $\mathbf{Y}^\star = f(\mathbf{X}^\star) + \epsilon^\star$
    return $\mathbf{X}, \mathbf{Y}, \sigma_\epsilon, \mathbf{X}^\star, \mathbf{Y}^\star$
function SINGLE-DESIGN-MSE($f, s_f, n, N$)
    Initialize average out-of-sample MSE: $avg$(MSE) $\leftarrow 0$
    for i in $[T]$ do
        $\mathbf{X}, \mathbf{Y}, \sigma_\epsilon, \mathbf{X}^\star, \mathbf{Y}^\star \leftarrow$ Sample-GP($f, s_f, n, N$)
        Fit a fixed noise GP $\hat{f}$ using training inputs $\mathbf{X}, \mathbf{Y}, \sigma_\epsilon$
        $\hat{f}(\mathbf{X}^\star) \leftarrow$ Predict on testing inputs $\mathbf{X}^\star$
        Compute out-of-sample MSE: MSE$_i = \|\hat{f}(\mathbf{X}^\star) - \mathbf{Y}^\star\|_2^2/m$
        Update $avg$(MSE) $\leftarrow avg$(MSE) + MSE$_i/T$
    return $avg$(MSE)
procedure SURROGATE-DATA(List of historical experiment data $\{\mathbf{X}_i, \mathbf{Y}_i, \sigma_i\}, i = 1, \ldots, E$)
    for i in $[E]$ do
        Initially empty list of surrogate data results: $SD = \emptyset$
        Fit a fixed noise GP $\hat{f}$ to $(\mathbf{X}_i, \mathbf{Y}_i, \sigma_i)$
        Estimate signal-to-noise ratio $\hat{s}_f$
        for $n$ in [min # of arms, max # of arms] do
            for $N$ in [min sample size, max sample size] do
                $avg$(MSE) $\leftarrow$ Single-Design-MSE($\hat{f}, \hat{s}_f, n, N$)
                Append ($\hat{s}_f, d, n, N, avg$(MSE)) to $SD$
    return $SD$
```

A.2 Results Evaluation Details

Figure 4 shows the output of the logistic regression used to model probability of a good model fit for ($s_f = 5, d = 8$). As a final step of the optimal design algorithm, we provide an *optimal design policy* recommending a design corresponding to the smallest sample size for which the predicted probability is above a given threshold (taken to be 0.8 in our algorithm). Table 1 in the Appendix shows the experiment design recommendations for other experiment dimensions.

Table 1 shows the recommended number of arms and the total experiment sample size as a percentage of population given by the proposed optimal design policy for a given d and with $s_f = 5$. For a brand new experiment optimizing a given objective metric, we use historical randomized experiment data tracking that metric to estimate its SNR. For a metric for which there is no prior SNR estimate one can use the overall median SNR of all historical metrics with data as an input into the logistic regression and thus get an optimal design policy for the new experiment with that objective metric.

Table 2 shows the median number of arms and the total experiment sample size across several designs from different experiment design algorithms. We see that our proposed algorithm needs 2.75 times

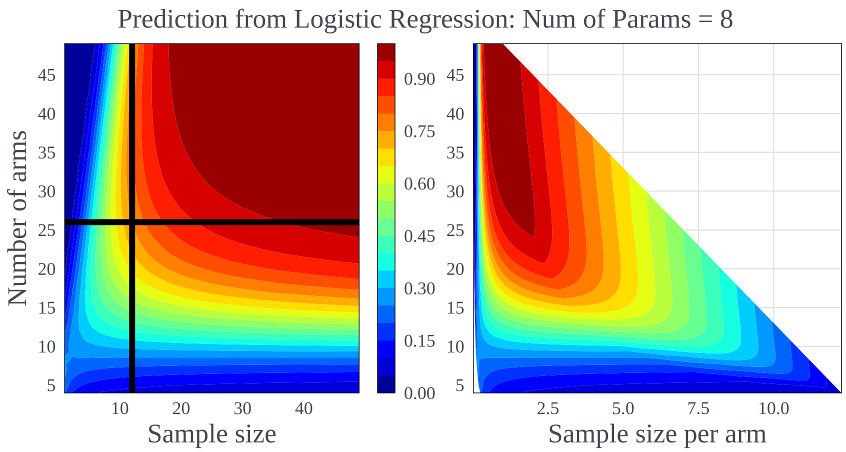

Figure 4: The predicted probabilities of a good model fit based on logistic regression used on surrogate data. The probabilities are depicted for varying number of arms and sample size and fixed number of parameters (eight) and fixed signal-to-noise metric ratio (two). The thick straight lines indicate the winning design.

d	n	N
1	9	5
2	12	6
3	15	7
4	17	8
5	19	9
6	22	10
7	24	11
8	26	12
9	29	13
10	31	14
11	33	15
12	35	16

Table 1: Optimal design policy recommendations for the number of arms (n) and the sample size (N) as a percentage of population given the number of parameters (d) and metric signal-to-noise ratio ($s_f = 5$).

less sample size than the other algorithms while the average MSE for each algorithm is kept below the set threshold $\tau = 0.2$, thus the proposed algorithm is the most resource-efficient.

Design	n	N
Recommended by the optimal sizing alg.	13	4
Observed design	22	11
Observed sample size, num. of arms $= 10 * d$	30	11

Table 2: The median of the number of arms and the total experiment sample size of different experiment designs.