ORTHOBO: Orthogonal Bayesian Hyperparameter Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bayesian optimization is widely used for hyperparameter optimization when model evaluations are expensive, but noisy acquisition estimates can lead to unstable decisions. We identify acquisition estimation noise as a distinct failure mode: even when the surrogate model and acquisition target are appropriate, finite-sample Monte Carlo error can perturb acquisition values, flip candidate rankings, and lead Bayesian optimization to evaluate suboptimal configurations. As a remedy, we propose an *orthogonal acquisition estimator*, that subtracts an optimally weighted score-function control variate, yielding an acquisition residual orthogonal to posterior score directions and thereby reducing Monte Carlo variance. Building on this estimator, we introduce ORTHOBO, a Bayesian optimization framework that combines orthogonalized acquisition estimates with ensemble surrogates for structural misspecification and an outer log transformation for numerical stability. Theoretically, we prove target preservation, variance reduction, and improved pairwise ranking stability. Empirically, ORTHOBO substantially reduces acquisition estimation variance, stabilizes candidate rankings, and achieves strong performance across synthetic benchmarks and downstream use cases, including vision-transformer fine-tuning on an industrial wafer-map classification task. These results show that stabilizing acquisition estimation can directly improve the reliability and sample efficiency of Bayesian hyperparameter optimization.

1 Introduction

Hyperparameter optimization (HPO) is a fundamental component of modern machine learning systems, and has been used across a wide range of domains, including biology [5, 47], chemistry [65], manufacturing [1], medicine [42], and physics [55]. Because evaluating hyperparameter candidates can require substantial computation, *Bayesian optimization (BO)* is frequently used to identify strong hyperparameter configurations under limited evaluation budgets [e.g., 19, 51, 53, 57, 63]. BO fits a probabilistic surrogate model of the objective and selects new configurations by maximizing an acquisition function such as expected improvement (EI) [34].[1]

In our work, we focus on a source of instability that is central to BO decisions but has received comparatively little attention: *the estimation of the acquisition value itself*. Acquisition functions are often treated as deterministic once a surrogate model has been fitted. In many practical BO pipelines, however, the acquisition value is obtained by marginalizing over uncertain surrogate parameters, aggregating across surrogate models, or using Monte Carlo (MC) approximations [e.g., 10, 30, 46, 53]. As a result, the quantity used to rank candidate configurations is itself a noisy estimate. This matters because BO acts on *rankings*: even small estimation errors can flip the order of two candidates with similar acquisition values and cause the algorithm to spend an expensive evaluation on a worse configuration.

[1] For a thorough introduction to BO see [e.g., 23, 51].

At a technical level, we treat the estimation of the acquisition function as a MC estimation problem [e.g., 10, 29, 30, 53] and propose a novel *orthogonalized estimator of the acquisition value* that serves as a variance-reduction technique, thus improving robustness to sampling errors and stabilizing acquisition rankings. Our notion of orthogonality stems from the orthogonal ML literature [e.g., 12, 20, 35, 37] in that our estimator offers robustness in terms of variance reduction guarantees. Note that our notion of robustness differs from that in existing 'robust' regression methods which aim to protect the surrogate model against corrupted or heavy-tailed observations [e.g., 38].

Example. Consider tuning a classification model for quality control on a large-scale industrial manufacturing dataset, where each evaluation requires training on many images and assessing performance under a highly skewed class distribution. Two hyperparameter candidate configurations may differ only slightly, e.g., a deeper network with stronger dropout versus a shallower network with weaker regularization. If acquisition values are estimated noisily, BO may spend an expensive evaluation on the worse configuration simply because it was ranked first due to sampling noise.

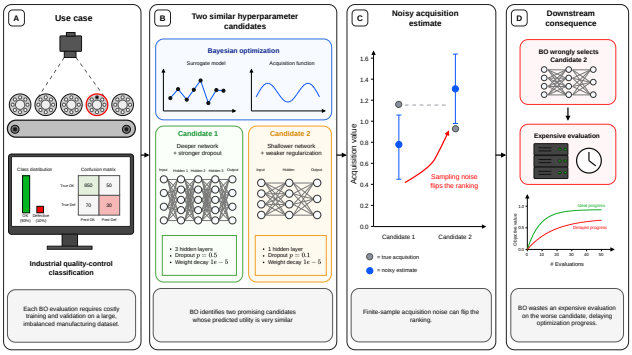

Figure 1: Motivational example of instabilities in Bayesian HPO.

Our contribution is complementary to existing work on robust BO. Prior methods have mainly focused on improving the surrogate model, for example by accounting for kernel misspecification, hyperparameter uncertainty, unreliable posterior uncertainty, trust-region localization, or ensemble surrogates [11, 8, 43, 16, 36, 46, 57]. Other work has improved the acquisition formula or its numerical optimization [2]. However, these approaches do **not** directly address the *variance of the estimated acquisition values* that drive BO decisions. In contrast, our method improves the statistical quality of the estimated acquisition values used for candidate selection through a variance-reduction techniques based on orthogonal scores.

Building on our orthogonal acquisition estimator, we introduce ORTHOBO, a BO framework for robust hyperparameter optimization. ORTHOBO combines three components: orthogonalized acquisition estimation to reduce MC acquisition noise, ensemble surrogates to mitigate structural surrogate misspecification, and an outer log transformation to improve numerical stability during acquisition maximization. The central idea is to separate these roles: surrogate ensembles improve the model class, orthogonalization improves acquisition estimation, and the outer log transformation improves numerical conditioning.

Our ORTHOBO offers several theoretical guarantees. The proposed orthogonalized estimator is unbiased for the marginal EI target and achieves lower variance than the corresponding naïve MC estimator ($\rightarrow$ Theorem 4.1). Additionally, the lower acquisition-estimation variance improves the stability of pairwise acquisition rankings ($\rightarrow$ Proposition 4.2), reducing the probability of falsely selecting a candidate based on MC noise. Empirically, we show that ORTHOBO improves robustness and sample efficiency in HPO tasks (Section 5).

Our **main contributions** are: **(1)** We propose a novel, orthogonalized estimator for hyperparameter acquisition, reducing estimation variance without changing the underlying acquisition target. **(2)** We develop ORTHOBO, a BO framework that explicitly targets acquisition estimation noise, structural surrogate misspecification, and numerical instabilities. **(3)** We provide theoretical guarantees for variance reduction and ranking stability, and demonstrate improved empirical performance on various HPO benchmarks. Importantly, ORTHOBO is also applicable to a wide range of other BO problems.

2 Related work[2]

BO for hyperparameter tuning. BO is a standard approach for HPO [e.g., 19, 53, 63], particularly when evaluations of the objective function are expensive. In practice, BO relies on a surrogate model to approximate the objective and guide the search. A common choice for the surrogate model are Gaussian processes (GPs) [27, 34, 53], while practical alternatives include tree-based and density-based surrogates such as tree-structured Parzen estimators (TPE) [6, 7, 60]. Subsequent work has addressed scaling challenges through multi-fidelity methods, batch BO, and trust-region approaches [e.g., 16, 17, 26, 61]. However, no method has addressed the misspecification and performance impact due to acquisition estimation noise.

Acquisition functions. The surrogate models are used by acquisition functions, which are heuristics to evaluate the utility of hyperparameter candidates before querying the (expensive) objective function [4]. The performance of the acquisition function thus directly influences the overall performance. In the literature, several types of acquisition functions or acquisition frameworks have been proposed, including qLogEI [2], GIBBON [40], UCB [62], MES [59], RPR [3], and TuRBO [16]. Our work is orthogonal to these works, as we focus on improving the *acquisition estimation* and not the acquisition function itself.[3]

Robustness to surrogate misspecification. A central weakness of BO is its reliance on a surrogate model which might be misspecified. Prior work has studied BO under misspecified kernel classes [11], unknown hyperparameters [8, 27], unreliable posterior uncertainty [43], and a surrogate complexity-efficiency trade-off [10]. Practical strategies for mitigating such failures include trust-region localization [16], robust likelihood models [3, 38], and model averaging or ensemble methods [36, 46]. However, these approaches address *structural* surrogate misspecification. In contrast, our method additionally targets misspecification due to *acquisition estimation* noise.

Research gap. We identify acquisition estimation errors under surrogate uncertainty as a crucial, yet previously overlooked, failure mode in Bayesian optimization. We propose an orthogonalized estimator that reduces variance and stabilizes acquisition-based decisions. To the best of our knowledge, this is the first work to explicitly formulate acquisition computation in BO as a variance-reduction problem and to develop an orthogonalized estimator for improving acquisition-based decisions.

3 Problem setting

Setup. We consider the standard setup for HPO over a search space $\Lambda \subseteq \mathbb{R}^d$ [3, 15]. For a hyperparameter configuration $\lambda \in \Lambda$, let $f(\lambda)$ denote the unknown performance of interest, e.g., the validation loss or validation error after training a model with hyperparameters λ. The objective f is expensive to evaluate and can only be observed through noisy evaluations $y = f(\lambda) + \epsilon$, where ϵ captures evaluation noise. At each iteration $t = 1, \dots, T$, the available data are $\mathcal{D}_t = \{(\lambda_i, y_i)\}_{i=1}^t$. We aim to sequentially choose hyperparameter configurations $\lambda_1, \lambda_2, \dots, \lambda_T$ to identify a high-performing configuration within an evaluation budget T.

Recap: BO. The central idea of BO is to maintain a probabilistic model of the unknown objective and to use it to guide future evaluations [24]. At iteration $t = 1, \dots, T$, BO fits a probabilistic *surrogate model* m to the current dataset and uses an *acquisition function* α to select the next evaluation point [3]. The surrogate provides two crucial quantities for decision-making: (i) a prediction of objective values in unexplored regions, and (ii) a measure of uncertainty about those predictions. BO then combines these two ingredients through α, which scores candidate points according to their potential utility for optimization. After selecting a new candidate point, the objective f is evaluated, and the new observation is added to the dataset $\mathcal{D}$. The surrogate is then updated, and the process repeats. In this way, BO adaptively concentrates evaluations in regions that are both informative and promising. As a result, the performance of BO depends on two components: the quality of the surrogate model and the stability of the acquisition values used to rank candidate configurations.

Surrogate model. Let m index a surrogate model, and θ_m denote its latent parameters. Conditional on θ_m, the surrogate induces a posterior predictive distribution for the objective $f(\lambda)$ given $\mathcal{D}_t$. In practice, the parameters θ_m are not known and must themselves be inferred from data.

[2] We provide an extended related work in Supplement A. Therein, we also provide a broader review of BO under different types of misspecification and orthogonal ML.

[3] We provide a generalization of ORTHOBO at the acquisition function level in Supplement B.

In our work, we focus on the *expected improvement (EI)* as our acquisition function. For a surrogate model m with parameters θ_m, we define

$$\mathrm{EI}_m(\lambda_t;\theta_m) = \mathbb{E}\big[(f^\star - f(\lambda_t))_+ \mid \mathcal{D}_t, \theta_m\big], \tag{1}$$

where $f^\star = \min_{i\le t} y_i$ is the currently best observed value up to iteration t and $(x)_+ := \max\{0, x\}$ [34, 53]. In practice, the model parameters are not fixed but follow an approximate posterior $\theta_m \sim q_{m,t}(\theta)$. The marginal EI over the parameters is then given by

$$\mathrm{EI}_m^{\mathrm{marg}}(\lambda_t) = \mathbb{E}_{\theta_m\sim q_{m,t}}[\mathrm{EI}_m(\lambda_t;\theta_m)], \tag{2}$$

which is commonly approximated through MC samples $s = 1,\dots,S$ given a specific MC sampling budget S [4, 53, 61], i.e.,

$$\widehat{\mathrm{EI}}_m^{\mathrm{MC}}(\lambda_t) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{EI}_m(\lambda_t;\theta_m^{(s)}), \quad \theta_m^{(s)} \sim q_{m,t}. \tag{3}$$

However, $\widehat{\mathrm{EI}}_m^{\mathrm{MC}}$ suffers from high variance for small budgets S and sensitivity to approximation errors in the posterior $q_{m,t}$, leading to instability in the estimated acquisition and thus the BO decisions.

Failure modes in BO. Several failure modes arise in a realistic HPO setting[e.g. 2, 8, 11, 43]. Bayesian HPO performance depends on the accuracy and stability of acquisition estimates, which are computed from a surrogate posterior that is often *misspecified and approximately inferred*, leading to instability in acquisition values. Specifically, BO suffers from the following failure modes:
(i) *Structural surrogate misspecification* arises when the surrogate model class does not adequately represent the unknown objective. For example, the assumed kernel, likelihood, prior smoothness, or stationarity structure may be incompatible with the true response surface. As a result, the posterior mean and posterior uncertainty estimates may be systematically biased.
(ii) *Posterior approximation and acquisition-estimation noise.* Marginalizing over surrogate uncertainty mitigates plug-in overconfidence, but the marginal acquisition value must be approximated from a finite number of posterior samples. This introduces *Monte Carlo variance* into the acquisition estimate, on top of any bias from the approximate posterior $q_{m,t}$ itself differing from the true posterior. Both effects can destabilize the ranking of candidate points.
(iii) *Numerical instability* can arise when acquisition values are very small, highly skewed, or nearly indistinguishable across candidates. Directly optimizing such quantities can lead to unstable gradients.

Prior work has recognized several of these failure modes. However, the finite-sample noise induced by estimating marginal acquisition values has not been addressed so far.

Our work. In this work, we address instability in BO at the level of the acquisition estimate itself. Even with a reasonable surrogate family, BO may still perform poorly if the acquisition estimator used to rank candidate points is noisy, sensitive to approximate inference, or numerically ill-conditioned.

We treat the EI-scale marginal acquisition obtained by integrating over surrogate uncertainty as an MC estimation problem. We propose an orthogonalized estimator that reduces variance induced by surrogate uncertainty, while preserving the underlying EI-scale target. To address structural surrogate misspecification and overcome numerical instabilities, we combine our estimator with an ensemble of surrogate models and adaptive model weighting optimized through an outer log transformation. Overall, ORTHOBO reduces acquisition-estimation variance, stabilizes candidate rankings, and improves BO decision quality.

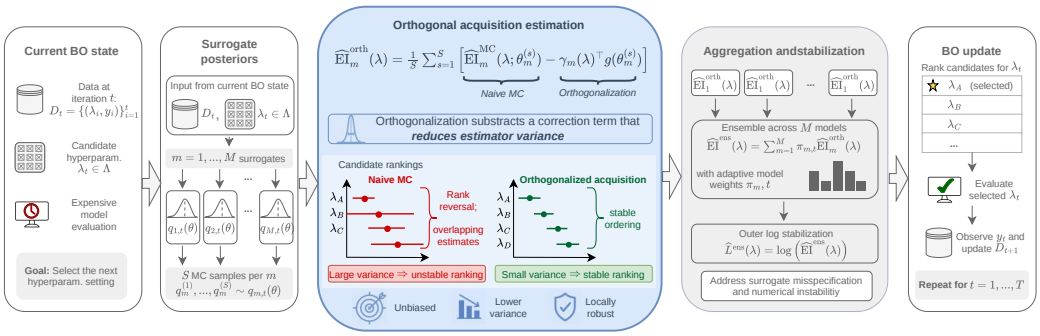

Figure 2: Our proposed ORTHOBO. It improves acquisition estimation through variance stabilization. For this, we orthogonalize the acquisition estimation without changing the acquisition target.

4 ORTHOBO

Below, we present ORTHOBO, our orthogonalized framework for robust HPO addressing the failure modes discussed above (an overview is shown in Fig. 2. We first present our *main contribution*, the general *orthogonalized acquisition estimation framework*. Of note, we do not change the acquisition target, but rather improve its estimation through variance stabilization. Then, we provide example instantiations of our orthogonal framework for two common surrogate classes. In the main text, we focus on EI-based acquisition functions in order to keep the notation and theoretical arguments simple. The key object throughout is the marginal EI under surrogate uncertainty and ensemble aggregation. We provide extensions to constrained and parallel acquisition functions in Supplement B.

4.1 Orthogonalized acquisition estimation

We now address the instability due to acquisition noise. Our key idea is to introduce an *orthogonalized acquisition functional* that shares the same marginal target as the original EI but has reduced variance under the posterior $q_{m,t}$. Let $g(\theta) := \nabla_\theta \log q_{m,t}(\theta)$ denote the posterior score function.

We define the *orthogonalized expected improvement* as

$$\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta) \;:=\; \mathrm{EI}_m(\lambda;\theta) - \gamma_m(\lambda)^\top g(\theta), \tag{4}$$

where $\gamma_m(\lambda) := \Sigma_g^{-1}\,\mathrm{Cov}_{q_{m,t}}\big(g,\,\mathrm{EI}_m(\lambda;\cdot)\big)$ and $\Sigma_g := \mathrm{Cov}_{q_{m,t}}(g,g)$. The corresponding marginal acquisition is

$$\mathrm{EI}_m^{\mathrm{orth,marg}}(\lambda) \;:=\; \mathbb{E}_{\theta\sim q_{m,t}}\Big[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\Big]. \tag{5}$$

The associated MC estimator is

$$\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta_m^{(s)}), \qquad \theta_m^{(s)} \sim q_{m,t}. \tag{6}$$

Of note, the construction is non-trivial because $g(\theta)$ is based on the *parameter posterior*, whereas common MC pipelines sample only from the predictive posterior over function values. We therefore draw parameter samples explicitly and evaluate both EI_m and g at the same $\theta_m^{(s)}$. In practice, $\gamma_m(\lambda)$ is unknown and we use the empirical plug-in $\hat{\gamma}_m(\lambda)$ estimated from the same MC samples; this introduces an in-sample bias of order $O(1/S)$ that is dominated by the $O(1/\sqrt{S})$ sampling standard deviation at the MC budgets used in our experiments.[4]

Properties of the orthogonalized acquisition. The orthogonalized EI satisfies three key properties: (i) *target preservation*: it shares the same marginal as the original EI, so the BO target is unchanged; (ii) *variance reduction*: it has smaller variance under $q_{m,t}$; and (iii) *local robustness*: it is first-order insensitive to score-tilt perturbations of $q_{m,t}$.

Regularity for the score function. Throughout our analysis, we assume that the surrogate-parameter distribution $q_{m,t}$ has a differentiable density on a support $\Theta \subseteq \mathbb{R}^{d_\theta}$ and that the corresponding boundary term vanishes, i.e. $\int_\Theta \nabla_\theta q_{m,t}(\theta)\,d\theta = 0$. This holds, for example, if $\Theta = \mathbb{R}^{d_\theta}$ and $q_{m,t}(\theta) \to 0$ sufficiently fast as $\|\theta\| \to \infty$.

Theorem 4.1. *Assume $\mathbb{E}_{q_{m,t}}[\mathrm{EI}_m(\lambda;\theta)^2] < \infty$ and $\mathbb{E}_{q_{m,t}}[\|g(\theta)\|_2^2] < \infty$, that Σ_g is nonsingular, and that the standard score-function regularity conditions hold for $q_{m,t}$. Then:*

(i) Target preservation. $\mathbb{E}[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)] = \mathrm{EI}_m^{\mathrm{marg}}(\lambda;\theta)$.

(ii) Variance reduction. *Under $q_{m,t}$,*

$$\mathrm{Var}\big(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\big) = \mathrm{Var}\big(\mathrm{EI}_m(\lambda;\theta)\big) - \mathrm{Cov}(g,\mathrm{EI}_m)^\top\Sigma_g^{-1}\,\mathrm{Cov}(g,\mathrm{EI}_m) \le \mathrm{Var}\big(\mathrm{EI}_m(\lambda;\theta)\big). \tag{7}$$

(iii) Local robustness in score-tilt directions. *For any $b \in \mathbb{R}^{d_\theta}$, consider the tilted family $q_{m,t}^{(\varepsilon)}(\theta) \propto q_{m,t}(\theta)\exp\big(\varepsilon\, b^\top g(\theta)\big)$ for small $|\varepsilon|$. Holding $\gamma_m(\lambda)$ fixed at its $\varepsilon = 0$ value,*

$$\frac{d}{d\varepsilon}\,\mathbb{E}_{q_{m,t}^{(\varepsilon)}}\Big[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\Big]\Big|_{\varepsilon=0} \;=\; 0. \tag{8}$$

[4] Sample-splitting (cross-fitting) would restore exact finite-sample unbiasedness without changing the leading asymptotic variance; see Supplement D.

Proof. We provide a proof for Theorem 4.1 in Supplement D. □

From acquisition to estimator. Theorem 4.1(i) shows that orthogonalization preserves the BO target. Since the MC estimator is an i.i.d. sample mean, $\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda)) = \mathrm{Var}(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta))/S$, Theorem 4.1(ii) directly implies $\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda)) \le \mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}}(\lambda))$ for any MC budget S, which improves the stability of acquisition-based decisions. Theorem 4.1(iii) provides local robustness in the orthogonal ML sense [12, 35]: small score-tilt perturbations of $q_{m,t}$ do not affect the orthogonalized acquisition to first order. Importantly, Theorem 4.1 is a *within-model* result: orthogonalization stabilizes acquisition estimation conditional on $q_{m,t}$, but does not directly correct structural misspecification of the surrogate family. We address the latter through ensemble modeling in Section 4.3.

Proposition 4.2 (Pairwise ranking stability)**.** *Let $\lambda, \lambda' \in \Lambda$ such that $\Delta(\lambda,\lambda') := \mathrm{EI}_m^{\mathrm{marg}}(\lambda) - \mathrm{EI}_m^{\mathrm{marg}}(\lambda') > 0$. Let $\widehat{\Delta}_{\mathrm{MC}}(\lambda,\lambda')$ and $\widehat{\Delta}_{\mathrm{orth}}(\lambda,\lambda')$ denote the corresponding Monte Carlo difference estimators. Then both estimators are unbiased for $\Delta(\lambda,\lambda')$, and*

$$\mathbb{P}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda,\lambda') \le 0\Big) \;\le\; \frac{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda,\lambda')\Big)}{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda,\lambda')\Big) + \Delta(\lambda,\lambda')^2} \;\le\; \frac{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{MC}}(\lambda,\lambda')\Big)}{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{MC}}(\lambda,\lambda')\Big) + \Delta(\lambda,\lambda')^2}. \tag{9}$$

Proof. We provide a proof for Proposition 4.2 in Supplement D. □

Proposition 4.2 shows how acquisition-estimation noise propagates into BO decisions. High estimator variance can flip the estimated sign of the acquisition gap between two candidates points, causing BO to select a suboptimal point. Orthogonalization reduces the variance without changing the acquisition target, and therefore directly reduces the probability of such ranking errors.

4.2 Instantiations for common surrogates

We instantiate ORTHOBO for two surrogate classes common in HPO: (A) Gaussian process and (B) tree-based density surrogates. The first corresponds to the standard BO setting, while the second captures widely used non-Gaussian and nonparametric approaches such as TPE-style optimization.

(A) Gaussian process surrogate

For GP surrogate models, acquisition noise is often driven by sensitivity of the mean $\mu(\cdot)$ and variance $\sigma(\cdot)$ function to lengthscale and noise hyperparameters. We show that orthogonalization reduces MC fluctuations induced by uncertainty in these directions. This is especially useful early in optimization when the approximation of the posterior over the GP hyperparameters is diffuse. Let

$$f \sim \mathcal{GP}(m_\theta(\cdot), k_\theta(\cdot,\cdot)), \tag{10}$$

where $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ denotes surrogate hyperparameters, such as kernel lengthscales, amplitude, and noise variance. Given data $\mathcal{D}_t = \{(\lambda_i, y_i)\}_{i=1}^t$, the predictive distribution at a candidate λ is

$$f(\lambda) \mid \mathcal{D}_t, \theta \sim \mathcal{N}\big(\mu_t(\lambda;\theta), \sigma_t^2(\lambda;\theta)\big), \tag{11}$$

with

$$\mu_t(\lambda;\theta) = m_\theta(\lambda) + k_\theta(\lambda, X_t)^\top \left(K_\theta(X_t,X_t) + \sigma_n^2 I_t\right)^{-1} (y_t - m_\theta(X_t)), \tag{12}$$

$$\sigma_t^2(\lambda;\theta) = k_\theta(\lambda,\lambda) - k_\theta(\lambda,X_t)^\top \left(K_\theta(X_t,X_t) + \sigma_n^2 I_t\right)^{-1} k_\theta(\lambda,X_t), \tag{13}$$

with where $X_t = (\lambda_1,\dots,\lambda_t)$ and $y_t = (y_1,\dots,y_t)^\top$. We aim to maximize the Expected Improvement

$$\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta) = (f^\star - \mu_t(\lambda;\theta))\,\Phi(z_t(\lambda;\theta)) + \sigma_t(\lambda;\theta)\,\phi(z_t(\lambda;\theta)), \tag{14}$$

where

$$z_t(\lambda;\theta) = \frac{f^\star - \mu_t(\lambda;\theta)}{\sigma_t(\lambda;\theta)}, \qquad f^\star = \min_{i\le t} y_i, \tag{15}$$

and Φ and ϕ denote the standard normal cumulative distribution and probability density function.

We assume an approximate posterior $q_t(\theta)$ over GP hyperparameters, obtained for example via a Laplace approximation or variational Gaussian approximation. The marginal acquisition is then

$$\mathrm{EI}^{\mathrm{GP,marg}}(\lambda) = \mathbb{E}_{\theta\sim q_t}\left[\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta)\right], \tag{16}$$

and the orthogonalized estimator is

$$\widehat{\mathrm{EI}}^{\mathrm{GP,orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\left[\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta^{(s)}) - \gamma^{\mathrm{GP}}(\lambda)^\top g_t(\theta^{(s)})\right], \qquad \theta^{(s)} \sim q_t, \tag{17}$$

where $g_t(\theta) = \nabla_\theta \log q_t(\theta)$ and $\gamma^{\mathrm{GP}}(\lambda) = \mathrm{Cov}(g_t, g_t)^{-1}\,\mathrm{Cov}\big(g_t, \mathrm{EI}^{\mathrm{GP}}(\lambda;\theta)\big)$.

(B) Tree-based / TPE-style surrogate

As a second instantiation, we consider tree-based or density-based surrogates as in Tree-structured Parzen Estimators (TPE). Rather than modeling $p(y \mid \lambda)$ directly, TPE-style methods model

$$p(\lambda \mid y) = \begin{cases} \ell_\theta(\lambda), & y < y^\star, \\ g_\theta(\lambda), & y \geq y^\star, \end{cases} \tag{18}$$

where $y^\star$ is a quantile threshold and θ denotes model parameters, such as kernel density bandwidths, mixture weights, or tree parameters. In this setting, the acquisition is proportional to a density ratio:

$$\mathrm{EI}^{\mathrm{TPE}}(\lambda;\theta) \propto \frac{\ell_\theta(\lambda)}{g_\theta(\lambda)}. \tag{19}$$

More generally, we denote the Expected Improvement as $h^{\mathrm{TPE}}(\lambda;\theta) := \mathrm{EI}^{\mathrm{TPE}}(\lambda;\theta)$, where h^{TPE} represents either the exact acquisition induced by the density model or a practical approximation.

Unlike the GP case, differentiability of $q_t(\theta)$ may not be possible. We therefore consider a more general control-variate construction. Let $c_t(\theta) \in \mathbb{R}^{d_c}$ denote any random vector satisfying $\mathbb{E}_{q_t}[c_t(\theta)] = 0$, such as centered bootstrap statistics, centered density ratio proxies, centered bandwidth or split summaries, or score functions when they exist. We then obtain our orthogonalized estimator

$$\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\left[h^{\mathrm{TPE}}(\lambda;\theta^{(s)}) - \gamma^{\mathrm{TPE}}(\lambda)^\top c_t(\theta^{(s)})\right], \qquad \theta^{(s)} \sim q_t, \tag{20}$$

with

$$\gamma^{\mathrm{TPE}}(\lambda) = \mathrm{Cov}(c_t, c_t)^{-1}\,\mathrm{Cov}\big(c_t, h^{\mathrm{TPE}}(\lambda;\theta)\big). \tag{21}$$

4.3 ORTHOBO optimization procedure

We present the full optimization in Alg. 1.

Ensemble surrogates. To mitigate structural surrogate misspecification, we consider an ensemble of surrogate models $m = 1, \ldots, M$. Such ensembles can consist of Gaussian processes with different kernels, models with varying likelihood assumptions, various tree-based or neural surrogates. For each model m, we obtain the respective orthogonalized acquisition estimate $\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda)$. We then ensemble the acquisition as a weighted average of weights $\pi_{m,t}$

$$\widehat{\mathrm{EI}}^{\mathrm{ens}}(\lambda) = \sum_{m=1}^{M} \pi_{m,t}\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda). \tag{22}$$

Exponentially weighted aggregation. We update ensemble weights through tempered exponentially weighted aggregation [e.g., 25], a widely used approach in model aggregation [18, 32, 52]. A minimum weight floor [28] prevents premature collapse onto a single surrogate

$$\pi_{m,t} \propto \max(\delta,\, \pi_{m,t-1}\exp(\ell_{m,t}/\tau)), \tag{23}$$

Algorithm 1: ORTHOBO

Input: Initial data size n_0, budget T, number of surrogate models M, candidate set generator $\mathcal{C}_t$, numerical floor $\alpha > 0$, objective function f

Initialize $\mathcal{D}_0 = \{(\lambda_i, y_i)\}_{i=1}^{n_0}$ using random hyperparameter configurations;

for $t = 0, 1, \ldots, T-1$ **do**

  for $m = 1, \ldots, M$ **do**

    Fit surrogate model m on $\mathcal{D}_t$;

    Obtain approximate posterior $q_{m,t}(\theta)$;

  end

  Update ensemble weights $\pi_{m,t}$ for $m = 1, \ldots, M$;

  foreach $\lambda \in \mathcal{C}_t$ **do**

    for $m = 1, \ldots, M$ **do**

      Orthogonalization: Compute orthogonalized marginal EI estimate $\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda)$ via Eq. (6);

    end

    Ensemble strategy:

$$\widehat{\mathrm{EI}}^{\mathrm{ens}}(\lambda) = \sum_{m=1}^{M} \pi_{m,t}\,\widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda)$$

  end

  Outer log transformation:

$$\lambda_{t+1} = \arg\max_{\lambda\in\mathcal{C}_t} \log\big(\max(\widehat{\mathrm{EI}}^{\mathrm{ens}}(\lambda), \alpha)\big)$$

  Evaluate $y_{t+1} = f(\lambda_{t+1})$; update $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\lambda_{t+1}, y_{t+1})\}$;

end

Output: Best observed configuration in $\mathcal{D}_T$

where $\ell_{m,t}$ is a predictive log score obtained from m_t using (λ_{t+1}, y_{t+1}), $\tau > 0$ a temperature parameter, and $\delta > 0$ enforces a minimum weight. This results in a robust aggregation rule that balances exploitation of well-performing surrogates with persistent exploration across model classes. Note, that our overall methodology is independent and can be combined with other update strategies.

Optimization based on outer log stabilization. To improve numerical stability, we follow recent literature [2] and optimize the *outer log transformed* acquisition

$$\widehat{L}^{\text{ens}}(\lambda_t) = \log\big(\max(\widehat{\text{EI}}^{\text{ens}}(\lambda_t), \alpha)\big), \tag{24}$$

where $\alpha > 0$ ensures that the logarithm is well defined. We select the next hyperparameter configuration $\lambda_{t+1} = \arg\max_{\lambda \in \mathcal{C}_t} \widehat{L}^{\text{ens}}(\lambda)$, and evaluate it on the true objective to obtain y_{t+1}.

5 Empirical results

Experimental setup. We evaluate OrthoBO against various baselines from the literature. Specifically, we compare OrthoBO, qLogEI [2] (main comparison method), UCB [62], TuRBO [16], and Sobol sampling [e.g. 3] as a random baseline. Our experimental settings follow the literature [e.g. 2, 3]: unless otherwise noted, all experiments use 16 replications, $S = 512$, and we start all methods from the same $n_0 = 32$ Sobol points. In optimizing the acquisition functions, we use 512 raw samples and allow 8 restarts to prevent getting stuck in local minima [56]. To prevent confounding, we use $M = 1$ throughout the experiments in the main paper. We consider four standard synthetic regression problems with known global optima and varying dimensionality: Hartmann6, Ackley8, Michalewicz10, and Levy16 [e.g., 2, 10, 15, 41]. Details of the benchmark functions and surrogate choices are provided in Supplement E. We compare methods using best-so-far regret and use 95% CI from standard normal approximation.

Evaluation. Our goal is to demonstrate the theoretical properties of OrthoBO and show its superiority on real-world training and fine-tuning tasks, answering four distinct research questions. We first isolate the effect of orthogonalization on acquisition estimation to RQ1) show the *variance reduction* and RQ2) *ranking stability* properties of OrthoBO. Then, we evaluate the resulting RQ3) *efficiency gain* in terms of MC sampling budget. Finally, we show RQ4) the improved *downstream utility*. We present further experiments on misspecified surrogate families, weakly fitted hyperparameters or surrogates, and ensemble experiments in Supplement F.

RQ1) Variance reduction. We probe the acquisition function with random Sobol samples and compare raw and orthogonalized acquisition estimate variance. We use a GP surrogate. We report results for a Matérn-$5/2$ kernel with ARD on Michalewicz10 and Levy16 in Table 1; results for a RBF kernel and a TPE surrogate are in Supplement F.1. ⇒ *Across all settings,* ***orthogonalization reduces the variance****, confirming that* OrthoBO *substantially stabilizes the acquisition estimate.*

Table 1: Variance reduction on Sobol probes through orthogonalization.

	Michalewicz10		Levy16	
S	$\text{Var}(\widehat{\text{EI}}_m^{\text{MC}})$	$\text{Var}(\widehat{\text{EI}}_m^{\text{ortho}})$	$\text{Var}(\widehat{\text{EI}}_m^{\text{MC}})$	$\text{Var}(\widehat{\text{EI}}_m^{\text{ortho}})$
8	9.51e-08	4.54e-08	0.00106	0.000709
32	1.67e-08	1.07e-09	0.000208	2.27e-05

RQ2) Ranking stability. We repeatedly evaluate the acquisition function on the same fixed Sobol probe set and compare the rankings across repetitions. We report: (i) *probe variance*: mean variance of the acquisition value over the probe points across repeated acquisition evaluations; (ii) *top1 agreement*: fraction of repeats that select the same top-ranked probe point, (iii) the *flip rate* of adjacent high-ranked candidate pairs, and (iv) *regret*. The flip rate is particularly close to Proposition 4.2, as it measures how often local pairwise orderings change due to acquisition-estimation noise. The main results are in Table 2; further results are in Supplement F.1. ⇒ *Our variance reduction yields more stable rankings.*

Table 2: Ranking stability ($\pm$ std), via Sobol probes. Lower variance and flip rate indicate more stable acquisition estimates; higher top-1 agreement indicates more stable candidate rankings across probes.

	Method	Probe variance ↓	Top1 agreement ↑	Flip rate ↓	Regret ↓
Michalewicz10	qLogEI	$168.603_{154.76}$	$0.925_{0.16}$	$0.148_{0.12}$	$7.140_{0.01}$
	OrthoBO (ours)	$0.019_{0.03}$	$0.988_{0.06}$	$0.014_{0.03}$	$6.961_{0.36}$
Levy16	qLogEI	$233.655_{287.67}$	$0.917_{0.12}$	$0.069_{0.09}$	$82.194_{6.05}$
	OrthoBO (ours)	$0.073_{0.21}$	$0.988_{0.06}$	$0.048_{0.04}$	$58.938_{0.00}$

RQ3) MC efficiency. We use an isotropic RBF kernel, a common smooth GP prior, combining moderate surrogate mismatch with reduced acquisition-estimation budgets. We show the main results

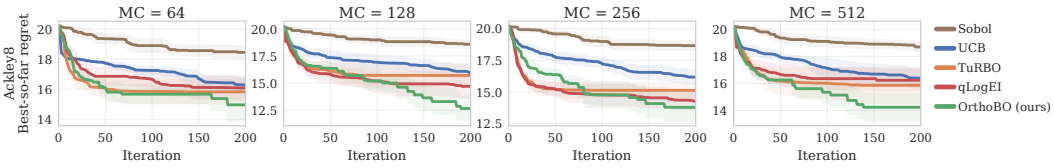

Figure 3: Best-so-far regret for different MC budgets $S \Rightarrow$ ORTHOBO *achieves the lowest regrets across all budgets.*

for varying S on Ackley8 in Fig. 3; further results are in Fig 12 in Supplement F. Reducing the MC budget makes the performance of acquisition-based methods more sensitive to estimation noise. Orthogonalization is most effective in earlier iterations, where MC estimation noise has a larger effect on the acquisition values. Across all benchmarks, our ORTHOBO remains competitive and frequently achieves the lowest regret.

RQ4) Downstream utility. We show how ORTHOBO improves hyperparameter optimization on two real-world tasks: (i) training a neural network and (ii) optimizing fine-tuning hyperparameters of a pre-trained vision transformer (ViT) [14] for manufacturing image classification.

Neural network. We employ the 5D CNN benchmark by Ament et al. [3] and include their RPR method as comparison. Details are in Supplement E.3. The corruption protocol prematurely stops training, providing unreliable objective values which can distort the surrogate fit, especially early in BO when only few observations are available. Results are in Fig. 4. Despite the corrupted observations, ORTHOBO improves steadily over the optimization horizon and attains the highest value.

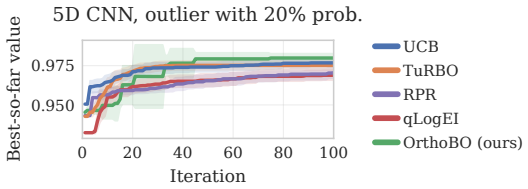

Figure 4: 5D CNN with corrupted evaluations. $\Rightarrow$ ORTHOBO *achieves the* ***strongest final performance****.*

Vision transformer. We optimize five fine-tuning hyperparameters on the industrial WM811K wafer-map dataset [33, 64] based on the F1 score. Details are provided in Appendix E.4. We present results in Fig. 5. ORTHOBO improves rapidly within the first few evaluations and maintains the highest best-so-far validation F1 score throughout most of the optimization horizon. Compared with qLogEI, ORTHOBO improves the best observed validation F1 score by up to 20 percentage points. This demonstrates that the proposed acquisition-stabilization mechanism also improves performance in a practical, high-cost HPO setting beyond synthetic benchmarks.

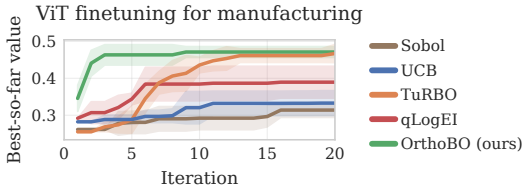

Figure 5: ViT. $\Rightarrow$ ORTHOBO ***improves faster than the baselines*** *and achieves the highest best-so-far validation score.*

6 Discussion

We introduced ORTHOBO, an orthogonalized framework for stabilizing acquisition estimation in Bayesian HPO. BO decisions depend on the ranking of estimated acquisition values, so MC noise in these estimates can directly induce ranking errors and lead to suboptimal evaluations. ORTHOBO addresses this failure mode by reducing variance while preserving the underlying marginal EI target. Our results show that this variance reduction improves ranking stability and translates into better downstream BO performance. ORTHOBO has limitations. First, it introduces additional computational overhead. However, this overhead is typically small compared with expensive objective evaluations in HPO (see Supplement C). Second, we focus only on EI-based acquisition functions. Note that the same variance-reduction principle can be extended to other acquisition functions but may require acquisition-specific orthogonalization schemes (see Supplement B.1). Overall, our work identifies acquisition-estimation noise as a distinct and practically relevant failure mode in BO. By reducing this noise, ORTHOBO stabilizes acquisition rankings and improves the quality of sequential hyperparameter decisions.

References

[1] A. AlBahar, I. Kim, and X. Yue. A robust asymmetric kernel function for Bayesian optimization, with application to image defect detection in manufacturing systems. *IEEE Transactions on Automation Science and Engineering*, 19(4):3222–3233, 2021.

[2] S. Ament, S. Daulton, D. Eriksson, M. Balandat, and E. Bakshy. Unexpected improvements to expected improvement for Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2023.

[3] S. Ament, E. Santorella, D. Eriksson, B. Letham, M. Balandat, and E. Bakshy. Robust Gaussian processes via relevance pursuit. In *Neural Information Processing Systems (NeurIPS)*, 2024.

[4] M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. Botorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[5] T. Ban, M. Ohue, and Y. Akiyama. Efficient hyperparameter optimization by using Bayesian optimization for drug-target interaction prediction. In *International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 2017.

[6] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Neural Information Processing Systems (NeurIPS)*, 2011.

[7] J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International Conference on Machine Learning (ICML)*, 2013.

[8] F. Berkenkamp, A. P. Schoellig, and A. Krause. No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019.

[9] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

[10] E. Bodin, M. Kaiser, I. Kazlauskaite, Z. Dai, N. Campbell, and C. H. Ek. Modulating surrogates for Bayesian optimization. In *International Conference on Machine Learning (ICML)*, 2020.

[11] I. Bogunovic and A. Krause. Misspecified Gaussian process bandit optimization. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[12] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] D. Eriksson and M. Poloczek. Scalable constrained Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[16] D. Eriksson, M. Pearce, J. Gardner, R. D. Turner, and M. Poloczek. Scalable global optimization via local Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[17] S. Falkner, A. Klein, and F. Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning (ICML)*, 2018.

[18] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[19] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Neural Information Processing Systems (NeurIPS)*, 2015.

[20] D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

[21] D. Frauen, M. Schröder, K. Hess, and S. Feuerriegel. Orthogonal survival learners for estimating heterogeneous treatment effects from time-to-event data. In *Neural Information Processing Systems (NeurIPS)*, 2025.

[22] D. Frauen, A. Deviyani, M. van der Schaar, and S. Feuerriegel. Nonparametric LLM evaluation from preference data. *arXiv preprint*, arXiv:2601.21816, 2026.

[23] P. I. Frazier. Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems*, pages 255–278. 2018.

[24] P. I. Frazier. A tutorial on Bayesian optimization. *arXiv preprint*, arXiv:1807.02811, 2018.

[25] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[26] J. González, Z. Dai, P. Hennig, and N. Lawrence. Batch Bayesian optimization via local penalization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[27] H. Ha, V. Nguyen, H. Tran-The, H. Zhang, X. Zhang, and A. v. d. Hengel. Provably efficient Bayesian optimization with unknown Gaussian process hyperparameter estimation. *arXiv preprint*, arXiv:2306.06844, 2023.

[28] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

[29] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Neural Information Processing Systems (NeurIPS)*, 2014.

[30] J. M. Hernández-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *International Conference on Machine Learning (ICML)*, 2015.

[31] K. Hess, D. Frauen, N. Kilbertus, and S. Feuerriegel. Debiased neural operators for estimating functionals. *arXiv preprint*, arXiv:2604.19296, 2026.

[32] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[33] J.-S. R. Jang. Mir-wm811k: Dataset for wafer map failure pattern recognition. `http://mirlab.org/dataset/public/`, 2015.

[34] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[35] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

[36] Q. Lu, K. D. Polyzos, B. Li, and G. B. Giannakis. Surrogate modeling for Bayesian optimization beyond a single Gaussian process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11283–11296, 2023.

[37] L. Mackey, V. Syrgkanis, and I. Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning (ICML)*, 2018.

[38] R. Martinez-Cantin, K. Tee, and M. McCourt. Practical Bayesian optimization in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[39] V. Melnychuk and S. Feuerriegel. GDR-learners: Orthogonal learning of generative models for potential outcomes. *arXiv preprint*, arXiv:2509.22953, 2025.

[40] H. B. Moss, D. S. Leslie, J. Gonzalez, and P. Rayson. GIBBON: General-purpose information-based Bayesian optimisation. *Journal of Machine Learning Research*, 22(235):1–49, 2021.

[41] H. B. Moss, S. W. Ober, and V. Picheny. Inducing point allocation for sparse Gaussian processes in high-throughput Bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[42] V. Nath, D. Yang, A. Hatamizadeh, A. A. Abidin, A. Myronenko, H. R. Roth, and D. Xu. The power of proxy data and proxy networks for hyper-parameter optimization in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.

[43] W. Neiswanger and A. Ramdas. Uncertainty quantification using martingales for misspecified Gaussian processes. In *International Conference on Algorithmic Learning Theory (ALT)*, 2021.

[44] X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[45] M. Oprescu, V. Syrgkanis, and Z. S. Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning (ICML)*, 2019.

[46] K. D. Polyzos, Q. Lu, and G. B. Giannakis. Bayesian optimization with ensemble learning models and adaptive expected improvement. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[47] A. Quitadadmo, J. Johnson, and X. Shi. Bayesian hyperparameter optimization for machine learning based eqtl analysis. In *International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017.

[48] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

[49] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[50] M. Schröder, V. Melnychuk, and S. Feuerriegel. Differentially private learners for heterogeneous treatment effects. *International Conference on Learning Representations (ICLR)*, 2025.

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

[52] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[53] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Neural Information Processing Systems (NeurIPS)*, 2012.

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

[55] L. Tani and C. Veelken. Comparison of Bayesian and particle swarm algorithms for hyperparameter optimisation in machine learning applications in high energy physics. *Computer Physics Communications*, 294:108955, 2024.

[56] A. Törn and A. Žilinskas. *Global optimization*, volume 350. Springer, 1989.

[57] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS Competition and Demonstration Track*, 2020.

[58] M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2, 2006.

[59] Z. Wang and S. Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International conference on machine learning*, pages 3627–3635. PMLR, 2017.

[60] S. Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint*, arXiv:2304.11127, 2023.

[61] J. Wilson, F. Hutter, and M. Deisenroth. Maximizing acquisition functions for Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[62] J. T. Wilson, R. Moriconi, F. Hutter, and M. P. Deisenroth. The reparameterization trick for acquisition functions. *arXiv preprint*, arXiv:1712.00424, 2017.

[63] J. Wu, S. Toscano-Palmerin, P. I. Frazier, and A. G. Wilson. Practical multi-fidelity Bayesian optimization for hyperparameter tuning. In *Uncertainty in Artificial Intelligence (UAI)*, 2020.

[64] M.-J. Wu, J.-S. R. Jang, and J.-L. Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1): 1–12, 2015.

[65] Y. Yuan, W. Wang, and W. Pang. A systematic comparison study on hyperparameter optimisation of graph neural networks for molecular property prediction. In *Proceedings of the genetic and evolutionary computation conference*, 2021.

A Extended related work

Below, we discuss more distant related literature, completing the discussion in Section 2. We first cover literature regarding the robustness of the three components of BO in our work, and finally present related work on orthogonal machine learning.

Marginalization over surrogate uncertainty. A standard source of instability in BO is uncertainty in the surrogate itself, especially in the hyperparameters such as kernel lengthscales. Prior work has emphasized that a fully Bayesian treatment of these hyperparameters can substantially improve robustness relative to empirical Bayes or plug-in optimization [53]. Related work has also incorporated hyperparameter uncertainty directly into acquisition design [29, 30]. While these approaches account for surrogate uncertainty by integrating over posterior uncertainty in the model, they do not directly address the additional MC instability that arises when the resulting acquisition value must itself be estimated.

Ensemble surrogates and model aggregation. A natural strategy for mitigating surrogate misspecification is to replace a single surrogate with an ensemble of surrogate models. Recent work has proposed GP ensembles and mixture-based surrogate models for BO, showing that combining multiple kernels or surrogate classes can improve robustness relative to committing to a single GP prior [36, 46]. These approaches address only structural misspecification at the model level and do not account for instability in the individual surrogates.

Acquisition functions and numerical stability. Acquisition functions are the core decision rule in BO, with classical choices including Probability of Improvement, Expected Improvement, and upper-confidence-based methods [34, 54]. Recent work has highlighted that vanilla EI and its variants can be numerically unstable, which can substantially degrade optimization performance [2]. This has motivated numerically stable reformulations based on the logarithmic scale, which are substantially easier to optimize in practice.

Orthogonal machine learning. Orthogonal machine learning is rooted in semi-parametric statistics [9], and is applied in many settings where the estimation target is defined based on unknown nuisance components [12, 20]. Very common examples are found in the missing data analysis [48] and causal inference literature [35, 44, 45, 49, 58]. Recently, orthogonal learning has also been expanded to other fields such as private machine learning [50], survival analysis [21], LLM evaluation [22], and general ML optimization topics such as neural operators or generative models [31, 39].

B Extension to constrained and parallel acquisitions

In Section 4, we develop an orthogonalized estimator for the standard Expected Improvement setting. The same construction extends to a broader class of acquisition functions, including constrained EI, parallel or batch EI, and their combinations. Below, we first provide a general extension to our standard EI setting in the main paper and then specialize it to constrained and parallel Bayesian optimization.

B.1 General acquisition-level extension

Let z denote a generic acquisition input. Depending on the setting, we write z to represent (i) a single candidate point λ in standard BO, (ii) a single candidate point λ in constrained BO, or (iii) a batch $\Lambda = (\lambda_1, \ldots, \lambda_q)$ in parallel BO.

For surrogate model m with parameter θ_m and approximate posterior $q_{m,t}(\theta)$, let $a_m(z; \theta_m)$ denote a generic acquisition function computed under parameter value θ_m. We define the marginal acquisition target

$$A_m(z) := \mathbb{E}_{\theta_m \sim q_{m,t}}[a_m(z; \theta_m)]\,. \tag{25}$$

As in Section 4, let $g_m(\theta_m) := \nabla_{\theta_m} \log q_{m,t}(\theta_m)$ denote the score function. With $\theta_m^{(s)} \overset{\text{i.i.d.}}{\sim} q_{m,t}$ the orthogonalized estimator is given by

$$\widehat{A}_m^{\text{orth}}(z) = \frac{1}{S} \sum_{s=1}^{S} \left[a_m(z; \theta_m^{(s)}) - \left(\text{Cov}(g_m, g_m)^{-1} \text{Cov}(g_m, a_m) \right)^\top g_m(\theta_m^{(s)}) \right]. \tag{26}$$

Proposition B.1 (Generic extension of the orthogonalized estimator)**.** *Fix an iteration t, a surrogate model m, and an acquisition input z. Assume that*

$$\mathbb{E}_{q_{m,t}}\left[\|g_m(\theta_m)\|_2^2 \right] < \infty, \qquad \mathbb{E}_{q_{m,t}}\left[a_m(\theta_m; z)^2 \right] < \infty, \tag{27}$$

and that $\text{Cov}(g_m, g_m)$ is nonsingular. Then:

1. *$\widehat{A}_m^{\text{orth}}(z)$ is unbiased for $A_m(z)$*
2. $$\text{Var}\Big(\widehat{A}_m^{\text{orth}}(z) \Big) = \frac{1}{S} \left(\text{Var}(h_m) - \text{Cov}(h_m, g_m)^\top \text{Cov}(g_m, g_m)^{-1} \text{Cov}(g_m, h_m) \right), \tag{28}$$

   and therefore

   $$\text{Var}\Big(\widehat{A}_m^{\text{orth}}(z) \Big) \le \text{Var}\Big(\widehat{A}_m^{\text{MC}}(z) \Big), \tag{29}$$

   where $\widehat{A}_m^{\text{MC}}(z) := \frac{1}{S} \sum_{s=1}^{S} h_m(\theta_m^{(s)}; z)$
3. *The residual $r_m(\theta_m; z) := a_m(\theta_m; z) - \left(\text{Cov}(g_m, g_m)^{-1} \text{Cov}(g_m, a_m) \right)^\top g_m(\theta_m)$ is orthogonal to the score directions.*

Proof. The proof is identical to that of the EI case in Supplement D after replacing $\text{EI}_m(\lambda; \theta_m)$ by the generic acquisition value $a_m(z; \theta_m)$. □

Remark. Proposition B.1 shows that the theoretical properties of our orthogonalized approach are acquisition-agnostic. We only require square integrability of the acquisition value under $q_{m,t}$ and the usual score-function regularity conditions.

B.2 Constrained Expected Improvement

Consider constrained BO with one objective f_1 and constraint functions $f_2, \ldots, f_M$, where, without loss of generality, the feasible set is defined by

$$f_i(\lambda) \le 0, \qquad i = 2, \ldots, M. \tag{30}$$

A standard constrained Expected Improvement (CEI) acquisition is

$$\mathrm{CEI}_m(\lambda;\theta_m) := \mathbb{E}\left[(f^\star - f_1(\lambda))_+ \prod_{i=2}^{M} \mathbf{1}\{f_i(\lambda)\le 0\} \,\middle|\, \mathcal{D}_t, \theta_m\right], \tag{31}$$

where $f^\star$ denotes the current best *feasible* objective value (i.e., the minimum of f_1 over feasible observations).

The corresponding marginal target is $\mathrm{CEI}_m^{\mathrm{marg}}(\lambda) := \mathbb{E}_{\theta_m\sim q_{m,t}}[\mathrm{CEI}_m(\lambda;\theta_m)]$. Defining $h_m^{\mathrm{CEI}}(\theta_m;\lambda) := \mathrm{CEI}_m(\lambda;\theta_m)$, the orthogonalized CEI estimator is given by

$$\widehat{\mathrm{CEI}}_m^{\mathrm{orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\left[\mathrm{CEI}_m(\lambda;\theta_m^{(s)}) - \left(\mathrm{Cov}(g_m,g_m)^{-1}\,\mathrm{Cov}\big(g_m, h_m^{\mathrm{CEI}}\big)\right)^\top g_m(\theta_m^{(s)})\right]. \tag{32}$$

Corollary B.2 (Orthogonalized constrained EI)**.** *Assume* $\mathbb{E}_{q_{m,t}}\left[\left(\mathrm{CEI}_m(\lambda;\theta_m)\right)^2\right] < \infty$. *Then Proposition B.1 applies with* $a_m(z;\theta_m) = \mathrm{CEI}_m(\lambda;\theta_m)$.

Proof. The indicator product is bounded by 1, so $0 \le \mathrm{CEI}_m(\lambda;\theta_m) \le \mathrm{EI}_{m,1}(\lambda;\theta_m) := \mathbb{E}[(f^\star - f_1(\lambda))_+ \mid \mathcal{D}_t, \theta_m]$. Square integrability of $\mathrm{EI}_{m,1}$ thus implies square integrability of CEI_m, and the corollary follows from Proposition B.1. □

B.3 Parallel or batch Expected Improvement

In parallel BO, we select a batch $\Lambda = (\lambda_1,\dots,\lambda_q)$ of q candidate points at each iteration. The standard batch expected improvement (qEI) is given by

$$\mathrm{qEI}_m(\Lambda;\theta_m) := \mathbb{E}\left[\max_{j=1,\dots,q}\big(f^\star - f(\lambda_j)\big)_+ \,\middle|\, \mathcal{D}_t, \theta_m\right], \tag{33}$$

with corresponding marginal target is $\mathrm{qEI}_m^{\mathrm{marg}}(\Lambda) := \mathbb{E}_{\theta_m\sim q_{m,t}}[\mathrm{qEI}_m(\Lambda;\theta_m)]$.

We define $h_m^{\mathrm{qEI}}(\theta_m;\Lambda) := \mathrm{qEI}_m(\Lambda;\theta_m)$. Then the orthogonalized qEI estimator is given by

$$\widehat{\mathrm{qEI}}_m^{\mathrm{orth}}(\Lambda) = \frac{1}{S}\sum_{s=1}^{S}\left[\mathrm{qEI}_m(\Lambda;\theta_m^{(s)}) - \left(\mathrm{Cov}(g_m,g_m)^{-1}\,\mathrm{Cov}\big(g_m, h_m^{\mathrm{qEI}}\big)\right)^\top g_m(\theta_m^{(s)})\right]. \tag{34}$$

Corollary B.3 (Orthogonalized batch EI)**.** *Assume* $\mathbb{E}_{q_{m,t}}\left[\left(\mathrm{qEI}_m(\Lambda;\theta_m)\right)^2\right] < \infty$. *Then Proposition B.1 applies with* $a_m(z;\theta_m) = \mathrm{qEI}_m(\Lambda;\theta_m)$.

Proof. Observe that for every fixed batch $X = (x_1,\dots,x_q)$ and parameter value θ, $0 \le \mathrm{qEI}_m(X;\theta) \le \sum_{j=1}^q \mathrm{EI}_m(x_j;\theta)$. Consequently, $\mathrm{qEI}_m(X;\theta)^2 \le \left(\sum_{j=1}^q \mathrm{EI}_m(x_j;\theta)\right)^2$. In particular, if $\mathbb{E}_{\theta\sim q_{m,t}}\left[\left(\sum_{j=1}^q \mathrm{EI}_m(x_j;\theta)\right)^2\right] < \infty$, then as well $\mathbb{E}_{\theta\sim q_{m,t}}\left[\mathrm{qEI}_m(X;\theta)^2\right] < \infty$. The corollary now directly follows from Proposition B.1. □

Unlike standard EI, qEI does not admit a closed-form expression. Therefore, we follow common practice and approximate it by inner MC sampling. Let

$$\widehat{\mathrm{qEI}}_m(\Lambda;\theta_m) = \frac{1}{N}\sum_{n=1}^{N}\max_{j=1,\dots,q}\big(f^\star - \xi^{(n)}(\lambda_j)\big)_+, \qquad \xi^{(n)} \sim p(\cdot \mid \mathcal{D}_t, \theta_m), \tag{35}$$

where the inner MC estimator is assumed to be unbiased for $\mathrm{qEI}_m(\Lambda;\theta_m)$.

The resulting orthogonalized MC estimator is then given by

$$\widehat{\mathrm{qEI}}_{m,\mathrm{nested}}^{\mathrm{orth}}(\Lambda) = \frac{1}{S}\sum_{s=1}^{S}\left[\widehat{\mathrm{qEI}}_m(\Lambda;\theta_m^{(s)}) - \left(\mathrm{Cov}(g_m,g_m)^{-1}\,\mathrm{Cov}(g_m, \widehat{\mathrm{qEI}}(\Lambda;\theta_m))\right)^\top g_m(\theta_m^{(s)})\right]. \tag{36}$$

Importantly, under the assumption that the inner estimator (35) is unbiased for $\mathrm{qEI}_m(X;\theta_m)$ for every fixed θ_m, and that all required second moments exist, then $\widehat{\mathrm{qEI}}^{\mathrm{orth}}_{m,\mathrm{nested}}(\Lambda)$ is unbiased for $\mathrm{qEI}_m(X;\theta_m)$ as well due to the law of iterated expectations.

C Theoretical properties

C.1 Computational Complexity

GP surrogate. For an exact GP surrogate, fitting requires $O(t^3)$ time due to kernel matrix inversion. Per candidate λ, computing $\mu_t(\lambda;\theta)$ and $\sigma_t^2(\lambda;\theta)$ for a fixed θ costs $O(t^2)$ in a naive implementation, or $O(t)$ if Cholesky factors are reused appropriately. For S posterior samples, marginal EI estimation costs $O(St^2)$ or $O(St)$, depending on the implementation.

The orthogonalization step introduces two types of cost. First, once per BO iteration, one computes the score covariance matrix $\widehat{\Sigma}_g := \widehat{\mathrm{Cov}}(g_t, g_t) \in \mathbb{R}^{d_\theta \times d_\theta}$, which depends only on the surrogate posterior samples and not on the candidate λ. This costs $O(Sd_\theta^2)$, followed by $O(d_\theta^3)$ to invert or factorize $\widehat{\Sigma}_g$. Second, for each candidate λ, one computes the cross-covariance vector $\widehat{c}_{gh}(\lambda) := \widehat{\mathrm{Cov}}(g_t, h(\cdot;\lambda))$, which costs $O(Sd_\theta)$, followed by $O(d_\theta^2)$ for a matrix–vector solve or multiplication with the precomputed factorization of $\widehat{\Sigma}_g^{-1}$. Hence, the additional per-candidate orthogonalization overhead is $O(Sd_\theta + d_\theta^2)$, and the total per-candidate cost is $O(St^2 + Sd_\theta + d_\theta^2)$, plus a one-time per-iteration preprocessing cost of $O(Sd_\theta^2 + d_\theta^3)$.

We compare the average per-step runtimes of ORTHOBO with qLogEI in Table 3. We observe that the runtimes vary with the dimensionality of the problem, but the overhead over MC-based qLogEI is often around 2 s. For our implementation, we used the `BoTorch` framework, which has made MC-sampling highly efficient. We thus expect further runtime reductions from a more improved orthogonalization implementation.

Table 3: Average per-step runtimes (in seconds) for ORTHOBO and qLogEI.

	Hartmann6	Ackley8	Michalewicz10	Levy16
qLogEI	0.696	0.156	0.325	0.582
ORTHOBO	1.607	0.492	2.362	2.563

Tree-based / TPE-style surrogate. Tree-based or density-based surrogates are typically cheaper to fit. A single fit often requires $O(t \log t)$ or comparable cost, depending on the underlying model. Per-candidate acquisition evaluation is usually $O(1)$ or logarithmic in t. As in the GP case, the covariance matrix of the control variate, $\widehat{\Sigma}_c := \widehat{\mathrm{Cov}}(c_t, c_t) \in \mathbb{R}^{d_c \times d_c}$, depends only on the posterior samples and can therefore be computed once per BO iteration at cost $O(Sd_c^2)$, followed by $O(d_c^3)$ for inversion or factorization. For each candidate, one then computes the cross-covariance vector $\widehat{c}_{ch}(\lambda) := \widehat{\mathrm{Cov}}(c_t, h(\cdot;\lambda))$, at cost $O(Sd_c)$, followed by $O(d_c^2)$ for the matrix–vector solve. Hence, the additional per-candidate orthogonalization overhead is $O(Sd_c + d_c^2)$, with one-time per-iteration preprocessing cost $O(Sd_c^2 + d_c^3)$.

We report average per-step runtimes for orthogonalized TPE versus MC-based TPE in Table 4. Runtime differences are generally small.

Table 4: Average per-step runtimes (in seconds) for ORTHOBO and MC-based TPE.

	Hartmann6	Ackley8	Michalewicz10	Levy16
MC-TPE	0.224	0.347	0.356	0.754
ORTHOBO	0.390	0.736	2.533	0.360

Ensemble overhead. For an ensemble of M surrogate models, the total cost scales linearly as $O\left(\sum_{m=1}^{M} \mathrm{cost}_m\right)$. In practice, the main computational advantage of debiasing through orthogonalization is not that it makes each acquisition evaluation cheaper, but that it can achieve comparable acquisition stability with a smaller MC budget S.

C.2 Variance reduction and ranking stability for GP and TPE

We provide versions of Theorem 4.1 and Proposition 4.2 for the two surrogate instantiations GP and TPE. The proofs follow directly from the proofs of the original theorem and proposition in Supplement D.

Theorem C.1 (Variance reduction for the GP surrogate)**.** *Fix an iteration t and a candidate $\lambda \in \Lambda$. Assume $\mathbb{E}_{q_t}\|g_t(\theta)\|_2^2 < \infty$, $\mathbb{E}_{q_t}[(\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta))^2] < \infty$, and that $\mathrm{Cov}(g_t, g_t)$ is nonsingular. Define*

$$\mathrm{EI}^{\mathrm{GP,orth}}(\lambda;\theta) := \mathrm{EI}^{\mathrm{GP}}(\lambda;\theta) - \gamma^{\mathrm{GP}}(\lambda)^\top g_t(\theta). \tag{37}$$

Then:

1. *Target preservation:* $\mathbb{E}_{q_t}[\mathrm{EI}^{\mathrm{GP,orth}}(\lambda;\theta)] = \mathrm{EI}^{\mathrm{GP,marg}}(\lambda)$.
2. *Variance reduction:*

   $$\mathrm{Var}\big(\mathrm{EI}^{\mathrm{GP,orth}}(\lambda;\theta)\big) = \mathrm{Var}\big(\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta)\big) - \mathrm{Cov}(\mathrm{EI}^{\mathrm{GP}}, g_t)^\top \mathrm{Cov}(g_t, g_t)^{-1}\,\mathrm{Cov}(g_t, \mathrm{EI}^{\mathrm{GP}}),$$

   and therefore $\mathrm{Var}(\mathrm{EI}^{\mathrm{GP,orth}}(\lambda;\theta)) \le \mathrm{Var}(\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta))$.

By the i.i.d. MC variance identity, the corresponding estimator satisfies $\mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{GP,orth}}(\lambda)) \le \mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{GP,MC}}(\lambda))$ *for any MC budget S.*

Proposition C.2 (Ranking stability for the GP surrogate)**.** *Let $\lambda, \lambda' \in \Lambda$ with $\Delta^{\mathrm{GP}}(\lambda,\lambda') := \mathrm{EI}^{\mathrm{GP,marg}}(\lambda) - \mathrm{EI}^{\mathrm{GP,marg}}(\lambda') > 0$. Let $\widehat{\Delta}^{\mathrm{GP,orth}}$ denote the MC estimator obtained by applying the orthogonalization construction in Theorem C.1 to the difference functional $\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta) - \mathrm{EI}^{\mathrm{GP}}(\lambda';\theta)$. Then*

$$\mathbb{P}\Big(\widehat{\Delta}^{\mathrm{GP,orth}} \le 0\Big) \;\le\; \frac{\mathrm{Var}(\widehat{\Delta}^{\mathrm{GP,orth}})}{(\Delta^{\mathrm{GP}}(\lambda,\lambda'))^2} \;\le\; \frac{\mathrm{Var}(\widehat{\Delta}^{\mathrm{GP,MC}})}{(\Delta^{\mathrm{GP}}(\lambda,\lambda'))^2}. \tag{38}$$

Theorem C.3 (Variance reduction for the tree-based / TPE-style surrogate)**.** *Fix an iteration t and a candidate $\lambda \in \Lambda$. Assume $\mathbb{E}_{q_t}[\|c_t(\theta)\|_2^2] < \infty$, $\mathbb{E}_{q_t}[(h^{\mathrm{TPE}}(\lambda;\theta))^2] < \infty$, $\mathbb{E}_{q_t}[c_t(\theta)] = 0$, and that $\mathrm{Cov}(c_t, c_t)$ is nonsingular. Define*

$$h^{\mathrm{TPE,orth}}(\lambda;\theta) := h^{\mathrm{TPE}}(\lambda;\theta) - \gamma^{\mathrm{TPE}}(\lambda)^\top c_t(\theta). \tag{39}$$

Then:

1. *Target preservation:* $\mathbb{E}_{q_t}[h^{\mathrm{TPE,orth}}(\lambda;\theta)] = \mathrm{EI}^{\mathrm{TPE,marg}}(\lambda)$.
2. *Variance reduction:*

   $$\mathrm{Var}\big(h^{\mathrm{TPE,orth}}(\lambda;\theta)\big) = \mathrm{Var}\big(h^{\mathrm{TPE}}(\lambda;\theta)\big) - \mathrm{Cov}(h^{\mathrm{TPE}}, c_t)^\top \mathrm{Cov}(c_t, c_t)^{-1}\,\mathrm{Cov}(c_t, h^{\mathrm{TPE}}),$$

   and therefore $\mathrm{Var}(h^{\mathrm{TPE,orth}}(\lambda;\theta)) \le \mathrm{Var}(h^{\mathrm{TPE}}(\lambda;\theta))$.

By the i.i.d. MC variance identity, the corresponding estimator satisfies $\mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}}(\lambda)) \le \mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{TPE,MC}}(\lambda))$ *for any MC budget S.*

Proposition C.4 (Ranking stability for the tree-based / TPE-style surrogate)**.** *Let $\lambda, \lambda' \in \Lambda$ with $\Delta^{\mathrm{TPE}}(\lambda,\lambda') := \mathrm{EI}^{\mathrm{TPE,marg}}(\lambda) - \mathrm{EI}^{\mathrm{TPE,marg}}(\lambda') > 0$. Let $\widehat{\Delta}^{\mathrm{TPE,orth}}$ denote the MC estimator obtained by applying the orthogonalization construction in Theorem C.3 to the difference functional $h^{\mathrm{TPE}}(\lambda;\theta) - h^{\mathrm{TPE}}(\lambda';\theta)$. Then*

$$\mathbb{P}\Big(\widehat{\Delta}^{\mathrm{TPE,orth}} \le 0\Big) \;\le\; \frac{\mathrm{Var}(\widehat{\Delta}^{\mathrm{TPE,orth}})}{(\Delta^{\mathrm{TPE}}(\lambda,\lambda'))^2} \;\le\; \frac{\mathrm{Var}(\widehat{\Delta}^{\mathrm{TPE,MC}})}{(\Delta^{\mathrm{TPE}}(\lambda,\lambda'))^2}. \tag{40}$$

D Proofs of the main theorems and propositions

D.1 Proof of Theorem 4.1

Theorem 4.1 Assume $\mathbb{E}_{q_{m,t}}[\mathrm{EI}_m(\lambda;\theta)^2] < \infty$ and $\mathbb{E}_{q_{m,t}}[\|g(\theta)\|_2^2] < \infty$, that Σ_g is nonsingular, and that the standard score-function regularity conditions hold for $q_{m,t}$. Then:

(i) Target preservation. $\mathrm{EI}_m^{\mathrm{orth,marg}}(\lambda) = \mathrm{EI}_m^{\mathrm{marg}}(\lambda)$.

(ii) Variance reduction. Under $q_{m,t}$,

$$\mathrm{Var}\big(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\big) = \mathrm{Var}\big(\mathrm{EI}_m(\lambda;\theta)\big) - \mathrm{Cov}(g,\mathrm{EI}_m)^\top \Sigma_g^{-1}\,\mathrm{Cov}(g,\mathrm{EI}_m) \le \mathrm{Var}\big(\mathrm{EI}_m(\lambda;\theta)\big). \tag{41}$$

(iii) Local robustness in score-tilt directions. For any $b \in \mathbb{R}^{d_\theta}$, consider the tilted family $q_{m,t}^{(\varepsilon)}(\theta) \propto q_{m,t}(\theta)\exp\big(\varepsilon\, b^\top g(\theta)\big)$ for small $|\varepsilon|$. Holding $\gamma_m(\lambda)$ fixed at its $\varepsilon = 0$ value,

$$\frac{d}{d\varepsilon}\,\mathbb{E}_{q_{m,t}^{(\varepsilon)}}\Big[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\Big]\Big|_{\varepsilon=0} \;=\; 0. \tag{42}$$

Proof. Let $h(\theta) := \mathrm{EI}_m(\lambda;\theta)$ and $g(\theta) := \nabla_\theta \log q_{m,t}(\theta)$, and write $\gamma := \gamma_m(\lambda) = \Sigma_g^{-1}\,\mathrm{Cov}(g,h)$. Then

$$\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta) = h(\theta) - \gamma^\top g(\theta). \tag{43}$$

(i) Target preservation. Under the assumed regularity conditions, the score function satisfies

$$\mathbb{E}_{q_{m,t}}[g(\theta)] = \int \nabla_\theta \log q_{m,t}(\theta)\, q_{m,t}(\theta)\, d\theta = \int \nabla_\theta q_{m,t}(\theta)\, d\theta = 0. \tag{44}$$

Hence

$$\mathrm{EI}_m^{\mathrm{orth,marg}}(\lambda) = \mathbb{E}_{q_{m,t}}[h(\theta)] - \gamma^\top \mathbb{E}_{q_{m,t}}[g(\theta)] = \mathbb{E}_{q_{m,t}}[h(\theta)] = \mathrm{EI}_m^{\mathrm{marg}}(\lambda). \tag{45}$$

(ii) Variance reduction. For any $a \in \mathbb{R}^{d_\theta}$,

$$\mathrm{Var}_{q_{m,t}}(h - a^\top g) = \mathrm{Var}(h) - 2a^\top\,\mathrm{Cov}(g,h) + a^\top \Sigma_g\, a. \tag{46}$$

Since Σ_g is positive definite by assumption, this quadratic in a is uniquely minimized at

$$a^\star = \Sigma_g^{-1}\,\mathrm{Cov}(g,h) = \gamma. \tag{47}$$

Substituting $a = \gamma$ gives

$$\mathrm{Var}\big(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\big) = \mathrm{Var}(h) - \mathrm{Cov}(g,h)^\top \Sigma_g^{-1}\,\mathrm{Cov}(g,h). \tag{48}$$

Because Σ_g is positive definite, the quadratic form $\mathrm{Cov}(g,h)^\top \Sigma_g^{-1}\,\mathrm{Cov}(g,h) \ge 0$, and therefore

$$\mathrm{Var}\big(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\big) \;\le\; \mathrm{Var}\big(\mathrm{EI}_m(\lambda;\theta)\big). \tag{49}$$

(iii) Local robustness. Throughout this part we hold γ fixed at its $\varepsilon = 0$ value. Consider the exponentially tilted family

$$q_{m,t}^{(\varepsilon)}(\theta) = \frac{q_{m,t}(\theta)\exp(\varepsilon\, b^\top g(\theta))}{Z(\varepsilon)}, \qquad Z(\varepsilon) = \int q_{m,t}(\theta)\exp(\varepsilon\, b^\top g(\theta))\, d\theta. \tag{50}$$

For any integrable test function φ, differentiation under the integral sign (justified by the square-integrability assumptions and dominated convergence in a neighborhood of $\varepsilon = 0$) yields

$$\frac{d}{d\varepsilon}\,\mathbb{E}_{q_{m,t}^{(\varepsilon)}}[\varphi(\theta)]\Big|_{\varepsilon=0} = \mathrm{Cov}_{q_{m,t}}\big(\varphi(\theta), b^\top g(\theta)\big). \tag{51}$$

Apply with $\varphi(\theta) = \mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta) = h(\theta) - \gamma^\top g(\theta)$:

$$\mathrm{Cov}\big(g,\, h - \gamma^\top g\big) = \mathrm{Cov}(g,h) - \Sigma_g\,\gamma = \mathrm{Cov}(g,h) - \Sigma_g\Sigma_g^{-1}\,\mathrm{Cov}(g,h) = 0. \tag{52}$$

Hence

$$\frac{d}{d\varepsilon}\,\mathbb{E}_{q_{m,t}^{(\varepsilon)}}\Big[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\Big]\Big|_{\varepsilon=0} = b^\top\,\mathrm{Cov}\big(g,\, h - \gamma^\top g\big) = 0 \tag{53}$$

for every $b \in \mathbb{R}^{d_\theta}$. The orthogonalized acquisition is therefore first-order insensitive to score-tilt perturbations of $q_{m,t}$. □

D.2 Proof of Proposition 4.2

Proposition 4.2 Let $\lambda, \lambda' \in \Lambda$ such that $\Delta(\lambda, \lambda') := \mathrm{EI}_m^{\mathrm{marg}}(\lambda) - \mathrm{EI}_m^{\mathrm{marg}}(\lambda') > 0$. Let $\widehat{\Delta}_{\mathrm{MC}}(\lambda, \lambda')$ and $\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')$ denote the corresponding Monte Carlo difference estimators. Then both estimators are unbiased for $\Delta(\lambda, \lambda')$, and

$$\mathbb{P}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda') \le 0\Big) \;\le\; \frac{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')\Big)}{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')\Big) + \Delta(\lambda, \lambda')^2} \;\le\; \frac{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{MC}}(\lambda, \lambda')\Big)}{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{MC}}(\lambda, \lambda')\Big) + \Delta(\lambda, \lambda')^2}. \tag{54}$$

Proof. By unbiasedness from Theorem 4.1,

$$\mathbb{E}\Big[\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')\Big] = \Delta(\lambda, \lambda'). \tag{55}$$

Hence

$$\mathbb{P}\big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda') \le 0\big) = \mathbb{P}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda') - \Delta(\lambda, \lambda') \le -\Delta(\lambda, \lambda')\Big). \tag{56}$$

Applying Cantelli's inequality to the centered random variable

$$Z := \widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda') - \Delta(\lambda, \lambda') \tag{57}$$

gives

$$\mathbb{P}(Z \le -a) \le \frac{\mathrm{Var}(Z)}{\mathrm{Var}(Z) + a^2}, \qquad a > 0. \tag{58}$$

With $a = \Delta(\lambda, \lambda')$, this yields

$$\mathbb{P}\big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda') \le 0\big) \le \frac{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')\Big)}{\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')\Big) + \Delta(\lambda, \lambda')^2}. \tag{59}$$

The second inequality follows from Theorem 4.1, which gives

$$\mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{orth}}(\lambda, \lambda')\Big) \le \mathrm{Var}\Big(\widehat{\Delta}_{\mathrm{MC}}(\lambda, \lambda')\Big). \tag{60}$$

Since the map $v \mapsto v/(v + \Delta^2)$ is increasing for $v \ge 0$, the result follows. □

E Implementation details

General implementation details All experiments were conducted on a proprietary compute cluster. Our experiments utilized 4 CPU cores, 16 GB of main memory, and a A100 MIG slice with 7GB of VRAM. We used `BoTorch` for all code implementations. In total, we use $\sim$ 1000 GPU hours including initial experimentation and bug fixing.

License details We used BoTorch, which is MIT licensed. We implemented all methods and benchmarks in this framework. The WM811K dataset [33, 64] is available at `http://mirlab.org/dataSet/public/` and allows *usage* under the following conditions (copied verbatim from the license document):

```
    2. Redistribution and use in any form must be
    accompanied by the following two citations:

 [1] Ming-Ju Wu, Jyh-Shing Roger Jang, and Jui-Long Chen,
    "Wafer Map Failure Pattern Recognition and
    Similarity Ranking for Large-Scale Data Sets,"
    in IEEE Transactions on Semiconductor Manufacturing,
    vol. 28, no. 1, pp. 1-12, Feb. 2015, doi: 10.1109/TSM.2014.2364237.

    [2] MIR-WM811K: Dataset for wafer map failure pattern recognition,
    2015 http://mirlab.org/dataset/public/
```

We included these citations at the appropriate places. Further, we received written permission by the authors to use the dataset for our research.

E.1 TPE implementation

TPE implementation. For the TPE-based experiments, we approximate the parameter distribution $q_t(\theta)$ by a nonparametric bootstrap over the current BO history. Concretely, for each sample $s = 1, \dots, S$, we resample the observed data with replacement, fit a TPE surrogate $\theta^{(s)}$, and evaluate the corresponding TPE acquisition

$$h_{\mathrm{TPE}}(\lambda; \theta^{(s)}) \propto \frac{\ell_{\theta^{(s)}}(\lambda)}{g_{\theta^{(s)}}(\lambda)}.$$

We use the empirical γ-quantile split with $\gamma = 0.2$ to define the good and bad sets.

As control variate, we use centered bootstrap summary statistics. For each bootstrap fit $\theta^{(s)}$, we form

$$\phi(\theta^{(s)}) = \left[\log b^{(s)}_{\mathrm{good},1}, \dots, \log b^{(s)}_{\mathrm{good},d}, \log b^{(s)}_{\mathrm{bad},1}, \dots, \log b^{(s)}_{\mathrm{bad},d}, y^{\star,(s)}\right]^\top,$$

where $b^{(s)}_{\mathrm{good},j}$ and $b^{(s)}_{\mathrm{bad},j}$ are the fitted KDE bandwidths in dimension j, and $y^{\star,(s)}$ is the bootstrap split threshold. The implemented control variate is then

$$c_t(\theta^{(s)}) = \phi(\theta^{(s)}) - \frac{1}{S}\sum_{r=1}^{S} \phi(\theta^{(r)}),$$

which is zero-mean by construction. The orthogonalized TPE estimator is

$$\widehat{\mathrm{EI}}^{c}_{\mathrm{TPE,orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\left(h_{\mathrm{TPE}}(\lambda; \theta^{(s)}) - \gamma_{\mathrm{TPE}}(\lambda)^\top c_t(\theta^{(s)})\right).$$

E.2 Misspecified kernel family experiments

E.3 CNN outlier experiment

We include the CNN outlier experiment used by Ament et al. [3]. Specifically, the goal is to optimize the test-set accuracy of a CNN trained on MNIST images. To model corrupted evaluations, we

Function	Qualitative structure	Kernel family		Reason for misspecification
		Mildly misspecified	**Strongly misspecified**	
Hartmann6	Smooth, anisotropic, interaction-heavy, multimodal	Isotropic RBF or isotropic Matérn	Linear	Isotropic kernel variants miss dimension-specific lengthscales. Linear kernels cannot capture the nonlinear multimodal structure.
Ackley8	Highly multimodal, oscillatory, with many local minima	Smooth RBF	Linear	Very smooth kernels tend to smear out Ackley's rugged local structure. Linear kernels are too simple for the landscape.
Michalewicz10	Sharp valleys, strong multimodality, narrow optima	Isotropic RBF or Matérn	Linear	Isotropic kernels blur the narrow valleys and local optima. Linear kernels are strongly misspecified.
Levy16	Rugged, oscillatory, with nonlinear interactions	Overly smooth isotropic RBF	Linear	Isotropic smooth kernels underfit sharper local variation and ignore anisotropy. Linear kernels cannot represent the nonlinear oscillatory structure.

Table 5: Synthetic benchmark functions, their qualitative structure, and examples of mildly and strongly misspecified kernel families.

prematurely stop training with probability $20\,\%$ after observing between 100 and 1000 training samples. We optimize the same parameters as Ament et al. [3]. For comparison, we also include their RPR method. To not implicitly mitigate the corrupted evaluations, we only use $n_0 = 2$ initial points. In initial experiments, we found that using larger n_0 made the problem *too easy for all methods*; all reached $\sim$100 % test accuracy after few iterations.

E.4 Case study experiment, additional details

Dataset. For our real-world case study, we use the large-scale industrial WM811K dataset [33, 64]. It contains $811\,457$ real-world wafer map images and annotations of common failure types. The images have pixel values $0, 1, 2$, where 0 is background, 1 are OK regions, and 2 are faulty regions. Depending on the clustering of the failures, one of eight class labels is assigned. Exemplary failure patterns are shown in Figure 6. Images without any failures are labelled with an additional class label. We resize all wafer maps to a common (48×48) resolution, rescale the data to $(0, 1)$ by dividing by $/2$, and then normalize using ImageNet statistics.

The dataset is strongly imbalanced, and we randomly split the data into a $80\,\%$ train and $20\,\%$ test set, stratifying on the labels to keep class distributions equal across the two splits. During fine-tuning, we use an (inverse) class-weighted cross-entropy loss.

Model. As a model, we use an ImageNet [13] pretrained ViT B16 [14] model. We freeze the backbone model and train only the penultimate output layers.

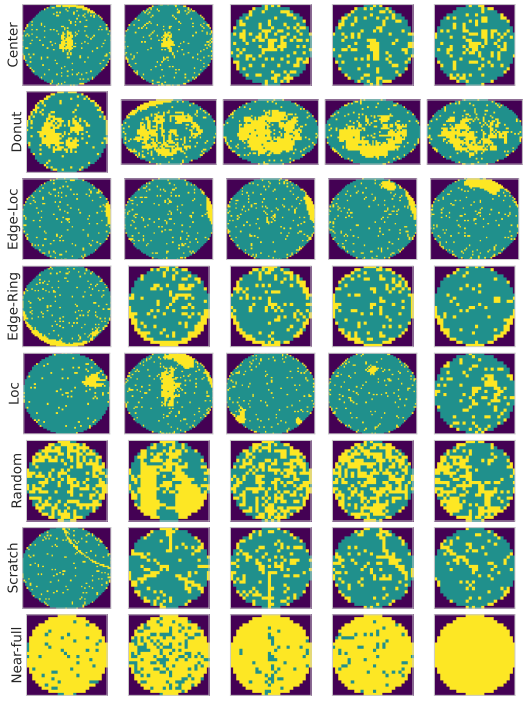

Figure 6: Examples of wafer-map manufacturing failures.

Experimental settings. For this experiment, due to computational constraints, we deviate from the settings used in the main paper. Specifically, we here use $S = 512$, $n_0 = 4$, and run four trials of 20 iterations. We use a RBF kernel.

Problem. The goal is to optimize the fine-tuning test accuracy of a pretrained ViT B16 model on the industrial WM811K dataset. The tunable parameters are: learning rate in $[1e-4, 1e-1]$, momentum in $[0, 1]$, weight decay in $[0, 1]$, step size (for the learning rate scheduling) in $[1, 100]$, and $\gamma \in [0, 1]$.[5]

[5]This experiment conceptually builds on Ament et al.'s experiment on CNN training.

F Further results

F.1 Variance reduction and ranking stability

We estimate the empirical macro variance of the acquisition estimator by fixing a BO state, drawing a fixed set of $n = 64$ Sobol probe samples, rebuilding the acquisition independently 16 times, and computing the sample variance across repeated estimator values at each probe point. We report the mean pointwise macro variance for the raw and orthogonalized estimators. (Note: this does *not* affect the fitted surrogate model, which stays untouched). We repeat this at fixed points throughout training. Due to space reasons, we only report abridged results in the main text. We report results for (i) a Matérn-$5/2$ kernel in section F.1.1, (ii) a RBF kernel in section F.1.2, and (iii) a TPE surrogate in section F.1.3.

F.1.1 Matérn-$5/2$ kernel

In Table 6, we give the variance reduction of ORTHOBO for all benchmark functions. For higher-dimensional Michalewicz10 and Levy16, we found that larger n_0 were required Across all benchmarks, we observe a strong reduction in variance, confirming that our proposed ORTHOBO stabilizes the acquisition estimate. Further, we report the ranking stabilities in Table 7, We observe that ORTHOBO again reduces the variance and consistently stabilizes the probe ranking.

Table 6: Variance reduction by orthogonalization.

		Hartmann6		Ackley8		Michalewicz10		Levy16	
n_0	S	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$
32	8	2e-08	1.61e-07	9.68e-09	5.24e-09	9.51e-08	4.54e-08	0.00106	0.000709
32	32	2.59e-09	3.02e-10	6.3e-09	4.17e-10	1.67e-08	1.07e-09	0.000208	2.27e-05

Table 7: Ranking stability. Lower probe variance and flip rate indicate more stable acquisition estimates; higher ranking correlation and top-1 agreement indicate more stable candidate rankings across probes.

Benchmark	Method	Probe variance	Top1 agr.	Flip rate	Regret
Hartmann6	qLogEI	$8.416_{15.43}$	$0.925_{0.16}$	$0.114_{0.12}$	$1.411_{0.42}$
	OrthoBO (ours)	$0.038_{0.04}$	$0.950_{0.10}$	$0.051_{0.04}$	$1.232_{0.67}$
Ackley8	qLogEI	$19.471_{39.29}$	$0.958_{0.11}$	$0.102_{0.12}$	$17.864_{2.08}$
	OrthoBO (ours)	$0.017_{0.04}$	$0.986_{0.06}$	$0.021_{0.05}$	$18.136_{0.78}$
Michalewicz10	qLogEI	$168.603_{154.76}$	$0.925_{0.16}$	$0.148_{0.12}$	$7.140_{0.01}$
	OrthoBO (ours)	$0.019_{0.03}$	$0.988_{0.06}$	$0.014_{0.03}$	$6.961_{0.36}$
Levy16	qLogEI	$233.655_{287.67}$	$0.917_{0.12}$	$0.069_{0.09}$	$82.194_{6.05}$
	OrthoBO (ours)	$0.073_{0.21}$	$0.988_{0.06}$	$0.048_{0.04}$	$58.938_{0.00}$

F.1.2 RBF kernel

We repeat our previous analysis, but use RBF kernel for the GP surrogate. We report the raw and orthogonalized variances in Table 8. Our observations are consistent: orthogonalization substantially reduces estimation variance.

Table 8: Variance reduction through orthogonalization, using a RBF kernel.

		Hartmann6		Ackley8		Michalewicz10		Levy16	
n_0	S	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$
32	8	5.91e-08	1.91e-07	1.28e-08	7.19e-09	1.15e-07	4.56e-08	0.00129	0.000758
32	32	8.62e-09	7.3e-10	3.71e-09	4e-10	2.61e-08	1.76e-09	0.000301	2.85e-05

Table 9: Ranking stability, **RBF kernel**. Lower probe variance and flip rate indicate more stable acquisition estimates; higher ranking correlation and top-1 agreement indicate more stable candidate rankings across probes.

Benchmark	Method	Probe variance	Top1 agr.	Flip rate	Regret
Hartmann6	qLogEI	$5.373_{11.64}$	$0.944_{0.15}$	$0.085_{0.10}$	$0.126_{0.00}$
	OrthoBO (ours)	$0.031_{0.04}$	$0.986_{0.08}$	$0.031_{0.04}$	$0.125_{0.00}$
Ackley8	qLogEI	$10.051_{31.34}$	$0.986_{0.06}$	$0.052_{0.10}$	$12.476_{2.06}$
	OrthoBO (ours)	$0.013_{0.04}$	$0.993_{0.04}$	$0.019_{0.04}$	$12.581_{4.03}$
Michalewicz10	qLogEI	$55.138_{100.54}$	$0.931_{0.15}$	$0.085_{0.09}$	$4.517_{0.20}$
	OrthoBO (ours)	$0.002_{0.01}$	$1.000_{0.00}$	$0.031_{0.04}$	$4.004_{0.52}$
Levy16	qLogEI	$82.913_{154.71}$	$0.889_{0.18}$	$0.201_{0.10}$	$3.109_{0.57}$
	OrthoBO (ours)	$0.034_{0.17}$	$1.000_{0.00}$	$0.049_{0.04}$	$2.819_{0.57}$

F.1.3 TPE-based surrogate

In Table 10, we show the variance reduction from OrthoBO for a TPE surrogate, using a MC budget of $S = 32$. Our observations are in line with the GP surrogate: OrthoBO can substantially reduce the variance.

Table 10: Variance reduction by orthogonalization, for a TPE surrogate.

Hartmann6		Ackley8		Michalewicz10		Levy16	
$\mathrm{Var}(\mathrm{EI}^{\mathrm{TPE}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}})$	$\mathrm{Var}(\mathrm{EI}^{\mathrm{TPE}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}})$	$\mathrm{Var}(\mathrm{EI}^{\mathrm{TPE}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}})$	$\mathrm{Var}(\mathrm{EI}^{\mathrm{TPE}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}})$
$11.222_{16.405}$	$1.214_{2.620}$	$13.182_{18.233}$	$1.284_{2.601}$	$21.712_{25.982}$	$2.197_{3.358}$	$20.944_{26.200}$	$1.399_{2.259}$

F.2 Misspecified surrogate families

We first study robustness to surrogate misspecification. In practical BO, the surrogate family is only an approximation to the unknown objective, and kernel hyperparameters are often estimated from limited data. As a result, the posterior quantities used inside MC acquisition estimates can be noisy or biased, which may destabilize candidate rankings. This experiment evaluates whether the variance reduction induced by OrthoBO improves BO decisions under such challenging, but realistic, surrogate conditions.

We consider four standard synthetic regression problems with known global optima and varying dimensionality: Hartmann6, Ackley8, Michalewicz10, and Levy16 [e.g., 2, 10, 15, 41]. Details of the benchmark functions and surrogate choices are provided in Table 5. We compare methods using best-so-far regret.

Mild misspecification: isotropic RBF kernel. We first use an isotropic RBF kernel, a common smooth GP prior. For anisotropic and multimodal objectives it is mildly misspecified because it uses a single lengthscale across all dimensions. (The dimension of the benchmark functions varies from 6 to 16.) The results are in Fig. 7.

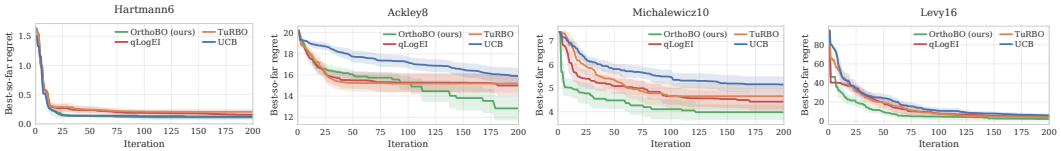

Figure 7: Best-so-far regret as a function of iterations on four widely used regression problems, using a **mildly misspecified isotropic RBF kernel**. Our proposed OrthoBO is competitive, and in higher-dimensional functions (Michalewicz10, Levy16) outperforms the baselines.

Results. Under mild surrogate misspecification, our proposed OrthoBO remains competitive across all benchmarks. On Hartmann6 and Ackley8, it improves over qLogEI and reaches the lowest regret. On the higher-dimensional Michalewicz10 and Levy16 functions, OrthoBO achieves the lowest

regret and improves substantially earlier than the baselines. For example, ORTHOBO reaches a regret of ~ 5 after 5 iterations, whereas the baseline methods require at least 50 to 60 *additional iterations*. Together, these results indicate that variance reduction at the acquisition-estimation level improves the reliability of BO decisions under realistic surrogate mismatch, with the largest gains appearing in higher-dimensional settings where candidate rankings are more sensitive to estimation noise.

Strong misspecification: linear kernel. We next consider a more challenging setting by using a linear kernel. This surrogate is unable to represent the nonlinear and multimodal structure of the benchmark functions, which amplifies the effect of posterior and acquisition-estimation errors. We therefore test whether stabilizing the acquisition estimate can still improve BO decisions when the surrogate family is substantially imperfect. The results are shown in Fig. 8.

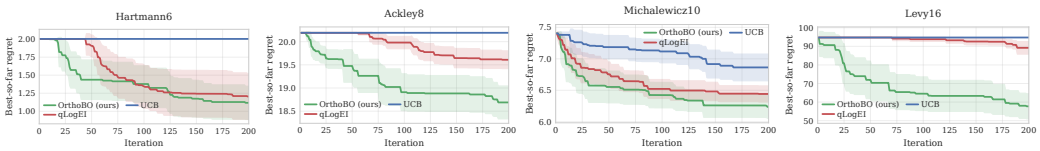

Figure 8: Best-so-far regret as a function of iterations on four widely used regression problems, using a **strongly misspecified linear kernel**. Our proposed ORTHOBO substantially outperforms other methods. The performance difference increases as the problem dimensionality grows. Notably, on Hartmann6, qLogEI takes five times more iterations than ORTHOBO to reach the same level of regret.

Results. We observe: (i) On all regression problems, our ORTHOBO outperforms the baselines. (ii) on Hartmann6, ORTHOBO reaches the same regret as qLogEI in half the time. (iii) On Levy16, ORTHOBO is the only method to substantially improve over time. Together, the results demonstrate the orthogonalization can successfully overcome strong misspecifications of the surrogate family.

With the strongly misspecified linear kernel, ORTHOBO outperforms the baselines on all four functions. On Hartmann6, it reaches the same regret level as qLogEI with substantially fewer evaluations. On Levy16, ORTHOBO is the only method that shows a clear improvement over the optimization horizon. Together, these results indicate that acquisition-estimation noise can be a significant failure mode when the surrogate is imperfect, and that orthogonalization helps by stabilizing the acquisition values used to rank candidates. Thus, ORTHOBO does not require a perfectly specified surrogate to improve BO performance; rather, it provides a variance-reduction mechanism that is especially useful when MC acquisition estimates become unreliable.

F.3 Weakly fitted hyperparameters

We next study a data-scarce regime in which the surrogate family is expressive, but its hyperparameters are weakly identified. This situation is common in expensive HPO problems: only a small initial design is available, yet the GP must already estimate lengthscales, output scale, and noise parameters that determine the posterior and hence the acquisition values. Uncertainty or instability in these hyperparameters can therefore propagate to the acquisition function and distort candidate rankings.

To isolate this effect, we use a GP with a Matérn-$5/2$ kernel and ARD. Unlike the isotropic kernels studied above, ARD assigns a separate lengthscale to each input dimension. This makes the surrogate more flexible, but also harder to fit when the initial design is small. We therefore initialize BO with only $n_0 = 4$ Sobol points and evaluate performance on Levy16, which is particularly challenging in this setting because the surrogate must estimate 16 lengthscales and a noise parameter from very limited data. The results are shown in Fig. 9.

Results. All methods initially exhibit high regret, reflecting the difficulty of fitting an ARD surrogate from only four initial observations. Nevertheless, ORTHOBO decreases regret steadily over the optimization horizon and reaches one of the lowest final regrets among the compared methods. These results show that orthogonalization remains useful even when the kernel class is flexible but its hyperparameters

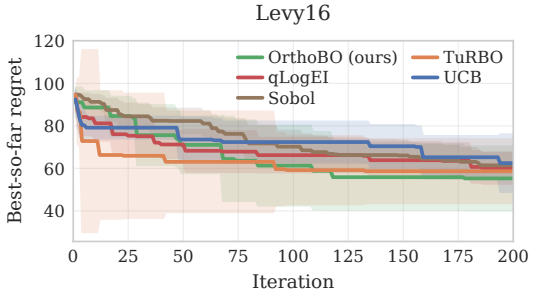

Figure 9: Best-so-far regret on Levy16 with a Matérn-$5/2$ ARD kernel and only $n_0 = 4$ initial Sobol points. ORTHOBO remains among the strongest methods and improves steadily despite weakly identified surrogate hyperparameters.

are poorly estimated. The gains are consistent with the proposed mechanism: reducing sensitivity of the acquisition estimate to surrogate-posterior perturbations yields more reliable candidate rankings in early, data-limited BO.

F.4 Ensembling

In the main experiments, we used $M = 1$ to avoid confounding the effect of orthogonalization with mere performance gains from a larger ensemble. We now use an ensemble of $M = 3$ models, each with a different kernel: linear, RBF, and Matérn-$5/2$. We increased the number of iterations by $25\,\%$ to account for additional fitting iterations for the models. We compute $\ell_{m,t}$ from the GP surrogate posteriors, set $\tau = 1$, and use $\delta = 0.001$ as a minimum weight floor.

The results for ensembling are in Figure 10.

Further, we compute (i) the variance reduction, (ii) ranking stability metrics, and (iii) weighting metrics. These additional measures are computed on a Sobol probe. We present the results in Table 11 and Figure 11.

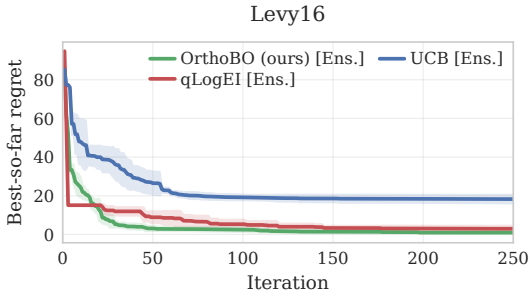

Figure 10: Results for ensembling.

Table 11: Variance reduction for ensembling on Levy16.

Method	Probe variance	Top1 agr.	Flip rate	Regret	Entropy	Weight 1	Weight 2	Weight 3
UCB [Ens.]	$39.255_{48.64}$	$1.000_{0.00}$	$0.026_{0.04}$	$18.267_{2.40}$	$0.163_{0.18}$	$0.008_{0.06}$	$0.062_{0.14}$	$0.933_{0.14}$
qLogEI [Ens.]	$390.598_{1033.83}$	$0.875_{0.20}$	$0.167_{0.10}$	$2.926_{1.66}$	$0.145_{0.16}$	$0.014_{0.07}$	$0.073_{0.21}$	$0.917_{0.21}$
OrthoBO (Ours) [Ens.]	$0.039_{0.12}$	$0.958_{0.14}$	$0.081_{0.08}$	$0.941_{0.54}$	$0.169_{0.23}$	$0.008_{0.06}$	$0.069_{0.14}$	$0.927_{0.14}$

Results. We observe: (1) OrthoBO achieves the lowest regret. (iii) OrthoBO has substantially lower Sobol probing variance. (iii) Weight are highest for the Matérn-$5/2$ kernel ("Weight 3") demonstrating that all methods primarily focus the "most correct" kernel. We plot the evolution of the entropy over time in Figure 11, which shows that the weight aggregation scheme balances exploration and exploitation.

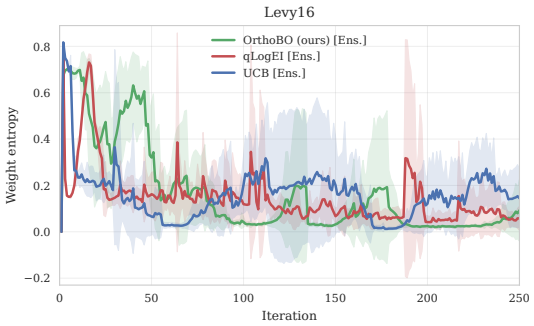

Figure 11: Evolution of the entropy across ensembling weights. For all methods, the used tempered exponentially weighted aggregation scheme is effective in balancing exploration and exploitation at the *ensemble* level.

F.5 Monte Carlo sampling ablation experiment

We here study the effect of the MC budget used to estimate the acquisition function. Here, finite-sample acquisition noise can change candidate rankings, with the effect becoming more pronounced for small sampling budgets S [cf. 2]. This setting is practically relevant because acquisition functions are evaluated many times during optimization, so increasing S can be computationally expensive. We therefore test whether the variance reduction of ORTHOBO improves BO performance when acquisition estimates must be computed from limited MC samples.

We use an isotropic RBF kernel, thereby combining moderate surrogate mismatch with reduced acquisition-estimation budgets. We vary $S \in \{64, 128, 256, 512\}$. The main results for Ackley8 and Michalewicz10 are shown in Fig. 3; additional results are provided in Fig. 12.

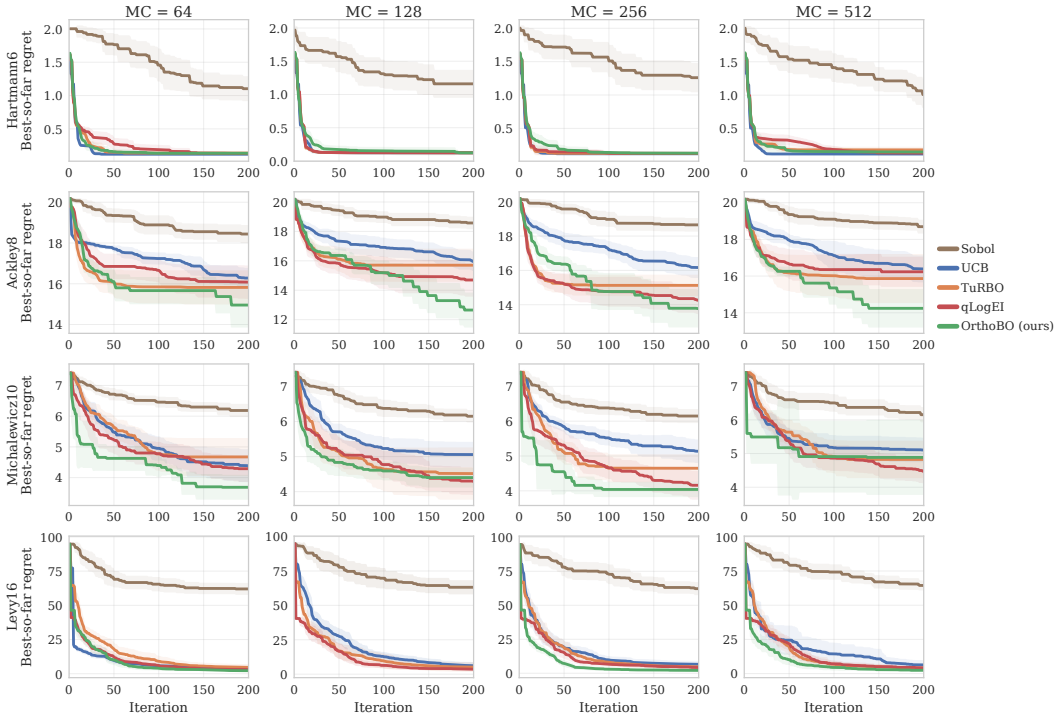

Figure 12: Different mc samples

Results Reducing the MC budget makes the performance of acquisition-based methods more sensitive to estimation noise. Across the tested budgets, ORTHOBO remains competitive on Ackley8 and consistently achieves low regret on Michalewicz10. The effect of orthogonalization are most visible

at earlier iterations, where MC estimation noise has a larger effect on the acquisition values and hence on the induced candidate rankings. Overall, these results support the central mechanism of ORTHOBO: orthogonalization improves the reliability of acquisition estimates when MC budgets are limited, leading to more stable BO decisions and thus lower regret.