Experimental Design for Optimization of Orthogonal Projection Pursuit Models

Mojmír Mutný
ETH Zurich
MOJMIR.MUTNY@INF.ETHZ.CH

Johannes Kirschner
ETH Zurich
JKIRSCHNER@INF.ETHZ.CH

Andreas Krause
ETH Zurich
KRAUSEA@ETHZ.CH

Abstract

Bayesian optimization and kernelized bandit algorithms are widely used techniques for sequential black box function optimization with applications in parameter tuning, control, robotics among many others. To be effective in high dimensional settings, previous approaches make additional assumptions, for example on low-dimensional subspaces or an additive structure. In this work, we go beyond the additivity assumption and use an orthogonal projection pursuit regression model, which strictly generalizes additive models. We present a two-stage algorithm motivated by experimental design to first decorrelate the additive components. Subsequently, the bandit optimization benefits from the statistically efficient additive model. Our method provably decorrelates the fully additive model and achieves optimal sublinear simple regret in terms of the number of function evaluations. To prove the rotation recovery, we derive novel concentration inequalities for linear regression on subspaces. In addition, we specifically address the issue of acquisition function optimization and present two domain dependent efficient algorithms. We validate the algorithm numerically on synthetic as well as real-world optimization problems.

1 Introduction

Experimental design (Chaloner and Verdinelli 1995) is a branch of statistics for optimally acquiring information in order to reduce uncertainty about a quantity of interest. The related fields of kernelized bandit algorithms and Bayesian optimization (Srinivas et al. 2010; Shahriari et al. 2016) seek to adaptively query a noisy function in order to identify an input with maximum value. Such a procedure can also be viewed as an adaptive experimental design for the optimal input. Bandit optimization and experimental design in general have been successfully used in many applications such as clinical trials (Lizotte 2008), parameter tuning (Kirschner et al. 2019) and reinforcement learning (Gopalan, Mannor, and Mansour 2014).

In this work, our objective is to maximize an unknown function that is assumed to be *additive up to a rotation*, using only noisy function evaluations. The additivity assumption has been placed in previous work in order to gain statistical and computational efficiency on high dimensional domains. We significantly generalize this setup by allowing the additivity to occur in an *unknown* rotated coordinate system.

A Bayesian approach to estimate the rotation is to treat the projection matrix as a hyper-parameter of a Gaussian process model and maximize the evidence (marginalized likelihood) to determine a candidate rotation (Rasmussen and Williams 2006). However, this approach has two drawbacks that severely limit its performance in practice. First, the evidence is a non-convex function on the space of rotation matrices, that can be heuristically optimized using manifold optimization techniques (Edelman, Arias, and Smith 1999). In practice we observed that this approach suffers from local optima and obtaining a good solution is difficult. Second, it is a largely open problem how to pick the initial data set in the first place, yet importantly, the sample set determines the quality of the solution significantly.

In this work we present a *principled* two-stage algorithm based on experimental design. We first derive a design to efficiently estimate the unknown rotation of the coordinate system by estimating the Hessian of the black box function at a single point with optimal experimental design. We control the error in the estimate of the rotation by evaluating the function at carefully chosen points and consequently *provably* de-correlate the parameters. Finally, we proceed to efficiently optimize the de-correlated function by using a bandit algorithm designed for additive functions.

Summary of Contributions

- We analyze a novel two stage bandit algorithm for estimating and optimizing the *orthogonal projection pursuit model*, which generalizes additive models.
- We reduce the estimation of the rotation to linear regression on a subspace; and we derive novel concentration inequalities for subspace estimation that are of independent interest. These allow us to carefully trade-off two counteracting errors - the local approximation error and the error due to noisy observations.
- We derive a sample complexity bound for simple regret that is asymptotically optimal in the number of queried points and scales polynomially in the dimension of the input parameter.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We explicitly address the problem of acquisition function maximization and provide tractable optimization algorithms.

Related Work Kernelized Bandits and Bayesian optimization are a family of algorithms that use frequentist confidence intervals or probabilistic models to determine the next evaluation point (Mockus 1982; Shahriari et al. 2016). Many variants appear in literature; including GP-UCB (Srinivas et al. 2010), Thompson Sampling (Chowdhury and Gopalan 2017), Expected Improvement (Mockus 1982); and recently information theoretic criteria such as MVES or IDS (Wang and Jegelka 2017; Kirschner and Krause 2018). Scaling Bayesian optimization to high dimensional setting has been considered recently, as many of the commonly used kernels suffer from the curse of dimensionality. Hence, to make the problem tractable, most approaches make structural assumptions on the function such as additivity (Rolland et al. 2018; Mutný and Krause 2018) or a low-dimensional active subspace (Djolonga, Krause, and Cevher 2013; Wang et al. 2016; Kirschner et al. 2019). Beyond additivity, a stronger assumption is made by Li et al. (2016) which assumes a version of projected pursuit regression (PPR). Their decorrelating algorithm is based on the EM algorithm of Saati, Cunningham, and Gilboa (2013) without provable guarantees. A provable algorithm for active learning is demonstrated in (Hemant and Cevher 2012) which considers a similar model to ours utilizing different techniques. More recently, Zhang, Li, and Su (2019) consider a more general model than PPR.

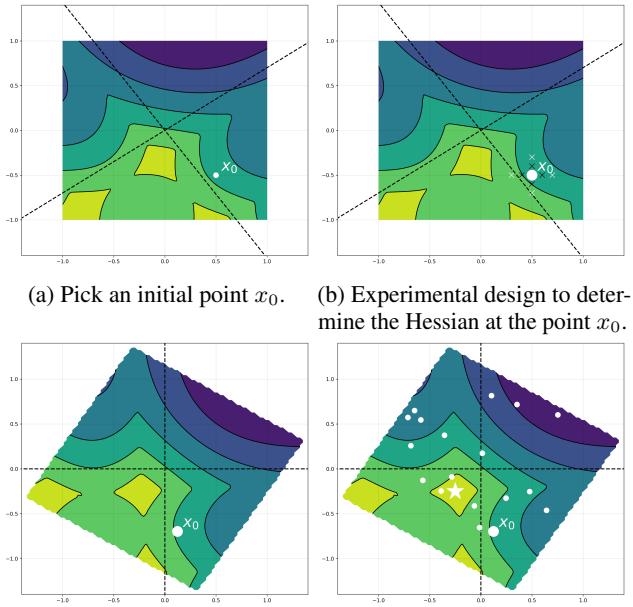
Problem Statement For positive integer k, we denote $[k] = \{1, \dots k\}$. Let $\mathcal{D} \subset \mathbb{R}^d$ be a bounded domain. Our objective is to maximize an unknown black box function $f : \mathcal{D} \to \mathbb{R}$ using only noisy point observations $y = f(x) + \epsilon$. The noise is assumed to be independent and σ^2-subgaussian. Simple regret of a point $x \in \mathcal{D}$ is defined as

$$r(x) \stackrel{\text{def}}{=} \max_{x' \in \mathcal{D}} f(x') - f(x) \tag{1}$$

The goal is to return a solution x_{final} with small simple regret $r_{\text{final}} = r(x_{\text{final}})$, which is the same as small optimization error. Clearly, the general problem is intractable without further smoothness assumptions on f. The central assumption used in the literature on kernelized bandits and Bayesian optimization is that f is a member of a known reproducing kernel Hilbert space (RKHS). We denote by $\mathcal{H}_k$ the RKHS with associated kernel k and Hilbert norm $\|\cdot\|_k$.

Assumption 1 (Global RKHS)**.** *There exists a known Mercer kernel κ, such that $f \in \mathcal{H}_\kappa$ and $\|f\|_\kappa \leq B$ and f is L-Lipschitz and k is twice differentiable.*

This assumption alone does not guarantee statistical efficiency, and in fact with isotropic kernels, sample complexity bounds are known to be exponential in the dimension (Scarlett, Bogunovic, and Cevher 2017). A powerful, yet tractable model is *projection pursuit regression* (PPR) (Friedman and Stuetzle 1981) which assumes that $f(x) = \sum_{i=1}^r f_i(\beta_i^\top x_i)$ for unknown vectors $\beta_j \in \mathbb{R}^d$ and one-dimensional component functions $f_i : \mathbb{R} \to \mathbb{R}$. For our approach we require that the vectors β_i are orthogonal, which allows for functions that are additive in an arbitrary rotated coordinate system.



(a) Pick an initial point x_0. (b) Experimental design to determine the Hessian at the point x_0.

(c) Rotate (decorrelate) the input space. (d) Perform additive bandit optimization.

Figure 1: The different stages of Algorithm 1. The dashed lines represent the axis over which the function is additive.

Assumption 2 (Orthogonal PPR)**.** *The function f can be written as $f(x) = \tilde{f}(\mathbf{R}x)$ for an unknown orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$ and an additive function $\tilde{f}(\tilde{x}) = \sum_{i=1}^d \tilde{f}_i(\tilde{x}_i)$. Further, $\tilde{f} \in \mathcal{H}_k$ for an additive RKHS with kernel $k(x, y) = \sum_{i=1}^d k_i(x_i, y_i)$ and $\|\tilde{f}\|_k \leq B$.*

One example of a function that satisfies our assumptions is a polynomial of fixed degree that is obtained from rotating another polynomial function without cross terms. In this case both kernels can be chosen as the corresponding polynomial kernel. Furthermore, our results hold more generally for maps $\mathbf{R} \in \mathbb{R}^{l \times d}$ with orthogonal columns $\mathbf{R}\mathbf{R}^\top = \mathbf{I}_{l \times l}$, where $l \leq d$ (also known as Stiefel manifold) and group additive function, as we discuss in more detail later.

2 A Two-Stage Algorithm

We present a two-state algorithm based on experimental design. In the first stage, we determine a design to estimate the Hessian of the function at an arbitrary starting point x_0 which allows us to identify the correlating matrix $\mathbf{R}$. At the second stage, we de-correlate the parameters and perform bandit optimization, where we benefit from the low sample complexity of additive models (Krause and Ong 2011). The stages of the algorithm are illustrated in Figure 1 and pseudo-code is provided in Algorithm 1.

Bayesian Optimization Bayesian optimization and kernelized bandits leverage the RKHS assumption by constructing a kernel regression estimate of the objective. Specifically,

Algorithm 1 Orthogonal PPR - Bandit Algorithm

Require: global kernel κ, kernels $\{k_i\}_{i=1}^{|d|}$, bound B, arbitrary starting point x_0, discretization of domain $\mathcal{D}$

1: Estimate Hessian $\tilde{\mathbf{H}}(x_0)$ of f at point x_0 via Experimental design (see Algorithm 3)
2: Perform eigendecomposition of $\mathbf{H} = \mathbf{Q}^\top \tilde{\mathbf{D}} \mathbf{Q}$.
3: Transform $\mathbf{Q}^\top \mathcal{D}$
4: Perform kernelized Bandit algorithm with additive structure on $\mathbf{Q}^\top \mathcal{D}$ (see Algorithm 2)

given data $\{(x_1, y_1), \ldots, (x_t, y_t)\}$, the mean estimate is

$$\mu_t = \arg\min_{\mu \in \mathcal{H}_k} \sum_{s=1}^{t} (\mu(x_s) - y_s)^2 + \lambda \|f\|_k^2 .$$

A closed-form solution can be obtain with representer theorem for infinite dimensional $\mathcal{H}_k$. Frequentist confidence bounds $\mu_t(x) \pm \beta_t(\delta)\sigma_t(x)$ that contain the true function value $f(x)$ are known also for adaptively collected data (Srinivas et al. 2010; Abbasi-Yadkori and Szepesvari 2012). Here $\sigma_t(x)^2$ is the posterior variance of the corresponding Gaussian process model $\mathrm{GP}(\mu_t(x), \sigma_t^2(x))$ and $\beta_t(\delta)$ is a scaling parameter to control the coverage probability. Bayesian optimization algorithms use the uncertainty $\sigma_t(x)$ associated to a prediction $\mu_t(x)$ to balance exploration between uncertain and promising parameters x. Many different methods are known. We focus on Thompson sampling (Chowdhury and Gopalan 2017) which samples points according to their posterior probability of being optimal in the Bayesian Gaussian process model (Algorithm 2). This approach was recently noted to have computational advantages when used with additive models (Mutný and Krause 2018).

Algorithm 2 Kernelized Thompson sampling

Require: Kernels $\{k_i\}_{i=1}^{d}$

for $t = 1 \ldots T$ **do**
  Compute GP estimates $\mu_t(x) \pm \beta_t(\delta)\sigma_t(x)$
  $\alpha_t \sim GP(\mu_t, \beta_t(\delta)^2 \sigma_t)$ ▷ Sample from a GP
  $x_t = \arg\max_{x \in \mathcal{D}} \alpha_t(x)$ ▷ Acquisition step
  $y_t = f(x_t) + \epsilon_t$ ▷ Noisy feedback
end for
return $x_{\text{final}} = \arg\max_{t \in [T]} \mu_t(x_t) - \beta_t(\delta)\sigma_t(x_t)$

3 Experimental Design for Rotations

In order to identify the unknown rotation matrix $\mathbf{R}$ we use the property that the Hessian $\nabla^2 \tilde{f}(\mathbf{R}x)$ of an additive function is diagonal. Namely, due to Assumption 2 we know that at any $x \in \mathcal{D}$,

$$\mathbf{H}(x) = \nabla^2 f(x) = \mathbf{R}^\top \nabla^2 \tilde{f}(\mathbf{R}x)\mathbf{R} . \tag{2}$$

As $\mathbf{H}(x)$ is symmetric and under the condition that all eigenvalues are distinct, we can recover $\mathbf{R}$ from the unique eigenspaces of $\mathbf{H}(x)$ up to a permutation of the coordinates. Hence, in order to determine $\mathbf{R}$, we first estimate the Hessian matrix at the initial point x_0 and subsequently perform an eigendecomposition to identify $\mathbf{R}$ up to a permutation.

Without noise we could resort to a simple finite difference scheme to estimate the Hessian up to machine precision by decreasing the step-size. With noisy observations, however, we cannot rely on the accuracy of a single measurement, and the closer we perform our estimation to x_0, the worse the signal to noise ratio gets. On the other hand, evaluations far apart from x_0 diminish the information content about the Hessian unless the true functions is quadratic. Hence, in order to estimate the Hessian well, it is crucial to query at points that balance both sources of error.

Experimental Design for the Hessian Formally, we would like to minimize the error of our estimate $\tilde{\mathbf{H}}$ of the Hessian $\mathbf{H}$ evaluated at x_0. For our further analysis, a bound with respect to the spectral norm is sufficient but cumbersome to optimize with optimal experimental design. Hence, we work with the Frobenius norm $\left\|\tilde{\mathbf{H}} - \mathbf{H}\right\|_F^2$ instead, which upper bounds the spectral norm.

Due to the Mercer's property (Assumption 2), f can be represented as an element of a finite dimensional RKHS, $f(\cdot) = \Phi(\cdot)^\top \theta$ up to arbitrary precision. In what follows we assume that $\theta \in \mathbb{R}^m$. Consequently, the Hessian at the point x_0 can be evaluated as

$$\mathbf{H}(x_0) = \sum_{k=1}^{m} \nabla_{x_0}^2 [\Phi_k] \theta_k .$$

In vectorized form, the Frobenius norm becomes the 2-norm, that can be minimized with experimental design. Optimal experimental design seeks a probability measure η over the domain $\mathcal{D}$ that minimizes the norm. In other words, η represents the budget spent on the points in $\mathcal{D}$ in order to maximize the statistical efficiency of estimating Hessian.

Let θ be the true parameter and $\hat{\theta}(\eta)$ be the least squares estimate with the design η. Let $\mathbf{C}_{kj} = \mathbf{vec}(\nabla_{x_0}^2 [\Phi_k])_j$ be the subspace that defines the Hessian. Note that knowing θ on this subspace fully specifies the Hessian at x_0. Hence our objective becomes

$$\min_{\eta} \left\|\tilde{\mathbf{H}}(\eta) - \mathbf{H}\right\|_F = \min_{\eta} \left\|\mathbf{C}(\tilde{\theta}(\eta) - \theta)\right\|_2 . \tag{3}$$

In the next theorem, we derive a novel concentration inequality on linear subspaces, which is the key result that allows to balance the error of local approximation and the statistical error in the estimation.

Theorem 1 (Fixed design on a subspace)**.** *Assume a linear regression problem with T data points in m dimensions (potentially $k \leq T$). Let $\hat{\theta}$ be regression estimate of $\theta \in \mathbb{R}^m$ s.t. $\|\theta\|_2 \leq B$. Let $\mathbf{C} \in \mathbb{R}^{k \times m}$ such that $\exists \mathbf{A} \in \mathbb{R}^{k \times T}$ full rank s.t.*

$$\|\mathbf{C} - \mathbf{A}\Phi(\mathbf{X})\|_2 \leq \varsigma \|\mathbf{A}\|_2 , \tag{4}$$

then the following holds,

$$\mathrm{P}\left(\left\|\mathbf{C}(\hat{\theta} - \theta)\right\|_{\mathbf{W}^{-1}} \geq \sigma \sqrt{\xi(\delta)} + \frac{\varsigma}{\sqrt{1 - \frac{\varsigma}{\|\Phi(\mathbf{X})\|_2}}} B \right) \leq \delta,$$

where $\zeta(\varsigma) := \frac{\varsigma}{\sqrt{1-\frac{\varsigma}{\|\Phi(\mathbf{X})\|_2}}}$, $\mathbf{W} = \mathbf{C}\mathbf{V}^{\dagger}\mathbf{C}^{\top}$ *invertible and* $\mathbf{V} = \Phi(\mathbf{X})^{\top}\Phi(\mathbf{X})$ *is the empirical covariance matrix,* $\mathbf{V}^{+}$ *is the pseudo-inverse of* $\mathbf{V}$ *and* $\xi(\delta) = k + 2\sqrt{k\log(\frac{1}{\delta})}$.

Specifically we can upper bound (3) by the largest eigenvalue of the matrix $\mathbf{W}$, and then apply the statement in the theorem,

$$\left\|\mathbf{C}(\tilde{\theta}(\eta)-\theta)\right\|_2 \leq \sqrt{\lambda_{max}(\mathbf{W})}\left\|\tilde{\theta}(\eta)-\theta\right\|_{\mathbf{W}^{-1}} \leq \sqrt{\lambda_{max}(\mathbf{W}(\eta))}(\sigma\sqrt{\xi(\delta)}+\zeta(\varsigma)B) \quad (5)$$

that holds with probability at least $1-\delta$, where $\xi(\delta) = k + 2\sqrt{k\log(\frac{1}{\delta})}$. Having this form we need to search for a design that minimizes the largest eigenvalue of $\mathbf{W}$ and satisfies the condition in (4). This condition captures whether the chosen design points $\Phi(\mathbf{X})$ can contain information about the vector θ in the subspace $\mathbf{C}$ to the precision ς ($\zeta(\varsigma)$ is close to ς for small values), and the condition can be checked by simply solving a linear system.

E-experimental design Minimizing the largest eigenvalue of $\mathbf{W}$ is known as E-experimental design in the literature, and it can be solved via a convex relaxation (Fedorov and Hackl 1997). This yields a probability measure η over a discrete subset of points from the domain $\mathcal{D}$. It is important that the subset is chosen to capture information about the subspace of interest (condition (4)) with $\zeta(\varsigma) = \epsilon$ in order to estimate the vector θ on the subspace up to precision ϵ. Having this, we could optimize the objective using Frank-Wolfe. The solution to the optimization problem leads to $\phi^* = \lambda_{\max}(\mathbf{W}(\eta^*))$. This value can be reduced by repeated evaluation of the design such that the overall accuracy of the estimated value on the subspace is below the desired ϵ. However, in general, the design η^* is supported at most d^2m design points (Chaloner 1984). As m can be arbitrarily large, this straightforward approach is not satisfactory.

Instead, to avoid the dependence on m, we resort to a specific experimental design for the Hessian. Namely, Taylor expansions of twice differentiable functions suggest a design based on finite differences (Quarteroni, Sacco, and Saleri 2007). These are called *stencils*, and one is demonstrated in Figure 2a and formalized in Definition 1. Stencils satisfy the condition (4) for sufficiently small h for twice differentiable kernels and are of size d^2+d+1. Informally, these evaluation points allow to estimate a local quadratic approximation of the function.

Definition 1 (Hessian stencil design)**.** *Let* $h \in \mathbb{R}_+$ *and* $x_0 \in \mathbb{R}^d$. *Let* $\{x_0, x_0 + he_1, \ldots, x_0 + he_d, x_0 - he_1, \ldots, x_0 - he_d\} \cup \{x_0 \pm h(e_i + e_j) | i < j\}$ *be a set of design points. We call such a set Hessian stencil design and associate to it the measure* $\eta(h)$.

This class of designs can be used to estimate a Hessian, and the parameter h determines the approximation-noise trade-off. Specifically, we choose designs $\eta(h)$ parametrized by a step size h. In contrast to the standard approach, our designs assures a small support size $\mathcal{O}(d^2)$. Let us define $\phi(h) = \lambda_{max}(\mathbf{W}(\eta(h)))$, the value of the design and $\phi^* = \min_h \phi(h)$ such that condition (4) is satisfied. The optimal h^* and its corresponding design η^* can be found by a one dimensional grid search prior to the experimental design phase as it depends only on the kernel, the noise level and the desired accuracy ϵ. Establishing the η^*, we can reduce the value of ϕ^* arbitrarily by repeating the point evaluations. This way we increase the accuracy of the least squares estimator on the subspace of interest.

Property 1. *Let* η *be a probability measure leading to a design with value* ϕ. *By performing* n *repeated evaluations according* η, *we reduce the value of the design* ϕ *by* $\frac{1}{n}$.

Specifically, in order to reduce $\left\|\tilde{\mathbf{H}}(\eta) - \mathbf{H}\right\|_F$ to ϵ with high probability, we need a design such that $\zeta(\varsigma) = \frac{\epsilon}{B}$ and perform $\frac{\sigma\phi^*\xi(\delta)}{\epsilon^2}$ repeated evaluation of the design supported on d^2+d+1 points. Inverting the expression, the error $\epsilon(T_R)$ of the Hessian estimate after T_R evaluations is

$$\epsilon(T_R) = \sqrt{\frac{\sigma(d^2+d+1)\xi(\delta)\phi^*}{T_R}}. \quad (6)$$

Algorithm 3 Experimental Design for Hessian Estimation

Require: global kernel κ, bound B, accuracy ϵ
1: Pick a point x_0 satisfying Assumption 3 ▷ Verifiable during runtime
2: Calculate $\mathbf{C}$, $\mathbf{C}_{kj} = \mathbf{vec}(\nabla^2_{x_0}[\Phi_k])_j$
3: Find $h^* = \arg\min_{h\in\mathbb{R}_+} \lambda_{max}(\mathbf{W}^{-1}(\eta(h)))$ where $\eta(h)$ is a Hessian stencil design and condition (4) is satisfied with $\zeta(\varsigma) = \epsilon$.
4: Repeat the design $\eta(h^*)$ to the desired accuracy ϵ.

Eigendecomposition Sensitivity Analysis In order to recover the matrix $\mathbf{R}$ via eigendecomposition, we make a weak assumption.

Assumption 3 (Non-degenerate starting point)**.** *Let* f *be as in Assumption 2 and let the eigenvalues* $\{\lambda_i\}_{i=1}^d$ *of the Hessian* $\mathbf{H}(x_0)$ *of* f *at* x_0 *be all different with* $\Delta = \min_{ij}|\lambda_i - \lambda_j| > 0$.

Assumption 3 states that x_0 needs to be a non-degenerate point that allows to recover the rotation. The condition can be verified in practice and if it is not satisfied a different point can be chosen. The next theorem states, that the rotation is recovered up to a permutation, which is sufficient for our approach.

Lemma 1. *Suppose* $\tilde{\mathbf{H}}$ *and* $\mathbf{H} \in \mathbb{R}^{d\times d}$ *be such that Assumption 3 holds, and have eigendecompositions* $\mathbf{H} = \mathbf{R}^{\top}\mathbf{D}\mathbf{R}$ *and* $\tilde{\mathbf{H}} = \mathbf{Q}^{\top}\tilde{\mathbf{D}}\mathbf{Q}$. *Also, let* $\left\|\mathbf{H} - \tilde{\mathbf{H}}\right\|_2 \leq \epsilon$, *then*

$$|(\mathbf{Q}^{\top}\mathbf{R})_{ij} - \mathbf{P}_{ij}| \leq \frac{2\epsilon d^2}{\Delta} \quad (7)$$

where $\Delta = \min_{ij}|\lambda_i - \lambda_j|$, $\{\lambda_i\}_{i=1}^d$ *are eigenvalues of* $\tilde{\mathbf{H}}$, *and* $\mathbf{P}$ *is a permutation matrix.*

As an immediate consequence of this lemma we obtain

$$\|\mathbf{R} - \mathbf{PQ}\|_2 \leq \frac{2\epsilon d^3}{\Delta}\,, \tag{8}$$

for details we refer to Lemma 8 in the Appendix.

4 Regret Analysis

We continue to analyze the regret of Algorithm 1. We make use of the fact that the estimated rotation $\mathbf{Q}$ is close to the true rotation in the sense that $\mathbf{RQ}^\top \approx \mathbf{P}$ is approximately a permutation. Therefore $f(\mathbf{R}x) = f(\mathbf{RQ}^\top \tilde{x})$ becomes *almost* additive in $\tilde{x}$. We make this precise below and show how to control the error induced by the miss-specification.

More formally, by (6) and (8) we can ensure that $\|\mathbf{RQ}^\top - \mathbf{P}\|_2 \leq 2\epsilon(T_R)d^3/\Delta$. As we show in Lemma 8 in the Appendix, this implies that

$$|f(\mathbf{RQ}^\top \tilde{x}) - f(\tilde{x})| \leq \frac{2BL\epsilon(T_R)d^3}{\Delta} = \omega\,. \tag{9}$$

We control how this bias affects the RKHS estimate in proposition in the Appendix. Effectively, the estimate of the function f satisfies $|\mu_t(\tilde{x}) - f(\tilde{x})| \leq \beta_t(\delta)\sigma_t(\tilde{x})$ with probability at least $1-\delta$, where $\beta_t(\delta) = \sqrt{\gamma_T + 2\log(1/\delta)} + \lambda B + \omega\sqrt{T}$. The quantity γ_T is called the maximum information gain and is a standard complexity measure in Bayesian optimization. It is equal to the log determinant $\gamma_T = \log \frac{\det(\mathbf{K}_t + \lambda\mathbf{I})}{\det(\lambda\mathbf{I})}$ of the kernel matrix $\mathbf{K}_t$.

The misspecification bias ω can be reduced by increasing the accuracy of the estimate of $\mathbf{Q}$. The accuracy increases with the number of data points T_R allocated to the first stage. The analysis above suggest that $\omega \leq \mathcal{O}(1/\sqrt{T_R})$. Consequently, we can rely on any Bayesian optimization algorithm for additive functions, that uses the rescaled confidence bounds. Specifically, the final solution returned by Thompson sampling satisfies the following bound

$$r_{\text{final}} \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{T}}\big(\gamma_T + B\ln\frac{1}{\delta} + \sqrt{T}\omega\big)\right) \tag{10}$$

where we suppress logarithmic factors with $\tilde{\mathcal{O}}(\cdot)$.

This result follows by substituting our modified confidence bounds into the analysis of Thompson sampling (Chowdhury and Gopalan 2017; Abeille and Lazaric 2016). Note that existing results typically bound cumulative regret, but a bound on the simple regret can be obtained for the final solution x_{final}. For further details see Appendix. By the same reasoning one can also use a different method to optimize the decorrelated problem such as GP-UCB (Srinivas et al. 2010) and SAFEOpt (Berkenkamp, Schoellig, and Krause 2016). However other methods do not use an additive acquisition function, which makes the acquisition step difficult; see also the discussion in the next section. We summarize the overall regret bound for our method in the next theorem.

Theorem 2 (Simple Regret)**.** *Suppose Assumptions 1, 2 and 3 hold. Further suppose all k_i are the same with maximum information gain $\tilde{\gamma}_T$. Let $\mathbf{Q}$ be the eigenvectors of the estimated Hessian $\tilde{\mathbf{H}}(x)$, then using bandit Algorithm 2 as part of Algorithm 1 for a fixed horizon T such that $\epsilon(T_R) \leq \sqrt{\frac{2\lambda}{T}\frac{\Delta}{d^3 4LBD}}$, where λ is the ridge regression constant, $D = \max_{x\in\mathcal{D}} \|x-y\|_2^2$ and T_R are the iterations of the first-stage. Then with probability at least $1-2\delta$*

$$r_{final} \leq \tilde{\mathcal{O}}\left(\frac{d^{3/2}\beta_T(\delta)\tilde{\gamma}_T^{1/2}}{\sqrt{T}}\right) \tag{11}$$

and $T_R = \Theta(T)$ is required to control the misspecification.

In particular, the analysis of Theorem 2 reveals that in order to achieve optimal simple regret, the first stage of the algorithm needs to use the same order of evaluations as the horizon of the optimization. Our theoretical result requires that all kernels are of the same complexity since the matrix $\mathbf{R}$ can be recovered only up to a permutation. However, a practitioner might approach the problem by first decorrelating the system, and only then making modeling assumptions on the individual additive components.

5 Optimizing the Acquisition Function

For optimizing the decorrelated objective, the evaluation point $x_t = \mathbf{Q}\tilde{x}_t$ is sequentially determined by maximizing the additive acquisition function $\alpha(\tilde{x}) = \sum_{i=1}^d \alpha_i(\tilde{x}_i)$ of Thompson sampling (Algorithm 2). Below we discuss two specific approaches, one for polyhedral domains (which includes commonly used box constraints) and one for spherical domains.

Polyhedral Domain If the domain is an axes-aligned rectangle without the rotation, one can find the maximizer by simply optimizing each coordinate separately, as also noted by Rolland et al. (2018) and Mutný and Krause (2018). In our case, however, the box constraints are rotated (Figure 1c) and the general problem might not admit an efficient solution. A practical approach assumes that the true maximizer is contained within a smaller, axes-aligned box that is inscribed in the rotated domain, or respectively enlarge the domain if permitted in the application as done by Li et al. (2016). To obtain an exact solution on the original domain, we propose an alternative approach that exploits the additivity via linear integer programming and is very efficient in practice.

Assume for simplicity that $\mathcal{D} = [-1,1]^d$, but the approach includes general polyhedral constraints. For the formulation, we require a one-dimensional discretization $\tau = \sqrt{2}[-N\ldots N]/N$ of each axis in the rotated domain, which can be thought as a vector in $\mathbb{R}^{2N+1}$. We introduce indicator variables $z^i \in \{0,1\}^{2N+1}$ for each $i \in [d]$. Each indicator variable satisfies the constraint $\sum_{i=1}^d (z^i)_j = 1$ to select exactly one of the grid points via $x_i = \tau^\top z^i$. Denote by $z = (z_1, \ldots, z_d)$ the concatenated vector and $x = \mathbf{T}z$ for an appropriate chosen matrix $\mathbf{T}$. The constraint on the domain $0 \leq \tilde{x} \leq 1$ can be modeled as $0 \leq \mathbf{Q}^\top\mathbf{T}z \leq 1$. Finally, we associate each component $f_j(x_i)$ with the vector $c_{(ij+(i-1)(2N+1))} = f_j(\tau_i)$. This leads to the integer pro-

gram

$$\begin{aligned} & \max_z c^\top z \\ \text{s.t.} \quad & 0 \leq \mathbf{Q}^\top \mathbf{T} z \leq 1 \\ & \textstyle\sum_{j=1}^{2N+1} z_{(i-1)(2N+1)+j} = 1 \text{ for each } i \in [d] \\ & z \in \{0,1\}^{d(2N+1)} \end{aligned}$$

The integer program can be solved by many established solvers, for instance by first relaxing the integer constraints and subsequently using a cutting plane algorithm. Importantly, if the optimal point is inside an axes-aligned box that can be inscribed inside the rotated domain, the LP relaxation has an integral solution and no discrete optimization is needed. Hence, if the optimization of the acquisition function is easy, this will be detected at the stage of the linear relaxation.

Due to Lipschitz continuity of f, we improve the performance of our solver by solving the problem at coarser levels of discretization and using these solutions as upper and lower bounds before solving the problem at the desired discretization.

Spherical Domain If the domain $\mathcal{D}$ is a sphere, e.g. $\mathcal{D} = \{x : \|x\|_2 \leq 1\}$ the domain remains unchanged after we apply the rotation. Note that optimizing each coordinate greedily still leads to suboptimal solutions in general because the range for a coordinate x_i depends on the values of all other coordinates. To obtain an exact solution we resort to dynamic programming which wield lied a problem of size $\frac{d}{\tau^2}$. For details please refer to the supplementary material.

6 Examples and Extensions

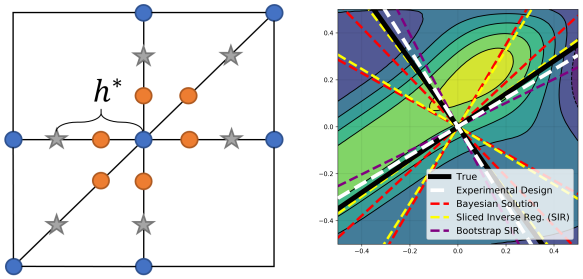
Groups of larger sizes The extension toward additive groups of larger size is straightforward. Let us consider $G = \{g_1 \dots g_{|G|}\}$ be the groups of variables where each group contains $|g_i|$ variables. Analogically to (2) the Hessian of the group additive function is block diagonal $\nabla^2 \tilde{f} = \bigoplus_{i=1}^{|G|} \mathbf{Y}_i$, where $\mathbf{Y}_i \in \mathbb{R}_{|g_i| \times |g_i|}$. One can diagonalize such matrix via $\bar{\mathbf{S}} = \bigoplus_{i=1}^{|G|} \mathbf{S}_i$, where each $\mathbf{S}_i$ is orthogonal on the respective subspace. Consequently, the eigenvectors of $\mathbf{H}$, similarly as in previous sections, recover an orthogonal matrix $\mathbf{Q}^\top = \mathbf{R}^\top \bar{\mathbf{S}}^\top \mathbf{P}$, where $\mathbf{P}$ is a permutation matrix. The following function becomes group additive as,

$$f(\mathbf{Q}^\top x) = \tilde{f}(\mathbf{R}\mathbf{R}^\top \bar{\mathbf{S}}^\top \mathbf{P}) = \tilde{f}(\bar{\mathbf{S}}^\top \mathbf{P} x) = \tilde{g}(\mathbf{P}x),$$

where the function g absorbs the rotations of the coordinates within the groups, and maintains the same additivity degree. The deficiency of this method is that because of the permutation we do not immediately get the correct group assignment of the decorrelated variables. As an example consider the function $\tilde{f}(x_1, x_2, x_3) = \tilde{f}_{12}(x_1, x_2) + \tilde{f}_3(x_3)$. Suppose that the permutation swaps x_1 and x_3. Consequently,

$$\tilde{f}(\bar{\mathbf{S}}\mathbf{P}x) = \underbrace{\tilde{f}_{12}(\mathbf{S}_1[x_3, x_2]^\top, \mathbf{S}_1[x_3, x_2]^\top)}_{\tilde{g}_{23}(x_2, x_3)} + \underbrace{f_3(\mathbf{S}_2 x_1)}_{\tilde{g}_1(x_1)},$$

which is group additive with groups $G = ((2,3),1)$. The group assignment can be estimated again by picking a point in the domain x and calculating Hessian at this point with respect to the decorrelated variables. The block structure of the new Hessian, namely the position of zero elements, reveals the assignment of the variables to different groups.



(a) (b)

Figure 2: a) Geometry of a Hessian stencil design ($d = 2$). The optimal design is has stared points (origin included). b) The alignment of additive components for the Michalewicz benchmark. Two best optima of the Bayesian method and SRI are reported in red and yellow respectively.

Invariant Subspaces Our method for decorrelation of the additive components is applicable for matrices $\mathbf{R} \in \mathbb{R}^{n \times l}$ from the Stiefel manifold $O(l, d)$, where $l < d$ satisfying $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_l$. With such $\mathbf{R}$, the estimated Hessian will have $d - l$ zero eigenvalues. The eigenvectors associated to these eigenvalues span the invariant subspace of the function $\tilde{f}$. The remaining eigenvectors of $\mathbb{R}$ can be recovered in the same fashion as previously if the distinct eigenvalue property holds among them.

Kernel Examples We briefly mention possible choices for the global and the additive kernel that can be used with our method. The perhaps most natural example is to choose a polynomial kernel for both global and additive RKHS. Another candidate is the finite basis approximation of stationary kernels such as in Quadrature Fourier Features (QFF) (Mutný and Krause 2018).

Additionally, for any kernel in the vicinity of the starting point x_0 one can calculate a finite dimensional approximation $f(x) \approx \Phi(x)^\top \theta$ where Φ can be calculated by SVD or Nyström approximation of the kernel matrix that contains points sampled around x_0. Such kernel can be used as global kernel for the first stage subject to the further constraint that the stencil size h is not larger than the area of approximation validity.

Lipschitz continuity in Assumption 1 can be assured if the finite dimensional feature map $\Phi(x)$ is Lipschitz continuous, or for more general cases if the induced metric $d_k(x, y) = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}$ is Lipschitz continuous. These assumptions are satisfied for all previously mentioned kernels.

7 Experimental Results

Implementation details In practice, we specify the value of $\epsilon = 10^{-3}$ in the first phase of the algorithm, and we model

T_R separately as our analysis suggests larger (but not unreasonable) values for T_R for short optimization horizons T. The value of the design depends significantly on the observation noise σ as in (5). On the other hand, from our experiments, the Lipschitz constant of f plays a significant role on some instances such as high-degree polynomials. In such circumstances due to the misspecification one can incur large regret, and in such circumstances a large number of design points is necessary - as suggested by the analysis.

For comparison, we implement an algorithm that maximizes the evidence given the rotation matrix:

$$\arg \min_{\mathbf{R}^\top \mathbf{R} = \mathbf{I}_k} \mathcal{L}(\{x_i, y_i\}_{i=1}^n | \mathbf{R}) \tag{12}$$

where $\mathcal{L}(\{x_i, y_i\}_{i=1}^n | \mathbf{R}) = y^\top (\mathbf{K}_\mathbf{R})^{-1} y + \log\det(\mathbf{K}_\mathbf{R})$. This is a common approach in fitting hyperparameters in Gaussian process regression (Rasmussen and Williams 2006). We call this the *Bayesian solution* in what follows. To solve it, we use a Stiefel manifold optimizer from the package pymanopt (Townsend, Koep, and Weichwald 2016). In our experiments we find that the optimization is erratic and yields multiple local minima with poor solutions.

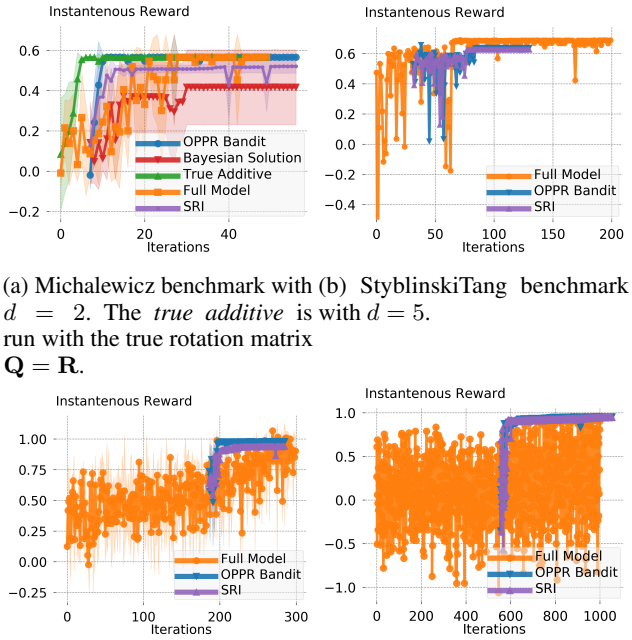
Additionally, we compare to Sliced Inverse Regreesion (SRI) of Li (1991) used by Zhang, Li, and Su (2019) to estimate $\mathbf{R}$. This method is implemented via conditioning on the response variables as $\mathbb{E}[X|Y = y]$, which is a curve in a low dimensional manifold. The conditioning is done by dividing values of y into non-overlapping slices in a randomized fashion. We define an empirical average of the algorithm randomization on the Stiefel manifold as Bootstrap SRI.

Benchmarks We validate our methods on standard benchmarks from the additive Bayesian optimization literature (Gardner et al. 2017). We first focus on an explanatory example in Figure 3a, where we optimize a two dimensional function. We see that the Bayesian solution converges to a suboptimal point due to the misspecification error which is evident from the Figure 2b. The SRI on itself performs eratically but the heuristic of averaging the SRI estimates (bootstrap) performs competitively. For the Bayesian solution and SRI, we use the same data points as in the stencil design for a fair comparison. For this example, we used QFF for global and local kernel with sufficiently large basis.

In Figure 2b, we optimize a 5 dimensional function, which is a sum of polynomials of degree 4, where the polynomial kernel was used globally but due to sensitivity of misspecification (large Lipschitz constant), the squared exponential kernel was used along the coordinates. We see that both estimates are misspecified in comparison to the full model and lead to suboptimal solutions.

In the last benchmark problem (Figures 3b and 3c), which models the performance of a real-world electron laser machine, we use the local inducing point approximation of the squared exponential kernel that we outlined in the previous section for the global kernel, and we use QFF for the additive components in the second stage. In both instances, the full model acquisition function is optimized only approximately using first-order optimization, along with polynomial or Random Fourier Feature approximation (Rahimi and Recht 2007) to efficiently optimize the posterior sample (which would not otherwise have an analytical form).

The one-time evaluation of $\mathbf{C}$ is the most costly operation of the algorithm. It requires second order derivatives of the feature map $\Phi(x)$. This can be potentially very large - even exponential in d as is the case for polynomial features.



(a) Michalewicz benchmark with $d = 2$. The *true additive* is run with the true rotation matrix $\mathbf{Q} = \mathbf{R}$.

(b) StyblinskiTang benchmark with $d = 5$.

(c) Electron Laser Simulator benchmark with $d = 5$

(d) Electron Laser Simulator benchmark with $d = 10$.

Figure 3: Numerical validation of the algorithm. To make the comparison fair, the OPPR algorithms are shifted to account for the Hessian estimation phase.

8 Conclusion

We presented a novel two-stage algorithm for black-box optimization of functions satisfying the orthogonal projection pursuit regression model, where each component function is modeled as a function in a RKHS. In the first stage, the algorithm uses experimental design to provably recover the correlating matrix such that the additive components can be efficiently optimized in the second stage. We specifically addressed how to optimize the acquisition function of such a model, presented extension to larger groups and invariant subspaces, and numerically validated the method.

Acknowledgments This research was supported by SNSF grant 407540 167212 through the NRP 75 Big Data program. Further, this project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme grant agreement No 815943.

References

Abbasi-Yadkori, Y., and Szepesvari, C. 2012. *Online learning for linearly parametrized control problems*. Ph.D. Dissertation, University of Alberta.

Abeille, M., and Lazaric, A. 2016. Linear thompson sampling revisited. *arXiv preprint arXiv:1611.06534*.

Berkenkamp, F.; Schoellig, A. P.; and Krause, A. 2016. Safe controller optimization for quadrotors with gaussian processes. In *ICRA*, 491–496. IEEE.

Chaloner, K., and Verdinelli, I. 1995. Bayesian experimental design: A review. *Statist. Sci.* 10(3):273–304.

Chaloner, K. 1984. Optimal Bayesian Experimental Design for Linear Models. *Annals of Statistics* 12(1):283–300.

Chowdhury, S. R., and Gopalan, A. 2017. On kernelized multi-armed bandits. In *International Conference on Machine Learning*.

Djolonga, J.; Krause, A.; and Cevher, V. 2013. High-dimensional Gaussian process bandits. In *Advances in Neural Information Processing Systems*, 1025–1033.

Edelman, A.; Arias, T. A.; and Smith, S. T. 1999. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* 20(2):303–353.

Fedorov, V. V., and Hackl, P. 1997. *Model-Oriented Design of Experiments | Valerii V. Fedorov | Springer*. Springer-Verlag New York.

Friedman, J. H., and Stuetzle, W. 1981. Projection Pursuit Regression. *J. Am. Stat. Assoc.* 76(376):817–823.

Gardner, J.; Guo, C.; Weinberger, K.; Garnett, R.; and Grosse, R. 2017. Discovering and exploiting additive structure for Bayesian optimization. In *Artificial Intelligence and Statistics*, 1311–1319.

Gopalan, A.; Mannor, S.; and Mansour, Y. 2014. Thompson sampling for complex online problems. In *ICML*, volume 14, 100–108.

Hemant, T., and Cevher, V. 2012. Active learning of multi-index function models. In *Advances in Neural Information Processing Systems*, 1466–1474.

Kirschner, J., and Krause, A. 2018. Information directed sampling and bandits with heteroscedastic noise. *arXiv preprint arXiv:1801.09667*.

Kirschner, J.; Mutný, M.; Hiller, N.; Ischebeck, R.; and Krause, A. 2019. Adaptive and safe bayesian optimization in high dimensions via one-dimensional subspaces. *ICML 2019*.

Krause, A., and Ong, C. S. 2011. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, 2447–2455.

Li, C.-L.; Kandasamy, K.; Poczos, B.; and Schneider, J. 2016. High Dimensional Bayesian Optimization via Restricted Projection Pursuit Models. *PMLR* 884–892.

Li, K.-C. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86(414):316–327.

Lizotte, D. J. 2008. *Practical bayesian optimization*. University of Alberta.

Mockus, J. 1982. The bayesian approach to global optimization. *System Modeling and Optimization* 473–481.

Mutný, M., and Krause, A. 2018. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Neural and Information Processing Systems (NeurIPS)*.

Quarteroni, A.; Sacco, R.; and Saleri, F. 2007. *Numerical Mathematics*. Springer, texts in applied mathematics edition.

Rahimi, A., and Recht, B. a. 2007. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 3, 5.

Rasmussen, C., and Williams, C. 2006. Gaussian processes for machine learning, vol. 1. *The MIT Press, Cambridge, doi* 10:S0129065704001899.

Rolland, P.; Scarlett, J.; Bogunovic, I.; and Cevher, V. 2018. High-dimensional Bayesian optimization via additive models with overlapping groups. *AISTATS*.

Saati, E. G.; Cunningham, J.; and Gilboa, E. 2013. Scaling multidimensional Gaussian processes using projected additive approximations. In *ICML*.

Scarlett, J.; Bogunovic, I.; and Cevher, V. 2017. Lower bounds on regret for noisy gaussian process bandit optimization. *arXiv preprint arXiv:1706.00090*.

Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; and de Freitas, N. 2016. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* 104(1):148–175.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. 2010. Gaussian process optimization in the bandit setting: No regret and experimental design. *ICML*.

Townsend, J.; Koep, N.; and Weichwald, S. 2016. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research* 17(137):1–5.

Wang, Z., and Jegelka, S. 2017. Max-value entropy search for efficient Bayesian optimization. *International Conference on Machine Learning*.

Wang, Z.; Hutter, F.; Zoghi, M.; Matheson, D.; and de Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research* 55:361–387.

Zhang, M.; Li, H.; and Su, S. 2019. High dimensional bayesian optimization via supervised dimension reduction. In *IJCAI*.