ORTHOBO: ORTHOGONAL BAYESIAN HYPERPARAMETER OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian optimization (BO) is widely used for hyperparameter optimization when model evaluations are expensive. However, noisy acquisition estimates can lead to unstable decisions. We identify acquisition estimation noise as a previously overlooked failure mode: even when the surrogate model and acquisition target are correctly specified, finite-sample Monte Carlo error can perturb acquisition values. This can flip candidate rankings and lead to suboptimal BO decisions. As a remedy, we adapt classical score-function control variates to acquisition estimation. Subtracting an optimally weighted control variate yields an acquisition residual *orthogonal* to posterior score directions, with reduced variance and an unchanged marginal acquisition target. We further introduce ORTHOBO: a BO framework that combines this orthogonalized acquisition estimation with ensemble surrogates and an outer log transformation. We prove that our estimator preserves the target, reduces variance, and stabilizes pairwise rankings. We further verify the theoretical properties of ORTHOBO through numerical experiments where our framework reduces acquisition estimation variance, stabilizes candidate rankings, and achieves strong performance. We also demonstrate the downstream utility of ORTHOBO in hyperparameter optimization for neural network training and fine-tuning.

1 INTRODUCTION

Hyperparameter optimization (HPO) is fundamental in modern machine learning (ML), and is used across many domains, including biology (Ban et al., 2017; Quitadadmo et al., 2017), chemistry (Yuan et al., 2021), manufacturing (AlBahar et al., 2021), medicine (Nath et al., 2021), and physics (Tani & Veelken, 2024). As evaluating hyperparameter candidates can require substantial computation, *Bayesian optimization (BO)* is frequently used to identify optimal hyperparameter configurations under limited evaluation budgets (e.g., (Feurer et al., 2015; Shahriari et al., 2016; Snoek et al., 2012; Turner et al., 2020; Wu et al., 2020)). BO fits a probabilistic surrogate model of the objective and selects new configurations by maximizing an *acquisition function* such as expected improvement (EI) (Jones et al., 1998).[1] We use HPO as our primary application because acquisition noise can change candidate rankings and waste expensive evaluations. However, the underlying problem is general, and our estimator applies whenever BO acquisition values are estimated by Monte Carlo integration.

In our work, we focus on a source of instability that is central to BO but has received little attention: *the acquisition value estimation*. 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., (Bodin et al., 2020; Hernández-Lobato et al., 2015; Polyzos et al., 2023; Snoek et al., 2012)). As a result, the quantity used to rank candidate configurations is itself a noisy estimate. This may hurt the performance of BO which acts on *rankings*. Even small estimation errors can flip the order of two candidates with similar acquisition values such that the BO algorithm evaluates many suboptimal configurations (see Fig. 1).

At a technical level, we treat the estimation of the acquisition function as a MC estimation problem (e.g., Bodin et al., 2020; Hernández-Lobato et al., 2014a; 2015; Snoek et al., 2012) and construct an *orthogonalized estimator of the acquisition value* by adapting score-function control variates

[1] For a thorough introduction to BO, see, e.g., (Frazier, 2018a; Shahriari et al., 2016).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Rubinstein & Marcus, 1985; Mira, 2013; Ranganath et al., 2014), thus improving robustness to sampling errors and stabilizing acquisition rankings. Our notion of orthogonality stems from the orthogonal ML literature (e.g., (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023; Kennedy, 2024; Mackey et al., 2018)) in that our estimator offers robustness due to *variance reduction* guarantees.[2]

Example. We tune a classification model on a large-scale image dataset, where each evaluation requires training on many images and assessing performance under an imbalanced class distribution. Two hyperparameter configurations may differ only slightly (e.g., a deeper network with stronger dropout versus a shallower network with weaker regularization). If estimates of the acquisition values are subject to noise, BO may spend a suboptimal configuration simply because it was ranked first due to sampling noise.

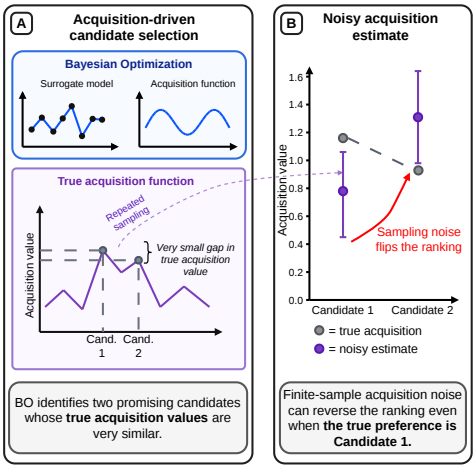

Figure 1: **Failure mode in Bayesian HPO:** acquisition value estimation noise can flip rankings leading to suboptimal updates.

Our contribution is complementary to existing work on robust BO. Prior methods have mainly focused on improving the surrogate model(Berkenkamp et al., 2019; Bogunovic & Krause, 2021; Eriksson et al., 2019; Lu et al., 2023; Neiswanger & Ramdas, 2021; Polyzos et al., 2023; Turner et al., 2020). Other work has improved the acquisition formula or its numerical optimization (Ament et al., 2023). 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 through score-function control variates on posterior scores (Rubinstein & Marcus, 1985; Mira, 2013; Ranganath et al., 2014).

ORTHOBO: We build on our orthogonal acquisition estimator and introduce ORTHOBO, a framework for robust HPO. ORTHOBO uses orthogonalized acquisition estimation as a core component and combines it with ensemble surrogates Lu et al. (2023); Polyzos et al. (2023) and an outer log transformation (Ament et al., 2023). Here, each component serves a distinct role: our orthogonalization stabilizes acquisition estimation, ensembling addresses structural misspecification, and the outer log transformation improves numerical conditioning during acquisition maximization.

Our ORTHOBO offers several theoretical guarantees. (i) Our orthogonalized acquisition estimator is unbiased for the marginal EI and achieves lower variance than the corresponding naïve MC estimator ($\rightarrow$ Theorem 4.1). (ii) The lower acquisition estimation variance improves the stability of pairwise acquisition rankings ($\rightarrow$ Proposition 4.3). Intuitively, this reduces the probability of falsely selecting a candidate based on MC noise, and leads thus to performance improvements in practice.

Main contributions[3]: **(1)** We identify finite-sample acquisition-estimation noise as a BO failure mode and formulate acquisition estimation as a variance-reduction problem, which we address via score-function control variates. The resulting orthogonalized estimator reduces estimation variance without changing the acquisition target. **(2)** We develop ORTHOBO, a BO framework that jointly targets acquisition noise, structural surrogate misspecification, and numerical instabilities. **(3)** We establish variance-reduction and ranking-stability guarantees and demonstrate strong empirical performance across generic BO benchmarks and HPO applications. Although motivated by HPO, ORTHOBO applies more generally whenever BO acquisition values are estimated by Monte Carlo integration.

2 RELATED WORK[4]

BO for hyperparameter tuning. BO is a standard approach for HPO (e.g., (Feurer et al., 2015; Snoek et al., 2012; Wu et al., 2020)), particularly when objective function evaluations are expensive. In practice, BO relies on a surrogate model to approximate the objective and guide the search. A

[2]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., (Martinez-Cantin et al., 2018)).

[3]Anonymous code for peer-review at `https://anonymous.4open.science/r/ortho_hpo`

[4]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.

common choice for the surrogate model are Gaussian processes (GPs) Ha et al. (2023); Jones et al. (1998); Snoek et al. (2012). Other alternatives include tree-based and density-based surrogates such as tree-structured Parzen estimators (TPE) Bergstra et al. (2011; 2013); Watanabe (2023). Subsequent work has addressed scaling challenges through multi-fidelity methods, batch BO, and trust-region approaches (e.g., (Eriksson et al., 2019; Falkner et al., 2018; González et al., 2016; Wilson et al., 2018)). *However, no method has explicitly studied acquisition estimation noise as a failure mode.*

Acquisition functions. Surrogate models are used by acquisition functions, which are heuristics to evaluate the utility of hyperparameter candidates before querying the (expensive) objective function (Balandat et al., 2020). 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 Ament et al. (2023; 2024); Eriksson et al. (2019); Moss et al. (2021); Wang & Jegelka (2017); Wilson et al. (2017). Prominent examples include, e.g., qLogEI (Ament et al., 2023), TuRBO (Eriksson et al., 2019), and UCB (Wilson et al., 2017). *Our work is orthogonal to these works, as we focus on improving the **estimation** and **not** the acquisition function itself.*[5]

Robustness to surrogate misspecification. A central weakness of BO is that it relies on a surrogate model which might be misspecified. Prior work has studied BO under misspecified kernel classes Bogunovic & Krause (2021), unknown hyperparameters Berkenkamp et al. (2019); Ha et al. (2023), unreliable posterior uncertainty Neiswanger & Ramdas (2021), and a surrogate complexity-efficiency trade-off Bodin et al. (2020). Several practical strategies have been developed to mitigate such failure modes (Ament et al., 2024; Eriksson et al., 2019; Lu et al., 2023; Martinez-Cantin et al., 2018; Polyzos et al., 2023) *However, these approaches address structural surrogate misspecification and thus model uncertainty. In contrast, our method additionally targets estimation noise.*

Control variates. Linear control-variate coefficients (Rubinstein & Marcus, 1985) and related score-based constructions have been used for zero-variance MCMC and stochastic variational inference (Assaraf & Caffarel, 1999; Mira, 2013; Ranganath et al., 2014). Control functionals (Oates et al., 2017) generalize this idea by constructing nonparametric controls in an RKHS. In contrast, we use a low-dimensional closed-form projection whose score covariance can be shared across acquisition candidates, and apply it to marginal acquisition estimation and the stability of BO candidate rankings.

Research gap. We identify acquisition estimation errors as a crucial failure mode in BO that has been previously overlooked. To address this, we propose an orthogonalized estimator for variance reduction that stabilizes acquisition-based decisions. To our knowledge, this is the first work to explicitly formulate acquisition value estimation in BO as a variance reduction problem and to develop an orthogonalized estimator to improve acquisition-based decisions.

3 Problem Setting

Setup. We consider the standard setup for HPO over a search space $\Lambda \subseteq \mathbb{R}^d$ (Ament et al., 2024; Eriksson & Poloczek, 2021). For a hyperparameter configuration $\lambda \in \Lambda$, let $f(\lambda)$ denote the unknown performance of interest, such as, 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. BO maintains a probabilistic model of the unknown objective and uses it to guide future evaluations (Frazier, 2018b). 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 (Ament et al., 2024). The surrogate provides two crucial quantities for decision-making: (i) a prediction of objective values in unexplored regions, and (ii) an uncertainty measure about those predictions. BO 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 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.

[5] We provide a generalization of OrthoBO at the acquisition function level in Supplement C.

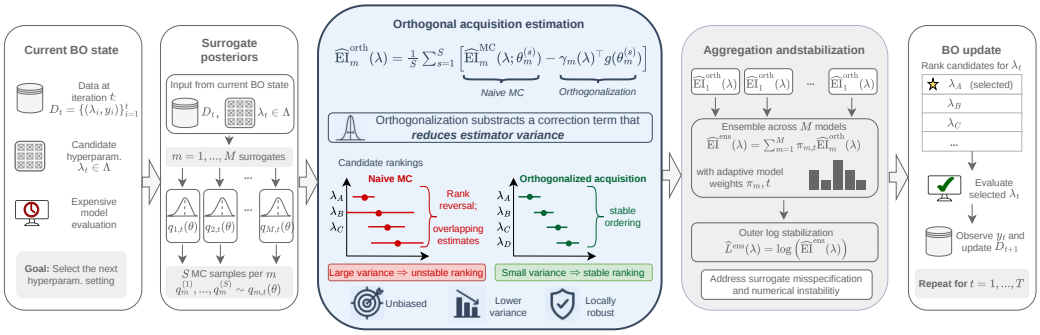

Figure 2: **Our proposed ORTHOBO.** It improves acquisition estimation through *variance reduction*. For this, we orthogonalize the acquisition estimation *without* changing the acquisition target.

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.

In our work, we focus on *expected improvement (EI)* (Jones et al., 1998) 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 $(x)_+ := \max\{0, x\}$ Jones et al. (1998); Snoek et al. (2012). 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 $s = 1, \ldots, S$ given a specific MC sampling budget S Balandat et al. (2020); Snoek et al. (2012); Wilson et al. (2018), 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 acquisition and thus suboptimal BO decisions.

Finite-sample acquisition estimation noise.[6] Bayesian HPO depends on the stable acquisition values to rank candidate configurations. In practice, these values are often obtained by marginalizing over surrogate uncertainty and must be approximated from a finite number of samples. Even when the surrogate model is reasonable, finite-sample MC error can perturb acquisition values, flip candidate rankings, and thus lead to suboptimal configuration evaluations. We identify *acquisition estimation noise* as an overlooked but practically important failure mode.

Our work. In this work, we address instability in BO at the acquisition-estimation level. 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 then propose an orthogonal acquisition estimator that reduces variance induced by surrogate uncertainty, while preserving the underlying EI-scale target. Therein, we subtract an optimally weighted score-function control variate, yielding residuals *orthogonal* to posterior score directions. In ORTHOBO, we combine our orthogonal acquisition estimator with an ensemble of surrogate models and adaptive model weighting optimized through an outer log transformation to obtain a full, state-of-the-art BO framework.

4 ORTHOGONAL BAYESIAN OPTIMIZATION

Below, we present ORTHOBO, our orthogonalized framework for BO that addresses the failure modes discussed above (an overview is shown in Fig. 2). We first present our *main contribution*: the

[6] We discuss other failure modes (e.g., misspecification of the surrogate, improper surrogate model classes, and numerical instabilities due to small or skewed acquisition values) in Supplement B.

orthogonal acquisition estimator (Section 4.1). Of note, we do *not* change the acquisition target, but rather improve the estimation through *variance reduction* techniques. Then, we provide example instantiations of our orthogonal framework for two common surrogate classes (Section 4.2). We then present ORTHOBO, where we embed our orthogonal acquisition estimator inside a state-of-the-art BO algorithm (Section 4.3). 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 C. Proofs are in Supplement E.

4.1 ORTHOGONAL ACQUISITION ESTIMATOR

We now address the instability due to acquisition estimation 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 $q_{m,t}$. Let $g(\theta) := \nabla_\theta \log q_{m,t}(\theta)$ denote the posterior score function.

We define the *orthogonalized EI* 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.[7]

Properties of the orthogonalized acquisition. The orthogonalized EI satisfies three key properties: (i) *preserved target*: it has 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 conditions 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., $\int_\Theta \nabla_\theta q_{m,t}(\theta)\,\mathrm{d}\theta = 0$. 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 (Variance reduction)**.** *Assume that* $\mathbb{E}_{q_{m,t}}[\mathrm{EI}_m(\lambda;\theta)^2] < \infty$ *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) Preserved target. $\mathbb{E}[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)] = \mathrm{EI}_m^{\mathrm{marg}}(\lambda;\theta)$

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

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

Theorem 4.1 is the classical optimal-linear-control-variate variance identity (e.g., (Rubinstein & Marcus, 1985)) instantiated for the marginal acquisition under $q_{m,t}$. We state and prove it in our notation for completeness. The BO-specific results are Corollary 4.2 and Proposition 4.3 below.

Corollary 4.2 (Local robustness.)**.** *Assume the conditions of Theorem 4.1. 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, it holds that*

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

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

Implications for acquisition estimation. 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) 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, improving the stability of acquisition-based decisions. Corollary 4.2 provides local robustness in an orthogonal ML sense (Chernozhukov et al., 2018; Kennedy, 2024): small score-tilt perturbations of $q_{m,t}$ do not affect the 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 correct structural misspecification of the surrogate family. We address the latter through ensemble modeling in Section 4.3.

Proposition 4.3 (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*

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

Proposition 4.3 shows how acquisition estimation noise propagates into BO decisions. High estimator variance can flip the estimated sign of the acquisition gap between two candidate points, thus 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 (GP) surrogate

For GP surrogate models Ha et al. (2023); Jones et al. (1998); Snoek et al. (2012), acquisition estimation 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 $f \sim \mathcal{GP}(m_\theta(\cdot), k_\theta(\cdot,\cdot))$, where here m_θ denotes the mean function (typically taken constant or zero) and k_θ the covariance kernel with $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ collecting the 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

$$f(\lambda) \mid \mathcal{D}_t, \theta \sim \mathcal{N}\big(\mu_t(\lambda;\theta), \sigma_t^2(\lambda;\theta)\big)\,, \tag{10}$$

with

$$\mu_t(\lambda;\theta) = m_\theta(\lambda) + k_\theta(\lambda,\Lambda_t)^\top \left(K_\theta(\Lambda_t,\Lambda_t) + \sigma_n^2 I_t\right)^{-1} (y_t - m_\theta(\Lambda_t))\,, \tag{11}$$

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

where $\Lambda_t = (\lambda_1,\dots,\lambda_t)$, $y_t = (y_1,\dots,y_t)^\top$, and K denotes the Gram matrix. We maximize the EI

$$\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta) = (f^\star - \mu_t(\lambda;\theta))\,\Phi(z_t(\lambda;\theta)) + \sigma_t(\lambda;\theta)\,\phi(z_t(\lambda;\theta)), \tag{13}$$

where $z_t(\lambda;\theta) = \frac{f^\star - \mu_t(\lambda;\theta)}{\sigma_t(\lambda;\theta)}$, $f^\star = \min_{i\le t} y_i$, 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, e.g., via a Laplace approximation or variational Gaussian approximation. The marginal acquisition is then

$$\mathrm{EI}^{\mathrm{GP,marg}}(\lambda) = \mathbb{E}_{\theta\sim q_t}\big[\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta)\big]\,. \tag{14}$$

With $g_t(\theta) = \nabla_\theta \log q_t(\theta)$, $\gamma^{\mathrm{GP}}(\lambda) = \mathrm{Cov}(g_t,g_t)^{-1}\,\mathrm{Cov}\big(g_t, \mathrm{EI}^{\mathrm{GP}}(\lambda;\theta)\big)$ the orthogonal estimator is

$$\widehat{\mathrm{EI}}^{\mathrm{GP,orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\Big[\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta^{(s)}) - \gamma^{\mathrm{GP}}(\lambda)^\top g_t(\theta^{(s)})\Big]\,, \qquad \theta^{(s)} \sim q_t. \tag{15}$$

B TREE-BASED / TPE-STYLE SURROGATE

We further consider tree-based or density-based surrogates as in Tree-structured Parzen Estimators (TPE) Bergstra et al. (2011; 2013); Watanabe (2023). Rather than modeling $p(y \mid \lambda)$ directly, TPE-style methods model

$$p(\lambda \mid y) = \begin{cases} \ell_\theta(\lambda), & y < y^\star, \\ g_\theta(\lambda), & y \geq y^\star, \end{cases} \tag{16}$$

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

$$\mathrm{EI}^{\mathrm{TPE}}(\lambda; \theta) \propto \frac{\ell_\theta(\lambda)}{g_\theta(\lambda)}. \tag{17}$$

More generally, we denote the EI 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 orthogonal acquisition estimator by

$$\widehat{\mathrm{EI}}^{\mathrm{TPE,orth}}(\lambda) = \frac{1}{S} \sum_{s=1}^{S} \left[h^{\mathrm{TPE}}(\lambda; \theta^{(s)}) - \gamma^{\mathrm{TPE}}(\lambda)^\top c_t(\theta^{(s)}) \right], \qquad \theta^{(s)} \sim q_t, \tag{18}$$

with

$$\gamma^{\mathrm{TPE}}(\lambda) = \mathrm{Cov}(c_t, c_t)^{-1} \, \mathrm{Cov}\big(c_t, h^{\mathrm{TPE}}(\lambda; \theta)\big). \tag{19}$$

4.3 ORTHOBO

We present the full ORTHOBO procedure in Alg. 1. Therein, we combine our orthogonalized acquisition estimator with ensemble surrogates and an outer log transformation.

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 $\pi_{m,t}$

$$\widehat{\mathrm{EI}}^{\mathrm{ens}}(\lambda) = \sum_{m=1}^{M} \pi_{m,t} \widehat{\mathrm{EI}}_m^{\mathrm{orth}}(\lambda). \tag{20}$$

Exponentially weighted aggregation. We update ensemble weights through tempered exponentially weighted aggregation Freund & Schapire (1997), a widely used approach in model aggregation Fedus et al. (2022); Jacobs et al. (1991); Shazeer et al. (2017). A minimum weight floor Herbster & Warmuth (1998) prevents premature collapse onto a single surrogate

$$\pi_{m,t} \propto \max(\delta, \, \pi_{m,t-1} \exp(\ell_{m,t}/\tau)), \tag{21}$$

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.

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 drawn from a Sobol distribution

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. equation 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 $\mathcal{D}_T$

Optimization based on outer log stabilization. To improve numerical stability, we follow recent literature Ament et al. (2023) and optimize the *outer log transformed* acquisition

$$\widehat{L}^{\text{ens}}(\lambda_t) = \log\big(\max(\widehat{\text{EI}}^{\text{ens}}(\lambda_t), \alpha)\big), \tag{22}$$

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. • *Baselines:* We evaluate OrthoBO against various baselines from the literature: **qLogEI** (Ament et al., 2023) (main comparison method), **UCB** (Wilson et al., 2017), **TuRBO** (Eriksson et al., 2019), and **Sobol** sampling (e.g., (Ament et al., 2024)). Our experimental settings follow prior literature (e.g., (Ament et al., 2023; 2024)): unless otherwise noted, all experiments use 32 replications, $S = 512$, and we start all methods from the same $n_0 = 32$ Sobol points. For optimizing the acquisition functions, we use 512 raw samples and allow 8 restarts to prevent getting stuck in local minima (Törn & Žilinskas, 1989). To isolate the effect of orthogonalization, we use $M = 1$ throughout the experiments in the main paper. • *Datasets:* We consider four standard BO benchmarking functions with known global optima (to compute the regret) and varying dimensionality: **Hartmann6**, **Ackley8**, **Michalewicz10**, and **Levy16** (e.g., (Ament et al., 2023; Bodin et al., 2020; Eriksson & Poloczek, 2021; Moss et al., 2023)). Details of the benchmark functions and surrogate choices are provided in Supplement G. • *Metrics:* We compare methods using best-so-far regret and report 95% confidence intervals (CIs) from standard normal approximation.

Evaluation. We first demonstrate the theoretical properties of OrthoBO by isolating the effect of orthogonalization of the acquisition estimation. For this, we show RQ1 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 demonstrate the effectiveness for HPO in real-world training and fine-tuning tasks where 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 H.

RQ1 **Variance reduction.** We probe the acquisition function with random Sobol samples and compare the variance of the raw and orthogonalized acquisition function using a GP surrogate. We report results for a Matérn-5/2 kernel with ARD on Michalewicz10 and Levy16 in Table 1. Results for an RBF kernel and a TPE surrogate are in Supplement H.1. ⇒ *Across all settings,* ***orthogonalization reduces the variance****, showing that* OrthoBO *stabilizes the acquisition.*

Table 1: **Variance reduction** through orthogonalization (with Sobol probes; $\times 10^{-8}$).

	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	9.50861	4.54151 (-52.24%)	105590.0	70935.8 (-32.82%)
32	1.66805	0.10681 (-93.60%)	20765.8	2266.7 (-89.08%)

RQ2 **Ranking stability.** We repeatedly evaluate the acquisition function on the same fixed Sobol probe set and compare the rankings across runs. We report: (i) *Top1 agreement*: fraction of repeats selecting the same top-ranked probe point, (ii) the *Flip rate* of adjacent high-ranked candidate pairs, and (iii) *Regret*. The flip rate mirrors the idea in Proposition 4.3, by measuring how often local pairwise orderings change due to noise in the acquisition-function. The main results are in Table 2. Further results are in Supplement H.1. ⇒ *Our variance reduction yields stabilized rankings.*

Table 2: **Ranking stability** (mean$\pm_{\text{std}}$), via Sobol probes. Lower flip rates indicate more stable acquisition estimates; higher top-1 agreement indicates more stable candidate rankings across probes..

	Method	Top1 agreement ↑	Flip rate ↓	Regret ↓
Michalewicz10	qLogEI	$0.96_{0.11}$	$0.06_{0.10}$	$5.96_{0.41}$
	OrthoBO (ours)	$0.99_{0.06}$	$0.02_{0.05}$	$5.87_{0.50}$
Levy16	qLogEI	$0.94_{0.13}$	$0.08_{0.10}$	$56.93_{8.12}$
	OrthoBO (ours)	$0.99_{0.07}$	$0.03_{0.05}$	$56.25_{8.87}$

RQ3 **MC efficiency.** We use an isotropic RBF kernel (as a common smooth GP prior) combined with a moderate surrogate mismatch and reduced acquisition budgets. We show the main results for all MC-sampling methods for varying S on Ackley8 in Fig. 3. Further results are in Fig. 16,

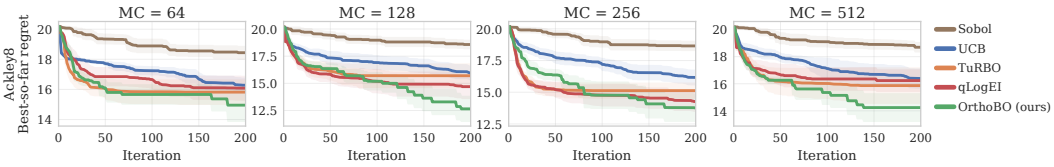

Figure 3: **MC efficiency.** Best-so-far regret for different MC budgets $S \Rightarrow$ ORTHOBO *achieves the lowest regrets across all budgets.*

Supplement H. Reducing the MC budget makes the acquisition-based methods more sensitive to estimation noise. Across all benchmarks, our ORTHOBO remains competitive and frequently achieves the lowest regret. We report cumulative regret in Supplement H.6. $\Rightarrow$ *Orthogonalization is especially effective in earlier iterations, where MC estimation noise has a larger effect on the acquisition values.*

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) (Dosovitskiy et al., 2020) for manufacturing image classification. We include the following additional baselines **PES** (Hernández-Lobato et al., 2014b), **JES** (Hvarfner et al., 2022), **GIBBON** (Moss et al., 2021), a stochastic variational GP (**SVGP**) (Hensman et al., 2013), and **SAASBO** (Eriksson & Jankowiak, 2021). Due to space reasons, results for all methods are in the respective supplementary materials.

(i) Training (neural network). We employ the 5D CNN benchmark by Ament et al. (2024). The goal is to optimize training hyperparameters for a CNN to reach higher accuracy (Acc). Results on FMNIST (Xiao et al., 2017) are in Fig. 4; more detailed and further results on with observational noise and on the MNIST (LeCun et al., 2010) are in Supplement G.3. There we also report AUC scores. $\Rightarrow$ ORTHOBO *improves early on and maintains its strong performance (higher AUC), reaching the highest final accuracy.*

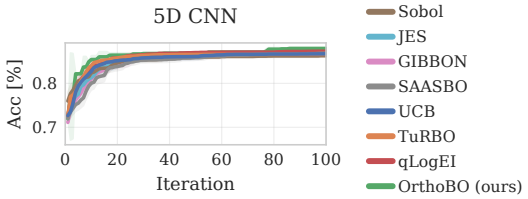

Figure 4: **Training (5D CNN).** $\Rightarrow$ ORTHOBO *achieves the* ***highest final accuracy****.* (Acc)

(ii) Fine-tuning (vision transformer). We optimize five fine-tuning hyperparameters on the WM811K wafer-map dataset (Jang, 2015; Wu et al., 2015) for a vision transformer (ViT) based on the F1 score (Fig. 5). Details and further results are in Appendix G.4.ORTHOBO improves rapidly within the first few evaluations and maintains the highest best-so-far F1 score throughout the optimization horizon. $\Rightarrow$ *Compared with qLogEI,* ORTHOBO *improves the best observed validation F1 score* ***by up to 20 percentage points****. This demonstrates that the* ORTHOBO *also improves performance in a practical, high-cost HPO setting beyond synthetic benchmarks.*

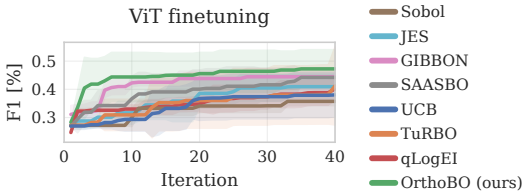

Figure 5: **Fine-tuning (ViT).** $\Rightarrow$ ORTHOBO ***improves faster than the baselines*** *and achieves the highest F1 score.*

6 DISCUSSION

• *Conclusion.* We introduce ORTHOBO, an orthogonalized framework for variance reduction in acquisition function estimation. Our work identifies a previously overlooked failure mode of BO: Decisions depend on acquisition values rankings, so MC noise can directly induce ranking errors and lead to suboptimal evaluations. ORTHOBO addresses this by reducing variance while preserving the underlying marginal EI target. • *Limitations.* First, ORTHOBO introduces additional computational overhead; however, the overhead is typically small compared with expensive evaluations of the objective function in HPO (see Supplement D). 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 C.1). We exemplarily demonstrate this for the PI acquisition function (e.g. (Jones et al., 1998) in Supplement H.7. • *Implications.* Our framework is widely applicable not only to HPO but essentially all BO problems.

AI USE STATEMENT

In this work, we used generative AI tools to help implement methods (coding agents), experimental utilities (e.g., data logging), and to give feedback on the overall research idea. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design the research methodology, clean and reformat datasets, interpret results, and assisting with translation, proposing or refining hypotheses, and supporting qualitative and thematic data analysis. Additionally, we used generative AI tools to design drafts for motivational figures, brainstorming, and grammar/spelling checks. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Areej AlBahar, Inyoung Kim, and Xiaowei 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.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Sebastian Ament, Elizabeth Santorella, David Eriksson, Ben Letham, Maximilian Balandat, and Eytan Bakshy. Robust Gaussian processes via relevance pursuit. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Roland Assaraf and Michel Caffarel. Zero-variance principle for monte carlo algorithms. *Physical review letters*, 83(23):4682, 1999.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Tomohiro Ban, Masahito Ohue, and Yutaka 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.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Neural Information Processing Systems (NeurIPS)*, 2011.

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

Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Erik Bodin, Markus Kaiser, Ieva Kazlauskaite, Zhenwen Dai, Neill Campbell, and Carl Henrik Ek. Modulating surrogates for Bayesian optimization. In *International Conference on Machine Learning (ICML)*, 2020.

Ilija Bogunovic and Andreas Krause. Misspecified Gaussian process bandit optimization. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

Samuel Daulton, Xingchen Wan, David Eriksson, Maximilian Balandat, Michael A Osborne, and Eytan Bakshy. Bayesian optimization over discrete and mixed spaces via probabilistic reparameterization. *Advances in Neural Information Processing Systems*, 35:12760–12774, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

David Eriksson and Martin Jankowiak. High-dimensional bayesian optimization with sparse axis-aligned subspaces. In *Uncertainty in artificial intelligence*, pp. 493–503. PMLR, 2021.

David Eriksson and Matthias Poloczek. Scalable constrained Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning (ICML)*, 2018.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Neural Information Processing Systems (NeurIPS)*, 2015.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Dennis Frauen, Maresa Schröder, Konstantin Hess, and Stefan Feuerriegel. Orthogonal survival learners for estimating heterogeneous treatment effects from time-to-event data. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Dennis Frauen, Athiya Deviyani, Mihaela van der Schaar, and Stefan Feuerriegel. Nonparametric LLM evaluation from preference data. *International Conference on Learning Representations (ICLR)*, 2026.

Peter I Frazier. Bayesian optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pp. 255–278. 2018a.

Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint*, arXiv:1807.02811, 2018b.

Yoav Freund and Robert 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.

Javier González, Zhenwen Dai, Philipp Hennig, and Neil Lawrence. Batch Bayesian optimization via local penalization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Huong Ha, Vu Nguyen, Hung Tran-The, Hongyu Zhang, Xiuzhen Zhang, and Anton van den Hengel. Provably efficient Bayesian optimization with unknown Gaussian process hyperparameter estimation. *arXiv preprint*, arXiv:2306.06844, 2023.

James Hensman, Nicolò Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

José M Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Neural Information Processing Systems (NeurIPS)*, 2014a.

José M Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014b.

José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *International Conference on Machine Learning (ICML)*, 2015.

Konstantin Hess, Dennis Frauen, Niki Kilbertus, and Stefan Feuerriegel. Debiased neural operators for estimating functionals. *arXiv preprint*, arXiv:2604.19296, 2026.

Carl Hvarfner, Frank Hutter, and Luigi Nardi. Joint entropy search for maximally-informed bayesian optimization. *Advances in Neural Information Processing Systems*, 35:11494–11506, 2022.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Jyh-Shing R. Jang. Mir-wm811k: Dataset for wafer map failure pattern recognition. `http://mirlab.org/dataset/public/`, 2015.

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

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp. 207–236, 2024.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Eric Lee, David Eriksson, David Bindel, Bolong Cheng, and Mike Mccourt. Efficient rollout strategies for bayesian optimization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 260–269. PMLR, 2020.

Qin Lu, Konstantinos D Polyzos, Bingcong Li, and Georgios 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.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning (ICML)*, 2018.

Ruben Martinez-Cantin, Kevin Tee, and Michael McCourt. Practical Bayesian optimization in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Valentyn Melnychuk and Stefan Feuerriegel. GDR-learners: Orthogonal learning of generative models for potential outcomes. *International Conference on Learning Representations (ICLR)*, 2026.

Antonietta Mira. Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 2013.

Henry B Moss, David S Leslie, Javier Gonzalez, and Paul Rayson. GIBBON: General-purpose information-based Bayesian optimisation. *Journal of Machine Learning Research*, 22(235):1–49, 2021.

Henry B Moss, Sebastian W Ober, and Victor Picheny. Inducing point allocation for sparse Gaussian processes in high-throughput Bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Vishwesh Nath, Dong Yang, Ali Hatamizadeh, Anas A Abidin, Andriy Myronenko, Holger R Roth, and Daguang 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.

Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified Gaussian processes. In *International Conference on Algorithmic Learning Theory (ALT)*, 2021.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718, 2017.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning (ICML)*, 2019.

Konstantinos D Polyzos, Qin Lu, and Georgios B Giannakis. Bayesian optimization with ensemble learning models and adaptive expected improvement. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Andrew Quitadadmo, James Johnson, and Xinghua Shi. Bayesian hyperparameter optimization for machine learning based eqtl analysis. In *International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822. PMLR, 2014.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Reuven Y Rubinstein and Ruth Marcus. Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33(3):661–677, 1985.

Maresa Schröder, Valentyn Melnychuk, and Stefan Feuerriegel. Differentially private learners for heterogeneous treatment effects. *International Conference on Learning Representations (ICLR)*, 2025.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Neural Information Processing Systems (NeurIPS)*, 2012.

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

Laurits Tani and Christian 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.

Aimo Törn and Antanas Žilinskas. *Global optimization*, volume 350. Springer, 1989.

Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle 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.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2, 2006.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pp. 3627–3635, 2017.

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

James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for Bayesian optimization. In *Neural Information Processing Systems (NeurIPS)*, 2018.

James T Wilson, Riccardo Moriconi, Frank Hutter, and Marc Peter Deisenroth. The reparameterization trick for acquisition functions. *arXiv preprint*, arXiv:1712.00424, 2017.

Jian Wu, Saul Toscano-Palmerin, Peter I Frazier, and Andrew Gordon Wilson. Practical multi-fidelity Bayesian optimization for hyperparameter tuning. In *Uncertainty in Artificial Intelligence (UAI)*, 2020.

Ming-Ju Wu, Jyh-Shing R. Jang, and Jui-Long Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1):1–12, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yingfang Yuan, Wenjun Wang, and Wei 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 how our ORTHOBO differs from (i) control-variates for acquisition estimation, (ii) score-based control variates, and (iii) more distant related literature to complement the discussion in Section 2. We first cover literature regarding the robustness of different BO components, and we finally present related work on orthogonal ML.

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 Snoek et al. (2012). Related work has also incorporated hyperparameter uncertainty directly into acquisition design Hernández-Lobato et al. (2014a; 2015). 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, thereby showing that combining multiple kernels or surrogate classes can improve robustness relative to committing to a single GP prior Lu et al. (2023); Polyzos et al. (2023). 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 Jones et al. (1998); Srinivas et al. (2010). Recent work has highlighted that vanilla EI and its variants can be numerically unstable, which can substantially degrade optimization performance Ament et al. (2023). This has motivated numerically stable reformulations based on the logarithmic scale, which are substantially easier to optimize in practice.

Control variates for acquisition estimation. Lee et al. (2020) combine QMC, common random numbers, and control variates for rollout acquisitions, using known-mean acquisition covariates (EI, PI). The distinction to our work is the noise source. They target randomness in sampled function values, while we target randomness from marginalizing the surrogate parameter posterior via a zero-mean score control variate requiring no known-mean covariate. Our implementation sits on BoTorch's QMC machinery, so our tables report reductions against an already state-of-the-art variance-reduced baseline

Score-based control variates Linear-in-score CVs are established in the MC/MCMC literature (Assaraf & Caffarel, 1999; Mira, 2013; Oates et al., 2017). We claim no novelty for the estimator class. What differs is that these works target a single posterior expectation, whereas we face a family of integrands indexed by λ, evaluated thousands of times per iteration, where the decision-relevant quantity is a **ranking**, *not* an expectation. This has two consequences specific to our setting, which, to our knowledge, are new. The score covariance Σ_g is computed once per iteration and amortized across all candidates (Supplement D), whereas control functionals would require per-candidate function-space regression, and Proposition 4.3 links variance reduction to fewer acquisition-ranking reversals (see Tab. 1). Closest in BO is Daulton et al. (2022), which uses a score control variate for gradient estimation over discrete and mixed spaces. We instead target the acquisition value under surrogate-parameter uncertainty.

Orthogonal machine learning. Orthogonal machine learning is rooted in semi-parametric statistics Bickel et al. (1993), and is applied in many settings where the estimation target is defined based on unknown nuisance components Chernozhukov et al. (2018); Foster & Syrgkanis (2023). Very common examples are found in the missing data analysis Robins et al. (1994) and causal inference literature Kennedy (2024); Nie & Wager (2021); Oprescu et al. (2019); Robins et al. (2000); Van Der Laan & Rubin (2006). Recently, orthogonal learning has also been expanded to other fields such as private machine learning Schröder et al. (2025), survival analysis Frauen et al. (2025), LLM evaluation Frauen et al. (2026), and general ML optimization topics such as neural operators or generative models Hess et al. (2026); Melnychuk & Feuerriegel (2026).

B FAILURE MODES IN BAYESIAN OPTIMIZATION

Failure modes in BO. Several failure modes arise in a realistic HPO setting (see, e.g., (Ament et al., 2023; Berkenkamp et al., 2019; Bogunovic & Krause, 2021; Neiswanger & Ramdas, 2021)). The reason is that 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. As a result, BO suffers from the following main 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 stationary structure may be incompatible with the true response surface. As a result, the posterior mean and posterior uncertainty estimates may be systematically biased.

(ii) *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.

(iii) *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 negatively affect the ranking of candidate points.

$\Rightarrow$ Our work contributes to (iii).

C 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 introduce separate versions for the constrained and parallel Bayesian optimization.

C.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

$$A_m(z) := \mathbb{E}_{\theta_m \sim q_{m,t}}[a_m(z; \theta_m)]\,. \tag{23}$$

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

$$\widehat{A}_m^{\text{orth}}(z) = \frac{1}{S} \sum_{s=1}^{S} \left[a_m(z; \theta_m^{(s)}) - \left(\text{Cov}(g_m, g_m)^{-1} \text{Cov}(g_m, a_m) \right)^\top g_m(\theta_m^{(s)}) \right]. \tag{24}$$

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

$$\mathbb{E}_{q_{m,t}}\left[\|g_m(\theta_m)\|_2^2 \right] < \infty, \qquad \mathbb{E}_{q_{m,t}}\left[a_m(\theta_m; z)^2 \right] < \infty, \tag{25}$$

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. $$\text{Var}\Big(\widehat{A}_m^{\text{orth}}(z) \Big) = \frac{1}{S} \left(\text{Var}(h_m) - \text{Cov}(h_m, g_m)^\top \text{Cov}(g_m, g_m)^{-1} \text{Cov}(g_m, h_m) \right), \tag{26}$$

   and therefore

   $$\text{Var}\Big(\widehat{A}_m^{\text{orth}}(z) \Big) \le \text{Var}\Big(\widehat{A}_m^{\text{MC}}(z) \Big), \tag{27}$$

   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 E after replacing $\text{EI}_m(\lambda; \theta_m)$ by the generic acquisition value $a_m(z; \theta_m)$. □

Remark. Proposition C.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.

C.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

$$f_i(\lambda) \le 0, \qquad i = 2, \ldots, M. \tag{28}$$

A standard constrained Expected Improvement (CEI) acquisition is

$$\mathrm{CEI}_m(\lambda;\theta_m) := \mathbb{E}\left[(f^\star - f_1(\lambda))_+ \prod_{i=2}^{M} \mathbf{1}\{f_i(\lambda) \le 0\} \,\middle|\, \mathcal{D}_t, \theta_m\right], \tag{29}$$

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

$$\widehat{\mathrm{CEI}}_m^{\mathrm{orth}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\left[\mathrm{CEI}_m(\lambda;\theta_m^{(s)}) - \left(\mathrm{Cov}(g_m,g_m)^{-1}\,\mathrm{Cov}\big(g_m, h_m^{\mathrm{CEI}}\big)\right)^\top g_m(\theta_m^{(s)})\right]. \tag{30}$$

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

C.3 Parallel or batch expected improvement

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

$$\mathrm{qEI}_m(\Lambda;\theta_m) := \mathbb{E}\left[\max_{j=1,\ldots,q}\left(f^\star - f(\lambda_j)\right)_+ \,\middle|\, \mathcal{D}_t, \theta_m\right], \tag{31}$$

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

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

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

Proof. Observe that for every fixed batch $X = (x_1, \ldots, x_q)$ 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 C.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

$$\widehat{\mathrm{qEI}}_m(\Lambda;\theta_m) = \frac{1}{N}\sum_{n=1}^{N}\max_{j=1,\ldots,q}\left(f^\star - \xi^{(n)}(\lambda_j)\right)_+, \qquad \xi^{(n)} \sim p(\cdot \mid \mathcal{D}_t, \theta_m), \tag{33}$$

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

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

Importantly, under the assumption that the inner estimator equation 33 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.

D THEORETICAL PROPERTIES

D.1 VARIANCE REDUCTION, ROBUSTNESS, AND RANKING STABILITY FOR GP AND TPE

We provide versions of Theorem 4.1 and Proposition 4.3 for the two surrogate instantiations GP and TPE. The proofs follow directly from the proofs of the original theorem and proposition in Supplement E.

Theorem D.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*

$$\mathrm{EI}^{\mathrm{GP,orth}}(\lambda;\theta) := \mathrm{EI}^{\mathrm{GP}}(\lambda;\theta) - \gamma^{\mathrm{GP}}(\lambda)^\top g_t(\theta). \tag{35}$$

Then:

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

(ii) ***Variance reduction:***

$$\mathrm{Var}\big(\mathrm{EI}^{\mathrm{GP,orth}}(\lambda;\theta)\big) = \mathrm{Var}\big(\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta)\big) - \mathrm{Cov}(\mathrm{EI}^{\mathrm{GP}}, g_t)^\top \mathrm{Cov}(g_t, g_t)^{-1}\,\mathrm{Cov}(g_t, \mathrm{EI}^{\mathrm{GP}}), \tag{36}$$

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 D.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 D.1 to the difference functional $\mathrm{EI}^{\mathrm{GP}}(\lambda;\theta) - \mathrm{EI}^{\mathrm{GP}}(\lambda';\theta)$. Then, it holds*

$$\mathbb{P}\Big(\widehat{\Delta}^{\mathrm{GP,orth}} \le 0\Big) \;\le\; \frac{\mathrm{Var}(\widehat{\Delta}^{\mathrm{GP,orth}})}{(\Delta^{\mathrm{GP}}(\lambda,\lambda'))^2} \;\le\; \frac{\mathrm{Var}(\widehat{\Delta}^{\mathrm{GP,MC}})}{(\Delta^{\mathrm{GP}}(\lambda,\lambda'))^2}. \tag{37}$$

Theorem D.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*

$$h^{\mathrm{TPE,orth}}(\lambda;\theta) := h^{\mathrm{TPE}}(\lambda;\theta) - \gamma^{\mathrm{TPE}}(\lambda)^\top c_t(\theta). \tag{38}$$

Then:

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

(ii) ***Variance reduction:***

$$\mathrm{Var}\big(h^{\mathrm{TPE,orth}}(\lambda;\theta)\big) = \mathrm{Var}\big(h^{\mathrm{TPE}}(\lambda;\theta)\big) - \mathrm{Cov}(h^{\mathrm{TPE}}, c_t)^\top \mathrm{Cov}(c_t, c_t)^{-1}\,\mathrm{Cov}(c_t, h^{\mathrm{TPE}}), \tag{39}$$

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 D.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 D.3 to the difference functional $h^{\mathrm{TPE}}(\lambda;\theta) - h^{\mathrm{TPE}}(\lambda';\theta)$. Then*

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

E PROOFS OF THE MAIN THEOREMS AND PROPOSITIONS

E.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) Preserved target. $\mathrm{EI}_m^{\mathrm{orth,marg}}(\lambda) = \mathrm{EI}_m^{\mathrm{marg}}(\lambda)$.

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

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

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

$$\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta) = h(\theta) - \gamma^\top g(\theta). \tag{42}$$

(i) Preserved target. Under the assumed regularity conditions, the score function satisfies

$$\mathbb{E}_{q_{m,t}}[g(\theta)] = \int \nabla_\theta \log q_{m,t}(\theta)\, q_{m,t}(\theta)\, d\theta = \int \nabla_\theta q_{m,t}(\theta)\, d\theta = 0. \tag{43}$$

Hence

$$\mathrm{EI}_m^{\mathrm{orth,marg}}(\lambda) = \mathbb{E}_{q_{m,t}}[h(\theta)] - \gamma^\top \mathbb{E}_{q_{m,t}}[g(\theta)] = \mathbb{E}_{q_{m,t}}[h(\theta)] = \mathrm{EI}_m^{\mathrm{marg}}(\lambda). \tag{44}$$

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

$$\mathrm{Var}_{q_{m,t}}(h - a^\top g) = \mathrm{Var}(h) - 2a^\top \mathrm{Cov}(g,h) + a^\top \Sigma_g\, a. \tag{45}$$

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

$$a^\star = \Sigma_g^{-1}\,\mathrm{Cov}(g,h) = \gamma. \tag{46}$$

Substituting $a = \gamma$ gives

$$\mathrm{Var}\big(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\big) = \mathrm{Var}(h) - \mathrm{Cov}(g,h)^\top \Sigma_g^{-1}\,\mathrm{Cov}(g,h). \tag{47}$$

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

$$\mathrm{Var}\big(\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\big) \;\le\; \mathrm{Var}\big(\mathrm{EI}_m(\lambda;\theta)\big). \tag{48}$$

□

E.2 PROOF OF COROLLARY 4.2

Corollary 4.2 Assume the conditions of Theorem 4.1. 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, then it holds that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\,\mathbb{E}_{q_{m,t}^{(\varepsilon)}}\Big[\mathrm{EI}_m^{\mathrm{orth}}(\lambda;\theta)\Big]\Big|_{\varepsilon=0} \;=\; 0. \tag{49}$$

Proof. Throughout, we hold γ fixed at its $\varepsilon = 0$ value. Consider the exponentially tilted family

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

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

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

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

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

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

E.3 Proof of Proposition 4.3

Proposition 4.3 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, we have

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

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. □

F COMPUTATIONAL COMPLEXITY

GP surrogate. For an exact GP surrogate, fitting has a complexity of $O(t^3)$ 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 naïve 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.

G IMPLEMENTATION DETAILS

General implementation details Our experiments utilized 4 CPU cores, 16 GB of main memory, and an 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 (Jang, 2015; Wu et al., 2015) 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.

G.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

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

G.2 KERNEL FAMILIES

We provide details about the synthetic benchmark functions and their characteristics in Table 5.

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

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.

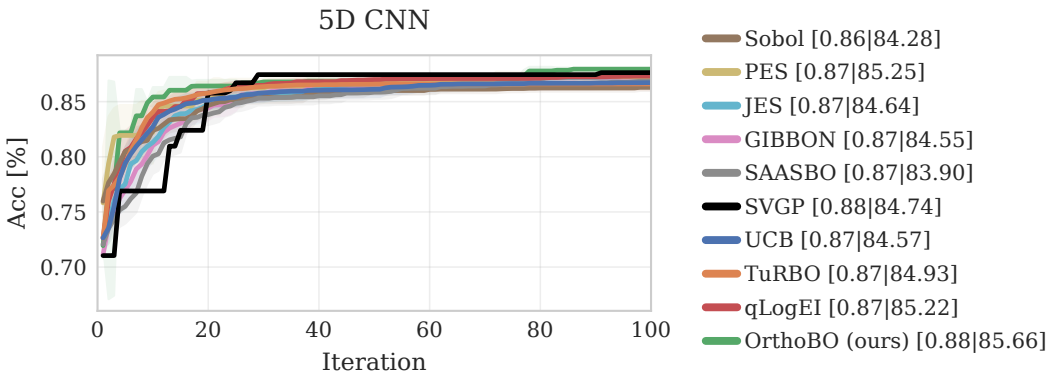

Figure 6: Results for HPO optimization for training on **FashionMNIST** from scratch. Numbers in the legend indicate the final performance and the area-under-curve; higher values are better.

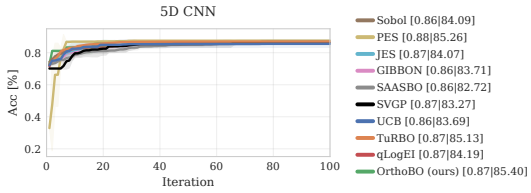

Figure 7: Results for HPO optimization for training on **FashionMNIST** from scratch, with random 20 % observational noise. Numbers in the legend indicate the final performance and the area-under-curve; higher values are better.

G.3 CNN EXPERIMENTS

FashionMNIST We include the CNN experiment used by Ament et al. (2024). Specifically, the goal is to optimize the test-set accuracy of a CNN trained on MNIST images. We optimize the same parameters as Ament et al. (2024). As we do not study robustness to outliers, we do **not** use corrupted evaluations. In Fig. 6, we replicate the figure from the main paper in larger resolution. We additionally report the area-under-curve (AUC). We observe: **OrthoBO reaches the highest final performance (together with SVGP), and has a higher AUC, indicating a steadily higher performance throughout the optimization.**

FashionMNIST with observational noise We repeat our previous experiment, but inject observational noise: we prematurely stop training with probability 20 % after observing between 100 and 1000 training samples. Results are in Fig. 7 => **OrthoBO reaches the highest AUC score**.

MNIST We conduct the CNN experiment with the MNIST (LeCun et al., 2010) dataset. Results are in Fig 8. Our observations are consistent with our previous findings: **OrthoBO reaches the same performance as baselines, but earlier (higher AUC)**.

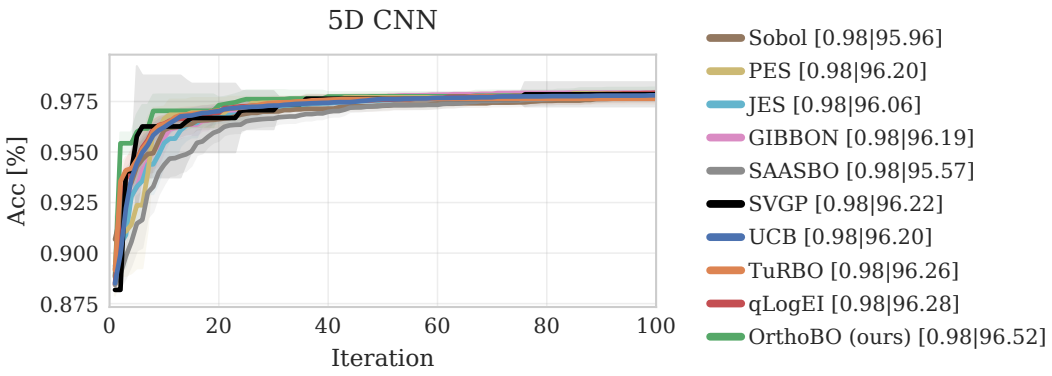

Figure 8: Results for HPO optimization for training on **MNIST** from scratch. Numbers in the legend indicate the final performance and the area-under-curve; higher values are better.

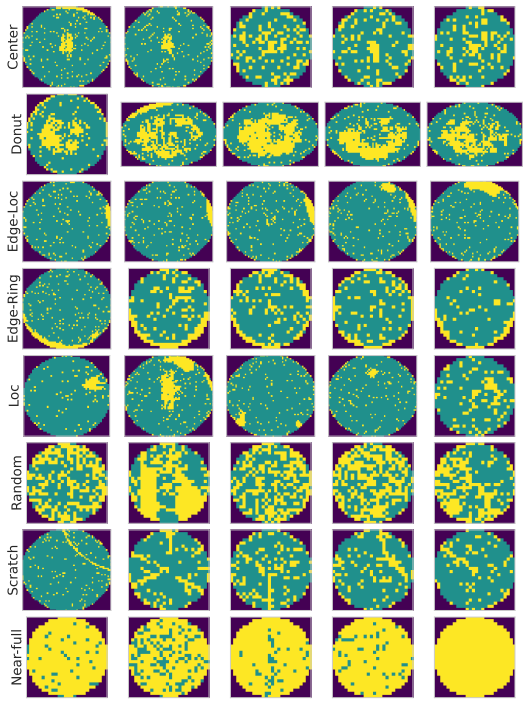

Figure 9: Examples of wafer-map manufacturing failures.

G.4 Case study experiment, additional details

Dataset. For our real-world case study, we use the large-scale industrial WM811K dataset (Jang, 2015; Wu et al., 2015). 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 9. 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.

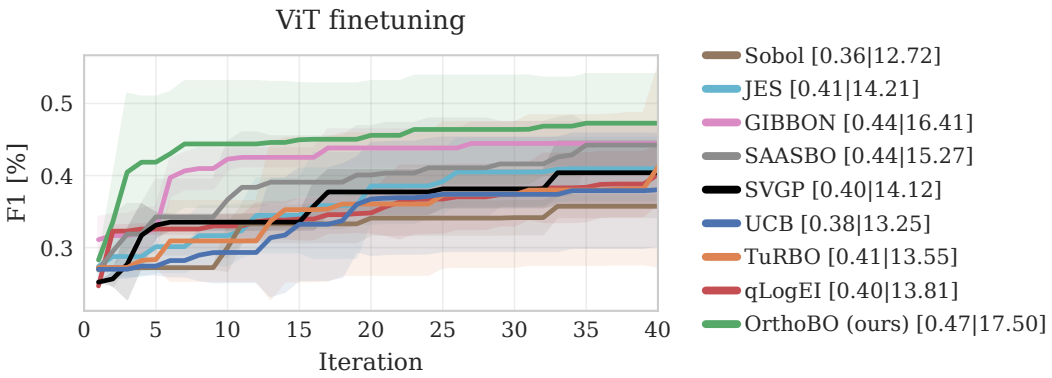

Figure 10: Results for HPO optimization of fine-tuning hyperparameters on WM811K. Numbers in the legend indicate the final performance and the area-under-curve; higher values are better.

Model. As a model, we use an ImageNet (Deng et al., 2009) pretrained ViT B16 (Dosovitskiy et al., 2020) model. We freeze the backbone model and train only the penultimate output layers.

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 $\gamma \in [0, 1]$.[8]

Results. We show detailed results, including final F1 and AUC scores, in Fig. 10. We observe: **OrthoBO reaches the highest F1 score very early and maintains it through optimization.**

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

H FURTHER RESULTS

H.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 summary results in the main text. We now report detailed results for (i) a Matérn-$5/2$ kernel in section H.1.1, (ii) a RBF kernel in section H.1.2, and (iii) a TPE surrogate in section H.1.3.

H.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)	$0.073_{0.21}$	$0.988_{0.06}$	$0.048_{0.04}$	$58.938_{0.00}$

H.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)	$0.034_{0.17}$	$1.000_{0.00}$	$0.049_{0.04}$	$2.819_{0.57}$

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

H.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., (Ament et al., 2023; Bodin et al., 2020; Eriksson & Poloczek, 2021; Moss et al., 2023)). 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. 11.

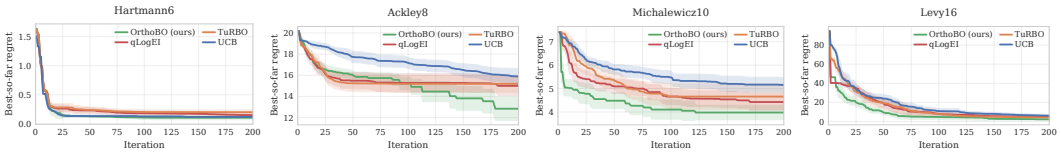

Figure 11: Best-so-far regret as a function of iterations on four widely used regression problems, using a **mildly misspecified isotropic RBF kernel**. Our proposed ORTHOBO is competitive, and in higher-dimensional functions (Michalewicz10, Levy16) outperforms the baselines.

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. 12.

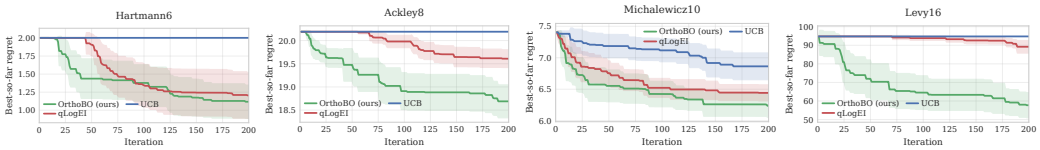

Figure 12: Best-so-far regret as a function of iterations on four widely used regression problems, using a **strongly misspecified linear kernel**. Our proposed ORTHOBO substantially outperforms other methods. The performance difference increases as the problem dimensionality grows. Notably, on Hartmann6, 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.

H.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. 13.

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 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.

H.4 ENSEMBLING

In the main experiments, we used $M = 1$ to avoid confounding the effect of orthogonalization with 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 14.

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 15.

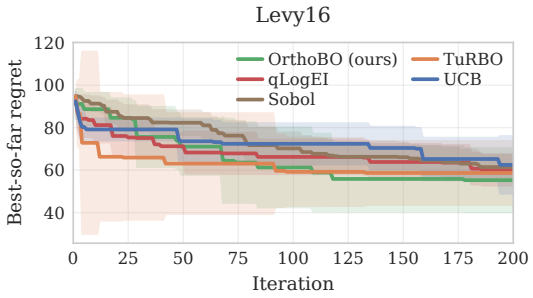

Figure 13: Best-so-far regret on Levy16 with a Matérn-$5/2$ ARD kernel and only $n_0 = 4$ initial Sobol points. ORTHOBO remains among the strongest methods and improves steadily despite weakly identified surrogate hyperparameters.

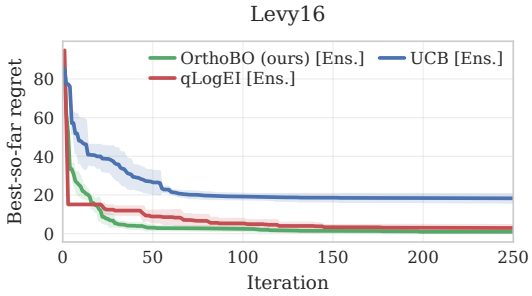

Figure 14: Results for ensembling.

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. (2) ORTHOBO has substantially lower Sobol probing variance. (3) The weights are highest for the Matérn-$5/2$ kernel ("Weight 3"), demonstrating that all methods primarily focus on the "most correct" kernel. We plot the evolution of the entropy over time in Figure 15, which shows that the weight aggregation scheme balances exploration and exploitation.

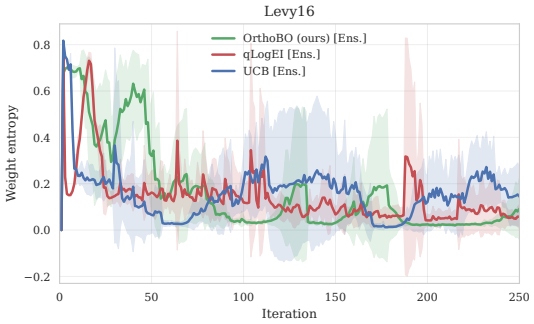

Figure 15: Evolution of the entropy across ensembling weights. For all methods, the used tempered exponentially weighted aggregation scheme is effective in balancing exploration and exploitation at the *ensemble* level.

H.5 Monte Carlo sampling ablation experiment

We now 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. Ament et al., 2023). 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. 16.

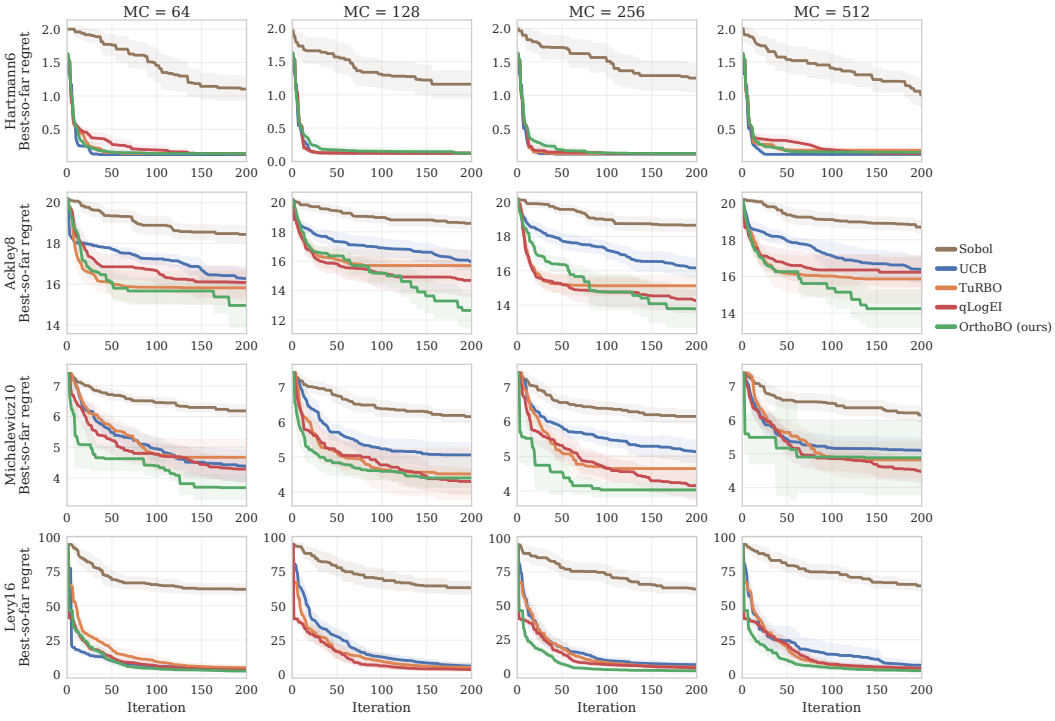

Figure 16: Different mc samples

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 is especially

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.

H.6 CUMULATIVE REGRET

An unconditional sublinear cumulative-regret bound is out of reach for our work, as no cumulative theory exists even for exact marginalized EI, and under the misspecification we target, linear worst-case terms are unavoidable Bodin et al. (2020). What exists is a regret-transfer result. With candidate sets as in Alg. 1, R_T exceeds the exact marginal EI one-step oracle's regret by at most $B \cdot N_T$, where the expected number of noise-flipped selections N_T is bounded by summing Proposition 4.3 over rounds and candidates. Under a mild budget schedule, this excess is O(1) a.s., and orthogonalization reduces the budget required for any target flip probability by exactly the explained-variance factor (9–16× in Table 1). This quantifies precisely how variance reduction affects the per-round candidate choice.

Beyond the theoretical discussion, we provide empirical regret below.

Table 12: Cumulative regret across iterations on the Ackley8 test function. Our ORTHOBO reaches the lowest cumulative regret.

Method	Iteration 25	Iteration 50	Iteration 100	Iteration 150	Iteration 200
Sobol	$498.32_{14.84}$	$988.71_{32.02}$	$1940.79_{72.25}$	$2878.04_{114.35}$	$3801.48_{151.77}$
UCB	$460.38_{12.89}$	$897.92_{30.89}$	$1749.24_{84.31}$	$2580.70_{135.65}$	$3392.71_{186.91}$
TuRBO	$440.34_{17.37}$	$849.17_{47.44}$	$1650.51_{127.76}$	$2448.54_{217.11}$	$3246.57_{307.19}$
qLogEI	$438.80_{16.93}$	$842.60_{43.13}$	$1634.16_{111.56}$	$2407.12_{203.17}$	$3174.39_{297.95}$
OrthoBO (ours)	$447.41_{22.00}$	$849.88_{47.79}$	$1619.74_{88.50}$	$2342.38_{128.19}$	$3026.99_{186.53}$

H.7 ORTHOGONALIZATION OF PI

The variance-reduction principle of ORTHOBO can be extended to other acquisition functions, but may require acquisition-specific orthogonalization schemes (see Supplement C.1). Below, we demonstrate the variance reduction for the PI acquisition function.

Table 13: Variance reduction for orthogonalized PI

Michalewicz10		Levy16	
$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{MC}})$	$\mathrm{Var}(\widehat{\mathrm{EI}}_m^{\mathrm{ortho}})$
81.7816	38.4443 (-52.99%%)	114.224	72.9921 (-36.10%)