A Hierarchical Expected Improvement Method for Bayesian Optimization

Zhehui Chen, Simon Mak & C. F. Jeff Wu

Taylor & Francis
Taylor & Francis Group

Check for updates

A Hierarchical Expected Improvement Method for Bayesian Optimization

Zhehui Chen*[a], Simon Mak*[b], and C. F. Jeff Wu[a]

[a]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA; [b]Department of Statistical Science, Duke University, Durham, NC

ABSTRACT

The Expected Improvement (EI) method, proposed by Jones, Schonlau, and Welch, is a widely used Bayesian optimization method, which makes use of a fitted Gaussian process model for efficient black-box optimization. However, one key drawback of EI is that it is overly greedy in exploiting the fitted Gaussian process model for optimization, which results in suboptimal solutions even with large sample sizes. To address this, we propose a new hierarchical EI (HEI) framework, which makes use of a hierarchical Gaussian process model. HEI preserves a closed-form acquisition function, and corrects the over-greediness of EI by encouraging exploration of the optimization space. We then introduce hyperparameter estimation methods which allow HEI to mimic a fully Bayesian optimization procedure, while avoiding expensive Markov-chain Monte Carlo sampling steps. We prove the global convergence of HEI over a broad function space, and establish near-minimax convergence rates under certain prior specifications. Numerical experiments show the improvement of HEI over existing Bayesian optimization methods, for synthetic functions and a semiconductor manufacturing optimization problem. Supplementary materials for this article are available online.

1. Introduction

Bayesian optimization (BO) is a widely-used optimization framework, which has broad applicability in a variety of problems, including rocket engine design (Mak et al. 2018), nanowire yield optimization (Dasgupta et al. 2008), and surgery planning (Chen et al. 2020). BO aims to solve the following black-box optimization problem:

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x} \in \Omega} f(\mathbf{x}). \tag{1}$$

Here, $\mathbf{x} \in \mathbb{R}^d$ are the input variables, and $\Omega \subset \mathbb{R}^d$ is the feasible domain for optimization. The key challenge in (1) is that the objective function $f(\cdot) : \Omega \to \mathbb{R}$ is assumed to be black-box: it admits no closed-form expression, and evaluations of f (which we presume to be noiseless in this work) typically require expensive simulations or experiments. For such problems, an optimization procedure should find a good solution to (1) given limited function evaluations. BO achieves this by first assigning to f a *prior model* capturing prior beliefs on the objective function, then sequentially querying f at points which maximize the *acquisition function*—the posterior expected utility of a new point. This provides a principled way to perform the so-called *exploration-exploitation tradeoff* (Kearns and Singh 2002): exploring the black-box function over Ω, and exploiting the fitted function when appropriate for optimization.

Much of the literature on BO can be categorized by (i) the prior stochastic model assumed on f, and (ii) the utility function

used for sequential sampling. For (i), the most popular stochastic model is the Gaussian process (GP) model (Santner et al. 2018). Under a GP model, several well-known BO methods have been derived using different utility functions for (ii). These include the expected improvement (EI) method (Mockus, Tiesis, and Zilinskas 1978; Jones, Schonlau, and Welch 1998), the upper confidence bound (UCB) method (Srinivas et al. 2010), and the Knowledge Gradient method (Frazier, Powell, and Dayanik 2008; Scott, Frazier, and Powell 2011). Of these, EI is arguably the most popular method, since it admits a simple *closed-form* acquisition function, which can be efficiently optimized to yield subsequent query points on f. EI has been subsequently developed for a variety of black-box optimization problems, including multi-fidelity optimization (Zhang, Han, and Zhang 2018), constrained optimization (Feliot, Bect, and Vazquez 2017), and parallel/batch-sequential optimization (Marmin, Chevalier, and Ginsbourger 2015).

Despite the popularity of EI, it does have key limitations. One such limitation is that it is too *greedy* (Qin, Klabjan, and Russo 2017): EI focuses nearly all sampling efforts near the optima of the *fitted* GP model, and does not sufficiently explore other regions. In terms of the exploration-exploitation tradeoff (Kearns and Singh 2002), EI over-exploits the fitted model on f, and under-explores the domain; this causes the procedure to get stuck in local optima and not converge to any global optimum $\mathbf{x}^*$ (Bull 2011). In recent work, an effective way to address this greediness is by *integrating uncer-*

tainty on model parameters within the EI acquisition function. Snoek, Larochelle, and Adams (2012) proposed a fully Bayesian EI, where GP model parameters are sampled using Markov chain Monte Carlo (MCMC); this incorporates parameter uncertainty within EI via a fully Bayesian framework, which enables improved optimization by encouraging exploration. Chen, Wang, and Wu (2017) proposed a variation of EI under an additive Bayesian model, which encourages exploration by increasing model uncertainty. However, by integrating parameter uncertainty, a *significant* bottleneck of these methods is that it requires *expensive* MCMC sampling, which can be very costly to integrate within the maximization of the acquisition function. This computational burden diminishes a key advantage of EI: efficient queries via a *closed-form* criterion.

Another remedy, proposed by Bull (2011), is to artificially inflate the maximum-likelihood estimator of the GP variance, and use this inflated estimate within the EI acquisition function for sequential sampling. This, in effect, encourages exploration of the black-box function by inflating the uncertainty of the fitted model. The work further employs an "ϵ-greedy" modification of the EI, where at each sampling iteration, one selects the next point uniformly-at-random with probability $\epsilon \in (0, 1)$. With a sufficiently large ϵ, this again forces the procedure to explore the feasible domain. While the EI with such modifications (we call this the ϵ-EI later) allows for appealing theoretical properties, such an approach may yield suboptimal optimization performance for black-box problems with limited function evaluations, as we show later in numerical experiments. One reason is that such adjustments, while indeed encouraging exploration, does so in a rather ad-hoc manner from a modeling perspective. As such, when evaluations are limited, the ϵ-EI can be suboptimal for striking a good balance between exploration and exploitation, particularly for higher-dimensional problems.

To address these limitations, we propose a novel Hierarchical EI (HEI) framework for effective Bayesian optimization with limited function evaluations. Similar to Snoek, Larochelle, and Adams (2012) and work thereafter, the HEI integrates parameter uncertainty on model parameters, but does so via the *hierarchical* GP model in Handcock and Stein (1993), which we show yields a *closed-form* acquisition function for Bayesian optimization. This closed-form enables *efficient* sequential sampling, without the need for integrating expensive MCMC samples in optimizing subsequent queries. With this hierarchical framework, we further introduce hyperparameter estimation methods, which allow the HEI to mimic a fully Bayesian optimization procedure (the "gold standard" for uncertainty quantification) while avoiding expensive MCMC steps. Under certain prior specifications, we then prove that the HEI indeed converges to a global optimum $\mathbf{x}^*$ over a broad function space for f and achieves a near-minimax convergence rate. These theoretical guarantees are similar to those achieved by the ϵ-EI (Bull 2011) and more recent results (see, e.g., Wynne, Briol, and Girolami 2020), but are achieved under a principled hierarchical Bayesian framework, without the need for ad-hoc variance inflation or ϵ-greedy sampling. As such, the proposed HEI provides a principled balance between exploration and exploitation guided by the underlying hierarchical GP model, which in turn yields improved practical performance given limited function evaluations. We finally demonstrate the sample-efficient performance of HEI over existing methods in a suite of numerical experiments and an application to semi-conductor manufacturing.

We note that a special case of HEI, called the Student EI (SEI), was proposed earlier in Benassi, Bect, and Vazquez (2011). The proposed HEI has several key advantages over the SEI: the HEI incorporates uncertainty on process nonstationarity, and can mimic a fully Bayesian optimization procedure via hyperparameter estimation. We also show that the HEI has provable global convergence and convergence rates for optimization, whereas the SEI (with the recommended hyperparameter specification) can fail to converge to a global optimum $\mathbf{x}^*$. Numerical experiments and a semiconductor manufacturing application show the improved performance of the HEI over existing BO methods, including the SEI approach in Benassi, Bect, and Vazquez (2011) and the ϵ-EI approach in Bull (2011).

The article is organized as follows: Section 2 reviews the GP model and the EI method. Section 3 presents the HEI method and contrasts it with existing methods. Section 4 provides methodological developments on hyperparameter specification and basis selection. Section 5 proves the global convergence for HEI and its associated convergence rates. Sections 6 and 7 compare HEI with existing methods in a suite of numerical experiments and for a semiconductor manufacturing problem, respectively. Concluding remarks are given in Section 8.

2. Background and Motivation

We first introduce the GP model, then review the EI method and its deficiencies, which will help motivate the proposed HEI method.

2.1. Gaussian Process Modeling

We first model the black-box objective function f as the following Gaussian process model:

$$f(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}), \quad \mu(\mathbf{x}) = \mathbf{p}^\top(\mathbf{x})\boldsymbol{\beta}, \quad Z(\mathbf{x}) \sim \text{GP}(0, \sigma^2 K), \tag{2}$$

where $\mu(\mathbf{x})$ is the mean function of the process, $\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}), \ldots, p_q(\mathbf{x})]^\top$ are the q basis functions for $\mu(\mathbf{x})$, and $\boldsymbol{\beta} \in \mathbb{R}^q$ are its corresponding coefficients. Here, we assume the residual term $Z(\mathbf{x})$ follows a stationary Gaussian process prior (Santner et al. 2018) with mean zero, process variance σ^2 and correlation function $K(\cdot, \cdot)$, which we denote as $Z(\mathbf{x}) \sim \text{GP}(0, \sigma^2 K)$. The model (2) is known as the *universal kriging* (UK) model in the geostatistics literature (Wackernagel 1995). When the modeler does not have prior information on appropriate basis functions to use, one can employ the so-called *ordinary kriging* (OK) model (Olea 2012), which sets a constant mean for $\mu(\mathbf{x})$, that is, $p_1(\mathbf{x}) = 1$ and $q = 1$. In what follows, we assume that the kernel K is radial with length-scale parameters $\boldsymbol{\theta}$; this is more formally defined later in Section 5 (see (24)).

Suppose noiseless function values $y_i = f(\mathbf{x}_i)$ are observed at inputs $\mathbf{x}_i$, yielding data $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Let $\mathbf{y}_n = (y_i)_{i=1}^n$ be the vector of observed function values, $\mathbf{k}_n(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_i))_{i=1}^n$ be the correlation vector between the unobserved response $f(\mathbf{x})$

and observed responses $\mathbf{y}_n$, $\mathbf{K}_n = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ be the correlation matrix for observed points, and $\mathbf{P}_n = [\mathbf{p}(\mathbf{x}_1), \ldots, \mathbf{p}(\mathbf{x}_n)]^\top$ be the model matrix for observed points. Then the posterior distribution of $f(\mathbf{x})$ at an unobserved input $\mathbf{x}$ has the closed form expression (Santner et al. 2018):

$$[f(\mathbf{x}) | \mathcal{D}_n] \sim \mathcal{N}\left(\hat{f}_n(\mathbf{x}), \sigma^2 s_n^2(\mathbf{x})\right). \tag{3}$$

Here, the posterior mean $\hat{f}_n(\mathbf{x})$ is given by

$$\hat{f}_n(\mathbf{x}) = \mathbf{p}^\top(\mathbf{x})\hat{\boldsymbol{\beta}}_n + \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\left(\mathbf{y}_n - \mathbf{P}_n\hat{\boldsymbol{\beta}}_n\right), \tag{4}$$

the posterior variance $\sigma^2 s_n^2(\mathbf{x})$ is given by

$$\sigma^2 s_n^2(\mathbf{x}) = \sigma^2\left(K(\mathbf{x}, \mathbf{x}) - \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}) + \mathbf{h}_n^\top(\mathbf{x})\mathbf{G}_n^{-1}\mathbf{h}_n(\mathbf{x})\right), \tag{5}$$

where $\hat{\boldsymbol{\beta}}_n = \mathbf{G}_n^{-1}\mathbf{P}_n^\top\mathbf{K}_n^{-1}\mathbf{y}_n$ is the maximum likelihood estimator (MLE) for $\boldsymbol{\beta}$, $\mathbf{G}_n = \mathbf{P}_n^\top\mathbf{K}_n^{-1}\mathbf{P}_n$ and $\mathbf{h}_n(\mathbf{x}) = \mathbf{p}(\mathbf{x}) - \mathbf{P}_n^\top\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})$. These expressions can be equivalently viewed as the best linear unbiased predictor of $f(\mathbf{x})$ under the presumed GP model and its variance (see sec. 3 of Santner et al. 2018 for further details).

2.2. Expected Improvement

The EI method (Jones, Schonlau, and Welch 1998) then makes use of the above closed-form equations to derive a closed-form acquisition function. Let $y_n^* = \min_{i=1}^n y_i$ be the current best objective value, and let $(y_n^* - f(\mathbf{x}))_+ = \max\{y_n^* - f(\mathbf{x}), 0\}$ be the *improvement* utility function. Given data $\mathcal{D}_n$, the expected improvement acquisition function can be written as

$$\mathbb{E}_{f|\mathcal{D}_n}(y_n^* - f(\mathbf{x}))_+ = I_n(\mathbf{x})\Phi\left(\frac{I_n(\mathbf{x})}{\sigma s_n(\mathbf{x})}\right) + \sigma s_n(\mathbf{x})\phi\left(\frac{I_n(\mathbf{x})}{\sigma s_n(\mathbf{x})}\right). \tag{6}$$

Here, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of the standard normal distribution, respectively, and $I_n(\mathbf{x}) = y_n^* - \hat{f}_n(\mathbf{x})$. For an unobserved point $\mathbf{x}$, $\mathbb{E}_{f|\mathcal{D}_n}(y_n^* - f(\mathbf{x}))_+$ can be interpreted as the expected improvement to the current best objective value over $\mathbf{x}$, if the next query is at point $\mathbf{x}$.

In order to compute the acquisition function in (6), we would need to know the process variance σ^2. In practice, however, this is typically unknown and needs to be estimated from data. A standard approach (Bull 2011) is to compute its MLE:

$$\hat{\sigma}_n^2 = \frac{1}{n}(\mathbf{y}_n - \mathbf{P}_n\hat{\boldsymbol{\beta}}_n)^\top\mathbf{K}_n^{-1}(\mathbf{y}_n - \mathbf{P}_n\hat{\boldsymbol{\beta}}_n), \tag{7}$$

then plug-in this into the acquisition function (6). This gives the following plug-in estimate of the EI acquisition function:

$$\mathrm{EI}_n(\mathbf{x}) = \underbrace{I_n(\mathbf{x})\Phi\left(\frac{I_n(\mathbf{x})}{\hat{\sigma}_n s_n(\mathbf{x})}\right)}_{\text{Exploitation}} + \underbrace{\hat{\sigma}_n s_n(\mathbf{x})\phi\left(\frac{I_n(\mathbf{x})}{\hat{\sigma}_n s_n(\mathbf{x})}\right)}_{\text{Exploration}}. \tag{8}$$

In practice, the length-scale parameters $\boldsymbol{\theta}$ of the GP are typically estimated via MLE and plugged into the estimated EI criterion (8) as well.

With (8) in hand, the next query point $\mathbf{x}_{n+1}$ is obtained by maximizing the EI acquisition function $\mathrm{EI}_n(\mathbf{x})$:

$$\mathbf{x}_{n+1} \leftarrow \underset{\mathbf{x} \in \Omega}{\operatorname{argmax}}\, \mathrm{EI}_n(\mathbf{x}). \tag{9}$$

This maximization of (8) implicitly captures the aforementioned *exploration-exploitation tradeoff*: exploration of the feasible region and exploitation near the current best solution. The maximization of the first term in (8) encourages *exploitation*, since larger values are assigned for points $\mathbf{x}$ with smaller predicted values $\hat{f}_n(\mathbf{x})$. The maximization of the second term in (8) encourages *exploration*, since larger values for points $\mathbf{x}$ indicate larger (estimated) predictive standard deviation $\hat{\sigma}_n s_n(\mathbf{x})$.

However, one drawback of EI is that it fails to capture the full uncertainty of model parameters within the acquisition function $\mathrm{EI}_n(\mathbf{x})$. This results in an *over-exploitation* of the fitted GP model for optimization, and an *under-exploration* of the black-box function over Ω. This over-greediness has been noted in several works, for example, Bull (2011) and Qin, Klabjan, and Russo (2017). In particular, Theorem 3 of Bull (2011) showed that, for a common class of correlation functions for K (see Assumption 1 later), there always exists some smooth function f within a function space $\mathcal{H}_{\boldsymbol{\theta}}(\Omega)$ (defined later in Section 5) such that EI fails to find a global optimum of f. This is stated formally below:

Proposition 1 (Theorem 3, Bull 2011). Suppose Assumption 1 holds with $\nu < \infty$. Suppose initial points are sampled according to some probability measure F over Ω. Let $(\mathbf{x}_i)_{i=1}^\infty$ be the points generated by maximizing EI_n in (8), with plug-in MLEs for GP length-scale parameters $\boldsymbol{\theta}$ satisfying Assumption 2 (introduced later). Then, for any $\epsilon > 0$, there exist some $f \in \mathcal{H}_{\boldsymbol{\theta}}(\Omega)$ and some constant $\delta > 0$ such that

$$\mathbb{P}_F\left(\lim_{n \to \infty} y_n^* - \min_{\mathbf{x} \in \Omega} f(\mathbf{x}) \geq \delta\right) > 1 - \epsilon.$$

This proposition shows that, even for relatively simple objective functions f, the EI may fail to converge to a global minimum due to its over-exploitation of the fitted GP model. This is concerning not only from an asymptotic perspective, but also in practical problems with limited samples. As we show later, this overexploitation can cause the EI to get stuck on suboptimal solutions, resulting in poor empirical performance compared to the HEI.

2.3. Existing Solutions and Limitations

As mentioned in the introduction, there are several notable methods in the literature which aim to tackle this over-exploitation. One approach which has proven effective is to *integrate uncertainty on GP model parameters* within the EI criterion. An early work in this direction is Snoek, Larochelle, and Adams (2012), which made use of a fully Bayesian formulation of the EI with a full quantification of uncertainty in all GP model parameters. This integration of parameter uncertainty has subsequently been developed in recent works; see, for example, Chen, Wang, and Wu (2017). While effective, a key limitation with these methods is that this integration of uncertainty requires not only MCMC sampling of model parameters, but also the use of such

samples within the optimization of the estimated EI criterion (9). Particularly in higher dimensions, where many MCMC samples are required, this can greatly slow down and even worsen the optimization performance of subsequent points via (9).

Another approach, as recommended in Bull (2011), is to encourage further exploration via an artificial inflation of the MLE for the variance parameter. In particular, in place of the MLE $\hat{\sigma}_n^2$, one instead uses the inflated estimator $n\hat{\sigma}_n^2$ within the estimated EI criterion (8). Along with an ϵ-greedy modification, where points are added uniformly-at-random with probability ϵ at each iteration, Bull (2011) proved this modified EI approach (which we call the ϵ-EI) can indeed achieve convergence and a near-minimax convergence rate for global optimization, thus, addressing the earlier limitation from Proposition 1. Despite its appealing theoretical properties, the ϵ-EI may yield a suboptimal tradeoff between exploration and exploitation, particularly with limited function evaluations. This may be due to the use of an artificially inflated variance parameter (which is ad-hoc from a modeling perspective), or the use of uniformly-random points within ϵ-greedy sampling (which may be inefficient). This then results in suboptimal optimization performance given limited function evaluations, which we show later in numerical experiments.

3. Hierarchical Expected Improvement

To address the aforementioned limitations, we present a new Hierarchical EI (HEI) method, which integrates parameter uncertainty within a *closed-form* acquisition function, thus, allowing for efficient optimization of sequential queries. In employing a hierarchical Bayesian framework for expected improvement, the HEI can be shown to enjoy similar theoretical convergence guarantees as the ϵ-EI in Bull (2011) for global optimization, thus, addressing the limitation from Proposition 1 for the standard EI. In contrast to the ϵ-EI, however, the HEI achieves this purely via a hierarchical Bayesian model, without a need for an ad-hoc variance inflation or ϵ-greedy sampling. As such, the HEI can provide an improved tradeoff between exploration and exploitation in practice, which then translates to better optimization performance given limited function evaluations, as we show later.

A key ingredient for the HEI is a *hierarchical* GP model on $f(\mathbf{x})$. Let us adopt the universal kriging model (2), but now with the following hierarchical priors assigned on two independent parameters $(\boldsymbol{\beta}, \sigma^2)$:

$$[\boldsymbol{\beta}] \propto \mathbf{1}, \qquad [\sigma^2] \sim \text{IG}(a, b). \tag{10}$$

In words, the coefficients $\boldsymbol{\beta}$ are assigned a flat improper (i.e., noninformative) prior over $\mathbb{R}^q$, and the process variance σ^2 is assigned a conjugate inverse-Gamma prior with shape and scale parameters a and b, respectively. The idea is to leverage this hierarchical structure on model parameters to account for estimation uncertainty, while preserving a closed-form criterion for efficient sequential sampling. In what follows, we first introduce the HEI for *fixed* GP length-scale parameters $\boldsymbol{\theta}$ for ease of exposition; Section 4.3 then presents the full HEI procedure with *estimated* length-scales, with corresponding theoretical analysis in Section 5.

The next lemma provides the posterior distribution of $f(\mathbf{x})$ under this hierarchical model:

Lemma 1. Assume the universal kriging model (2) with hierarchical priors (10) and $n > q$. Given data $\mathcal{D}_n$, we have

$$[\sigma^2|\mathcal{D}_n] \sim \text{IG}(a_n, b_n) \quad \text{and} \quad [\boldsymbol{\beta}|\mathcal{D}_n] \sim T_q(2a_n, \hat{\boldsymbol{\beta}}_n, \tilde{\sigma}_n^2 \mathbf{G}_n^{-1}), \tag{11}$$

where $a_n = a + (n-q)/2$, $b_n = b + n\hat{\sigma}_n^2/2$, $\tilde{\sigma}_n^2 = b_n/a_n$, and $T_q(v, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a q-dimensional non-standardized t-distribution with degrees of freedom v, location vector $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$. Furthermore, the posterior distribution of $f(\mathbf{x})$ is

$$[f(\mathbf{x})|\mathcal{D}_n] \sim T_1(2a + n - q, \hat{f}_n(\mathbf{x}), \tilde{\sigma}_n s_n(\mathbf{x})). \tag{12}$$

The proof of this lemma follows from chap. 4.4 of Santner et al. (2018). Lemma 1 shows that under the universal kriging model (2) with hierarchical priors (10), the posterior distribution of $f(\mathbf{x})$ is now a non-standarized t-distribution, with closed-form expressions for its location and scale parameters $\hat{f}_n(\mathbf{x})$ and $\tilde{\sigma}_n s_n(\mathbf{x})$.

Comparing the predictive distributions in (3) and (12), there are several differences which highlight the increased predictive uncertainty from the hierarchical GP model. First, the new posterior (12) is now t-distributed, whereas the earlier posterior (3) is normally distributed, which suggests that the hierarchical model imposes heavier tails. Second, the scale term $\tilde{\sigma}_n^2$ in (12) can be decomposed as

$$\tilde{\sigma}_n^2 = (2b + n\hat{\sigma}_n^2)/(2a + (n-q)) > n/(2a + (n-q)) \cdot \hat{\sigma}_n^2. \tag{13}$$

When $a < q/2$ (which is satisfied via a weakly informative prior on σ^2), $\tilde{\sigma}_n^2$ is larger than the MLE $\hat{\sigma}_n^2$, which again increases predictive uncertainty.

Similar to the EI criterion (6), we now define the HEI acquisition function as

$$\text{HEI}_n(\mathbf{x}) = \mathbb{E}_{f|\mathcal{D}_n}(y_n^* - f(\mathbf{x}))_+, \tag{14}$$

where the conditional expectation over $[f(\mathbf{x})|\mathcal{D}_n]$ is under the hierarchical GP model. The proposition below gives a *closed-form* expression for $\text{HEI}_n(\mathbf{x})$:

Proposition 2. Assume the universal kriging model (2) with hierarchical priors (10) and $n > q$. Then:

$$\text{HEI}_n(\mathbf{x}) = \underbrace{I_n(\mathbf{x})\Phi_{v_n}\left(\frac{I_n(\mathbf{x})}{\tilde{\sigma}_n s_n(\mathbf{x})}\right)}_{\textbf{Exploitation}} \tag{15}$$

$$+ \underbrace{m_n \tilde{\sigma}_n s_n(\mathbf{x})\phi_{v_n-2}\left(\frac{I_n(\mathbf{x})}{m_n \tilde{\sigma}_n s_n(\mathbf{x})}\right)}_{\textbf{Exploration}},$$

where $m_n = \sqrt{v_n/(v_n - 2)}$, $v_n = 2a_n$, and $\phi_{v_n}(x)$, $\Phi_{v_n}(x)$ denote the pdf and cdf of a Student's t-distribution with v_n degrees of freedom, respectively.

Proposition 2 shows that the HEI criterion preserves the desirable properties of original EI criterion (8): it has an easily-computable, closed-form expression, which allows for efficient optimization of the next query point. This HEI criterion also has an equally interpretable exploration-exploitation tradeoff. Similar to the EI criterion, the first term encourages exploitation near the current best solution $\mathbf{x}_n^*$, and the second term encourages exploration of regions with high predictive variance.

More importantly, the differences between the HEI (15) and the EI (8) acquisition functions highlight how our approach addresses the over-greediness issue. There are three notable differences. First, the HEI exploration term depends on the t-pdf ϕ_{ν_n-2}, whereas the EI exploration term depends on the normal pdf ϕ. Since the former has heavier tails, the HEI exploration term is inflated, which encourages exploration. Second, the larger scale term $\tilde{\sigma}_n^2$ (see (13)) also inflates the HEI exploration term and encourages exploration. Third, the HEI contains an additional adjustment factor $\sqrt{\nu_n/(\nu_n - 2)}$ in its exploration term. Since this factor is larger than 1, HEI again encourages exploration. This adjustment is most prominent for small sample sizes, since the factor $\sqrt{\nu_n/(\nu_n - 2)} \to 1$ as sample size $n \to \infty$. All three differences *correct* the over-exploitation of EI via a principled hierarchical Bayesian framework. We will show later that these modifications for the HEI address the aforementioned theoretical and empirical limitations of the EI.

Finally, we note that the Student EI, proposed by Benassi, Bect, and Vazquez (2011), can be viewed as a special case of the HEI criterion, with (i) a constant mean function $\mu(\mathbf{x}) = \mu$, and (ii) "fixed" hyperparameters a and b (in that it does not scale with sample size n) for the inverse-Gamma prior in (10). We will show later that the HEI, by generalizing (i) and (ii), can yield improved theoretical and empirical performance over the SEI. For (i), instead of the *stationary* GP model used in the SEI, the HEI instead considers a broader *nonstationary* GP model with mean function $\mu(\mathbf{x}) = \mathbf{p}^\top(\mathbf{x})\boldsymbol{\beta}$, and factors in the uncertainty on coefficients $\boldsymbol{\beta}$ for optimization. This allows HEI to integrate uncertainty on GP nonstationarity to encourage more exploration in sequential sampling. For (ii), Benassi, Bect, and Vazquez (2011) recommended a "fixed" hyperparameter setting for the SEI, where the hyperparameters a and b do not scale with sample size n. However, the following proposition shows that the SEI (under such a setting) can fail to find the global optimum, under mild regularity conditions (see Assumptions 1 and 2 later).

Proposition 3. Suppose Assumptions 1 and 2 hold with $\nu < \infty$. Suppose initial points are sampled according to some probability measure F over Ω. Given fixed hyperparameters a and b, let $(\mathbf{x}_i)_{i=1}^\infty$ be the points returned by the SEI procedure. Then, for any $\epsilon > 0$, there exist some $f \in \mathcal{H}_{\boldsymbol{\theta}}(\Omega)$ and some constant $\delta > 0$ such that

$$\mathbb{P}_F\left(\lim_{n\to\infty} y_n^* - \min_{\mathbf{x}\in\Omega} f(\mathbf{x}) \geq \delta\right) > 1 - \epsilon.$$

The proof is provided in Appendix A.2, supplementary material. This proposition shows that the SEI with "fixed" hyperparameters has the same limitation as the EI: it can fail to converge to a global minimum for relatively smooth objective functions f. We will show later in Section 5 that, under a more general prior

specification which allows the hyperparameter b to scale with sample size n (more specifically, $b = \Theta(n)$), the HEI not only has the desired global convergence property for optimization, but also a near-minimax convergence rate.

4. Methodology and Algorithm

Using the HEI acquisition function (15), we now present a methodology for integrating this for effective black-box optimization. We first introduce hyperparameter estimation techniques which allow the HEI to mimic a fully Bayesian optimization procedure, then present an algorithmic framework for implementing the HEI. For ease of exposition, Sections 4.1 and 4.2 are presented with *fixed* GP length-scale parameters $\boldsymbol{\theta}$; Section 4.3 then presents the full algorithm with *estimated* length-scales, with corresponding theoretical analysis in Section 5.

4.1. Hyperparameter Specification

We present below several plausible specifications for the hyperparameters (a, b) in the hierarchical prior $[\sigma^2] \sim \text{IG}(a, b)$ in (10), and discuss when certain specifications may yield better optimization performance.

(i) Weakly Informative. Consider first a *weakly informative* specification of the hyperparameters (a, b), with $a = b = \epsilon$ for a small choice of ϵ, for example, $\epsilon = 0.1$. This reflects weak information on the variance parameter σ^2, and provides regularization for parameter inference. The limiting case of $\epsilon \to 0$ yields the non-informative Jeffreys' prior for variance parameters.

While weakly informative (and noninformative) priors are widely used in Bayesian analysis (Gelman 2006), we have found that such priors can result in poor optimization performance for HEI (Section 6 provides further details). One reason is that, for many black-box problems, only a small sample size can be afforded on the objective function f, since each evaluation is expensive. One can perhaps address this with a carefully elicited subjective prior, but such informative priors are typically not available when the objective f is black-box. We present next two specifications which may offer improved optimization performance, both in theory (Section 5) and in practice (Sections 6 and 7).

(ii) Empirical Bayes. Consider next an empirical Bayes (EB, Carlin and Louis 2000) approach, which uses the observed data on f to estimate the hyperparameters (a, b). This is achieved by maximizing the following marginal likelihood for (a, b):

$$p(\mathbf{y}_n; a, b) = \int \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_n)\pi(\boldsymbol{\beta})\pi(\sigma^2; a, b) \, d\boldsymbol{\beta} d\sigma^2. \quad (16)$$

Here, $\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_n)$ is the likelihood function of the universal kriging model (2) (see Santner et al. 2018 for the full expression), and $\pi(\boldsymbol{\beta})$ and $\pi(\sigma^2; a, b)$ are the prior densities of $\boldsymbol{\beta}$ and σ^2 given hyperparameters a and b. The model with estimated hyperparameters via EB provides a close approximation to a fully hierarchical Bayesian model (Carlin and Louis 2000), where additional hyperpriors are assigned on a and b. The latter

can be viewed as a "gold standard" quantification of model uncertainty, but typically requires MCMC sampling, which can be more expensive than optimization. Here, an EB estimate of hyperparameters (a, b) would allow the HEI to closely *mimic* a fully Bayesian optimization procedure (the "gold standard"), while avoiding expensive MCMC sampling via a *closed-form* acquisition function.

Unfortunately, the proposition below shows that a direct application of EB for the HEI yields unbounded hyperparameter estimates:

Proposition 4. The marginal likelihood for the universal kriging model (2) with priors (10) is given by:

$$p(\mathbf{y}_n; a, b) = \det(\mathbf{G}_n \mathbf{K}_n)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + (n-q)/2)}{(b + w_n)^{a + \frac{n-q}{2}}}, \quad (17)$$

where $w_n = (\mathbf{y}_n^\top \mathbf{K}_n^{-1} \mathbf{y}_n - \hat{\boldsymbol{\beta}}_n^\top \mathbf{G}_n \hat{\boldsymbol{\beta}}_n)/2$. Furthermore, the maximization problem:

$$\underset{a>0, b>0}{\operatorname{argmax}} \; p(\mathbf{y}_n; a, b) \quad (18)$$

is unbounded for all values of $\mathbf{y}_n$.

The proof of Proposition 4 is provided in Appendix A.3.

To address this issue of unboundedness, one can instead use a modified EB approach, called the marginal maximum a posteriori estimator (MMAP, Doucet, Godsill, and Robert 2002). The MMAP adds an additional level of hyperpriors $\pi(a, b)$ to the marginal likelihood maximization problem, yielding the modified formulation:

$$\underset{a>0, b>0}{\operatorname{argmax}} \; \tilde{p}(\mathbf{y}_n; a, b) := \underset{a>0, b>0}{\operatorname{argmax}} \; p(\mathbf{y}_n; a, b) \pi(a, b). \quad (19)$$

The MMAP approach for hyperparameter specification has been used in a variety of problems, for example, scalable training of large-scale Bayesian networks (Liu and Ihler 2013). The next proposition shows that the MMAP indeed yields finite solutions for a general class of hyperpriors on (a, b):

Proposition 5. Assume the following independent hyperpriors on (a, b):

$$[a] \sim \operatorname{Gamma}(\zeta, \iota), \qquad [b] \propto \mathbf{1}, \quad (20)$$

where ζ and ι are the shape and scale parameters, respectively. Then the maximization of $\tilde{p}(\mathbf{y}_n; a, b)$ is always finite for $(a, b) \in \mathbb{R}_+^2$.

The proof of Proposition 5 is provided in Appendix A.4.

In practice, we recommend a weakly-informative specification of the hyperparameters (ζ, ι) (i.e., with $\zeta = \iota$ set to be small), which appears to yield robust optimization performance for the HEI-MMAP. We note that the specification (20) is simply one which works well in our implementation; given prior knowledge, a modeler has the flexibility of specifying an alternate prior which captures such information. By mimicking a fully Bayesian optimization procedure, this MMAP approach can consistently outperform the weakly informative specification for HEI; we will show this later in numerical studies.

(iii) Data-Size-Dependent (DSD). Finally, we consider the so-called "data-size-dependent" (DSD) hyperparameter specification. This is motivated from the prior specification needed for global optimization convergence of the HEI, which we present and justify in the following section. The DSD specification requires the shape parameter a to be constant, and the scale parameter b to grow at the same order as the sample size n, that is, $b = \kappa n$ for some constant $\kappa > 0$. One appealing property of this specification is that it ensures the HEI converges to a global optimum $\mathbf{x}^*$ (see Theorem 1 later).

For the DSD specification, we can similarly use the MMAP to estimate hyperparameters (a, κ), to mimic a fully Bayesian EI procedure. Suppose data $\mathcal{D}_{n_{\text{ini}}}$ are collected from n_{ini} initial design points (more on this in Section 4.3). Then the hyperparameters a and κ can be estimated via the MMAP optimization:

$$(a^*, \kappa^*) = \underset{a>0, \kappa>0}{\operatorname{argmax}} \left\{ p(\mathbf{y}_{n_{\text{ini}}}; a, \kappa n_{\text{ini}}) \pi(a, \kappa) \right\}, \quad (21)$$

where $\pi(a, \kappa)$ is the hyperprior density on a and κ. One possible setting for $\pi(a, \kappa)$ is a Gamma hyperprior on a and a noninformative hyperprior $[\kappa] \propto 1$ (independent of a). By Proposition 5, this specification again yields a finite optimization problem for MMAP. Using these estimated hyperparameters, subsequent points are then queried using HEI with $a = a^*$ and $b = \kappa^* n$, where n is the current sample size.

4.2. Order Selection for Basis Functions

In addition to hyperparameter estimation, the choice of basis functions in $\mathbf{p}(\mathbf{x})$ and the order selection of such bases are also important for an effective implementation of the HEI. In our experiments, we take these bases to be complete polynomials up to a certain order l. Letting $\mathcal{M}^{(l)}$ denote the polynomial model with maximum order l, we have $\mathbf{p}(\mathbf{x}) = 1$ for model $\mathcal{M}^{(0)}$ (a constant model), $\mathbf{p}(\mathbf{x}) = [1, x_1, \ldots, x_d]^\top$ for model $\mathcal{M}^{(1)}$ (a linear model), $\mathbf{p}(\mathbf{x}) = [1, x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, x_1 x_2, \ldots, x_1 x_d, x_2 x_3, \ldots, x_{d-1} x_d]^\top$ for model $\mathcal{M}^{(2)}$ (a second-order interaction model), etc. One can also make use of other basis functions (e.g., orthogonal polynomials; Xiu 2010) depending on the problem at hand.

A careful selection of order l is also important: an overly small estimate of l results in over-exploitation of a poorly-fit model, whereas an overly large estimate results in variance inflation and over-exploration of the domain. We found that the standard Bayesian Information Criterion (BIC) (Schwarz 1978) provides good order selection performance for the HEI. Given initial data $\mathcal{D}_{n_{\text{ini}}}$, the BIC selects the model $\mathcal{M}^{(l^*)}$ with order:

$$l^* = \underset{l \in \mathbb{N}}{\operatorname{argmin}} \left\{ -2 \log \mathcal{L}(\mathcal{M}^{(l)}) + q_l \log(n_{\text{ini}}) \right\}. \quad (22)$$

Here, $\mathcal{L}(\mathcal{M}^{(l)})$ denotes the likelihood of model $\mathcal{M}^{(l)}$ (this likelihood expression can be found in Santner et al. 2018), and q_l denotes the number of basis functions in model $\mathcal{M}^{(l)}$. With this optimal order selected, subsequent samples are then obtained using HEI with mean function following this polynomial order.

Algorithm 1 Hierarchical Expected Improvement for Bayesian Optimization

---

Initialization

- Generate n_{ini} space-filling design points $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n_{\text{ini}}}\}$ on Ω.
- Evaluate function points $y_i = f(\mathbf{x}_i)$, yielding the initial dataset $\mathcal{D}_{n_{\text{ini}}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{ini}}}$.

Model selection

- Select model order via BIC using (22).
- Estimate hyperparameters (a, b) via MMAP using (19).

Optimization
for $n \leftarrow n_{\text{ini}}$ *to* $n_{\text{tot}} - 1$ **do**

- Given $\mathcal{D}_n$, estimate length-scale parameters $\boldsymbol{\theta}$ via MAP and compute $\text{HEI}_n(\mathbf{x})$.
- Obtain the next evaluation point $\mathbf{x}_{n+1}$ by maximizing $\text{HEI}_n(\mathbf{x})$:

$$\mathbf{x}_{n+1} \leftarrow \underset{\mathbf{x} \in \Omega}{\arg\max}\, \text{HEI}_n(\mathbf{x}). \qquad (23)$$

- Evaluate $y_{n+1} = f(\mathbf{x}_{n+1})$, and update data $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{(\mathbf{x}_{n+1}, y_{n+1})\}$.

Return: The best observed solution $\mathbf{x}_{i^*}$, where $i^* = \arg\min_{i=1}^{n_{\text{tot}}} f(\mathbf{x}_i)$.

4.3. Algorithm Statement

Algorithm 1 summarizes the above steps for HEI. First, initial data on the black-box function f are collected on a "space-filling" design, which provides good coverage of the feasible space Ω. For the unit hypercube $\Omega = [0, 1]^d$, we have found that the maximin Latin hypercube design (MmLHD, Morris and Mitchell 1995) works quite well in practice. For non-hypercube domains, more elaborate design methods on non-hypercube regions (e.g., Lekivetz and Jones 2015; Mak and Joseph 2018; Joseph, Gul, and Ba 2019) can be used. The number of initial points is set as $n_{\text{ini}} = 10d$, as recommended in Loeppky, Sacks, and Welch (2009). Using this initial data, the model order for the hierarchical GP is selected using (22). The hyperparameters a and b are also estimated from data (if necessary) using the methods described in Section 4.1. Next, the following two steps are repeated until the sample size budget n_{tot} is exhausted: (i) the GP length-scale parameters $\boldsymbol{\theta}$ are fitted via maximum a posteriori (MAP) estimation[1] using the observed data points, (ii) a new sample $f(\mathbf{x})$ is collected at the point $\mathbf{x}$ which maximizes the HEI criterion (15).

5. Convergence Analysis

We present next the global optimization convergence result for the HEI, then provide a near-minimax optimal convergence rate

---
[1] In numerical experiments, we use an independent uniform prior $\theta_l \overset{\text{iid}}{\sim} U[0, 100]$ for this MAP estimate.

for the proposed method. In what follows, we will assume that the domain Ω is convex and compact.

Let us first adopt the following shift-invariant form for the kernel K:

$$K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) := C\left(\frac{x_1 - z_1}{\theta_1}, \ldots, \frac{x_d - z_d}{\theta_d}\right), \qquad (24)$$

where C is a stationary correlation function with $C(\mathbf{0}) = 1$ and length-scale parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. From this, we can then define a function space—the reproducing kernel Hilbert space (RKHS, Wendland 2004)—for the objective function f. Given kernel $K_{\boldsymbol{\theta}}$ (which is symmetric and positive definite), define the linear space

$$\mathcal{F}_{\boldsymbol{\theta}}(\Omega) = \left\{ \sum_{i=1}^N \alpha_i K_{\boldsymbol{\theta}}(\cdot, \mathbf{x}_i) : N \in \mathbb{N}_+, \mathbf{x}_i \in \Omega, \alpha_i \in \mathbb{R} \right\}, \quad (25)$$

and equip this space with the bilinear form

$$\left\langle \sum_{i=1}^N \alpha_i K_{\boldsymbol{\theta}}(\cdot, \mathbf{x}_i), \sum_{j=1}^M \gamma_j K_{\boldsymbol{\theta}}(\cdot, \mathbf{y}_j) \right\rangle_{K_{\boldsymbol{\theta}}} := \sum_{i=1}^N \sum_{j=1}^M \alpha_i \gamma_j K_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_j).$$
$$(26)$$

The RKHS $\mathcal{H}_{\boldsymbol{\theta}}(\Omega)$ of kernel $K_{\boldsymbol{\theta}}$ is defined as the closure of $\mathcal{F}_{\boldsymbol{\theta}}(\Omega)$ under $\langle \cdot, \cdot \rangle_{K_{\boldsymbol{\theta}}}$, with its inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\boldsymbol{\theta}}}$ induced by $\langle \cdot, \cdot \rangle_{K_{\boldsymbol{\theta}}}$ (Wendland 2004).

Next, we make the following two regularity assumptions. The first is a smoothness assumption on the correlation function C:

Assumption 1. C is continuous, integrable, and satisfies:

$$|C(\mathbf{x}) - Q_r(\mathbf{x})| = \mathcal{O}\left(\|\mathbf{x}\|_2^{2\nu}(-\log\|\mathbf{x}\|_2)^{2\alpha}\right) \quad \text{as} \quad \|\mathbf{x}\|_2 \to 0,$$

for some constants $\nu > 0$ and $\alpha \geq 0$. Here, $r = \lfloor 2\nu \rfloor$ and $Q_r(\mathbf{x})$ is the rth order Taylor approximation of $C(\mathbf{x})$. Furthermore, its Fourier transform

$$\widehat{C}(\boldsymbol{\xi}) := \int_{\mathbf{x}} e^{-2\pi i \langle \boldsymbol{\xi}, \mathbf{x} \rangle} C(\mathbf{x}) d\mathbf{x}$$

is isotropic, radially nonincreasing and satisfies either: as $\|\mathbf{x}\|_2 \to \infty$

$$\widehat{C}(\mathbf{x}) = \Theta\left(\|\mathbf{x}\|_2^{-2\nu-d}\right) \quad \text{or} \quad \widehat{C}(\mathbf{x}) = \mathcal{O}\left(\|\mathbf{x}\|_2^{-2\lambda-d}\right) \text{ for}$$
any $\lambda > 0$.

Note that in Assumption 1, since we assume C is continuous and integrable, its Fourier transform $\hat{C}$ must exists. A widely-used correlation function which satisfies this assumption is the Matérn correlation function (see Cressie 1991 and Bull 2011).

The second assumption is a bounded assumption on the prior $\pi(\boldsymbol{\theta})$ for the GP length-scale parameters $\boldsymbol{\theta}$.

Assumption 2. We assume that the prior $\pi(\boldsymbol{\theta})$ is bounded and bounded away from $\mathbf{0}$. Thus, given data $\mathcal{D}_n$, if we let $\tilde{\boldsymbol{\theta}}_n$ be the MAP of $\boldsymbol{\theta}$ under prior $\pi(\boldsymbol{\theta})$, it follows that for any $n > q$:

$$\boldsymbol{\theta}^L \leq \tilde{\boldsymbol{\theta}}_n \leq \boldsymbol{\theta}^U \quad \text{for some constants } \boldsymbol{\theta}^L, \boldsymbol{\theta}^U \in \mathbb{R}_+^d. \qquad (27)$$

Note that the above equation indicates component-wise inequalities. This is similar to Definition 2 in Bull (2011).

Under these two regularity assumptions, we can then prove the global optimization convergence of the HEI method (more specifically, for HEI-DSD).

Theorem 1. Suppose Assumptions 1 and 2 hold. Further suppose the hyperparameter a is a constant (in n) and $b = \Theta(n)$, with basis functions $p_i(x) \in \mathcal{H}_{\theta^U}(\Omega)$. Let $(\mathbf{x}_i)_{i=1}^{\infty}$ be the points generated by maximizing HEI_n in (15), with iterative plug-in MAP estimates $\tilde{\boldsymbol{\theta}}_n$. Then, for any $f \in \mathcal{H}_{\theta^U}(\Omega)$ and any choice of initial points $\{\mathbf{x}_i\}_{i=1}^{n_{\text{ini}}}$, we have:

$$y_n^* - \min_{\mathbf{x} \in \Omega} f(\mathbf{x}) = \begin{cases} \mathcal{O}(n^{-\nu/d}(\log n)^{\alpha}), & \nu \leq 1, \\ \mathcal{O}(n^{-1/d}), & \nu > 1. \end{cases} \quad (28)$$

The proof of this theorem is given in Appendix A.5, where its dependence on other parameters (e.g., $\boldsymbol{\theta}^L$ and $\boldsymbol{\theta}^U$) are made explicit. The key idea is to upper bound the prediction gap $f(\mathbf{x}) - \hat{f}_n(\mathbf{x})$ by the posterior variance term $s_n^2(\mathbf{x})$ in (5), which is a generalization of the power function used in the function approximation literature (see, e.g., Theorem 11.4 of Wendland 2004). We then show that the hyperparameter assumption $b = \Theta(n)$ prevents the estimator $\tilde{\sigma}_n^2$ from collapsing to 0, and allows us to apply approximation bounds on $s_n^2(\mathbf{x})$ to obtain the desired global convergence result. This proof is inspired by Theorem 4 of Bull (2011).

Theorem 1 shows that, for all objective functions f in the RKHS $\mathcal{H}_{\theta^U}(\Omega)$, the HEI indeed has the desired convergence property for global optimization. This addresses the lack of convergence for the EI from Proposition 1 and the SEI from Proposition 3. It is worth nothing that, when C is the Matérn correlation with smoothness parameter ν (Cressie 1991), the RKHS $\mathcal{H}_{\theta}(\Omega)$ consists of functions f with continuous derivatives of order $\nu' < \nu$ (Santner et al. 2018). Hence, for the Matérn correlation, the HEI achieves a global convergence rate of $\mathcal{O}(n^{-1/d})$ for objective functions $f \in \mathcal{H}_{\theta}(\Omega)$ with $\nu > 1$, and $\mathcal{O}(n^{-\nu/d}(\log n)^{\alpha})$ for objective functions $f \in \mathcal{H}_{\theta^U}(\Omega)$ with $\nu \leq 1$.

At first glance, the prior specification in Theorem 1 may appear slightly peculiar, since the hyperparameter $b = \Theta(n)$ depends on the sample size n. However, such *data-size-dependent* priors have been studied extensively in the context of high-dimensional Bayesian linear regression, particularly in its connection to optimal minimax estimation (see, e.g., Castillo, Schmidt-Hieber, and Van der Vaart 2015). The data-size-dependent prior in Theorem 1 can be interpreted in a similar way: the hyperparameter condition $b = \Theta(n)$ is sufficient in encouraging exploration in the sequential sampling points, so that HEI converges to a global optimum for all f in the RKHS $\mathcal{H}_{\theta^U}(\Omega)$.

If an additional γ-stability condition (Wynne, Briol, and Girolami 2020) holds, we can further show that the HEI achieves the minimax convergence rate for Bayesian optimization (we provide discussion on this minimax rate at the end of the section). This condition is stated below:

Condition 1. Let $(\mathbf{x}_i)_{i=1}^{\infty}$ be the sequence of points generated by the HEI. We assume that

$$s_n(\mathbf{x}_{n+1}) \geq \gamma \|s_n(\mathbf{x})\|_{\infty} \quad \text{for all } n = 1, 2, \ldots, \quad (29)$$

for some constant $\gamma \in (0, 1]$.

In words, this requires that every sequential point $\mathbf{x}_{n+1}$ has a posterior standard deviation term $s_n(\mathbf{x}_{n+1})$ (from (5)) at least

as large as $\gamma \|s_n(\mathbf{x})\|_{\infty}$, where $\|s_n(\mathbf{x})\|_{\infty}$ is the maximum posterior standard deviation term over domain Ω. If this condition holds, we can then show a quicker convergence rate for global optimization:

Theorem 2. Suppose Assumptions 1 and 2 hold, along with Condition 1. Further suppose the hyperparameter a is a constant (in n) and $b = \Theta(n)$, with basis functions $p_i(x) \in \mathcal{H}_{\theta^U}(\Omega)$. Let $(\mathbf{x}_i)_{i=1}^{\infty}$ be the points generated by maximizing HEI_n in (15) with iterative plug-in MAP estimates $\tilde{\boldsymbol{\theta}}_n$. Then, for any $f \in \mathcal{H}_{\theta^U}(\Omega)$ and any initial points, we have:

$$y_n^* - \min_{\mathbf{x} \in \Omega} f(\mathbf{x}) = O(n^{-\nu/d}). \quad (30)$$

Further details on this theorem are provided in Appendix A.6.

Condition 1 is unfortunately quite difficult to guarantee theoretically, but appears to be satisfied empirically for the recommended HEI-DSD and HEI-MMAP methods in all of our later numerical experiments. Figure 1(a) shows the log-ratio $\log\{s_n(\mathbf{x}_{n+1})/\|s_n(\mathbf{x})\|_{\infty}\}$ for HEI-DSD using the Branin function (in our later simulation study). As can be seen, the log ratio appears to be bounded away from zero as n increases, which suggests that the HEI with data-size-dependent hyperparameter specification indeed satisfies Condition 1 for a sufficiently small $\gamma > 0$. This is not surprising, since while a sensible optimization algorithm would place more points around the optimum, the hierarchical nature of the HEI encourages further exploration, thus, ensuring the sequential points explore the domain from time to time such that no point has overly high predictive variance. The same γ-stability assumption was used in Wynne, Briol, and Girolami (2020) for proving convergence of existing BO methods.

It is crucial to note that, while these rates provides a reassuring check for HEI convergence, *such asymptotic analysis does not tell the full story on the effectiveness of a Bayesian optimization method.* Bull (2011) proved that, of all optimization strategies for minimizing $f \in \mathcal{H}_{\theta}(\Omega)$ under Assumption 1, the minimax rate for the optimization gap $y_n^* - \min_{\mathbf{x} \in \Omega} f(\mathbf{x})$ is $\mathcal{O}(n^{-\nu/d})$, that is, there does not exist an optimization strategy with a quicker asymptotic rate. The HEI rate in Theorem 2, in this sense, is precisely the minimax rate. However, Bull (2011) also showed that the simple (nonadaptive) strategy of optimization via a quasi-uniform sequence (see, e.g., Niederreiter 1992) can also achieve this minimax rate! Such a strategy, however, typically performs terribly in practice and is not competitive with existing BO methods (Bull 2011), since it is nonadaptive to observed function evaluations. This shows that *such asymptotic analysis, while providing a reassuring check, cannot be used as a sole metric for gauging the practical effectiveness of different methods, particularly given the current setting of limited sample sizes.*

We further note that analogous rates to Theorems 1 and 2 have been also shown for various Bayesian optimization methods in the literature. In particular, similar rates to Theorem 1 were proved for the ϵ-EI (Bull 2011), and our results leverage an adaptation of their proof techniques to show convergence for the HEI. Similarly, the γ-stability assumption (Condition 1) was used in Wynne, Briol, and Girolami (2020) to establish similar optimization rates as Theorem 2 for certain Bayesian optimization methods. The novelty for the HEI is thus not

(a) Log-ratio $\log \frac{s_n(\mathbf{x}_{n+1})}{\|s_n(\mathbf{x})\|_\infty}$ for HEI-DSD.

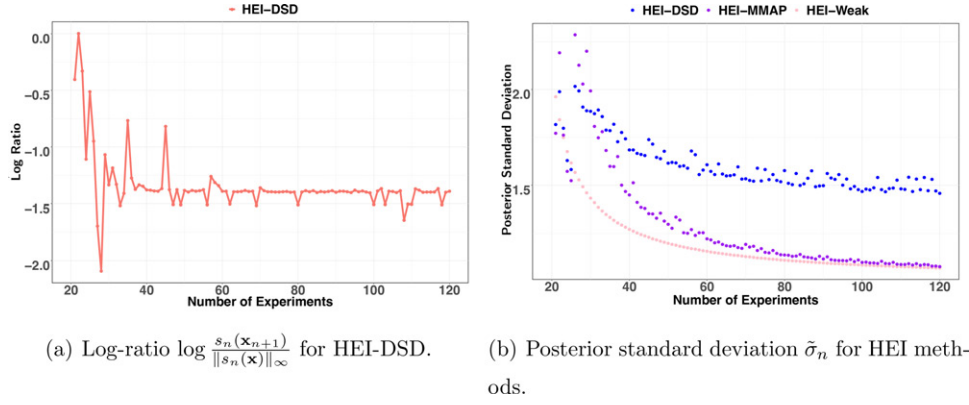(b) Posterior standard deviation $\tilde{\sigma}_n$ for HEI methods.

Figure 1. Visualizing select statistics for HEI methods on the Branin numerical experiment.

in terms of improved asymptotic rates over existing methods; such rates primarily serve to provide theoretical footing for our approach. Instead, the key contribution of the HEI is methodological: the proposed hierarchical framework provides a principled exploitation-exploration tradeoff via a closed-form acquisition function, which as we show later, allows for improved optimization performance with limited samples.

6. Numerical Experiments

We now investigate the numerical performance of HEI in comparison to existing BO methods, for a suite of test optimization functions. We consider the following five test functions, taken from Surjanovic and Bingham (2015):

- *Branin* (two-dimensional function on domain $\Omega = [0, 1]^2$):

$$f(\mathbf{x}) = (x_2 - 5.1/(4\pi^2) \cdot x_1^2 + 5/\pi \cdot x_1 - 6)^2$$
$$+ 10(1 - 1/(8\pi)) \cos(x_1) + 10,$$

- *Three-Hump Camel* (two-dimensional function on domain $\Omega = [-2, 2]^2$):

$$f(\mathbf{x}) = 2x_1^2 - 1.05x_1^4 + x_1^6/6 + x_1 x_2 + x_2^2,$$

- *Six-Hump Camel* (two-dimensional function on domain $\Omega = [-2, 2]^2$):

$$f(\mathbf{x}) = (4 - 2.1x_1^2 + x_1^4/3)x_1^2 + x_1 x_2 + (-4 + 4x_2^2)x_2^2,$$

- *Levy Function* (six-dimensional function on domain $\Omega = [-10, 10]^6$):

$$f(\mathbf{x}) = \sin^2(\pi\omega_1) + \sum_{i=1}^{5}(\omega_i - 1)^2[1 + 10\sin^2(\pi\omega_i + 1)]$$
$$+ (\omega_6 - 1)^2[1 + \sin^2(2\pi\omega_6)],$$

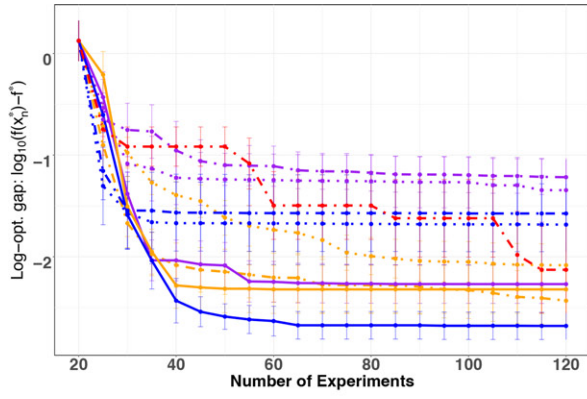where $\omega_i = 1 + (x_i - 1)/4$ for $i = 1, \ldots, 6$,

- *Ackley Function* (10-dimensional function on domain $\Omega = [-5, 5]^{10}$):

$$f(\mathbf{x}) = -20 \exp\left(-\frac{0.2}{\sqrt{10}}\|\mathbf{x}\|_2\right)$$
$$- \exp\left(\frac{1}{10}\sum_{i=1}^{10}\cos(2\pi x_i)\right) + 20 + \exp(1).$$
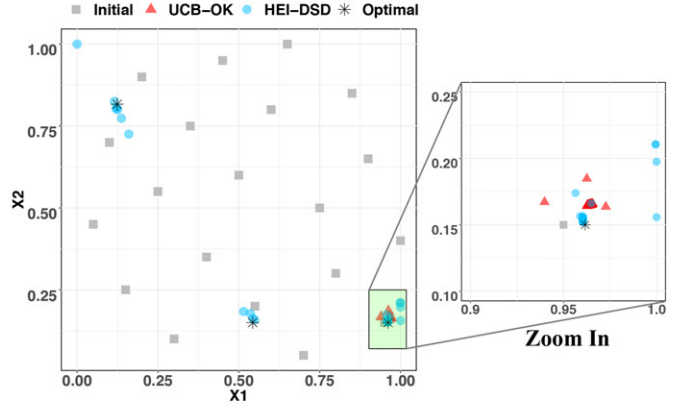
The simulation set-up is as follows. We compare the proposed HEI method under different hyperparameter specifications (HEI-Weak, HEI-MMAP, HEI-DSD), with the EI method under ordinary kriging (EI-OK) and universal kriging (EI-UK), the Student EI (SEI) method with fixed hyperparameters $(0.2, 12)$ as recommended in Benassi, Bect, and Vazquez (2011), the UCB approach under ordinary kriging (UCB-OK, Srinivas et al. 2010) with default exploration parameter 2.96, the ϵ-greedy EI approach (Bull 2011) under ordinary kriging (ϵ-EI-OK) and universal kriging (ϵ-EI-UK) with $\epsilon = 0.1$ as suggested in Sutton and Barto (2018), and the γ-stabilized EI method (Wynne, Briol, and Girolami 2020) under universal kriging (Stab-EI-UK) with $\gamma = \min(0.1d, 0.8)^2$. For HEI-Weak, the hyperparameters (a, b) are set as $a = b = 0.1$; for HEI-MMAP and HEI-DSD, the hyperparameters (ζ, ι) are set as $\zeta = \iota = 2$. All methods use the Matérn correlation with smoothness parameter 2.5, and are run for a total of $T = 120$ function evaluations. Here, the kriging model is fitted using the R package kergp (Deville, Ginsbourger, and Roustant 2019). As mentioned in Section 4.3, all methods are initialized using maximin Latin hypercube designs (Morris and Mitchell 1995). Simulation results are averaged over 20 replications except the Ackley function, due to the heavy computation burden for fitting the high-dimensional GP model.

Figure 2(a)–(d) and (f) show the log-optimality gap $\log_{10}(f(\mathbf{x}_n^*) - f(\mathbf{x}^*))$ against the number of samples n for the first three functions, and Figure 2(e) shows the optimality gap $f(\mathbf{x}_n^*) - f(\mathbf{x}^*)$ for the Levy function. We see that the three HEI methods outperform the existing Bayesian optimization methods: the optimality gap for the latter methods stagnates for larger sample sizes, whereas the former enjoys steady improvements as n increases. This shows that the proposed method indeed corrects the over-greediness of EI, and provides a more effective correction of this via hierarchical modeling, compared to the ϵ-greedy and Stab-EI-UK methods. Furthermore, of the HEI methods, HEI-MMAP and HEI-DSD appear to greatly outperform HEI-Weak. This is in line with the earlier observation that weakly informative priors may yield poor optimization for HEI; the MMAP and DSD specifications give better performance by mimicking a fully Bayesian optimization procedure. This is further supported by
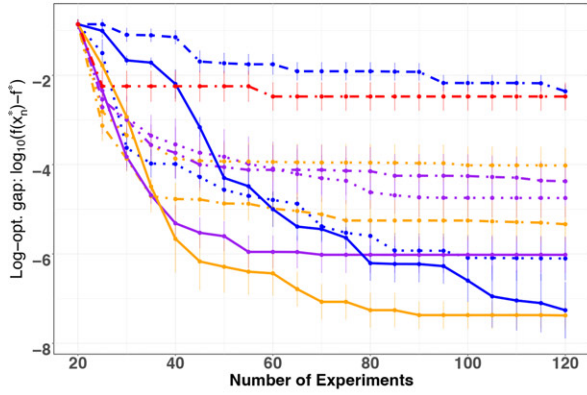
---

[2] Stab-EI-UK requires the next query point x_{n+1} to satisfy $s_n(x_{n+1}) \geq \gamma\|s_n(x)\|_\infty$. In our implementation, we randomly sample 10^{d+2} points to find $\|s_n(x)\|_\infty$ and set $\gamma = \min(0.1d, 0.8)$.
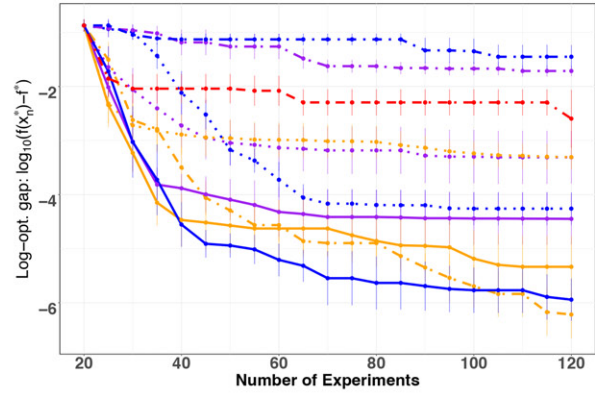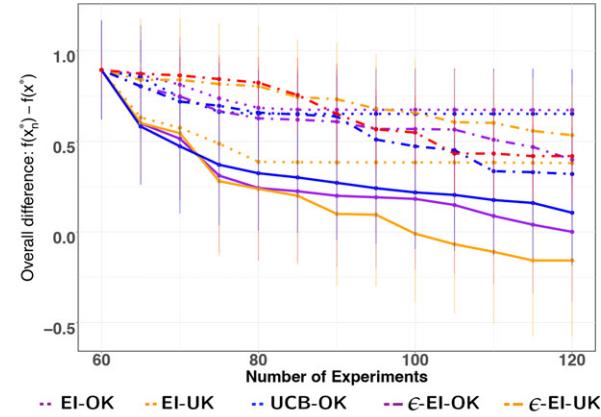
(a) Branin (2-d)
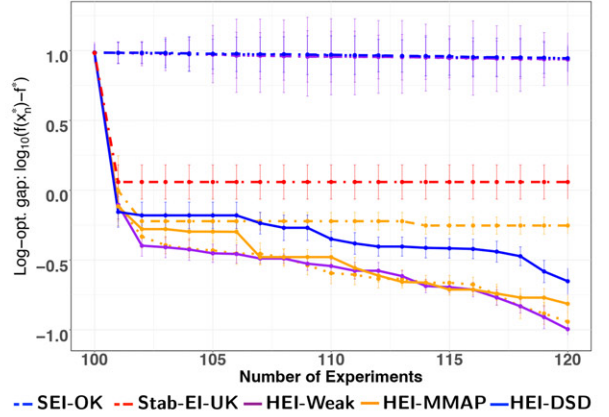
(b) A visualization of sampled points

(c) Three-Hump Camel (2-d)

(d) Six-Hump Camel (2-d)

(e) Levy (6-d)

(f) Ackley (10-d)

Figure 2. Numerical results for synthetic functions. (a) and (c)–(f) show the average optimality gap over 10 replications (dotted lines: EI-OK, EI-UK, and UCB-OK; dashed lines: ϵ-EI-OK, ϵ-EI-UK, SEI-OK, Stab-EI-UK; solid lines: HEI-Weak, HEI-MMAP, HEI-DSD). (b) presents a visualization of sampled points for the Branin function (grey squares: initial points, black stars: global optima, red triangles: UCB-OK points, blue circles: HEI-DSD points).

Figure 1(b), which shows the posterior estimate $\tilde{\sigma}_n$ of the scale parameter as a function of sample size n for the HEI methods in the Branin experiment. We see that both the HEI-DSD and HEI-MMAP provide larger estimates than HEI-DSD, which shows that the former methods are indeed integrating further

uncertainty for exploration. The steady improvement of HEI-DSD also supports the data-size-dependent prior condition needed for global convergence in Theorems 1 and 2.

Figure 2(b) shows the sampled points from HEI-DSD and UCB-OK for one run of the Branin function. The points for
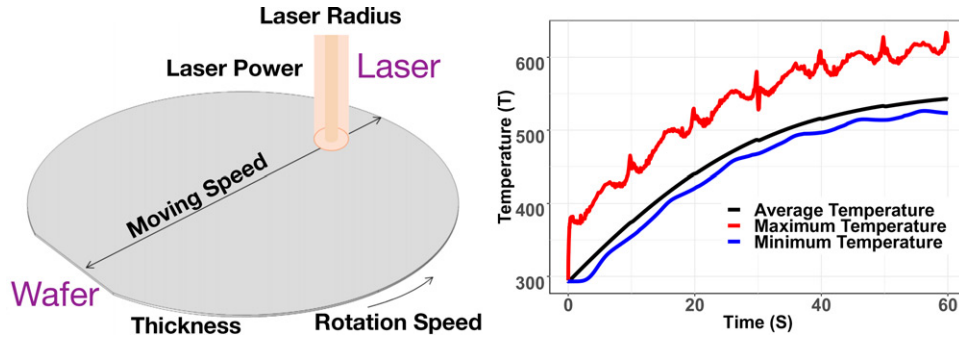
Figure 3. A visualization of the laser heating process of a silicon wafer (left), and the simulated temperature profile from COMSOL Multiphysics (right).

HEI-Weak and HEI-MMAP are quite similar to HEI-DSD, and the points for EI-OK, EI-UK, and SEI are quite similar to UCB-OK, so we only plot one of each for easy visualization. We see that HEI indeed encourages more exploration in optimization: it successfully finds *all* three global optima for f, whereas existing methods cluster points near only one optimum. The need to identify multiple global optima often arises in multiobjective optimization. For example, a company may wish to offer multiple product lines to suit different customer preferences (Mak and Wu 2019). For such problems, HEI can provide more practical solutions over existing methods.

Lastly, we compare the performance of HEI with the SEI method (Benassi, Bect, and Vazquez 2011). From Figure 2(e), we see that the SEI performs quite well for the Levy function: it is slightly worse than HEI methods, but better than the other methods. However, from Figure 2, the SEI achieves only comparable performance with EI-OK for the Branin function (which is in line with the results reported in Benassi, Bect, and Vazquez 2011), and is one of the worst-performing methods. This shows that the performance of SEI can vary greatly for different problems.

Finally, we note that for higher-dimensional problems, the HEI (and indeed, all GP-based Bayesian optimization methods) will likely require more sophisticated GP models to work well. In particular, such GP models should ideally be able to learn low-dimensional embeddings of the objective function over the high-dimensional domain; see, for example, recent work in Seshadri, Yuchi, and Parks (2019) and Zhang, Mak, and Dunson (2022). Integrating such low-dimensional structure within the HEI will be the topic of future work.

7. Semiconductor Manufacturing Optimization

We now investigate the performance of HEI in a process optimization problem in semiconductor wafer manufacturing. In semiconductor manufacturing (Jin, Chang, and Shi 2012), thin silicon wafers undergo a series of refinement stages. Of these, thermal processing is one of the most important stage, since it facilitates necessary chemical reactions and allows for surface oxidation (Singh, Fakhruddin, and Poole 2000). Figure 3(a) visualizes a typical thermal processing procedure: a laser beam is moved radially in and out across the wafer, while the wafer itself is rotated at a constant speed. There are two objectives here. First, the wafer should be heated to a target temperature to facilitate the desired chemical reactions. Second, temperature

Table 1. Design ranges of the five control parameters, where rpm (revolutions per minute) measures the rotation speed of the wafer.

Thickness	Rotation speed	Laser period	Laser radius	Power
$[160, 300]\mu$s	$[2, 50]$rpm	$[5, 15]$s	$[2, 10]$mm	$[10, 20]$W

fluctuations over the wafer surface should be made as small as possible, to reduce unwanted strains and improve wafer fabrication (Brunner et al. 2013). The goal is to find an "optimal" setting of the manufacturing process which achieves these two objectives.
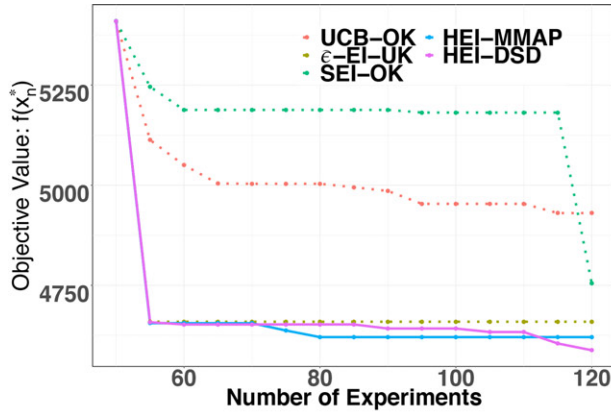
We consider five control parameters: wafer thickness, rotation speed, laser period, laser radius, and laser power (a full specification is given in Table 1). The heating is performed over 60 sec, and a target temperature of $\mathcal{T}^* = 600$ F is desired over this timeframe. We use the following objective function:

$$f(\mathbf{x}) := \sum_{t=1}^{60} \max_{\mathbf{s} \in \mathcal{S}} |\mathcal{T}_t(\mathbf{s}; \mathbf{x}) - \mathcal{T}^*|. \tag{31}$$
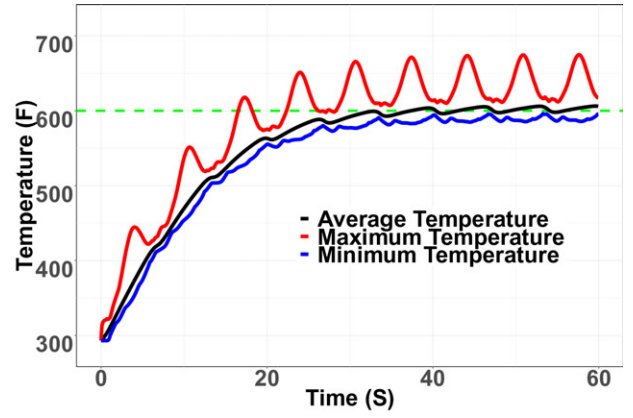
Here, $\mathbf{s}$ denotes a spatial location on the wafer domain $\mathcal{S}$, $t = 1, \dots, 60$ denotes the heating time (in seconds), and $\mathcal{T}_t(\mathbf{s}; \mathbf{x})$ denotes the wafer temperature at location $\mathbf{s}$ and time t, using control setting $\mathbf{x} \in \mathbb{R}^5$. Note that $f(\mathbf{x})$ captures both objectives of the study: wafer temperatures $\mathcal{T}_t$ close to $\mathcal{T}^*$ results in smaller values of $f(\mathbf{x})$, and the same is true when $\mathcal{T}_t(\mathbf{s}; \mathbf{x})$ is stable over $\mathbf{s} \in \mathcal{S}$.

Clearly, each evaluation of $f(\mathbf{x})$ is expensive, since it requires a full run of wafer heating process. We will simulate each run using COMSOL Multiphysics (COMSOL 2018), a reliable finite-element analysis software for solving complex systems of partial differential equations (PDEs). COMSOL models the incident heat flux from the moving laser as a spatially distributed heat source on the surface, then computes the transient thermal response by solving the coupled heat transfer and surface-to-ambient radiation PDEs. Figure 3(b) visualizes the simulation output from COMSOL: the average, maximum, and minimum temperature over the wafer domain at every time step. Experiments are performed on a desktop computer with quad-core Intel I7-8700K processors, and take around 5 min per run.
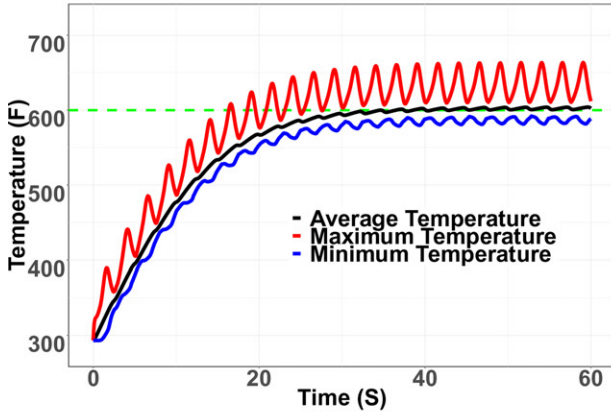
Figure 4(a) shows the best objective values $f(\mathbf{x}_n^*)$ for HEI-MMAP and HEI-DSD (the best performing HEI methods from simulations), and for the UCB-OK, SEI, and ϵ-greedy EI methods. We see that UCB-OK and SEI perform noticeably poorly, whereas the proposed HEI-MMAP and HEI-DSD methods provide the best optimization performance, with the ϵ-greedy-EI
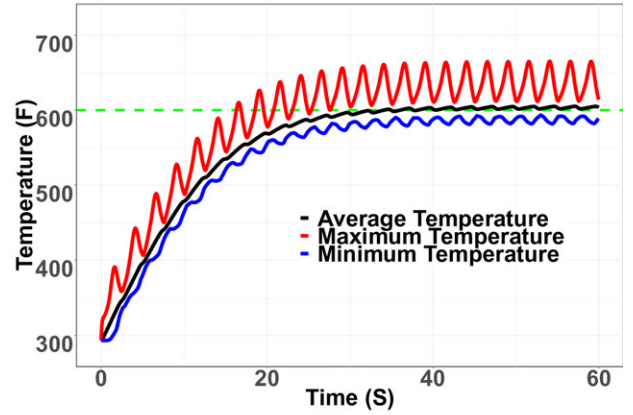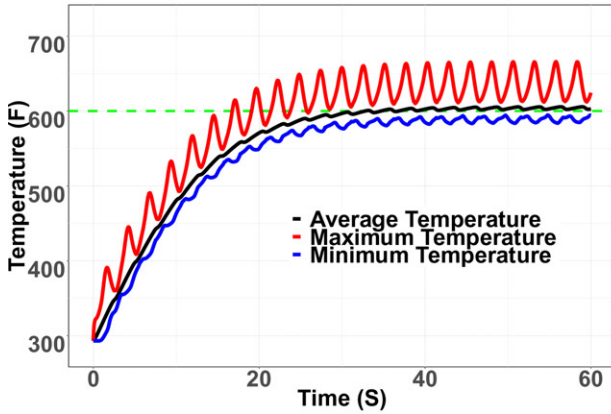
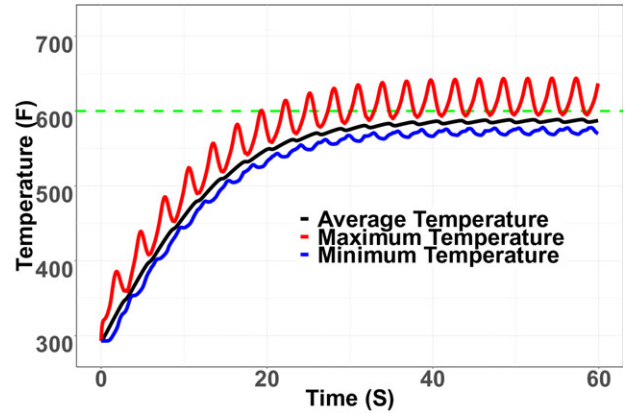**Figure 4.** (a) shows the best objective value $f(x_n^*)$ for the five compared methods. (b)–(f) show the average, maximum, and minimum temperature of the wafer over time, for each of the tested BO methods. The dotted green line marks the target temperature of $T^* = 600$ F.

method slightly worse. This again shows that the proposed HEI can provide a principled correction to the over-greediness of EI via hierarchical modeling, which translates to more effective optimization performance over the compared existing methods.

The remaining plots in Figure 4 show the average, maximum, and minimum temperature over the wafer surface, as a function of time. For HEI-DSD and HEI-MMAP, the average temperature

quickly hits 600 F, with a slight temperature oscillation over the wafer. For SEI, the average temperature reaches the target temperature slowly, but the temperature fluctuation is much higher than for HEI-DSD and HEI-MMAP. For UCB-OK, the average temperature does not even reach the target temperature. The two proposed HEI methods (and ϵ-EI-UK, although its performance is slightly worse) return noticeably improved settings compared

to the two earlier methods, thereby providing engineers with an effective and robust wafer heating process for semiconductor manufacturing.

8. Conclusion

In this article, we presented a hierarchical expected improvement (HEI) framework for Bayesian optimization of a black-box objective f. HEI aims to correct a key limitation of the expected improvement (EI) method: its over-exploitation of the fitted GP model, which results in a lack of convergence to a global solution even for smooth objective functions. HEI addresses this via a hierarchical GP model, which integrates parameter uncertainty of the fitted model within a closed-form acquisition function. This provides a principled way for correcting over-exploitation by encouraging exploration of the optimization space. We then introduce several hyperparameter specification methods, which allow HEI to efficiently approximate a fully Bayesian optimization procedure. Under certain prior specifications, we prove the global convergence of HEI over a broad function class for f, and derive near-minimax convergence rates. In numerical experiments, HEI provides improved optimization performance over existing Bayesian optimization methods, for both simulations and a process optimization problem in semiconductor manufacturing.

Given these promising results, there are several intriguing avenues for future work. One direction is to explore potential extensions of the HEI for the high-dimensional setting of $d > n$, where the dimension of the problem may exceed the number of function evaluations. This presents interesting challenges for both theory and methodology, and may require more sophisticated GP models which can learn low-dimensional embeddings in high dimensions (see, e.g., Zhang, Mak, and Dunson 2022). Another interesting direction is to study the dependence of the optimization rates on the constant γ in Condition 1.

Supplementary Materials

The online supplementary materials provide proofs of technical results in the paper.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

References

Benassi, R., Bect, J., and Vazquez, E. (2011), "Robust Gaussian Process-based Global Optimization Using a Fully Bayesian Expected Improvement Criterion," in *International Conference on Learning and Intelligent Optimization*, pp. 176–190, Springer. [2,5,9,11]

Brunner, T. A., Menon, V. C., Wong, C. W., Gluschenkov, O., Belyansky, M. P., Felix, N. M., Ausschnitt, C. P., Vukkadala, P., Veeraraghavan, S., and Sinha, J. K. (2013), "Characterization of Wafer Geometry and

Overlay Error on Silicon Wafers with Nonuniform Stress," *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 12, 1–13–13. [11]

Bull, A. D. (2011), "Convergence Rates of Efficient Global Optimization Algorithms," *Journal of Machine Learning Research*, 12, 2879–2904. [1,2,3,4,7,8,9]

Carlin, B. P., and Louis, T. A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* (Vol. 88), Boca Raton: Chapman & Hall/CRC. [5]

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015), "Bayesian Linear Regression with Sparse Priors," *The Annals of Statistics*, 43, 1986–2018. [8]

Chen, J., Mak, S., Joseph, V. R., and Zhang, C. (2020), "Function-on-Function Kriging, with Applications to Three-Dimensional Printing of Aortic Tissues," *Technometrics*, 63, 384–395. [1]

Chen, R.-B., Wang, W., and Wu, C.-F. J. (2017), "Sequential Designs based on Bayesian Uncertainty Quantification in Sparse Representation Surrogate Modeling," *Technometrics*, 59, 139–152. [2,3]

COMSOL. (2018). COMSOL Multiphysics® v. 5.4. [11]

Cressie, N. (1991). *Statistics for Spatial Data*, New York: Wiley. [7,8]

Dasgupta, T., Ma, C., Joseph, V. R., Wang, Z., and Wu, C. F. J. (2008), "Statistical Modeling and Analysis for Robust Synthesis of Nanostructures," *Journal of the American Statistical Association*, 103, 594–603. [1]

Deville, Y., Ginsbourger, D., and Roustant, O. (2019), "Kergp: Gaussian Process Laboratory," https://cran.r-project.org/web/packages/kergp. [9]

Doucet, A., Godsill, S. J., and Robert, C. P. (2002), "Marginal Maximum a Posteriori Estimation using Markov Chain Monte Carlo," *Statistics and Computing*, 12, 77–84. [6]

Feliot, P., Bect, J., and Vazquez, E. (2017), "A Bayesian Approach to Constrained Single-and Multi-Objective Optimization," *Journal of Global Optimization*, 67, 97–133. [1]

Frazier, P. I., Powell, W. B., and Dayanik, S. (2008), "A Knowledge-Gradient Policy for Sequential Information Collection," *SIAM Journal on Control and Optimization*, 47, 2410–2439. [1]

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–534. [5]

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410. [2]

Jin, R., Chang, C.-J., and Shi, J. (2012), "Sequential Measurement Strategy for Wafer Geometric Profile Estimation," *IIE Transactions*, 44, 1–12. [11]

Jones, D. R., Schonlau, M., and Welch, W. J. (1998), "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, 13, 455–492. [1,3]

Joseph, V. R., Gul, E., and Ba, S. (2019), "Designing Computer Experiments with Multiple Types of Factors: The MaxPro Approach," *Journal of Quality Technology*, 52, 343–354. [7]

Kearns, M., and Singh, S. (2002), "Near-Optimal Reinforcement Learning in Polynomial Time," *Machine Learning*, 49, 209–232. [1]

Lekivetz, R., and Jones, B. (2015), "Fast Flexible Space-Filling Designs for Nonrectangular Regions," *Quality and Reliability Engineering International*, 31, 829–837. [7]

Liu, Q., and Ihler, A. (2013), "Variational Algorithms for Marginal MAP," *Journal of Machine Learning Research*, 14, 3165–3200. [6]

Loeppky, J. L., Sacks, J., and Welch, W. J. (2009), "Choosing the Sample Size of a Computer Experiment: A Practical Guide," *Technometrics*, 51, 366–376. [7]

Mak, S., and Joseph, V. R. (2018), "Minimax and Minimax Projection Designs Using Clustering," *Journal of Computational and Graphical Statistics*, 27, 166–178. [7]

Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., Yang, V., and Wu, C.-F. J. (2018), "An Efficient Surrogate Model for Emulation and Physics Extraction of Large Eddy Simulations," *Journal of the American Statistical Association*, 113, 1443–1456. [1]

Mak, S., and Wu, C.-F. J. (2019), "Analysis-of-Marginal-Tail-Means (ATM): A Robust Method for Discrete Black-Box Optimization," *Technometrics*, 61, 545–559. [11]

Marmin, S., Chevalier, C., and Ginsbourger, D. (2015), "Differentiating the Multipoint Expected Improvement for Optimal Batch Design," in *International Workshop on Machine Learning, Optimization and Big Data*, pp. 37–48, Springer. [1]

Mockus, J., Tiesis, V., and Zilinskas, A. (1978), "The Application of Bayesian Methods for Seeking the Extremum," *Towards Global Optimization*, 2, 117–129. [1]

Morris, M. D., and Mitchell, T. J. (1995), "Exploratory Designs for Computational Experiments," *Journal of Statistical Planning and Inference*, 43, 381–402. [7,9]

Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, Philadelphia, PA: SIAM. [8]

Olea, R. A. (2012). *Geostatistics for Engineers and Earth Scientists*, New York: Springer. [2]

Qin, C., Klabjan, D., and Russo, D. (2017), "Improving the Expected Improvement Algorithm," in *Advances in Neural Information Processing Systems*, pp. 5381–5391. [1,3]

Santner, T. J., Williams, B. J., Notz, W., and Williams, B. J. (2018), *The Design and Analysis of Computer Experiments* (2nd ed.) New York: Springer. [1,2,3,4,5,6,8]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [6]

Scott, W., Frazier, P., and Powell, W. (2011), "The Correlated Knowledge Gradient for Simulation Optimization of Continuous Parameters Using Gaussian Process Regression," *SIAM Journal on Optimization*, 21, 996–1026. [1]

Seshadri, P., Yuchi, S., and Parks, G. T. (2019), "Dimension Reduction via Gaussian Ridge Functions," *SIAM/ASA Journal on Uncertainty Quantification*, 7, 1301–1322. [11]

Singh, R., Fakhruddin, M., and Poole, K. (2000), "Rapid Photothermal Processing as a Semiconductor Manufacturing Technology for the 21st Century," *Applied Surface Science*, 168, 198–203. [11]

Snoek, J., Larochelle, H., and Adams, R. P. (2012), "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems*, pp. 2951–2959. [2,3]

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010), "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, USA, pp. 1015–1022. Omnipress. [1,9]

Surjanovic, S., and Bingham, D. (2015), "Virtual Library of Simulation Experiments: Test Functions and Datasets," Retrieved August 16, 2019, from *http://www.sfu.ca/~ssurjano*. [9]

Sutton, R. S., and Barto, A. G. (2018), *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press. [9]

Wackernagel, H. (1995), *Multivariate Geostatistics: An Introduction with Applications*, Berlin: Springer. [2]

Wendland, H. (2004), *Scattered Data Approximation. Cambridge Monographs on Applied and Computational Mathematics* (Vol. 17), Cambridge: Cambridge University Press. [7,8]

Wynne, G., Briol, F., and Girolami. M. (2020), "Convergence Guarantees for Gaussian Process Means with Misspecified Likelihoods and Smoothness," arXiv preprint arXiv:2001.10818. [2,8,9]

Xiu, D. (2010), *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton: Princeton University Press. [6]

Zhang, R., Mak, S., and Dunson, D. (2022), "Gaussian Process Subspace Prediction for Model Reduction," *SIAM Journal on Scientific Computing*, 44, A1428–A1449. [11,13]

Zhang, Y., Han, Z.-H., and Zhang, K.-S. (2018), "Variable-Fidelity Expected Improvement Method for Efficient Global Optimization of Expensive Functions. *Structural and Multidisciplinary Optimization*, 58, 1431–1451. [1]