Journal of Machine Learning Research 12 (2011) 2879-2904 Submitted 1/11; Revised 7/11; Published 10/11

Convergence Rates of Efficient Global Optimization Algorithms

Adam D. Bull A.BULL@STATSLAB.CAM.AC.UK
Statistical Laboratory
University of Cambridge
Cambridge, CB3 0WB, UK

Editor: Manfred Opper

Abstract

In the efficient global optimization problem, we minimize an unknown function f, using as few observations $f(x)$ as possible. It can be considered a continuum-armed-bandit problem, with noiseless data, and simple regret. Expected-improvement algorithms are perhaps the most popular methods for solving the problem; in this paper, we provide theoretical results on their asymptotic behaviour.

Implementing these algorithms requires a choice of Gaussian-process prior, which determines an associated space of functions, its reproducing-kernel Hilbert space (RKHS). When the prior is fixed, expected improvement is known to converge on the minimum of any function in its RKHS. We provide convergence rates for this procedure, optimal for functions of low smoothness, and describe a modified algorithm attaining optimal rates for smoother functions.

In practice, however, priors are typically estimated sequentially from the data. For standard estimators, we show this procedure may never find the minimum of f. We then propose alternative estimators, chosen to minimize the constants in the rate of convergence, and show these estimators retain the convergence rates of a fixed prior.

Keywords: convergence rates, efficient global optimization, expected improvement, continuum-armed bandit, Bayesian optimization

1. Introduction

Suppose we wish to minimize a continuous function $f : X \to \mathbb{R}$, where X is a compact subset of $\mathbb{R}^d$. Observing $f(x)$ is costly (it may require a lengthy computer simulation or physical experiment), so we wish to use as few observations as possible. We know little about the shape of f; in particular we will be unable to make assumptions of convexity or unimodality. We therefore need a *global* optimization algorithm, one which attempts to find a global minimum.

Many standard global optimization algorithms exist, including genetic algorithms, multistart, and simulated annealing (Pardalos and Romeijn, 2002), but these algorithms are designed for functions that are cheap to evaluate. When f is expensive, we need an *efficient* algorithm, one which will choose its observations to maximize the information gained.

We can consider this a continuum-armed-bandit problem (Srinivas et al., 2010, and references therein), with noiseless data, and loss measured by the simple regret (Bubeck et al., 2009). At time n, we choose a design point $x_n \in X$, make an observation $z_n = f(x_n)$, and then report a point x_n^* where we believe $f(x_n^*)$ will be low. Our goal is to find a strategy for choosing the x_n and x_n^*, in terms of previous observations, so as to minimize $f(x_n^*)$.

We would like to find a strategy which can guarantee convergence: for functions f in some smoothness class, $f(x_n^*)$ should tend to $\min f$, preferably at some fast rate. The simplest method

would be to fix a sequence of x_n in advance, and set $x_n^* = \arg\min \hat{f}_n$, for some approximation $\hat{f}_n$ to f. We will show that if $\hat{f}_n$ converges in supremum norm at the optimal rate, then $f(x_n^*)$ also converges at its optimal rate. However, while this strategy gives a good worst-case bound, on average it is clearly a poor method of optimization: the design points x_n are completely independent of the observations z_n.

We may therefore ask if there are more efficient methods, with better average-case performance, that nevertheless provide good guarantees of convergence. The difficulty in designing such a method lies in the trade-off between *exploration* and *exploitation*. If we exploit the data, observing in regions where f is known to be low, we will be more likely to find the optimum quickly; however, unless we explore every region of X, we may not find it at all (Macready and Wolpert, 1998).

Initial attempts at this problem include work on Lipschitz optimization (summarized in Hansen et al., 1992) and the DIRECT algorithm (Jones et al., 1993), but perhaps the best-known strategy is expected improvement. It is sometimes called Bayesian optimization, and first appeared in Močkus (1974) as a Bayesian decision-theoretic solution to the problem. Contemporary computers were not powerful enough to implement the technique in full, and it was later popularized by Jones et al. (1998), who provided a computationally efficient implementation. More recently, it has also been called a knowledge-gradient policy by Frazier et al. (2009). Many extensions and alterations have been suggested by further authors; a good summary can be found in Brochu et al. (2010).

Expected improvement performs well in experiments (Osborne, 2010, §9.5), but little is known about its theoretical properties. The behaviour of the algorithm depends crucially on the Gaussian process prior π chosen for f. Each prior has an associated space of functions $\mathcal{H}$, its reproducing-kernel Hilbert space. $\mathcal{H}$ contains all functions $X \to \mathbb{R}$ as smooth as a posterior mean of f, and is the natural space in which to study questions of convergence.

Vazquez and Bect (2010) show that when π is a fixed Gaussian process prior of finite smoothness, expected improvement converges on the minimum of any $f \in \mathcal{H}$, and almost surely for f drawn from π. Grunewalder et al. (2010) bound the convergence rate of a computationally infeasible version of expected improvement: for priors π of smoothness ν, they show convergence at a rate $O^*(n^{-(\nu \wedge 0.5)/d})$ on f drawn from π. We begin by bounding the convergence rate of the feasible algorithm, and show convergence at a rate $O^*(n^{-(\nu \wedge 1)/d})$ on all $f \in \mathcal{H}$. We go on to show that a modification of expected improvement converges at the near-optimal rate $O^*(n^{-\nu/d})$.

For practitioners, however, these results are somewhat misleading. In typical applications, the prior is not held fixed, but depends on parameters estimated sequentially from the data. This process ensures the choice of observations is invariant under translation and scaling of f, and is believed to be more efficient (Jones et al., 1998, §2). It has a profound effect on convergence, however: Locatelli (1997, §3.2) shows that, for a Brownian motion prior with estimated parameters, expected improvement may not converge at all.

We extend this result to more general settings, showing that for standard priors with estimated parameters, there exist smooth functions f on which expected improvement does not converge. We then propose alternative estimates of the prior parameters, chosen to minimize the constants in the convergence rate. We show that these estimators give an automatic choice of parameters, while retaining the convergence rates of a fixed prior.

Table 1 summarizes the notation used in this paper. We say $f : \mathbb{R}^d \to \mathbb{R}$ is a bump function if f is infinitely differentiable and of compact support, and $f : \mathbb{R}^d \to \mathbb{C}$ is Hermitian if $\overline{f(x)} = f(-x)$. We use the Landau notation $f = O(g)$ to denote $\limsup |f/g| < \infty$, and $f = o(g)$ to denote $f/g \to 0$. If $g = O(f)$, we say $f = \Omega(g)$, and if both $f = O(g)$ and $f = \Omega(g)$, we say $f = \Theta(g)$. If further

$f/g \to 1$, we say $f \sim g$. Finally, if f and g are random, and $\mathbb{P}(\sup|f/g| \leq M) \to 1$ as $M \to \infty$, we say $f = O_p(g)$.

In Section 2, we briefly describe the expected-improvement algorithm, and detail our assumptions on the priors used. We state our main results in Section 3, and discuss implications for further work in Section 4. Finally, we give proofs in Appendix A.

2. Expected Improvement

Suppose we wish to minimize an unknown function f, choosing design points x_n and estimated minima x_n^* as in the introduction. If we pick a prior distribution π for f, representing our beliefs about the unknown function, we can describe this problem in terms of decision theory. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, equipped with a random process f having law π. A strategy u is a collection of random variables (x_n), (x_n^*) taking values in X. Set $z_n := f(x_n)$, and define the filtration $\mathcal{F}_n := \sigma(x_i, z_i : i \leq n)$. The strategy u is valid if x_n is conditionally independent of f given $\mathcal{F}_{n-1}$, and likewise x_n^* given $\mathcal{F}_n$. (Note that we allow random strategies, provided they do not depend on unknown information about f.)

When taking probabilities and expectations we will write $\mathbb{P}_\pi^u$ and $\mathbb{E}_\pi^u$, denoting the dependence on both the prior π and strategy u. The average-case performance at some future time N is then given by the expected loss,

$$\mathbb{E}_\pi^u[f(x_N^*) - \min f],$$

and our goal, given π, is to choose the strategy u to minimize this quantity.

2.1 Bayesian Optimization

For $N > 1$ this problem is very computationally intensive (Osborne, 2010, §6.3), but we can solve a simplified version of it. First, we restrict the choice of x_n^* to the previous design points $x_1, \ldots, x_n$. (In practice this is reasonable, as choosing an x_n^* we have not observed can be unreliable.) Secondly, rather than finding an optimal strategy for the problem, we derive the myopic strategy: the strategy which is optimal if we always assume we will stop after the next observation. This strategy is suboptimal (Ginsbourger et al., 2008, §3.1), but performs well, and greatly simplifies the calculations involved.

In this setting, given $\mathcal{F}_n$, if we are to stop at time n we should choose $x_n^* := x_{i_n^*}$, where $i_n^* := \arg\min_{1,\ldots,n} z_i$. (In the case of ties, we may pick any minimizing i_n^*.) We then suffer a loss $z_n^* - \min f$, where $z_n^* := z_{i_n^*}$. Were we to observe at x_{n+1} before stopping, the expected loss would be

$$\mathbb{E}_\pi^u[z_{n+1}^* - \min f \mid \mathcal{F}_n],$$

so the myopic strategy should choose x_{n+1} to minimize this quantity. Equivalently, it should maximize the expected improvement over the current loss,

$$EI_n(x_{n+1}; \pi) := \mathbb{E}_\pi^u[z_n^* - z_{n+1}^* \mid \mathcal{F}_n] = \mathbb{E}_\pi^u[(z_n^* - z_{n+1})^+ \mid \mathcal{F}_n], \tag{1}$$

where $x^+ = \max(x, 0)$.

So far, we have merely replaced one optimization problem with another. However, for suitable priors, EI_n can be evaluated cheaply, and thus maximized by standard techniques. The expected-improvement algorithm is then given by choosing x_{n+1} to maximize (1).

Section 1	
f	unknown function $X \to \mathbb{R}$ to be minimized
X	compact subset of $\mathbb{R}^d$ to minimize over
d	number of dimensions to minimize over
x_n	points in X at which f is observed
z_n	observations $z_n = f(x_n)$ of f
x_n^*	estimated minimum of f, given $z_1, \ldots, z_n$
Section 2.1	
π	prior distribution for f
u	strategy for choosing x_n, x_n^*
$\mathcal{F}_n$	filtration $\mathcal{F}_n = \sigma(x_i, z_i : i \leq n)$
z_n^*	best observation $z_n^* = \min_{i=1,\ldots,n} z_i$
EI_n	expected improvement given $\mathcal{F}_n$
Section 2.2	
μ, σ^2	global mean and variance of Gaussian-process prior π
K	underlying correlation kernel for π
K_θ	correlation kernel for π with length-scales θ
ν, α	smoothness parameters of K
$\hat{\mu}_n$, $\hat{f}_n$, s_n^2, $\hat{R}_n^2$	quantities describing posterior distribution of f given $\mathcal{F}_n$
Section 2.3	
$EI(\pi)$	expected improvement strategy with fixed prior
$\hat{\sigma}_n^2$, $\hat{\theta}_n$	estimates of prior parameters σ^2, θ
c_n	rate of decay of $\hat{\sigma}_n^2$
θ^L, θ^U	bounds on $\hat{\theta}_n$
$EI(\hat{\pi})$	expected improvement strategy with estimated prior
Section 3.1	
$\mathcal{H}_\theta(S)$	reproducing-kernel Hilbert space of K_θ on S
$H^s(D)$	Sobolev Hilbert space of order s on D
Section 3.2	
L_n	loss suffered over an RKHS ball after n steps
Section 3.3	
$EI(\tilde{\pi})$	expected improvement strategy with robust estimated prior
Section 3.4	
$EI(\cdot, \varepsilon)$	ε-greedy expected improvement strategies

Table 1: Notation

2.2 Gaussian Process Models

We still need to choose a prior π for f. Typically, we model f as a stationary Gaussian process: we consider the values $f(x)$ to be jointly Gaussian, with mean and covariance

$$\mathbb{E}_\pi[f(x)] = \mu, \quad \mathbb{C}\mathrm{ov}_\pi[f(x), f(y)] = \sigma^2 K_\theta(x-y). \tag{2}$$

$\mu \in \mathbb{R}$ is the global mean of f; we place a flat prior on μ, reflecting our uncertainty over the location of f.

$\sigma > 0$ is the global scale of variation of f, and $K_\theta : \mathbb{R}^d \to \mathbb{R}$ its correlation kernel, governing the local properties of f. In the following, we will consider kernels

$$K_\theta(t_1, \ldots, t_d) \coloneqq K(t_1/\theta_1, \ldots, t_d/\theta_d), \tag{3}$$

for an underlying kernel K with $K(0) = 1$. (Note that we can always satisfy this condition by suitably scaling K and σ.) The $\theta_i > 0$ are the length-scales of the process: two values $f(x)$ and $f(y)$ will be highly correlated if each $x_i - y_i$ is small compared with θ_i. For now, we will assume the parameters σ and θ are fixed in advance.

For (2) and (3) to define a consistent Gaussian process, K must be a symmetric positive-definite function. We will also make the following assumptions.

Assumption 1. *K is continuous and integrable.*

K thus has Fourier transform

$$\widehat{K}(\xi) \coloneqq \int_{\mathbb{R}^d} e^{-2\pi i \langle x, \xi \rangle} K(x)\, dx,$$

and by Bochner's theorem, $\widehat{K}$ is non-negative and integrable.

Assumption 2. *$\widehat{K}$ is isotropic and radially non-increasing.*

In other words, $\widehat{K}(x) = \widehat{k}(\|x\|)$ for a non-increasing function $\widehat{k} : [0, \infty) \to [0, \infty)$; as a consequence, K is isotropic.

Assumption 3. *As $x \to \infty$, either:*

(i) $\widehat{K}(x) = \Theta(\|x\|^{-2\nu-d})$ *for some* $\nu > 0$*; or*

(ii) $\widehat{K}(x) = O(\|x\|^{-2\nu-d})$ *for all* $\nu > 0$ *(we will then say that* $\nu = \infty$*).*

Note the condition $\nu > 0$ is required for $\widehat{K}$ to be integrable.

Assumption 4. *K is C^k, for k the largest integer less than 2ν, and at the origin, K has k-th order Taylor approximation P_k satisfying*

$$|K(x) - P_k(x)| = O\left(\|x\|^{2\nu}(-\log\|x\|)^{2\alpha}\right)$$

as $x \to 0$, for some $\alpha \geq 0$.

When $\alpha = 0$, this is just the condition that K be 2ν-Hölder at the origin; when $\alpha > 0$, we instead require this condition up to a log factor.

The rate ν controls the smoothness of functions from the prior: almost surely, f has continuous derivatives of any order $k < \nu$ (Adler and Taylor, 2007, §1.4.2). Popular kernels include the Matérn class,

$$K^\nu(x) \coloneqq \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\|x\|\right)^\nu k_\nu \left(\sqrt{2\nu}\|x\|\right), \quad \nu \in (0,\infty),$$

where k_ν is a modified Bessel function of the second kind, and the Gaussian kernel,

$$K^\infty(x) \coloneqq e^{-\frac{1}{2}\|x\|^2},$$

obtained in the limit $\nu \to \infty$ (Rasmussen and Williams, 2006, §4.2). Between them, these kernels cover the full range of smoothness $0 < \nu \leq \infty$. Both kernels satisfy Assumptions 1–4 for the ν given; $\alpha = 0$ except for the Matérn kernel with $\nu \in \mathbb{N}$, where $\alpha = \frac{1}{2}$ (Abramowitz and Stegun, 1965, §9.6).

Having chosen our prior distribution, we may now derive its posterior. We find

$$f(x) \mid z_1, \ldots, z_n \sim N\left(\hat{f}_n(x;\theta), \sigma^2 s_n^2(x;\theta)\right),$$

where

$$\hat{\mu}_n(\theta) \coloneqq \frac{1^T V^{-1} z}{1^T V^{-1} 1}, \tag{4}$$

$$\hat{f}_n(x;\theta) \coloneqq \hat{\mu}_n + v^T V^{-1}(z - \hat{\mu}_n 1), \tag{5}$$

and

$$s_n^2(x;\theta) \coloneqq 1 - v^T V^{-1} v + \frac{(1 - 1^T V^{-1} v)^2}{1^T V^{-1} 1}, \tag{6}$$

for $z = (z_i)_{i=1}^n$, $V = (K_\theta(x_i - x_j))_{i,j=1}^n$, and $v = (K_\theta(x - x_i))_{i=1}^n$ (Santner et al., 2003, §4.1.3). Equivalently, these expressions are the best linear unbiased predictor of $f(x)$ and its variance, as given in Jones et al. (1998, §2). We will also need the reduced sum of squares,

$$\hat{R}_n^2(\theta) \coloneqq (z - \hat{\mu}_n 1)^T V^{-1} (z - \hat{\mu}_n 1). \tag{7}$$

2.3 Expected Improvement Strategies

Under our assumptions on π, we may now derive an analytic form for (1), as in Jones et al. (1998, §4.1). We obtain

$$EI_n(x_{n+1};\pi) = \rho\left(z_n^* - \hat{f}_n(x_{n+1};\theta), \sigma s_n(x_{n+1};\theta)\right), \tag{8}$$

where

$$\rho(y,s) \coloneqq \begin{cases} y\Phi(y/s) + s\varphi(y/s), & s > 0, \\ \max(y,0), & s = 0, \end{cases} \tag{9}$$

and Φ and φ are the standard normal distribution and density functions respectively.

For a prior π as above, expected improvement chooses x_{n+1} to maximize (8), but this does not fully define the strategy. Firstly, we must describe how the strategy breaks ties, when more than one $x \in X$ maximizes EI_n. In general, this will not affect the behaviour of the algorithm, so we allow any choice of x_{n+1} maximizing (8).

Secondly, we must say how to choose x_1, as the above expressions are undefined when $n = 0$. In fact, Jones et al. (1998, §4.2) find that expected improvement can be unreliable given few data points, and recommend that several initial design points be chosen in a random quasi-uniform arrangement. We will therefore assume that until some fixed time k, points $x_1, \ldots, x_k$ are instead chosen by some (potentially random) method independent of f. We thus obtain the following strategy.

Definition 1. *An $EI(\pi)$ strategy chooses:*

(i) initial design points $x_1, \ldots, x_k$ independently of f; and

(ii) further design points x_{n+1} $(n \geq k)$ from the maximizers of (8).

So far, we have not considered the choice of parameters σ and θ. While these can be fixed in advance, doing so requires us to specify characteristic scales of the unknown function f, and causes expected improvement to behave differently on a rescaling of the same function. We would prefer an algorithm which could adapt automatically to the scale of f.

A natural approach is to take maximum likelihood estimates of the parameters, as recommended by Jones et al. (1998, §2). Given θ, the MLE $\hat{\sigma}_n^2 = \hat{R}_n^2(\theta)/n$; for full generality, we will allow any choice $\hat{\sigma}_n^2 = c_n \hat{R}_n^2(\theta)$, where $c_n = o(1/\log n)$. Estimates of θ, however, must be obtained by numerical optimization. As θ can vary widely in scale, this optimization is best performed over $\log\theta$; as the likelihood surface is typically multimodal, this requires the use of a global optimizer. We must therefore place (implicit or explicit) bounds on the allowed values of $\log\theta$. We have thus described the following strategy.

Definition 2. *Let $\hat{\pi}_n$ be a sequence of priors, with parameters $\hat{\sigma}_n$, $\hat{\theta}_n$ satisfying:*

(i) $\hat{\sigma}_n^2 = c_n \hat{R}_n^2(\hat{\theta}_n)$ for constants $c_n > 0$, $c_n = o(1/\log n)$; and

(ii) $\theta^L \leq \hat{\theta}_n \leq \theta^U$ for constants $\theta^L, \theta^U \in \mathbb{R}_+^d$.

An $EI(\hat{\pi})$ strategy satisfies Definition 1, replacing π with $\hat{\pi}_n$ in (8).

3. Convergence Rates

To discuss convergence, we must first choose a smoothness class for the unknown function f. Each kernel K_θ is associated with a space of functions $\mathcal{H}_\theta(X)$, its reproducing-kernel Hilbert space (RKHS) or native space. $\mathcal{H}_\theta(X)$ contains all functions $X \to \mathbb{R}$ as smooth as a posterior mean of f, and is the natural space to study convergence of expected-improvement algorithms, allowing a tractable analysis of their asymptotic behaviour.

3.1 Reproducing-Kernel Hilbert Spaces

Given a symmetric positive-definite kernel K on $\mathbb{R}^d$, set $k_x(t) = K(t-x)$. For $S \subseteq \mathbb{R}^d$, let $\mathcal{E}(S)$ be the space of functions $S \to \mathbb{R}$ spanned by the k_x, for $x \in S$. Furnish $\mathcal{E}(S)$ with the inner product defined by

$$\langle k_x, k_y \rangle \coloneqq K(x-y).$$

The completion of $\mathcal{E}(S)$ under this inner product is the reproducing-kernel Hilbert space $\mathcal{H}(S)$ of K on S. The members $f \in \mathcal{H}(S)$ are abstract objects, but we can identify them with functions $f : S \to \mathbb{R}$ through the reproducing property,

$$f(x) = \langle f, k_x \rangle,$$

which holds for all $f \in \mathcal{E}(S)$. See Aronszajn (1950), Berlinet and Thomas-Agnan (2004), Wendland (2005) and van der Vaart and van Zanten (2008).

We will find it convenient also to use an alternative characterization of $\mathcal{H}(S)$. We begin by describing $\mathcal{H}(\mathbb{R}^d)$ in terms of Fourier transforms. Let $\widehat{f}$ denote the Fourier transform of a function $f \in L^2$. The following result is stated in Parzen (1963, §2), and proved in Wendland (2005, §10.2); we give a short proof in Appendix A.

Lemma 1. *$\mathcal{H}(\mathbb{R}^d)$ is the space of real continuous $f \in L^2(\mathbb{R}^d)$ whose norm*

$$\|f\|^2_{\mathcal{H}(\mathbb{R}^d)} := \int \frac{|\widehat{f}(\xi)|^2}{\widehat{K}(\xi)}\, d\xi$$

is finite, taking $0/0 = 0$.

We may now describe $\mathcal{H}(S)$ in terms of $\mathcal{H}(\mathbb{R}^d)$.

Lemma 2 (Aronszajn, 1950, §1.5)**.** *$\mathcal{H}(S)$ is the space of functions $f = g|_S$ for some $g \in \mathcal{H}(\mathbb{R}^d)$, with norm*

$$\|f\|_{\mathcal{H}(S)} := \inf_{g|_S = f} \|g\|_{\mathcal{H}(\mathbb{R}^d)},$$

and there is a unique g minimizing this expression.

These spaces are in fact closely related to the Sobolev Hilbert spaces of functional analysis. Say a domain $D \subseteq \mathbb{R}^d$ is Lipschitz if its boundary is locally the graph of a Lipschitz function (see Tartar, 2007, §12, for a precise definition). For such a domain D, the Sobolev Hilbert space $H^s(D)$ is the space of functions $f : D \to \mathbb{R}$, given by the restriction of some $g : \mathbb{R}^d \to \mathbb{R}$, whose norm

$$\|f\|^2_{H^s(D)} := \inf_{g|_D = f} \int \frac{|\widehat{g}(\xi)|^2}{(1 + \|\xi\|^2)^{s/2}}\, d\xi$$

is finite. Thus, for the kernel K with Fourier transform $\widehat{K}(\xi) = (1 + \|\xi\|^2)^{s/2}$, this is just the RKHS $\mathcal{H}(D)$. More generally, if K satisfies our assumptions with $\nu < \infty$, these spaces are equivalent in the sense of normed spaces: they contain the same functions, and have norms $\|\cdot\|_1, \|\cdot\|_2$ satisfying

$$C\|f\|_1 \le \|f\|_2 \le C'\|f\|_1,$$

for constants $0 < C \le C'$.

Lemma 3. *Let $\mathcal{H}_\theta(S)$ denote the RKHS of K_θ on S, and $D \subseteq \mathbb{R}^d$ be a Lipschitz domain.*

(i) If $\nu < \infty$, $\mathcal{H}_\theta(\bar{D})$ is equivalent to the Sobolev Hilbert space $H^{\nu + d/2}(D)$.

(ii) If $\nu = \infty$, $\mathcal{H}_\theta(\bar{D})$ is continuously embedded in $H^s(D)$ for all s.

Thus if $\nu < \infty$, and X is, say, a product of intervals $\prod_{i=1}^d [a_i, b_i]$, the RKHS $\mathcal{H}_\theta(X)$ is equivalent to the Sobolev Hilbert space $H^{\nu+d/2}(\prod_{i=1}^d (a_i, b_i))$, identifying each function in that space with its unique continuous extension to X.

3.2 Fixed Parameters

We are now ready to state our main results. Let $X \subset \mathbb{R}^d$ be compact with non-empty interior. For a function $f : X \to \mathbb{R}$, let $\mathbb{P}_f^u$ and $\mathbb{E}_f^u$ denote probability and expectation when minimizing the fixed function f with strategy u. (Note that while f is fixed, u may be random, so its performance is still probabilistic in nature.) We define the loss suffered over the ball B_R in $\mathcal{H}_\theta(X)$ after n steps by a strategy u,

$$L_n(u, \mathcal{H}_\theta(X), R) := \sup_{\|f\|_{\mathcal{H}_\theta(X)} \leq R} \mathbb{E}_f^u[f(x_n^*) - \min f].$$

We will say that u converges on the optimum at rate r_n, if

$$L_n(u, \mathcal{H}_\theta(X), R) = O(r_n)$$

for all $R > 0$. Note that we do not allow u to vary with R; the strategy must achieve this rate without prior knowledge of $\|f\|_{\mathcal{H}_\theta(X)}$.

We begin by showing that the minimax rate of convergence is $n^{-\nu/d}$.

Theorem 1. *If $\nu < \infty$, then for any $\theta \in \mathbb{R}_+^d$, $R > 0$,*

$$\inf_u L_n(u, \mathcal{H}_\theta(X), R) = \Theta(n^{-\nu/d}),$$

and this rate can be achieved by a strategy u not depending on R.

The upper bound is provided by a naive strategy as in the introduction: we fix a quasi-uniform sequence x_n in advance, and take x_n^* to minimize a radial basis function interpolant of the data. As remarked previously, however, this naive strategy is not very satisfying; in practice it will be outperformed by any good strategy varying with the data. We may thus ask whether more sophisticated strategies, with better practical performance, can still provide good worst-case bounds.

One such strategy is the $EI(\pi)$ strategy of Definition 1. We can show this strategy converges at least at rate $n^{-(\nu \wedge 1)/d}$, up to log factors.

Theorem 2. *Let π be a prior with length-scales $\theta \in \mathbb{R}_+^d$. For any $R > 0$,*

$$L_n(EI(\pi), \mathcal{H}_\theta(X), R) = \begin{cases} O(n^{-\nu/d}(\log n)^\alpha), & \nu \leq 1, \\ O(n^{-1/d}), & \nu > 1. \end{cases}$$

For $\nu \leq 1$, these rates are near-optimal. For $\nu > 1$, we are faced with a more difficult problem; we discuss this in more detail in Section 3.4.

3.3 Estimated Parameters

First, we consider the effect of the prior parameters on $EI(\pi)$. While the previous result gives a convergence rate for any fixed choice of parameters, the constant in that rate will depend on the parameters chosen; to choose well, we must somehow estimate these parameters from the data. The $EI(\hat{\pi})$ strategy, given by Definition 2, uses maximum likelihood estimates for this purpose. We can show, however, that this may cause the strategy to never converge.

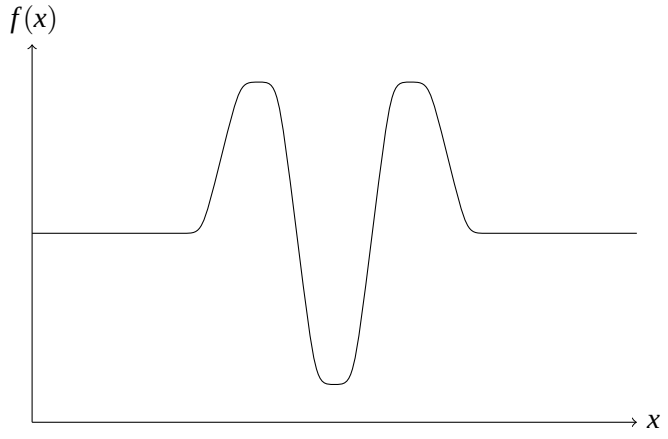

Figure 1: A counterexample from Theorem 3

Theorem 3. *Suppose* $\nu < \infty$. *Given* $\theta \in \mathbb{R}_+^d$, $R > 0$, $\varepsilon > 0$, *there exists* $f \in \mathcal{H}_\theta(X)$ *satisfying* $\|f\|_{\mathcal{H}_\theta(X)} \leq R$, *and for some fixed* $\delta > 0$,

$$\mathbb{P}_f^{EI(\hat{\pi})}\left(\inf_n f(x_n^*) - \min f \geq \delta\right) > 1 - \varepsilon.$$

The counterexamples constructed in the proof of the theorem may be difficult to minimize, but they are not badly-behaved (Figure 1). A good optimization strategy should be able to minimize such functions, and we must ask why expected improvement fails.

We can understand the issue by considering the constant in Theorem 2. Define

$$\tau(x) := x\Phi(x) + \varphi(x).$$

From the proof of Theorem 2, the dominant term in the convergence rate has constant

$$C(R+\sigma)\frac{\tau(R/\sigma)}{\tau(-R/\sigma)}, \tag{10}$$

for $C > 0$ not depending on R or σ. In Appendix A, we will prove the following result.

Corollary 1. $\hat{R}_n(\theta)$ *is non-decreasing in* n, *and bounded above by* $\|f\|_{\mathcal{H}_\theta(X)}$.

Hence for fixed θ, the estimate $\hat{\sigma}_n^2 = \hat{R}_n^2(\theta)/n \leq R^2/n$, and thus $R/\hat{\sigma}_n \geq n^{1/2}$. Inserting this choice into (10) gives a constant growing exponentially in n, destroying our convergence rate.

To resolve the issue, we will instead try to pick σ to minimize (10). The term $R+\sigma$ is increasing in σ, and the term $\tau(R/\sigma)/\tau(-R/\sigma)$ is decreasing in σ; we may balance the terms by taking $\sigma = R$. The constant is then proportional to R, which we may minimize by taking $R = \|f\|_{\mathcal{H}_\theta(X)}$. In practice, we will not know $\|f\|_{\mathcal{H}_\theta(X)}$ in advance, so we must estimate it from the data; from Corollary 1, a convenient estimate is $\hat{R}_n(\theta)$.

Suppose, then, that we make some bounded estimate $\hat{\theta}_n$ of θ, and set $\hat{\sigma}_n^2 = \hat{R}_n^2(\hat{\theta}_n)$. As Theorem 3 holds for any $\hat{\sigma}_n^2$ of faster than logarithmic decay, such a choice is necessary to ensure convergence.

(We may also choose θ to minimize (10); we might then pick $\hat{\theta}_n$ minimizing $\hat{R}_n(\theta)\prod_{i=1}^{d}\theta_i^{-\nu/d}$, but our assumptions on $\hat{\theta}_n$ are weak enough that we need not consider this further.)

If we believe our Gaussian-process model, this estimate $\hat{\sigma}_n$ is certainly unusual. We should, however, take care before placing too much faith in the model. The function in Figure 1 is a reasonable function to optimize, but as a Gaussian process it is highly atypical: there are intervals on which the function is constant, an event which in our model occurs with probability zero. If we want our algorithm to succeed on more general classes of functions, we will need to choose our parameter estimates appropriately.

To obtain good rates, we must add a further condition to our strategy. If $z_1 = \cdots = z_n$, $EI_n(\cdot;\hat{\pi}_n)$ is identically zero, and all choices of x_{n+1} are equally valid. To ensure we fully explore f, we will therefore require that when our strategy is applied to a constant function $f(x) = c$, it produces a sequence x_n dense in X. (This can be achieved, for example, by choosing x_{n+1} uniformly at random from X when $z_1 = \cdots = z_n$.) We have thus described the following strategy.

Definition 3. *An $EI(\tilde{\pi})$ strategy satisfies Definition 2, except:*

(i) we instead set $\hat{\sigma}_n^2 = \hat{R}_n^2(\hat{\theta}_n)$; and

(ii) we require the choice of x_{n+1} maximizing (8) *to be such that, if f is constant, the design points are almost surely dense in X.*

We cannot now prove a convergence result uniform over balls in $\mathcal{H}_\theta(X)$, as the rate of convergence depends on the ratio $R/\hat{R}_n$, which is unbounded. (Indeed, any estimator of $\|f\|_{\mathcal{H}_\theta(X)}$ must sometimes perform poorly: f can appear from the data to have arbitrarily small norm, while in fact having a spike somewhere we have not yet observed.) We can, however, provide the same convergence rates as in Theorem 2, in a slightly weaker sense.

Theorem 4. *For any $f \in \mathcal{H}_{\theta^U}(X)$, under $\mathbb{P}_f^{EI(\tilde{\pi})}$,*

$$f(x_n^*) - \min f = \begin{cases} O_p(n^{-\nu/d}(\log n)^{\alpha}), & \nu \leq 1, \\ O_p(n^{-1/d}), & \nu > 1. \end{cases}$$

3.4 Near-Optimal Rates

So far, our rates have been near-optimal only for $\nu \leq 1$. To obtain good rates for $\nu > 1$, standard results on the performance of Gaussian-process interpolation (Narcowich et al., 2003, §6) then require the design points x_i to be quasi-uniform in a region of interest. It is unclear whether this occurs naturally under expected improvement, but there are many ways we can modify the algorithm to ensure it.

Perhaps the simplest, and most well-known, is an ε-greedy strategy (Sutton and Barto, 1998, §2.2). In such a strategy, at each step with probability $1-\varepsilon$ we make a decision to maximize some greedy criterion; with probability ε we make a decision completely at random. This random choice ensures that the short-term nature of the greedy criterion does not overshadow our long-term goal.

The parameter ε controls the trade-off between global and local search: a good choice of ε will be small enough to not interfere with the expected-improvement algorithm, but large enough to prevent it from getting stuck in a local minimum. Sutton and Barto (1998, §2.2) consider the values $\varepsilon = 0.1$ and $\varepsilon = 0.01$, but in practical work ε should of course be calibrated to a typical problem set.

We therefore define the following strategies.

Definition 4. *Let $\cdot$ denote π, $\hat{\pi}$ or $\tilde{\pi}$. For $0 < \varepsilon < 1$, an $EI(\cdot, \varepsilon)$ strategy:*

(i) chooses initial design points $x_1, \cdots, x_k$ independently of f;

(ii) with probability $1 - \varepsilon$, chooses design point x_{n+1} $(n \geq k)$ as in $EI(\cdot)$; or

(iii) with probability ε, chooses x_{n+1} $(n \geq k)$ uniformly at random from X.

We can show that these strategies achieve near-optimal rates of convergence for all $\nu < \infty$.

Theorem 5. *Let $EI(\cdot, \varepsilon)$ be one of the strategies in Definition 4. If $\nu < \infty$, then for any $R > 0$,*

$$L_n(EI(\cdot, \varepsilon), \mathcal{H}_{\theta^U}(X), R) = O((n/\log n)^{-\nu/d}(\log n)^{\alpha}),$$

while if $\nu = \infty$, the statement holds for all $\nu < \infty$.

Note that unlike a typical ε-greedy algorithm, we do not rely on random choice to obtain global convergence: as above, the $EI(\pi)$ and $EI(\tilde{\pi})$ strategies are already globally convergent. Instead, we use random choice simply to improve upon the worst-case rate. Note also that the result does not in general hold when $\varepsilon = 1$; to obtain good rates, we must combine global search with inference about f.

4. Conclusions

We have shown that expected improvement can converge near-optimally, but a naive implementation may not converge at all. We thus echo Diaconis and Freedman (1986) in stating that, for infinite-dimensional problems, Bayesian methods are not always guaranteed to find the right answer; such guarantees can only be provided by considering the problem at hand.

We might ask, however, if our framework can also be improved. Our upper bounds on convergence were established using naive algorithms, which in practice would prove inefficient. If a sophisticated algorithm fails where a naive one succeeds, then the sophisticated algorithm is certainly at fault; we might, however, prefer methods of evaluation which do not consider naive algorithms so successful.

Vazquez and Bect (2010) and Grunewalder et al. (2010) consider a more Bayesian formulation of the problem, where the unknown function f is distributed according to the prior π, but this approach can prove restrictive: as we saw in Section 3.3, placing too much faith in the prior may exclude functions of interest. Further, Grunewalder et al. find the same issues are present also within the Bayesian framework.

A more interesting approach is given by the continuum-armed-bandit problem (Srinivas et al., 2010, and references therein). Here the goal is to minimize the cumulative regret,

$$R_n := \sum_{i=1}^{n} (f(x_i) - \min f),$$

in general observing the function f under noise. Algorithms controlling the cumulative regret at rate r_n also solve the optimization problem, at rate r_n/n (Bubeck et al., 2009, §3). The naive algorithms above, however, have poor cumulative regret. We might, then, consider the cumulative regret to be a better measure of performance, but this approach too has limitations. Firstly, the cumulative regret

is necessarily increasing, so cannot establish rates of optimization faster than n^{-1}. (This is not an issue under noise, where typically $r_n = \Omega(n^{1/2})$, see Kleinberg and Slivkins, 2010.) Secondly, if our goal is optimization, then minimizing the regret, a cost we do not incur, may obscure the problem at hand.

Bubeck et al. (2010) study this problem with the additional assumption that f has finitely many minima, and is, say, quadratic in a neighbourhood of each. This assumption may suffice in practice, and allows the authors to obtain impressive rates of convergence. For optimization, however, a further weakness is that these rates hold only once the algorithm has found a basin of attraction; they thus measure local, rather than global, performance. It may be that convergence rates alone are not sufficient to capture the performance of a global optimization algorithm, and the time taken to find a basin of attraction is more relevant. In any case, the choice of an appropriate framework to measure performance in global optimization merits further study.

Finally, we should also ask how to choose the smoothness parameter ν (or the equivalent parameter in similar algorithms). Van der Vaart and van Zanten (2009) show that Bayesian Gaussian-process models can, in some contexts, automatically adapt to the smoothness of an unknown function f. Their technique requires, however, that the estimated length-scales $\hat{\theta}_n$ to tend to 0, posing both practical and theoretical challenges. The question of how best to optimize functions of unknown smoothness remains open.

Acknowledgments

We would like to thank the referees, as well as Richard Nickl and Steffen Grunewalder, for their valuable comments and suggestions.

Appendix A. Proofs

We now prove the results in Section 3.

A.1 Reproducing-Kernel Hilbert Spaces

Proof of Lemma 1. Let V be the space of functions described, and W be the closed real subspace of Hermitian functions in $L^2(\mathbb{R}^d, \widehat{K}^{-1})$. We will show $f \mapsto \widehat{f}$ is an isomorphism $V \to W$, so we may equivalently work with W. Given $\widehat{f} \in W$, by Cauchy-Schwarz and Bochner's theorem,

$$\int |\widehat{f}| \leq \left(\int \widehat{K} \right)^{1/2} \left(\int |\widehat{f}|^2 / \widehat{K} \right)^{1/2} < \infty,$$

and as $\|\widehat{K}\|_\infty \leq \|K\|_1$,

$$\int |\widehat{f}|^2 \leq \|\widehat{K}\|_\infty \int |\widehat{f}|^2 / \widehat{K} < \infty,$$

so $\widehat{f} \in L^1 \cap L^2$. $\widehat{f}$ is thus the Fourier transform of a real continuous $f \in L^2$, satisfying the Fourier inversion formula everywhere.

$f \mapsto \widehat{f}$ is hence an isomorphism $V \to W$. It remains to show that $V = \mathcal{H}(\mathbb{R}^d)$. W is complete, so V is. Further, $\mathcal{E}(\mathbb{R}^d) \subset V$, and by Fourier inversion each $f \in V$ satisfies the reproducing property,

$$f(x) = \int e^{2\pi i \langle x,\xi\rangle} \widehat{f}(\xi)\, d\xi = \int \frac{\widehat{f}(\xi)\overline{\widehat{k}_x(\xi)}}{\widehat{K}(\xi)}\, d\xi = \langle f, k_x\rangle,$$

so $\mathcal{H}(\mathbb{R}^d)$ is a closed subspace of V. Given $f \in \mathcal{H}(\mathbb{R}^d)^\perp$, $f(x) = \langle f, k_x\rangle = 0$ for all x, so $f = 0$. Thus $V = \mathcal{H}(\mathbb{R}^d)$. □

Proof of Lemma 3. By Lemma 1, the norm on $\mathcal{H}_\theta(\mathbb{R}^d)$ is

$$\|f\|^2_{\mathcal{H}_\theta(\mathbb{R}^d)} = \int \frac{|\widehat{f}(\xi)|^2}{\widehat{K}_\theta(\xi)}\, d\xi,$$

and K_θ has Fourier transform

$$\widehat{K}_\theta(\xi) = \frac{\widehat{K}(\xi_1/\theta_1, \ldots, \xi_d/\theta_d)}{\prod_{i=1}^d \theta_i}.$$

If $\nu < \infty$, by assumption $\widehat{K}(\xi) = \widehat{k}(\|\xi\|)$, for a finite non-increasing function $\widehat{k}$ satisfying $\widehat{k}(\|\xi\|) = \Theta(\|\xi\|^{-2\nu-d})$ as $\xi \to \infty$. Hence

$$C(1+\|\xi\|^2)^{-(\nu+d/2)} \le \widehat{K}_\theta(\xi) \le C'(1+\|\xi\|^2)^{-(\nu+d/2)},$$

for constants $C, C' > 0$, and we obtain that $\mathcal{H}_\theta(\mathbb{R}^d)$ is equivalent to the Sobolev space $H^{\nu+d/2}(\mathbb{R}^d)$.

From Lemma 2, $\mathcal{H}_\theta(D)$ is given by the restriction of functions in $\mathcal{H}_\theta(\mathbb{R}^d)$; as D is Lipschitz, the same is true of $H^{\nu+d/2}$. $\mathcal{H}_\theta(D)$ is thus equivalent to $H^{\nu+d/2}(D)$. Finally, functions in $\mathcal{H}_\theta(\bar{D})$ are continuous, so uniquely identified by their restriction to D, and

$$\mathcal{H}_\theta(\bar{D}) \simeq \mathcal{H}_\theta(D) \simeq H^{\nu+d/2}(D).$$

If $\nu = \infty$, by a similar argument $\mathcal{H}_\theta(\bar{D})$ is continuously embedded in all $H^s(D)$. □

From Lemma 1, we can derive results on the behaviour of $\|f\|_{\mathcal{H}_\theta(S)}$ as θ varies. For small θ, we obtain the following result.

Lemma 4. *If $f \in \mathcal{H}_\theta(S)$, then $f \in \mathcal{H}_{\theta'}(S)$ for all $0 < \theta' \le \theta$, and*

$$\|f\|^2_{\mathcal{H}_{\theta'}(S)} \le \left(\prod_{i=1}^d \theta_i/\theta'_i\right) \|f\|^2_{\mathcal{H}_\theta(S)}.$$

Proof. Let $C = \prod_{i=1}^d (\theta'_i/\theta_i)$. As $\widehat{K}$ is isotropic and radially non-increasing,

$$\widehat{K}_{\theta'}(\xi) = C\widehat{K}_\theta((\theta'_1/\theta_1)\xi_1, \ldots, (\theta'_d/\theta_d)\xi_d) \ge C\widehat{K}_\theta(\xi).$$

Given $f \in \mathcal{H}_\theta(S)$, let $g \in \mathcal{H}_\theta(\mathbb{R}^d)$ be its minimum norm extension, as in Lemma 2. By Lemma 1,

$$\|f\|^2_{\mathcal{H}_{\theta'}(S)} \le \|g\|^2_{\mathcal{H}_{\theta'}(\mathbb{R}^d)} = \int \frac{|\widehat{g}|^2}{\widehat{K}_{\theta'}} \le \int \frac{|\widehat{g}|^2}{C\widehat{K}_\theta} = C^{-1}\|f\|^2_{\mathcal{H}_\theta(S)}.$$ □

Likewise, for large θ, we obtain the following.

Lemma 5. *If $\nu < \infty$, $f \in \mathcal{H}_\theta(S)$, then $f \in \mathcal{H}_{t\theta}(S)$ for $t \geq 1$, and*

$$\|f\|^2_{\mathcal{H}_{t\theta}(S)} \leq C'' t^{2\nu} \|f\|^2_{\mathcal{H}_\theta(S)},$$

for a $C'' > 0$ depending only on K and θ.

Proof. As in the proof of Lemma 3, we have constants $C, C' > 0$ such that

$$C(1+\|\xi\|^2)^{-(\nu+d/2)} \leq \widehat{K}_\theta(\xi) \leq C'(1+\|\xi\|^2)^{-(\nu+d/2)}.$$

Thus for $t \geq 1$,

$$\begin{aligned}\widehat{K}_{t\theta}(\xi) = t^d \widehat{K}_\theta(t\xi) &\geq C t^d (1+t^2\|\xi\|^2)^{-(\nu+d/2)} \\ &\geq C t^{-2\nu} (1+\|\xi\|^2)^{-(\nu+d/2)} \\ &\geq C C'^{-1} t^{-2\nu} \widehat{K}_\theta(\xi),\end{aligned}$$

and we may argue as in the previous lemma. □

We can also describe the posterior distribution of f in terms of $\mathcal{H}_\theta(S)$; as a consequence, we may deduce Corollary 1.

Lemma 6. *Suppose $f(x) = \mu + g(x)$, $g \in \mathcal{H}_\theta(S)$.*

(i) $\hat{f}_n(x;\theta) = \hat{\mu}_n + \hat{g}_n(x)$ solves the optimization problem

$$\text{minimize } \|\hat{g}\|^2_{\mathcal{H}_\theta(S)}, \quad \text{subject to } \hat{\mu} + \hat{g}(x_i) = z_i, \quad 1 \leq i \leq n,$$

with minimum value $\hat{R}_n^2(\theta)$.

(ii) The prediction error satisfies

$$|f(x) - \hat{f}_n(x;\theta)| \leq s_n(x;\theta)\|g\|_{\mathcal{H}_\theta(S)}$$

with equality for some $g \in \mathcal{H}_\theta(S)$.

Proof.

(i) Let $W = \text{span}(k_{x_1}, \ldots, k_{x_n})$, and write $\hat{g} = \hat{g}^\| + \hat{g}^\perp$ for $\hat{g}^\| \in W$, $\hat{g}^\perp \in W^\perp$. $\hat{g}^\perp(x_i) = \langle \hat{g}^\perp, k_{x_i} \rangle = 0$, so $\hat{g}^\perp$ affects the optimization only through $\|\hat{g}\|$. The minimal $\hat{g}$ thus has $\hat{g}^\perp = 0$, so $\hat{g} = \sum_{i=1}^n \lambda_i k_{x_i}$. The problem then becomes

$$\text{minimize } \lambda^T V \lambda, \quad \text{subject to } \hat{\mu} 1 + V\lambda = z.$$

The solution is given by (4) and (5), with value (7).

(ii) By symmetry, the prediction error does not depend on μ, so we may take $\mu = 0$. Then

$$f(x) - \hat{f}_n(x;\theta) = g(x) - (\hat{\mu}_n + \hat{g}_n(x)) = \langle g, e_{n,x} \rangle,$$

for $e_{n,x} = k_x - \sum_{i=1}^n \lambda_i k_{x_i}$, and

$$\lambda = \frac{V^{-1}1}{1^T V^{-1} 1} + \left(I - \frac{V^{-1}1}{1^T V^{-1}1} 1^T \right) V^{-1} \nu.$$

Now, $\|e_{n,x}\|^2_{\mathcal{H}_\theta(S)} = s_n^2(x;\theta)$, as given by (6); this is a consequence of Loève's isometry, but is easily verified algebraically. The result then follows by Cauchy-Schwarz. □

A.2 Fixed Parameters

Proof of Theorem 1. We first establish the lower bound. Suppose we have $2n$ functions ψ_m with disjoint supports. We will argue that, given n observations, we cannot distinguish between all the ψ_m, and thus cannot accurately pick a minimum x_n^*.

To begin with, assume $X = [0,1]^d$. Let $\psi : \mathbb{R}^d \to [0,1]$ be a C^∞ function, supported inside X and with minimum -1. By Lemma 3, $\psi \in \mathcal{H}_\theta(\mathbb{R}^d)$. Fix $k \in \mathbb{N}$, and set $n = (2k)^d/2$. For vectors $m \in \{0,\dots,2k-1\}^d$, construct functions $\psi_m(x) = C(2k)^{-\nu}\psi(2kx-m)$, where $C > 0$ is to be determined. ψ_m is given by a translation and scaling of ψ, so by Lemmas 1, 2 and 5, for some $C' > 0$,

$$\|\psi_m\|_{\mathcal{H}_\theta(X)} \le \|\psi_m\|_{\mathcal{H}_\theta(\mathbb{R}^d)} = C(2k)^{-\nu}\|\psi\|_{\mathcal{H}_{2k\theta}(\mathbb{R}^d)} \le CC'\|\psi\|_{\mathcal{H}_\theta(\mathbb{R}^d)}.$$

Set $C = R/C'\|\psi\|_{\mathcal{H}_\theta(\mathbb{R}^d)}$, so that $\|\psi_m\|_{\mathcal{H}_\theta(X)} \le R$ for all m and k.

Suppose $f = 0$, and let x_n and x_n^* be chosen by any valid strategy u. Set $\chi = \{x_1,\dots,x_{n-1},x_{n-1}^*\}$, and let A_m be the event that $\psi_m(x) = 0$ for all $x \in \chi$. There are n points in χ, and the $2n$ functions ψ_m have disjoint support, so $\sum_m \mathbb{I}(A_m) \ge n$. Thus

$$\sum_m \mathbb{P}_0^u(A_m) = \mathbb{E}_0^u\left[\sum_m \mathbb{I}(A_m)\right] \ge n,$$

and we have some fixed m, depending only on u, for which $\mathbb{P}_0^u(A_m) \ge \frac{1}{2}$. On the event A_m,

$$\psi_m(x_{n-1}^*) - \min\psi_m = C(2k)^{-\nu},$$

but on that event, u cannot distinguish between 0 and ψ_m before time n, so

$$C^{-1}(2k)^\nu \mathbb{E}_{\psi_m}^u[f(x_{n-1}^*) - \min f] \ge \mathbb{P}_{\psi_m}^u(A_m) = \mathbb{P}_0^u(A_m) \ge \tfrac{1}{2}.$$

As the minimax loss is non-increasing in n, for $(2(k-1))^d/2 \le n < (2k)^d/2$ we conclude

$$\begin{aligned}\inf_u L_n(u,\mathcal{H}_\theta(X),R) &\ge \inf_u L_{(2k)^d/2-1}(u,\mathcal{H}_\theta(X),R)\\ &\ge \inf_u \sup_m \mathbb{E}_{\psi_m}^u\left[f\left(x_{(2k)^d/2-1}^*\right) - \min f\right]\\ &\ge \tfrac{1}{2}C(2k)^{-\nu} = \Omega(n^{-\nu/d}).\end{aligned}$$

For general X having non-empty interior, we can find a hypercube $S = x_0 + [0,\varepsilon]^d \subseteq X$, with $\varepsilon > 0$. We may then proceed as above, picking functions ψ_m supported inside S.

For the upper bound, consider a strategy u choosing a fixed sequence x_n, independent of the z_n. Fit a radial basis function interpolant $\hat{f}_n$ to the data, and pick x_n^* to minimize $\hat{f}_n$. Then if x^* minimizes f,

$$\begin{aligned}f(x_n^*) - f(x^*) &\le f(x_n^*) - \hat{f}_n(x_n^*) + \hat{f}_n(x^*) - f(x^*)\\ &\le 2\|\hat{f}_n - f\|_\infty,\end{aligned}$$

so the loss is bounded by the error in $\hat{f}_n$.

From results in Narcowich et al. (2003, §6) and Wendland (2005, §11.5), for suitable radial basis functions the error is uniformly bounded by

$$\sup_{\|f\|_{\mathcal{H}_\theta(X)} \le R} \|\hat{f}_n - f\|_\infty = O(h_n^{-\nu}),$$

where the mesh norm

$$h_n := \sup_{x \in X} \min_{i=1}^{n} \|x - x_i\|.$$

(For $\nu \notin \mathbb{N}$, this result is given by Narcowich et al. for the radial basis function K^ν, which is ν-Hölder at 0 by Abramowitz and Stegun, 1965, §9.6; for $\nu \in \mathbb{N}$, the result is given by Wendland for thin-plate splines.) As X is bounded, we may choose the x_n so that $h_n = O(n^{-1/d})$, giving

$$L_n(u, H_\theta(X), R) = O(n^{-\nu/d}). \qquad \square$$

To prove Theorem 2, we first show that some observations z_n will be well-predicted by past data.

Lemma 7. *Set*

$$\beta := \begin{cases} \alpha, & \nu \leq 1, \\ 0, & \nu > 1. \end{cases}$$

Given $\theta \in \mathbb{R}^d_+$, there is a constant $C' > 0$ depending only on X, K and θ which satisfies the following. For any $k \in \mathbb{N}$, and sequences $x_n \in X$, $\theta_n \geq \theta$, the inequality

$$s_n(x_{n+1}; \theta_n) \geq C' k^{-(\nu \wedge 1)/d} (\log k)^\beta$$

holds for at most k distinct n.

Proof. We first show that the posterior variance s_n^2 is bounded by the distance to the nearest design point. Let π_n denote the prior with variance $\sigma^2 = 1$, and length-scales θ_n. Then for any $i \leq n$, as $\hat{f}_n(x; \theta_n) = \mathbb{E}_{\pi_n}[f(x) \mid \mathcal{F}_n]$,

$$\begin{aligned} s_n^2(x; \theta_n) &= \mathbb{E}_{\pi_n}[(f(x) - \hat{f}_n(x; \theta_n))^2 \mid \mathcal{F}_n] \\ &= \mathbb{E}_{\pi_n}[(f(x) - f(x_i))^2 - (f(x_i) - \hat{f}_n(x; \theta_n))^2 \mid \mathcal{F}_n] \\ &\leq \mathbb{E}_{\pi_n}[(f(x) - f(x_i))^2 \mid \mathcal{F}_n] \\ &= 2(1 - K_{\theta_n}(x - x_i)). \end{aligned}$$

If $\nu \leq \frac{1}{2}$, then by assumption

$$|K(x) - K(0)| = O\left(\|x\|^{2\nu} (-\log \|x\|)^{2\alpha}\right)$$

as $x \to 0$. If $\nu > \frac{1}{2}$, then K is differentiable, so as K is symmetric, $\nabla K(0) = 0$. If further $\nu \leq 1$, then

$$|K(x) - K(0)| = |K(x) - K(0) - x \cdot \nabla K(0)| = O\left(\|x\|^{2\nu} (-\log \|x\|)^{2\alpha}\right).$$

Similarly, if $\nu > 1$, then K is C^2, so

$$|K(x) - K(0)| = |K(x) - K(0) - x \cdot \nabla K(0)| = O(\|x\|^2).$$

We may thus conclude

$$|1 - K(x)| = |K(x) - K(0)| = O\left(\|x\|^{2(\nu \wedge 1)} (-\log \|x\|)^{2\beta}\right),$$

and

$$s_n^2(x;\theta_n) \le C^2\|x-x_i\|^{2(\nu\wedge 1)}(-\log\|x-x_i\|)^{2\beta},$$

for a constant $C>0$ depending only on X, K and θ.

We next show that most design points x_{n+1} are close to a previous x_i. X is bounded, so can be covered by k balls of radius $O(k^{-1/d})$. If x_{n+1} lies in a ball containing some earlier point x_i, $i \le n$, then we may conclude

$$s_n^2(x_{n+1};\theta_n) \le C'^2 k^{-2(\nu\wedge 1)/d}(\log k)^{2\beta},$$

for a constant $C'>0$ depending only on X, K and θ. Hence as there are k balls, at most k points x_{n+1} can satisfy

$$s_n(x_{n+1};\theta_n) \ge C' k^{-(\nu\wedge 1)/d}(\log k)^{\beta}. \qquad \square$$

Next, we provide bounds on the expected improvement when f lies in the RKHS.

Lemma 8. *Let $\|f\|_{\mathcal{H}_\theta(X)} \le R$. For $x \in X$, $n \in \mathbb{N}$, set $I = (f(x_n^*) - f(x))^+$, and $s = s_n(x;\theta)$. Then for*

$$\tau(x) := x\Phi(x) + \phi(x),$$

we have

$$\max\left(I - Rs, \frac{\tau(-R/\sigma)}{\tau(R/\sigma)}I\right) \le EI_n(x;\pi) \le I + (R+\sigma)s.$$

Proof. If $s=0$, then by Lemma 6, $\hat{f}_n(x;\theta) = f(x)$, so $EI_n(x;\pi) = I$, and the result is trivial. Suppose $s>0$, and set $t = (f(x_n^*) - f(x))/s$, $u = (f(x_n^*) - \hat{f}_n(x;\theta))/s$. From (8) and (9),

$$EI_n(x;\pi) = \sigma s \tau(u/\sigma),$$

and by Lemma 6, $|u-t| \le R$. As $\tau'(z) = \Phi(z) \in [0,1]$, τ is non-decreasing, and $\tau(z) \le 1+z$ for $z \ge 0$. Hence

$$EI_n(x;\pi) \le \sigma s \tau\left(\frac{t^+ + R}{\sigma}\right) \le \sigma s\left(\frac{t^+ + R}{\sigma} + 1\right) = I + (R+\sigma)s.$$

If $I=0$, then as EI is the expectation of a non-negative quantity, $EI \ge 0$, and the lower bounds are trivial. Suppose $I>0$. Then as $EI \ge 0$, $\tau(z) \ge 0$ for all z, and $\tau(z) = z + \tau(-z) \ge z$. Thus

$$EI_n(x;\pi) \ge \sigma s \tau\left(\frac{t-R}{\sigma}\right) \ge \sigma s\left(\frac{t-R}{\sigma}\right) = I - Rs.$$

Also, as τ is increasing,

$$EI_n(x;\pi) \ge \sigma\tau\left(\frac{-R}{\sigma}\right)s.$$

Combining these bounds, and eliminating s, we obtain

$$EI_n(x;\pi) \ge \frac{\sigma\tau(-R/\sigma)}{R + \sigma\tau(-R/\sigma)}I = \frac{\tau(-R/\sigma)}{\tau(R/\sigma)}I. \qquad \square$$

We may now prove the theorem. We will use the above bounds to show that there must be times n_k when the expected improvement is low, and thus $f(x_{n_k}^*)$ is close to $\min f$.

Proof of Theorem 2. From Lemma 7 there exists $C > 0$, depending on X, K and θ, such that for any sequence $x_n \in X$ and $k \in \mathbb{N}$, the inequality

$$s_n(x_{n+1};\theta) > Ck^{-(\nu\wedge 1)/d}(\log k)^\beta$$

holds at most k times. Furthermore, $z_n^* - z_{n+1}^* \geq 0$, and for $\|f\|_{\mathcal{H}_\theta(X)} \leq R$,

$$\sum_n z_n^* - z_{n+1}^* \leq z_1^* - \min f \leq 2\|f\|_\infty \leq 2R,$$

so $z_n^* - z_{n+1}^* > 2Rk^{-1}$ at most k times. Since $z_n^* - f(x_{n+1}) \leq z_n^* - z_{n+1}^*$, we have also $z_n^* - f(x_{n+1}) > 2Rk^{-1}$ at most k times. Thus there is a time n_k, $k \leq n_k \leq 3k$, for which $s_{n_k}(x_{n_k+1};\theta) \leq Ck^{-(\nu\wedge 1)/d}(\log k)^\beta$ and $z_{n_k}^* - f(x_{n_k+1}) \leq 2Rk^{-1}$.

Let f have minimum z^* at x^*. For k large, x_{n_k+1} will have been chosen by expected improvement (rather than being an initial design point, chosen at random). Then as z_n^* is non-increasing in n, for $3k \leq n < 3(k+1)$ we have by Lemma 8,

$$\begin{aligned}
z_n^* - z^* &\leq z_{n_k}^* - z^* \\
&\leq \frac{\tau(R/\sigma)}{\tau(-R/\sigma)} EI_{n_k}(x^*;\pi) \\
&\leq \frac{\tau(R/\sigma)}{\tau(-R/\sigma)} EI_{n_k}(x_{n_k+1};\pi) \\
&\leq \frac{\tau(R/\sigma)}{\tau(-R/\sigma)} \left(2Rk^{-1} + C(R+\sigma)k^{-(\nu\wedge 1)/d}(\log k)^\beta\right).
\end{aligned}$$

This bound is uniform in f with $\|f\|_{\mathcal{H}_\theta(X)} \leq R$, so we obtain

$$L_n(EI(\pi), \mathcal{H}_\theta(X), R) = O(n^{-(\nu\wedge 1)/d}(\log n)^\beta). \qquad \square$$

A.3 Estimated Parameters

To prove Theorem 3, we first establish lower bounds on the posterior variance.

Lemma 9. *Given $\theta^L, \theta^U \in \mathbb{R}_+^d$, pick sequences $x_n \in X$, $\theta^L \leq \theta_n \leq \theta^U$. Then for open $S \subset X$,*

$$\sup_{x\in S} s_n(x;\theta_n) = \Omega(n^{-\nu/d}),$$

uniformly in the sequences x_n, θ_n.

Proof. S is open, so contains a hypercube T. For $k \in \mathbb{N}$, let $n = \frac{1}{2}(2k)^d$, and construct $2n$ functions ψ_m on T with $\|\psi_m\|_{\mathcal{H}_{\theta^U}(X)} \leq 1$, as in the proof of Theorem 1. Let $C^2 = \prod_{i=1}^d (\theta_i^U/\theta_i^L)$; then by Lemma 4, $\|\psi_m\|_{\mathcal{H}_{\theta_n}(X)} \leq C$.

Given n design points $x_1, \ldots, x_n$, there must be some ψ_m such that $\psi_m(x_i) = 0$, $1 \leq i \leq n$. By Lemma 6, the posterior mean of ψ_m given these observations is the zero function. Thus for $x \in T$ minimizing ψ_m,

$$s_n(x;\theta_n) \geq C^{-1} s_n(x;\theta_n)\|\psi_m\|_{\mathcal{H}_{\theta_n}(X)} \geq C^{-1}|\psi_m(x) - 0| = \Omega(k^{-\nu}).$$

As $s_n(x;\theta)$ is non-increasing in n, for $\frac{1}{2}(2(k-1))^d < n \leq \frac{1}{2}(2k)^d$ we obtain

$$\sup_{x\in S} s_n(x;\theta_n) \geq \sup_{x\in S} s_{\frac{1}{2}(2k)^d}(x;\theta_n) = \Omega(k^{-\nu}) = \Omega(n^{-\nu/d}). \qquad \square$$

Next, we bound the expected improvement when prior parameters are estimated by maximum likelihood.

Lemma 10. *Let $\|f\|_{\mathcal{H}_{\theta^U}(X)} \leq R$, $x_n, y_n \in X$. Set $I_n(x) = z_n^* - f(x)$, $s_n(x) = s_n(x;\hat{\theta}_n)$, and $t_n(x) = I_n(x)/s_n(x)$. Suppose:*

(i) for some $i < j$, $z_i \neq z_j$;

(ii) for some $T_n \to -\infty$, $t_n(x_{n+1}) \leq T_n$ whenever $s_n(x_{n+1}) > 0$;

(iii) $I_n(y_{n+1}) \geq 0$; and

(iv) for some $C > 0$, $s_n(y_{n+1}) \geq e^{-C/c_n}$.

Then for $\hat{\pi}_n$ as in Definition 2, eventually $EI_n(x_{n+1};\hat{\pi}_n) < EI_n(y_{n+1};\hat{\pi}_n)$. If the conditions hold on a subsequence, so does the conclusion.

Proof. Let $\hat{R}_n^2(\theta)$ be given by (7), and set $\hat{R}_n^2 = \hat{R}_n^2(\hat{\theta}_n)$. For $n \geq j$, $\hat{R}_n^2 > 0$, and by Lemma 4 and Corollary 1,

$$\hat{R}_n^2 \leq \|f\|^2_{\mathcal{H}_{\hat{\theta}_n}(X)} \leq S^2 = R^2 \prod_{i=1}^{d} (\theta_i^U/\theta_i^L).$$

Thus $0 < \hat{\sigma}_n^2 \leq S^2 c_n$. Then if $s_n(x) > 0$, for some $|u_n(x) - t_n(x)| \leq S$,

$$EI_n(x;\hat{\pi}_n) = \hat{\sigma}_n s_n(x) \tau(u_n(x)/\hat{\sigma}_n),$$

as in the proof of Lemma 8.

If $s_n(x_{n+1}) = 0$, then $x_{n+1} \in \{x_1, \ldots, x_n\}$, so

$$EI_n(x_{n+1};\hat{\pi}_n) = 0 < EI_n(y_{n+1};\hat{\pi}_n).$$

When $s_n(x_{n+1}) > 0$, as τ is increasing we may upper bound $EI_n(x_{n+1};\hat{\pi}_n)$ using $u_n(x_{n+1}) \leq T_n + S$, and lower bound $EI_n(y_{n+1};\hat{\pi}_n)$ using $u_n(y_{n+1}) \geq -S$. Since $s_n(x_{n+1}) \leq 1$, and $\tau(x) = \Theta(x^{-2}e^{-x^2/2})$ as $x \to -\infty$ (Abramowitz and Stegun, 1965, §7.1),

$$\begin{aligned}
\frac{EI_n(x_{n+1};\hat{\pi}_n)}{EI_n(y_{n+1};\hat{\pi}_n)} &\leq \frac{\tau((T_n+S)/\hat{\sigma}_n)}{e^{-C/c_n}\tau(-S/\hat{\sigma}_n)} \\
&= O\left((T_n+S)^{-2} e^{C/c_n - (T_n^2+2ST_n)/2\hat{\sigma}_n^2}\right) \\
&= O\left((T_n+S)^{-2} e^{-(T_n^2+2ST_n-2CS^2)/2S^2c_n}\right) \\
&= o(1).
\end{aligned}$$

If the conditions hold on a subsequence, we may similarly argue along that subsequence. □

Finally, we will require the following technical lemma.

Lemma 11. *Let $x_1, \ldots, x_n$ be random variables taking values in $\mathbb{R}^d$. Given open $S \subseteq \mathbb{R}^d$, there exist open $U \subseteq S$ for which $\mathbb{P}(\bigcup_{i=1}^n \{x_i \in U\})$ is arbitrarily small.*

Proof. Given $\varepsilon > 0$, fix $m \geq n/\varepsilon$, and pick disjoint open sets $U_1, \ldots, U_m \subset S$. Then

$$\sum_{j=1}^{m} \mathbb{E}[\#\{x_i \in U_j\}] \leq \mathbb{E}[\#\{x_i \in \mathbb{R}^d\}] = n,$$

so there exists U_j with

$$\mathbb{P}\left(\bigcup_i \{x_i \in U_j\}\right) \leq \mathbb{E}[\#\{x_i \in U_j\}] \leq n/m \leq \varepsilon. \qquad \square$$

We may now prove the theorem. We will construct a function f on which the $EI(\hat{\pi})$ strategy never observes within a region W. We may then construct a function g, agreeing with f except on W, but having different minimum. As the strategy cannot distinguish between f and g, it cannot successfully find the minimum of both.

Proof of Theorem 3. Let the $EI(\hat{\pi})$ strategy choose initial design points $x_1, \ldots, x_k$, independently of f. Given $\varepsilon > 0$, by Lemma 11 there exists open $U_0 \subseteq X$ for which $\mathbb{P}^{EI(\hat{\pi})}(x_1, \ldots, x_k \in \overline{U_0}) \leq \varepsilon$; we may choose U_0 so that $V_0 = X \setminus U_0$ has non-empty interior. Pick open U_1 such that $V_1 = \overline{U_1} \subset U_0$, and set f to be a C^∞ function, 0 on V_0, 1 on V_1, and everywhere non-negative. By Lemma 1, $f \in \mathcal{H}_{\theta^U}(X)$.

We work conditional on the event A, having probability at least $1-\varepsilon$, that $z_k^* = 0$, and thus $z_n^* = 0$ for all $n \geq k$. Suppose $x_n \in V_1$ infinitely often, so the z_n are not all equal. By Lemma 7, $s_n(x_{n+1}; \hat{\theta}_n) \to 0$, so on a subsequence with $x_{n+1} \in V_1$, we have

$$t_n = (z_n^* - f(x_{n+1}))/s_n(x_{n+1}; \hat{\theta}_n) = -s_n(x_{n+1}; \hat{\theta}_n)^{-1} \to -\infty$$

whenever $s_n(x_{n+1}; \hat{\theta}_n) > 0$. However, by Lemma 9, there are points $y_n \in V_0$ with $z_n^* - f(y_{n+1}) = 0$, and $s_n(y_{n+1}; \hat{\theta}_n) = \Omega(n^{-\nu/d})$. Hence by Lemma 10, $EI_n(x_{n+1}; \hat{\pi}_n) < EI_n(y_{n+1}; \hat{\pi}_n)$ for some n, contradicting the definition of x_{n+1}.

Hence, on A, there is a random variable T taking values in $\mathbb{N}$, for which $n > T \implies x_n \notin V_1$. Hence there exists a constant $t \in \mathbb{N}$ for which the event $B = A \cap \{T \leq t\}$ has $\mathbb{P}_f^{EI(\hat{\pi})}$-probability at least $1-2\varepsilon$. By Lemma 11, we thus have an open set $W \subset V_1$ for which the event

$$C = B \cap \{x_n \notin W : n \in \mathbb{N}\} = B \cap \{x_n \notin W : n \leq t\}$$

has $\mathbb{P}_f^{EI(\hat{\pi})}$-probability at least $1-3\varepsilon$.

Construct a smooth function g by adding to f a C^∞ function which is 0 outside W, and has minimum -2. Then $\min g = -1$, but on the event C, $EI(\hat{\pi})$ cannot distinguish between f and g, and $g(x_n^*) \geq 0$. Thus for $\delta = 1$,

$$\mathbb{P}_g^{EI(\hat{\pi})}\left(\inf_n g(x_n^*) - \min g \geq \delta\right) \geq \mathbb{P}_g^{EI(\hat{\pi})}(C) = \mathbb{P}_f^{EI(\hat{\pi})}(C) \geq 1 - 3\varepsilon.$$

As the behaviour of $EI(\hat{\pi})$ is invariant under rescaling, we may scale g to have norm $\|g\|_{\mathcal{H}_\theta(X)} \leq R$, and the above remains true for some $\delta > 0$. $\square$

Proof of Theorem 4. As in the proof of Theorem 2, we will show there are times n_k when the expected improvement is small, so $f(x_{n_k})$ must be close to the minimum. First, however, we must control the estimated parameters $\hat{\sigma}_n^2$, $\hat{\theta}_n$.

If the z_n are all equal, then by assumption the x_n are dense in X, so f is constant, and the result is trivial. Suppose the z_n are not all equal, and let T be a random variable satisfying $z_T \neq z_i$ for some $i < T$. Set $U = \inf_{\theta^L \leq \theta \leq \theta^U} \hat{R}_T(\theta)$. $\hat{R}_T(\theta)$ is a continuous positive function, so $U > 0$. Let $S^2 = R^2 \prod_{i=1}^d (\theta_i^U / \theta_i^L)$. By Lemma 4, $\|f\|_{\mathcal{H}_{\hat{\theta}_n}(X)} \leq S$, so by Corollary 1, for $n \geq T$,

$$U \leq \hat{R}_T(\hat{\theta}_n) \leq \hat{\sigma}_n \leq \|f\|_{\mathcal{H}_{\hat{\theta}_n}(X)} \leq S.$$

As in the proof of Theorem 2, we have a constant $C > 0$, and some n_k, $k \leq n_k \leq 3k$, for which $z^*_{n_k} - f(x_{n_k+1}) \leq 2Rk^{-1}$ and $s_{n_k}(x_{n_k+1}; \hat{\theta}_{n_k}) \leq Ck^{-\alpha}(\log k)^\beta$. Then for $k \geq T$, $3k \leq n < 3(k+1)$, arguing as in Theorem 2 we obtain

$$\begin{aligned} z^*_n - z^* &\leq z^*_{n_k} - z^* \\ &\leq \frac{\tau(S/\hat{\sigma}_{n_k})}{\tau(-S/\hat{\sigma}_{n_k})} \left(2Rk^{-1} + C(S + \hat{\sigma}_{n_k})k^{-(\nu \wedge 1)/d}(\log k)^\beta\right) \\ &\leq \frac{\tau(S/U)}{\tau(-S/U)} \left(2Rk^{-1} + 2CSk^{-(\nu \wedge 1)/d}(\log k)^\beta\right). \end{aligned}$$

We thus have a random variable C' satisfying $z^*_n - z^* \leq C' n^{-(\nu \wedge 1)/d}(\log n)^\beta$ for all n, and the result follows. □

A.4 Near-Optimal Rates

To prove Theorem 5, we first show that the points chosen at random will be quasi-uniform in X.

Lemma 12. *Let x_n be i.i.d. random variables, distributed uniformly over X, and define their mesh norm,*

$$h_n := \sup_{x \in X} \min_{i=1}^n \|x - x_i\|.$$

For any $\gamma > 0$, there exists $C > 0$ such that

$$\mathbb{P}(h_n > C(n/\log n)^{-1/d}) = O(n^{-\gamma}).$$

Proof. We will partition X into n regions of size $O(n^{-1/d})$, and show that with high probability we will place an x_i in each one. Then every point x will be close to an x_i, and the mesh norm will be small.

Suppose $X = [0,1]^d$, fix $k \in \mathbb{N}$, and divide X into $n = k^d$ sub-cubes $X_m = \frac{1}{k}(m + [0,1]^d)$, for $m \in \{0, \ldots, k-1\}^d$. Let I_m be the indicator function of the event

$$\{x_i \notin X_m : 1 \leq i \leq \lfloor \gamma n \log n \rfloor\},$$

and define

$$\mu_n = \mathbb{E}\left[\sum_m I_m\right] = n\mathbb{E}[I_0] = n(1 - 1/n)^{\lfloor \gamma n \log n \rfloor} \sim ne^{-\gamma \log n} = n^{-(\gamma - 1)}.$$

For n large, $\mu_n \leq 1$, so by the generalized Chernoff bound of Panconesi and Srinivasan (1997, §3.1),

$$\mathbb{P}\left(\sum_m I_m \geq 1\right) \leq \left(\frac{e^{(\mu_n^{-1} - 1)}}{\mu_n^{-\mu_n^{-1}}}\right)^{\mu_n} \leq e\mu_n \sim en^{-(\gamma - 1)}.$$

On the event $\sum_m I_m < 1$, $I_m = 0$ for all m. For any $x \in X$, we then have $x \in X_m$ for some m, and $x_j \in X_m$ for some $1 \le j \le \lfloor \gamma n \log n \rfloor$. Thus

$$\min_{i=1}^{\lfloor \gamma n \log n \rfloor} \|x - x_i\| \le \|x - x_j\| \le \sqrt{d} k^{-1}.$$

As this bound is uniform in x, we obtain $h_{\lfloor \gamma n \log n \rfloor} \le \sqrt{d} k^{-1}$. Thus for $n = k^d$,

$$\mathbb{P}(h_{\lfloor \gamma n \log n \rfloor} > \sqrt{d} k^{-1}) = O(k^{-d(\gamma-1)}),$$

and as h_n is non-increasing in n, this bound holds also for $k^d \le n < (k+1)^d$. By a change of variables, we then obtain

$$\mathbb{P}(h_n > C(n/\gamma \log n)^{-1/d}) = O((n/\gamma \log n)^{-(\gamma-1)}),$$

and the result follows by choosing γ large. For general X, as X is bounded it can be partitioned into n regions of measure $\Theta(n^{-1/d})$, so we may argue similarly. □

We may now prove the theorem. We will show that the points x_n must be quasi-uniform in X, so posterior variances must be small. Then, as in the proofs of Theorems 2 and 4, we have times when the expected improvement is small, so $f(x_n^*)$ is close to $\min f$.

Proof of Theorem 5. First suppose $\nu < \infty$. Let the $EI(\cdot, \varepsilon)$ choose k initial design points independent of f, and suppose $n \ge 2k$. Let A_n be the event that $\lfloor \frac{\varepsilon}{4} n \rfloor$ of the points $x_{k+1}, \ldots, x_n$ are chosen uniformly at random, so by a Chernoff bound,

$$\mathbb{P}^{EI(\cdot,\varepsilon)}(A_n^c) \le e^{-\varepsilon n/16}.$$

Let B_n be the event that one of the points $x_{n+1}, \ldots, x_{2n}$ is chosen by expected improvement, so

$$\mathbb{P}^{EI(\cdot,\varepsilon)}(B_n^c) = \varepsilon^n.$$

Finally, let C_n be the event that A_n and B_n occur, and further the mesh norm $h_n \le C(n/\log n)^{-1/d}$, for the constant C from Lemma 12. Set $r_n = (n/\log n)^{-\nu/d} (\log n)^{\alpha}$. Then by Lemma 12, since $C_n \subset A_n$,

$$\mathbb{P}_f^{EI(\cdot,\varepsilon)}(C_n^c) \le C' r_n,$$

for a constant $C' > 0$ not depending on f.

Let $EI(\cdot, \varepsilon)$ have prior π_n at time n, with (fixed or estimated) parameters σ_n, θ_n. Suppose $\|f\|_{\mathcal{H}_{\theta^U}(X)} \le R$, and set $S^2 = R^2 \prod_{i=1}^d (\theta_i^U / \theta_i^L)$, so by Lemma 4, $\|f\|_{\mathcal{H}_{\theta_n}(X)} \le S$. If $\alpha = 0$, then by Narcowich et al. (2003, §6),

$$\sup_{x \in X} s_n(x; \theta) = O(M(\theta) h_n^{\nu})$$

uniformly in θ, for $M(\theta)$ a continuous function of θ. Hence on the event C_n,

$$\sup_{x \in X} s_n(x; \theta_n) \le \sup_{x \in X} \sup_{\theta^L \le \theta \le \theta^U} s_n(x; \theta) \le C'' r_n,$$

for a constant $C'' > 0$ depending only on X, K, C, θ^L and θ^U. If $\alpha > 0$, the same result holds by a similar argument.

On the event C_n, we have some x_m chosen by expected improvement, $n < m \leq 2n$. Let f have minimum z^* at x^*. Then by Lemma 8,

$$\begin{aligned} z^*_{m-1} - z^* &\leq EI_{m-1}(x^*; \cdot) + C'' S r_{m-1} \\ &\leq EI_{m-1}(x_m; \cdot) + C'' S r_{m-1} \\ &\leq (f(x_{m-1}) - f(x_m))^+ + C''(2S + \sigma_{m-1}) r_{m-1} \\ &\leq z^*_{m-1} - z^*_m + C'' T r_n, \end{aligned}$$

for a constant $T > 0$. (Under $EI(\pi, \varepsilon)$, we have $T = 2S + \sigma$; otherwise $\sigma_{m-1} \leq S$ by Corollary 1, so $T = 3S$.) Thus, rearranging,

$$z^*_{2n} - z^* \leq z^*_m - z^* \leq C'' T r_n.$$

On the event C_n^c, we have $z^*_{2n} - z^* \leq 2\|f\|_\infty \leq 2R$, so

$$\begin{aligned} \mathbb{E}_f^{EI(\cdot,\varepsilon)}[z^*_{2n+1} - z^*] &\leq \mathbb{E}_f^{EI(\cdot,\varepsilon)}[z^*_{2n} - z^*] \\ &\leq 2R \mathbb{P}_f^{EI(\cdot,\varepsilon)}(C_n^c) + C'' T r_n \\ &\leq (2C'R + C'' T) r_n. \end{aligned}$$

As this bound is uniform in f with $\|f\|_{\mathcal{H}_{\theta^U}(X)} \leq R$, the result follows. If instead $\nu = \infty$, the above argument holds for any $\nu < \infty$. □

References

Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1965.

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, New York, 2007.

Nachman Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, Massachusetts, 2004.

Eric Brochu, Mike Cora, and Nando de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Arxiv preprint arXiv:1012.2599, 2010.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Proc. 20th International Conference on Algorithmic Learning Theory (ALT '09)*, pages 23–37, Porto, Portugal, 2009.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. X-armed bandits. Arxiv preprint arXiv:1001.4475, 2010.

Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1): 1–26, 1986.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS J. Comput.*, 21(4):599–613, 2009.

David Ginsbourger, Rodolphe le Riche, and Laurent Carraro. A multi-points criterion for deterministic parallel global optimization based on Gaussian processes. HAL preprint hal-00260579, 2008.

Steffen Grunewalder, Jean-Yves Audibert, Manfred Opper, and John Shawe-Taylor. Regret bounds for Gaussian process bandit problems. In *Proc. 13th International Conference on Artificial Intelligence and Statistics (AISTATS '10)*, pages 273–280, Sardinia, Italy, 2010.

Pierre Hansen, Brigitte Jaumard, and Shi-Hui Lu. Global optimization of univariate Lipschitz functions: I. survey and properties. *Math. Program.*, 55(1):251–272, 1992.

Donald R. Jones, Cary D. Perttunen, and Bruce E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *J. Optim. Theory Appl.*, 79(1):157–181, 1993.

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

Robert Kleinberg and Aleksandrs Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA '10)*, pages 827–846, Austin, Texas, 2010.

Marco Locatelli. Bayesian algorithms for one-dimensional global optimization. *J. Global Optim.*, 10(1):57–76, 1997.

William G. Macready and David H. Wolpert. Bandit problems and the exploration / exploitation tradeoff. *IEEE Trans. Evol. Comput.*, 20(1):2–22, 1998.

Jonas Močkus. On Bayesian methods for seeking the extremum. In *Proc. IFIP Technical Conference*, pages 400–404, Novosibirsk, Russia, 1974.

Francis J. Narcowich, Joseph D. Ward, and Holger Wendland. Refined error estimates for radial basis function interpolation. *Constr. Approx.*, 19(4):541–564, 2003.

Michael Osborne. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. DPhil thesis, University of Oxford, Oxford, UK, 2010.

Alessandro Panconesi and Aravind Srinivasan. Randomized distributed edge coloring via an extension of the Chernoff-Hoeffding bounds. *SIAM J. Comput.*, 26(2):350–368, 1997.

Panos M. Pardalos and H. Edwin Romeijn, editors. *Handbook of Global Optimization, Volume 2*. Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht, the Netherlands, 2002.

Emanuel Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In *Proc. Symposium on Time Series Analysis*, pages 155–169, Providence, Rhode Island, 1963.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006.

Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer, New York, 2003.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proc. 27th International Conference on Machine Learning (ICML '10)*, Haifa, Israel, 2010.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: an Introduction*. MIT Press, Cambridge, Massachusetts, 1998.

Luc Tartar. *An Introduction to Sobolev Spaces and Interpolation Spaces*, volume 3 of *Lecture Notes of the Unione Matematica Italiana*. Springer, New York, 2007.

Aad W. van der Vaart and J. Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3 of *Institute of Mathematical Statistics Collections*, pages 200–222. Institute of Mathematical Statistics, Beachwood, Ohio, 2008.

Aad W. van der Vaart and J. Harry van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

Emmanuel Vazquez and Julien Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *J. Statist. Plann. Inference*, 140(11):3088–3095, 2010.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, UK, 2005.