Optimization Under Unknown Constraints

Robert B. Gramacy & Herbert K. H. Lee

University of Cambridge, UK   University of California, Santa Cruz, USA

bobby@statslab.cam.ac.uk   herbie@ams.ucsc.edu

Summary

Optimization of complex functions, such as the output of computer simulators, is a difficult task that has received much attention in the literature. A less studied problem is that of optimization under unknown constraints, i.e., when the simulator must be invoked both to determine the typical real-valued response *and* to determine if a constraint has been violated, either for physical or policy reasons. We develop a statistical approach based on Gaussian processes and Bayesian learning to *both* approximate the unknown function and estimate the probability of meeting the constraints. A new integrated improvement criterion is proposed to recognize that responses from inputs that violate the constraint may still be informative about the function, and thus could potentially be useful in the optimization. The new criterion is illustrated on synthetic data, and on a motivating optimization problem from health care policy.

Keywords and Phrases: Constrained Optimization, Surrogate Model, Gaussian Process, Sequential Design, Expected Improvement

1. INTRODUCTION

A common optimization problem that arises in fields ranging from applied engineering to public policy is to find $x^* = \arg\min_{x \in \mathcal{X}} f(x)$, subject to constraints: $x^* \in C$, where we may only learn about the relationship between x and $f(x) : \mathcal{X} \to \mathbb{R}$ and the constraint region C through expensive evaluations of the noisy joint process

$$Z(x) = f(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \eta^2) \tag{1}$$
$$C(x) = c(x + \varepsilon_c) = \mathbb{I}_{\{x + \varepsilon_c \in C\}} \in \{0, 1\}.$$

The real-valued noise variance, η^2, is unknown but may be zero, and ε_c indicates that the constraint mapping may be random. In particular, the constraint region

$C \subset \mathcal{X}$ is well-defined but often non-trivial. Although it will typically be deterministic ($\varepsilon_c = 0$), this is not required by our treatment. Finally, we suppose that observing the joint response $(Z, C)(x)$ is expensive. So we wish to keep the number of evaluations, $(x_1, z_1, c_1), \ldots (x_N, z_N, c_N)$, small. One way to do this is to build regression and classification models $f_N(x)$ for $f(x)$ and $c_N(x)$ for $c(x)$ based on the data. The surrogate surfaces may be searched to find x' yielding a small objective in expectation, and satisfying the constraint with high probability. We can then repeat the process with $N + 1$ points, including $(x', Z(x'), C(x'))$, stopping when convergence in the location of x^* is achieved, or when resources are exhausted.

To shed light upon the difficulty in solving this problem, and to thereby suggest possible points of attack, consider the following simplification where the constraint region C is known at the outset (i.e., there is no need to estimate c_N). In this case a sensible approach is as follows. Obtain realizations $z(x)$ of $Z(x)$ only for $x \in C$ with the largest *expected improvement* (EI, Jones et al., 1998) under f_N [more on this in Section 2] and proceed to construct f_{N+1} by adding in the $(x, z(x))$ pair into the design. This presumes that evaluating $f(x)$ for $x \in \mathcal{X} \setminus C$ is a waste of resources. But this need not be so, since $Z(x)$, for *any* x, contains information about f, and therefore about promising location(s) for $x^* \in C$. It could even be that $x' \notin C$ is best at reducing the overall uncertainty in the location of $x^* \in C$, through an improved new surrogate f_{N+1}. When this is the case [e.g., see Section 3.3] it makes sense to sample $Z(x')$ for $x' \notin C$ despite the constraint violation.

Assessing when this odd maneuver is advantageous requires a more global notion of improvement; EI cannot directly quantify the extent to which $x' \notin C$ improves our information at $x \in C$. Finally, when C is not known *a priori*, new evaluations $(x', z' = z(x'))$ provide information about both f and c through their surrogates f_N and c_N. Thus incremental decisions toward solving the constrained optimization problem must incorporate uncertainty from both surrogates. We propose a new integrated improvement statistic to fit the bill.

The rest of the paper is outlined as follows. In Section 2 we outline EI for (unconstrained) optimization and the GP surrogate models upon which it is based. In Section 3 we develop the conditional and integrated expected improvement statistic(s) for the case of known constraints, with an illustration. We extend the method to unknown constraints in Section 4, and demonstrate the resulting constrained optimization algorithm on synthetic data. In Section 5 we consider a motivating problem from health care policy research, and conclude with some discussion and extensions in Section 6.

2. PREVIOUS WORK

2.1. *Surrogate Modeling*

The canonical choice of surrogate model for computer experiments is the stationary Gaussian process (GP, Sacks et al., 1989; O'Hagan et al., 1999; Santner et al., 2003), which is one way of characterizing a zero mean random process where covariance $C(x, x') = \sigma^2 K(x, x')$ varies spatially depending upon the x locations. Let $Z_N = (z_1, \ldots, z_N)^T$ be the vector of observed responses at the design points $x_1, \ldots, x_N$ collected (row-wise) in X_N. Conditional on this data $D_N = \{X_N, Z_N\}$, the (posterior) predictive distribution of $Z(x)$ at a new point x under the GP is normal with

$$
\begin{array}{rcll}
\text{mean} & \hat{z}_N(x) & = & k_N^T K_N^{-1} Z_N, \\
\text{and variance} & \hat{\sigma}_N^2(x) & = & \sigma^2[K(x,x) - k_N^T(x)K_N^{-1}k_N(x)],
\end{array}
\tag{2}
$$

where $k_N^T(x)$ is the N-vector whose i^{th} component is $K(x,x_i)$, and K_N is the $N \times N$ matrix with i,j element $K(x_i,x_j)$. These are sometimes called the *kriging equations*. Joint prediction at a collection of points X is multivariate normal with mean vector $\hat{z}_N(X)$ and covariance matrix $\hat{\Sigma}_N(X)$ which are defined by the straightforward matrix extension of $k_N(X)$ and $K(X,X)$. We follow Gramacy and Lee (2008) in specifying that $K(\cdot,\cdot)$ have the form $K(x,x'|g) = K^*(x,x') + \eta\delta_{x,x'}$, where $\delta_{.,.}$ is the Kronecker delta function, and K^* is a *true* correlation function. The η term, referred to as the *nugget*, is positive ($\eta > 0$) and provides a mechanism for introducing measurement error into the stochastic process—implementing $\eta^2 > 0$ in Eq. (1) (Gramacy, 2005, appendix). It causes the predictive equations (2) to smooth rather than interpolate the data (X_N, Z_N). It is common to take $K^*(\cdot,\cdot)$ from a parametric family, such as the separable Matérn or power families (e.g., Abrahamsen, 1997), which roughly model $K^*(\cdot,\cdot)$ as an inverse function of coordinate-wise Euclidean distance. We prefer the power family, which is standard for computer experiments.

2.2. Optimization by Expected Improvement

Conditional on a GP surrogate f_N, a step towards finding the minimum may be based upon the *expected improvement* (EI) statistic (Jones et al., 1998). For a deterministic function ($\eta = 0$), the current minimum $f_{\min} = \min\{z_1, \ldots, z_N\}$ is deterministic. In this case, the improvement is defined as $I(x) = \max\{f_{\min} - Z(x), 0\}$. The next location is chosen as

$$
x' = \arg\max_{x \in \mathcal{X}} \mathbb{E}\{I(x)\},
\tag{3}
$$

where the expectation is taken over $Z(x) \sim F_N(x)$, the predictive distribution (2) implied by f_N evaluated at x. Jones et al. (1998) give an analytical expression for the EI:

$$
\mathbb{E}\{I(x)\} = (f_{\min} - \hat{z}_N(x))\Phi\left(\frac{f_{\min} - \hat{z}_N(x)}{\hat{\sigma}_N(x)}\right) + \hat{\sigma}_N(x)\phi\left(\frac{f_{\min} - \hat{z}_N(x)}{\hat{\sigma}_N(x)}\right).
\tag{4}
$$

Basically, the EI is the cumulative distribution of the predictive density that lies "underneath" $f_{\min}$. A relevant diagram illustrating EI appears in Figure 1 in Section 3.1.1. Jones et al. (1998) also provide a branch and bound algorithm for performing the maximization over $\mathcal{X}$ to find x'. Once x' is chosen it is added into the design as $(x_{N+1}, z_{N+1}) = (x', f(x'))$ and the procedure repeats with f_{N+1}. Jones et al. (1998) use maximum likelihood inference to set the parameters for f_N, i.e., d only since $\eta = 0$, and call the resulting iterative procedure the *efficient global optimization* (EGO) algorithm. The above choice of $f_{\min}$ is sensible but somewhat arbitrary. Another reasonable choice that we promote in this paper is $f_{\min} = \min \hat{z}_N(x)$, the minimum of the (posterior) mean predictive surface.

The situation is more complicated for noisy responses. We must then estimate the nugget, η, and extend the Jones et al. (1998) definition of $f_{\min}$ to be a random variable: the first order statistic of $Z_1, \ldots, Z_N$. Calculating the EI would thus

require integrating over $f_{\min}$ in Eq. (3). This breaks the analytical tractability of EGO algorithm, however one can always proceed by Monte Carlo methods. Once in the Monte Carlo framework, extensions abound. For example, it is trivial to take a Bayesian approach and thereby factor parameter uncertainty into the EI calculation. Conditional on the parameters however, choosing $f_{\min} \min \hat{z}_N(x)$ is still deterministic. So this choice allows an analytical approach to proceed when point-estimates (i.e., MLEs) of parameters are used, or it leads to a more efficient Monte Carlo algorithm when sampling from the Bayesian posterior. The downside of the Monte Carlo approach, whether taken for Bayesian or $f_{\min}$ considerations, is that the branch and bound algorithm for determining x' in Eq. (3) is no longer available. However, proceeding with a discrete set of space-filling candidates, and leveraging direct optimization methods in tandem, has proved fruitful (Taddy et al., 2009b).

2.3. *Towards Constrained Optimization*

Ours in not the first attempt at tackling the constrained optimization problem via surrogate modeling. Schonlau et al. (1998) consider deterministic responses ($\eta = 0$) where the known constraint region can be written as $a_k \leq c_k(x) \leq b_k$, for $k = 1, \ldots, K$. They then treat the $c_k(x)$ as additional response variables that co-vary with $f(x)$. This breaks the analytical tractability of the EI calculation. Assuming that the $K+1$ responses are independent the calculation is again tractable, otherwise a Monte Carlo approach is needed. We are not aware of any previous literature addressing our more general problem: where the function f may not be deterministic, and when there are unknown constraints of arbitrary form. Even in simpler settings, like the one above, it may be advantageous to sample outside the constraint region. This requires a new improvement statistic—one that weighs the overall expected improvement of the next sequentially chosen design point in aggregate.

3. INTEGRATED EXPECTED CONDITIONAL IMPROVEMENT

Here we generalize the EI framework to accommodate the drawbacks outlined above. To start with, we assume that constraints are deterministic, and known (with trivial computation) in advance. Section 4 provides extensions for unknown constraints.

Define the *conditional improvement* as

$$I(y|x) = \max\{f_{\min} - Z(y|x), 0\}, \tag{5}$$

where $Z(y|x) \sim F_N(y|x)$, which is the predictive distribution of the response $Z(y)$ at a *reference* input location y under the surrogate model f_N given that the *candidate* location x is added into the design. We do not use an $N + 1$ subscript for the posterior predictive distribution because the realization of the response $z(x)$ is not yet available.

The *expected conditional improvement* (ECI) at the reference point y is then $\mathbb{E}\{I(y|x)\}$. Here the expectation is over all of the random quantities: the distribution of $Z(y|x)$, and perhaps of $f_{\min}$ depending upon how it is defined. The ECI may be evaluated at all pairs of inputs $(x, y) \in \mathcal{X}$. The potential to generalize EI, which accounts for improvement at the point x alone, comes by integrating over the choices for y. Let $g(y)$ denote a density over $y \in \mathcal{X}$ which may be uniform in a bounded region. Then the *integrated expected conditional improvement* (IECI) is defined as

$$\mathbb{E}_g\{I(x)\} = -\int_{\mathcal{X}} \mathbb{E}\{I(y|x)\}g(y)\, dy. \tag{6}$$

This suggests using $x' = \arg\max_{x \in \mathcal{X}} \mathbb{E}_g\{I(x)\}$ as the next adaptively sampled point. As long as $\mathbb{E}\{I(y|x)\} \leq \mathbb{E}\{I(y)\}$ for all $x \in \mathcal{X}$, this statistic (6) is defensible. Defining $f_{\min}$ carefully [see Section 3.1.1] ensures that this *monotonicity condition* holds.

The negation in Eq. (6) keeps IECI in line with the convention of maximizing, i.e., of preferring large EI statistics over small ones. To explain, consider how $I(y|x)$ "looks ahead". We wish to measure an improvement at x, but in a roundabout way we assess that improvement at a reference point y instead, supposing x has been added into the design. If y still has high improvement potential after x has been added in, then x must not have had much influence on the improvement at y. If x is influential at y, then the improvement at y should be small after x is added in, not large.

We can alternatively define IECI as the expected reduction in improvement at the reference location, y, when x is added into the design:

$$\mathbb{E}_g\{I(x)\} = \int_{\mathcal{X}} (\mathbb{E}\{I(y)\} - \mathbb{E}\{I(y|x)\})g(y)\, dy, \tag{7}$$

which is guaranteed to be positive under our monotonicity assumption. We would then take the x' which gave the largest reduction. But clearly this is within an additive constant (the weighted-average EI over $g(y)$) of the definition given in Eq. (6), and is thus equivalent.

The integrated approach allows constraints to be handled through $g(y)$. E.g., $g(y)$ can be uniform for $y \in C$ and zero otherwise. Or, [as we discuss in Section 4] it can give higher weight to y with a greater chance of satisfying the constraint. When there are no constraints, choosing $g(y)$ uniform on $y \in \mathcal{X}$ yields an aggregated statistic that will offer a more global search, compared to EI, in a manner similar to how the expected reduction in variance generalizes the predictive variance for sequential design by active learning (Seo et al., 2000; Gramacy and Lee, 2009).

3.1. *Expected Conditional Improvement*

The key ingredient in calculating the ECI is an assumption about how $Z(y|x)$ behaves relative to $Z(y)$. Let $F_N(y|x)$ denote the distribution of $Z(y|x)$. Overloading the notation somewhat, let $f_N(z(x))$ denote the density of $Z(x)$ under F_N, and likewise $f_N(z(y)|x)$ for $Z(y|x)$. By the law of total probability,

$$f_N(z(y)|x) = \int f_N(z(y), z(x)|x)\, dz(x) \tag{8}$$

$$= \int f_{N+1}(z(y)|x, z(x))f_N(z(x))\, dz(x),$$

where $f_{N+1}(z(y)|x, z(x))$ is the predictive density of $Z(y)$ when the design matrix and response vector are augmented by $(x, z(x))$. Note that the above expressions involving $z(y)$ have an implicit conditioning upon y. For an arbitrary surrogate, computing the integral in Eq. (8) analytically would present a serious challenge. However, under a GP surrogate it is trivial since F_N and F_{N+1} are both (univariate) normal distributions (2), and a convolution of normals is also normal. Trivially, the mean and variance of the (normal) predictive density $f_{N+1}(z(y)|x, z(x))$ is unchanged after integrating out $Z(x)$ since the GP is not dynamic, so there is no update from f_N without observing $z(x_{N+1})$.

But at the same time, the predictive variance (2) does not depend upon the responses, Z_N or $z(x)$ via Z_{N+1}. So we can deduce what variance of the predictive density $f_{N+1}(z(y)|x, z(x))$ *will be* once $z(x)$ arrives. We will have $\hat{\sigma}^2_{N+1}(y|x, z(x)) = \hat{\sigma}^2_{N+1}(y|x)$ under the *assumption* that the evidence in $z(x)$ does not update/change parameters of the GP (which it can't if it is not observed!). Now, $\hat{\sigma}^2_{N+1}(y|x, z(x))$ depends upon $K^{-1}_{N+1}(x)$ whose $N + 1^{\text{st}}$ row and column are populated with $K(x_i, x)$ for $i = 1, \ldots, N$ and with $K(x, x)$ appearing in the bottom right-hand corner. So $K^{-1}_{N+1}(x)$ can then be obtained in terms of K^{-1}_N via partitioned inverse equations. If

$$K_{N+1}(x) = \begin{bmatrix} K_N & k_N(x) \\ k_N^T(x) & K(x, x) \end{bmatrix}, \quad \text{then}$$

$$K^{-1}_{N+1}(x) = \begin{bmatrix} [K^{-1}_N + g(x)g^T(x)\mu^{-1}(x)] & g(x) \\ g^T(x) & \mu(x) \end{bmatrix},$$

where $g(x) = -\mu(x)K^{-1}_N k_N(x)$ and $\mu^{-1}(x) = K(x, x) - k_N^T(x)K^{-1}_N k_N(x)$. This saves us from performing any additional $O(N^3)$ matrix operations. So $\hat{\sigma}^2_{N+1}(y|x) = \sigma^2[K(y, y) - k_{N+1}^T(x; y)K^{-1}_{N+1}(x)k_{N+1}(x; y)]$ where $k_{N+1}^T(x; y)$ is an $(N + 1)$-vector whose first N entries are identical to $k_N(y)$ and with an $N + 1^{\text{st}}$ entry of $K(y, x)$. The amount by which $\hat{\sigma}^2_{N+1}(y|x, z(x))$ is reduced compared to $\hat{\sigma}^2_N(y)$ is then readily available. Let $G(x) \equiv g(x)g^T(x)$. Then,

$$\hat{\sigma}^2_{N+1}(y|x) = \hat{\sigma}^2_N(y) - \sigma^2[k_N^T(y)G(x)\mu^{-1}k_N(y) \tag{9}$$
$$+ 2k_N^T(y)g(x)K(x, y) + K(x, y)^2\mu].$$

So we can see that the *deduced* predictive variance at y will be reduced when $z(x)$ is observed by an amount that depends upon how far apart y and x are. This is not only sensible, but will also be helpful for determining the influence of x in improvement calculations.

To sum up, we propose to *define* $F_N(y|x)$, for the purposes of sequential design, to be a normal distribution with (true) mean $\hat{z}_N(y|x) = \hat{z}_N(y)$ and deduced variance $\hat{\sigma}^2_N(y|x) \equiv \hat{\sigma}^2_{N+1}(y|x, z(x)) = \hat{\sigma}^2_{N+1}(y|x)$ as given in Eq. (9), above. As with the kriging equations (2), joint sampling for a collection of (M) reference inputs Y_M is possible via the appropriate matrix extensions to $k_N(Y_M)$ and $K(Y_M, Y_M)$ in order to derive $\hat{z}_N(Y_M|x)$ and $\hat{\Sigma}_N(Y_M|x)$.

Now, with an appropriate definition of a deterministic $f_{\min}$, the same analytic expression for the EI from Section 2 can be extended to the ECI:

$$\mathbb{E}\{I(y|x)\} = \tag{10}$$
$$(f_{\min} - \hat{z}_N(y|x))\Phi\left(\frac{f_{\min} - \hat{z}_N(y|x)}{\hat{\sigma}_N(y|x)}\right) + \hat{\sigma}_N(y|x)\phi\left(\frac{f_{\min} - \hat{z}_N(y|x)}{\hat{\sigma}_N(y|x)}\right).$$

If we cared only about the ECI (without integration (6)), the branch and bound algorithm given by Jones et al. (1998) would apply leading to a *conditional* EGO algorithm.

3.1.1. *Choosing* $f_{\min}$

Figure 1 illustrates how a deterministic choice of $f_{\min}$ can influence the ECI. Consider two cases ((a) and (b)), which pertain to the choices for $f_{\min}$ introduced in

Section 2.2 (represented by horizontal lines): (a) uses only the observed locations and (b) uses the whole predictive curve. We will return to details of these choices shortly. In the figure, the solid parabolic curve represents the predictive mean surface $\mathbb{E}\{Z(\cdot)\}$. The EI is the area of the predictive density drawn as a solid line, plotted vertically and centered at $\hat{z}(y)$, which lies underneath the horizontal line(s), representing choices of $f_{\min}$. The ECI is likewise the area of the predictive den-
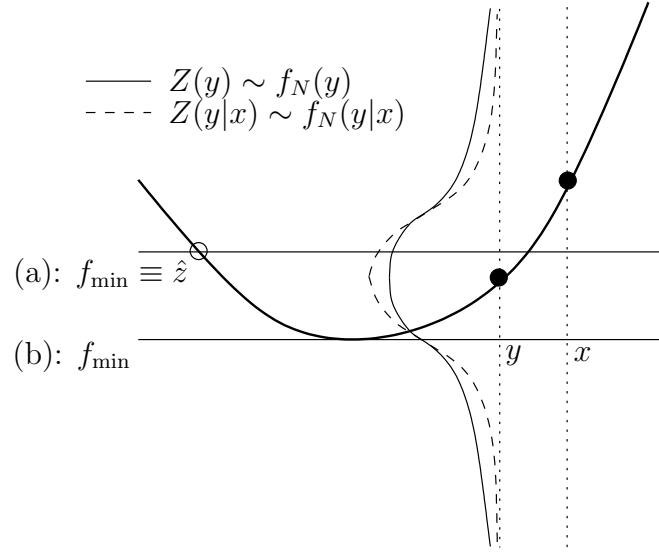
Figure 1: Illustrating how the choice of $f_{\min}$ influences the ECI. The solid curve represents the mean-predictive $\mathbb{E}\{Z(\cdot)\}$. The densities of $Z(y)$ and $Z(y|x)$ are shown as solid and dashed "bell-curves", respectively. In (a) $f_{\min}$ is taken to be the mean predictive at the N input locations whereas in (b) it is taken to be the minimum of predictive-mean surface. The respective improvements are the areas of the densities underneath $f_{\min}$.

sity drawn as a dashed line lying below the horizontal line(s). This dashed density has the same mean/mode as the solid one, but it is more sharply peaked by the influence of x. If we suppose that the densities, drawn as bell-curves in the figure, are symmetric (as they are for a GP), then it is clear that the relationship between ECI and EI depends upon $f_{\min}$. As the dashed line is more peaked, the left-tail cumulative distributions have the property that $F_N(f_{\min}|x) \geq F_N(f_{\min})$ for all $f_{\min} \geq \mathbb{E}\{Z(y|x)\} = \mathbb{E}\{Z(y)\}$, to which choice (a) for $f_{\min}$ corresponds. Therefore $\mathbb{E}\{I(y|x)\} \geq \mathbb{E}\{I(y)\}$ in this case, violating our desired monotonicity property. But for choice (b) the ECI represents a reduction compared to the EI, since $f_{\min} \leq \mathbb{E}\{Z(y|x)\}$, thus satisfying the monotonicity property.

Case (a) in Figure 1 is meant to represent taking $f_{\min} = \min\{z_1, \ldots, z_N\}$, deterministically. It may similarly represent the minimum of the mean-predictive at

the X_N locations, which would coincide with the minimum of the Z_N values in the no-noise ($\eta = 0$) case. In the noisy case ($\eta > 0$) $f_{\min}$ in Eq. (5) is a random variable whose distribution can be approximated by simulation from F_N. But this extra computational effort would be in vain because the monotonicity property is not guaranteed. Case (b) corresponds to taking $f_{\min} = \min \mathbb{E}\{Z(\cdot)\}$, the minimum of the posterior mean-predictive—another deterministic choice. In this case it is clear that $f_{\min}$ will always cut through the density of $Z(y|x)$ at or below its mean/mode $\mathbb{E}\{Z(y|x)\} = \mathbb{E}\{Z(y)\}$ and ensure that the monotonicity property is satisfied. Accordingly, we shall use this choice throughout the remainder of the paper.

3.1.2. *A Monte Carlo Approach for Calculating the ECI*

The following Monte Carlo procedure may be used to obtain samples of the ECI via the GP surrogate posterior predictive f_N, taking full account of uncertainty in the parameters $\theta = (\sigma^2, d, \eta)$. The procedure is borne out via Monte Carlo sampling for θ in Figure 2. If θ is considered known, or has been estimated offline, e.g., via

---

For $t = 1, \ldots, T$, current design $D_N = (X_N, Z_N)$, repeat:

 (i) Sample $\theta^{(t)}$ from the posterior distribution conditional upon f_N, priors, and D_N, and possibly conditional upon $\theta^{(t-1)}$ in an MCMC setup

 (ii) Calculate $f_{\min} = \min \hat{z}_N(\cdot|\theta^{(t)})$

 (iii) Obtain the tth sample from $\mathbb{E}\{I(y|x)\}$ as
 $\mathbb{E}^{(t)}\{I(y|x)\} = \mathbb{E}\{I(x|y)|\theta^{(t)}\}$, following Eq. (**??**) with $\hat{z}_N(y|x, \theta^{(t)})$
 and $\hat{\sigma}_N^2(y|x, \theta^{(t)})$

Figure 2: Monte Carlo approximation of the ECI statistic.

maximum likelihood, then we may skip the loop (and Step 1), taking $T = 1$ with $\theta^{(1)} = \theta$. In either case, an estimate of the ECI is obtained by ergodic averaging:

$$\mathbb{E}\{I(y|x)\} \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^{(t)}\{I(y|x)\}. \tag{11}$$

3.2. *Integrated Expected Conditional Improvement Algorithm*

Calculating the IECI (6) from the ECI requires integrating over $y \in \mathcal{X}$ according to $g(y)$, which may be uniform in a bounded (constraint) region. It will not generally be possible to integrate analytically, so we propose to augment the Monte Carlo procedure from Section 3.1.2. Given a large number of sampled reference locations $Y_M \equiv y^{(1)}, \ldots, y^{(M)} \overset{\text{iid}}{\sim} g$, the IECI may be approximated with T Monte Carlo

samples from the ECI as follows.

$$\mathbb{E}_g\{I(x)\} \approx - \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} \mathbb{E}^{(t)}\{I(y^{(m)}|x)\} \tag{12}$$

When the parameters θ are known, $T = 1$ as before. With larger M (and T) we obtain an improved approximation, and in the limit we have equality. In the case where g is uniform over a convex region, a grid or maximum entropy design may be preferred (Santner et al., 2003, Section 6.2.1). When the marginals of g are known, a Latin Hypercube Design (LHD, Santner et al., 2003, Section 5.2.2) may be more generally appropriate.

If we choose (or are required) to work with a size M grid, design, or LHD of reference locations $y \in \mathcal{X}$, we may view g as discrete and finite measure. An alternate approach in this case is to forgo (re-)sampling from g and compute a weighted average instead:

$$\mathbb{E}_g\{I(x)\} \approx -\frac{1}{T} \sum_{t=1}^{T} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}^{(t)}\{I(y^{(m)}|x)\}g(y^{(m)}). \tag{13}$$

This has the disadvantage that the ECI may be evaluated at many reference locations $y^{(m)}$ with low (or zero) probability under g. But it has the advantage of an implementation that is easily adapted to the unknown constraint situations described shortly.

3.3. *Illustrating IECI*

To illustrate IECI consider the following process $\mathbb{E}\{Z(x)\} = f(x) = \sin(x) + 2.55\phi_{0.45}(x-3)$, observed for $x \in [0,7]$. As a mixture of a sinusoid and normal density function (with $\mu = 3$ and $\sigma = 0.45$) it has two local minima in this region. To make things interesting, realizations of the process are observed with i.i.d. noise so that $\text{Var}\{Z(x)\} = 0.15^2$. The *top-left* panel of Figure 3 shows one random realization of this process at LHD inputs. The predictive mean and 90% interval obtained by sampling from the posterior under the GP is also shown. A visual inspection of the surface(s) reveals that, indeed, there are two local minima.

Below that panel, on the *bottom-left*, the EI (solid black) and IECI (dashed red) surfaces are plotted, normalized to appear on the same $[0,1]$ scale. As a further visual aid, the design X_N is also shown, and the vertical lines crossing the x-axis intersect with the curves at their maxima. We took a uniformly spaced set of 100 candidate locations in $[0,7]$, our $\mathcal{X}$, and calculated the EI and IECI at $x \in \mathcal{X}$. Likewise, we took the same $M = 100$ points as reference locations $Y_M = \mathcal{X}$ for the IECI calculations via Eq. (12). EI recommends taking a sample from the left local minima, although the relative heights of the two disparate regions of highest EI betrays that this decision is indeed a "close call". In contrast, IECI suggests taking the next sample from the right local minima, and with much greater decisiveness. The lower concentration of samples nearby this right-minima lead to higher variance in that region which may be pooled by the more globally-scoped IECI.

The *right-hand* panels in Figure 3 show a similar sequence of plots in the presence of a known constraint $C = [0,2] \cup [4,7]$. To illustrate EI and IECI in this scenario, consider the random realization and corresponding posterior predictive surface in
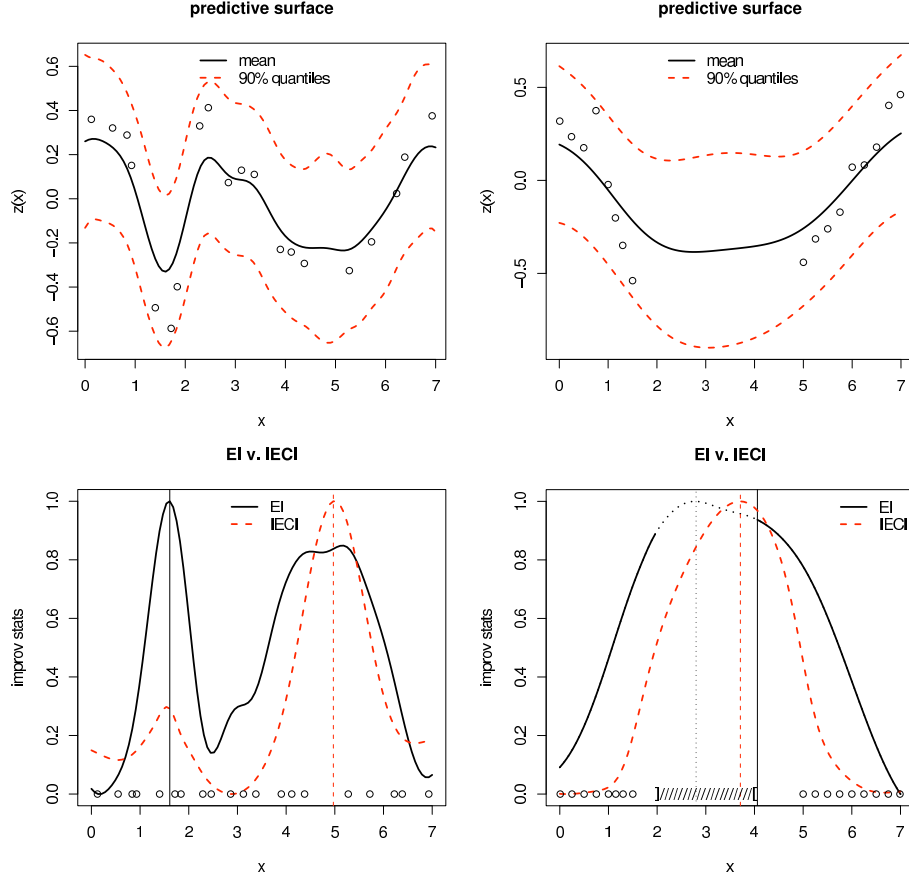
Figure 3: Comparing EI and IECI. The *top* panels show the design and posterior predictive surface. The *bottom* panels show EI and IECI statistics for the corresponding surfaces above. In the case of constrained optimization, in the *right* panels, the constraint violation region C^c is shown with slashes.

the *top-right* panel. Here the X_N design locations all reside inside C. The *bottom-right* panel shows the EI statistic over the entire (discrete) range for $x \in \mathcal{X}$, as above. Those parts of the EI curve corresponding to inputs which violate the constraint are dotted. The EI is maximized outside of the constraint region near $x = 2.75$, with the maximal value inside C at the $x = 4$ boundary. The IECI statistic is also shown over the entire range, but the $y^{(m)}$ locations are restricted to C. I.e., $Y_M = \mathcal{X} \cap C$. This is so that we may consider the extent to which every location $x \in \mathcal{X}$ reduces the average conditional improvement $y \in C$. Observe that the maximal IECI point is $x = 3.75$. This point gives the greatest reduction in improvement averaged over

the constraint region, even though it does not, itself, satisfy the constraint.

4. DEALING WITH UNKNOWN CONSTRAINTS

Here we extend the IECI to unknown constraints. Much of the necessary scaffolding has already been built into the IECI via $g(y)$, e.g., $g(y) = \mathbb{P}(C(y) = 1)$. It remains for us to flesh out the Monte Carlo by incorporating the surrogate c_N for $C(y)$. We extend the parameter vector θ to contain parameters for both surrogates: $\theta = (\theta_f, \theta_c)$; and the data to include the class/constraint labels: $D_N = (X_N, Z_N, C_N)$. Inference for unknown $\theta|D_N$ is via samples from the joint posterior. An appropriate choice of c_N is discussed in Section 4.1.

For now, overload the generic classification surrogate notation to let $c_N(y^{(m)}|\theta_c^{(t)})$ denote the probability input $y^{(m)}$ satisfies the constraint given parameters $\theta_c^{(t)}$. Then,

$$\mathbb{E}_c\{I(x)\} \approx -\frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}^{(t)}\{I(y^{(m)}|x)\} \cdot c_N(y^{(m)}|\theta_c^{(t)}). \tag{14}$$

Note that in $\mathbb{E}^{(t)}\{I(y^{(m)}|x)\}$ there is an implicit dependence upon $\theta_f^{(t)}$, unless these parameters are taken as known. In that case we may drop the (t) superscript from the ECI expression in Eq. (14), and re-arrange the order of summation to avoid unnecessarily re-calculating the ECI for each t. Observe that Eq. (14) extends Eq. (13) rather than (12). Sampling from the surrogate g_N, rather than simply evaluating the related quantity c_N, would not generally be straightforward, and so we prefer to work with design-based candidates $y \in \mathcal{X}$.

4.1. *An Appropriate Constraint Surrogate, and Sequential Inference*

An appropriate partner to the canonical GP (regression) surrogate f_N for f is a classification GP (CGP) surrogate c_N for c. For details on CGP specification and corresponding Monte Carlo inference based on MCMC, see Neal (1998). As in the regression case, the CGP model is highly flexible and competitive with, or better than, the most modern models for non-parametric classification. However, batch inference methods based on MCMC are at odds with the sequential nature of the design strategy. Except to guide the initialization of the new Markov chain, it is not clear how fits from earlier iterations may re-used in search of the next design point. So after each sequential design step the MCMC must be re-started and iterated until convergence. The result is a slow algorithm.

So instead of taking the traditional, established, MCMC approach to C/GP inference we follow a new, promising, sequential Monte Carlo (SMC) approach outlined by Gramacy and Polson (2010). They show how GP and CGP models can be implemented in an online setting, by efficiently updating a discrete approximation to the posterior via *particle learning* (Carvalho et al., 2008). This approach leads to fast online—and in some cases statistically superior (i.e., lower MC error)—posterior summaries compared to MCMC. Gramacy and Polson (2010) go on to describe to how EI for optimization and entropy based boundary exploration for classification can proceed efficiently with particles. This is easy to extend to IECI by coupling the regression and classification models (f_N and c_N) via the Monte Carlo approximations described earlier in this paper.

4.2. *Illustrations and Examples*

We provide two synthetic data examples where the constraint region is unknown. In both cases we take the candidate and reference locations (identically: $Y_m = \mathcal{X}$) as a LHD randomly generated at the beginning of each round and then augmented with an *oracle point* (Taddy et al., 2009b). We follow Gramacy and Taddy (2010) in taking the oracle point as the local maximum obtained via numerical non-derivative minimization initialized at the last added design point and searched over the the MAP predictive surface inferred in the previous round. An implementation via particles is described by (Gramacy and Polson, 2010).

The objective function and constraint region for the first example was presented in Section 3.3. We initialize the optimization with a size 20 LHD, and then collect 60 points by IECI with 100 fresh candidates/reference locations as described above. Figure 4 summarizes the results after the 80 total samples were gathered. Observe from the plots in the *top* row that most samples (after the 20 initial ones) were gathered in the two local minima, with a few taken outside C. The oracle candidates (solid circles) indicate the most likely locations of minima according to the posterior predictive distribution. The *bottom* panes show an estimate of c_N via the posterior mean probability of violating the constraint ($\hat{P}(c_N(x) = 1)$), and a progress meter showing the largest (log) expected reduction in average improvement (7) at each round. Observe how the ability to improve upon the current minimum decreases over time, giving a heuristic indication of convergence.

In our second example, the objective function for 2-d inputs $x = (x_1, x_2)$ is given by

$$f(x_1, x_2) = -w(x_1)w(x_2), \quad \text{where} \tag{15}$$
$$w(x) = \exp\left(-(x-1)^2\right) + \exp\left(-0.8(x+1)^2\right) - 0.05\sin\left(8(x+0.1)\right)$$

and observed without noise. The constraint (satisfaction) region is the interior of an ellipse defined by the 95% contour of a bivariate normal distribution centered at the origin, with correlation -0.5 and variance 0.75^2. The true global minimum is at $(x_1, x_2) = (-1.408, -1.408)$, which does not satisfy the constraint. There are, however, three other local minima—two of which satisfy the constraint. The setup is as described above for the 1-d example except that the optimization is initialized with 25 LHD samples, after which 100 are gathered by IECI with 100 fresh candidates in each round. Figure 5 summarizes the results after the 125 total samples were gathered. Observe that very few samples were gathered outside the unknown constraint region, except near the local minima. It is sensible to sample heavily on the boundary of the constraint region where the response is quickly changing and local minima are likely to occur. This is in case the global minimum is on the boundary, and also helps to extract the GP parameters in regions of highest importance. Notice that large concentrations of samples occur for two minima well inside the constraint region. But the *bottom-right* plot indicates that further progress can be made by additional sampling.

5. HEALTH POLICY OPTIMIZATION

Our motivating example involves a simulation of health care policy in the United States. The COMPARE simulator (Girosi et al., 2009) was developed at the RAND Corporation to predict the effect of various health care policies in terms of individual choices of health insurance and the associated costs. It is an agent-based microsimulation model that uses a maximum utility approach to predict the health insurance
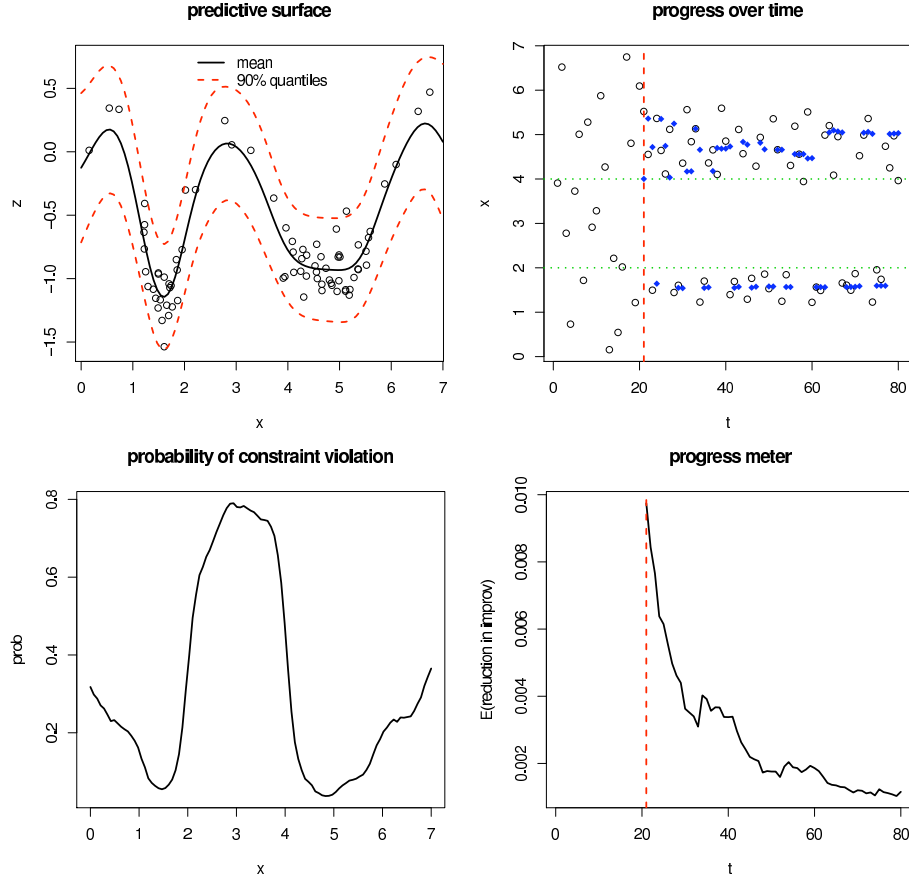
Figure 4: Progress in 1-d optimization after 80 samples: *top-left* shows the posterior mean predictive surface (of f_N); *top-right* shows sampled x-values (open circles) and oracle candidates (closed) before and after the initial design, as separated by the vertical bar; horizontal lines indicate the unknown constraint region; *bottom-left* posterior mean of constraint (violation) surface (c_N); *bottom-right* the maximum of the log expected reduction in average improvement (7) over time.

decisions of individuals, families, and firms as a function of a wide range of inputs on available types of policies, and on taxes, penalties, and regulations. The population is simulated based on Census Bureau data. Additional datasets provide values for many of the parameters in the simulation, and other parameters are set as part of the possible policy interventions. However, there are several calibration parameters
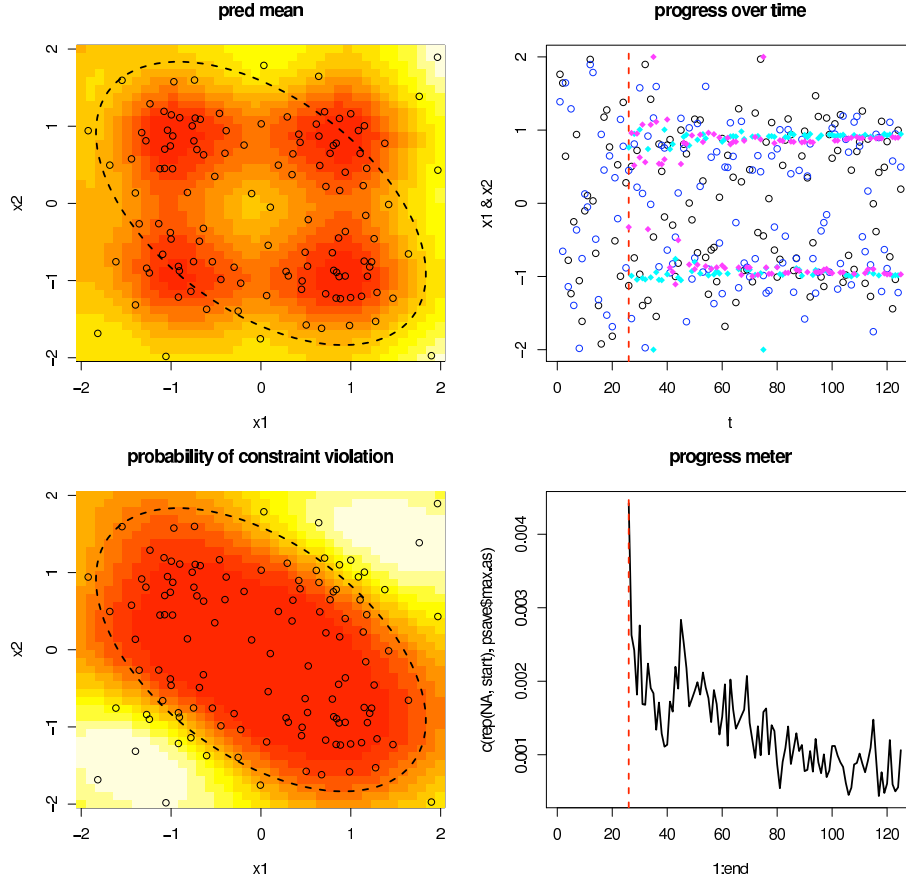
Figure 5: Progress in 2-d optimization after 125 samples: *top-left*: posterior mean predictive surface; *top-right*: sampled (x_1, x_2)-values (open) and oracle candidates (closed); *bottom-left*: posterior mean of constraint surface; *bottom-right*: the progress meter (7).

that are tuned so that when the simulator is run on current policies, it makes predictions as close as possible to the current observable situation in the United States. Such a calibration can be viewed as a minimization problem, choosing the values of the calibration parameters to minimize the discrepancy between predictions and reality. This setup is common for computer simulators and has been investigated in the unconstrained setting (e.g., Kennedy and O'Hagan, 2001). What differs from the standard setup here is the presence of unknown constraints.

The simulator has a number of inputs and outputs, here we focus on a subset

deemed most important by our collaborators, the designers of the simulator. The inputs over which we optimize are a set of six calibration parameters: utility tuning parameters for adults on ESI programs, adults on individual programs, and adults on public programs, and an analogous set of three parameters for children. The outputs of interest are the predicted counts in each type of insurance (or the uninsured category) and the elasticities of response for the key categories of adults in individual plans, adults in restricted individual plans, uninsured adults, children in individual plans, children in restricted individual plans, and uninsured children. The objective function specified by our collaborators is a combination of the absolute errors in the predicted counts and the squares of the predicted elasticities:

$$Z(\mathbf{x}) = \alpha_1 \sum_{j=1}^{4} |y_{aj} - \hat{y}_{aj}| + \alpha_2 \sum_{j=1}^{4} |y_{cj} - \hat{y}_{cj}| + \sum_{k=1}^{4} \alpha_{3k} y_{ek}^2 \mathbb{I}_{\{|y_{ek}|>1\}}$$

where α_1, α_2, and α_{3k} are constants specified by our collaborators that weight the pieces appropriately. Our goal is to minimize this objective function under the constraint that the elasticities for the insured are negative and the elasticities for the uninsured are positive. The elasticities can only be found by running the simulator, so this set of constraints fits under our unknown constraints regime.
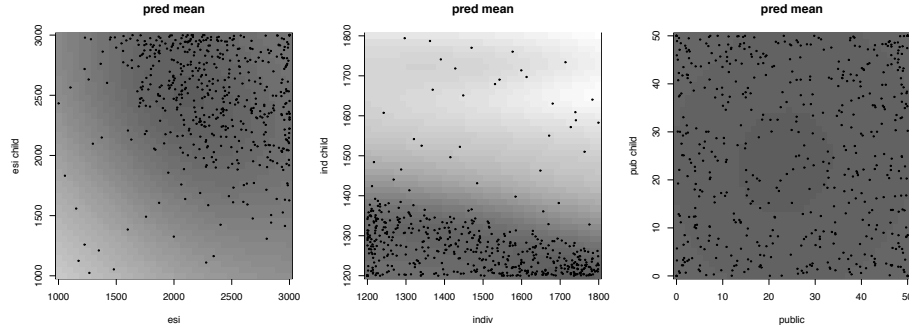
**Figure** 6: Slices of the fitted response surface; dark shades are lower values.

Figure 6 shows pairwise slices of the fitted response surface. The *left* panel shows how the fitted predicted surface varies as a function of the parameters for adult and child ESI, when the other four parameters are held fixed at a value around that which produces the minimum response. The *middle* and *right* panels vary by the parameters for individual programs and public programs respectively. Dark shades are lower values, so it can be seen that both ESI parameters need to be relatively high, the child individual parameter needs to be low, and the other three parameters are relatively less important. The points plotted in the figure are the 550 total inputs sampled projected into the each of the three pairs of input coordinates.

Figure 7 shows the fitted probability of a constraint violation over the portions of the space which were routinely sampled. As seen in Figure 6, some regions are not well-sampled because they do not help in finding the minimum, the goal of the problem. These sparsely sampled regions do not provide much information for
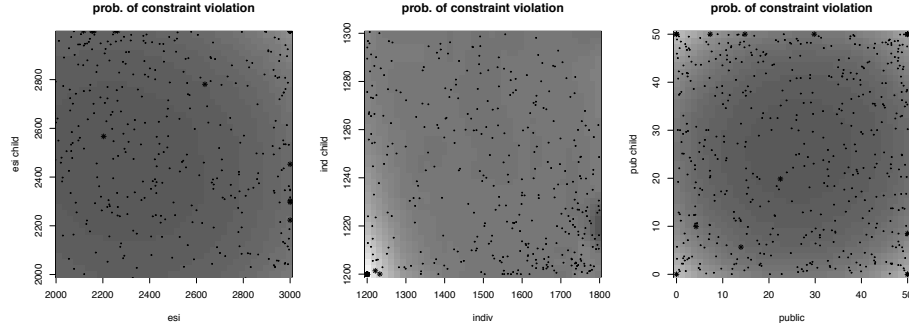
Figure 7: Slices of the fitted probability of constraint violation; dark shades are lower values; sampled points violating the constraint are shown with asterisks.

estimating the probability of a constraint violation (which is not the primary goal of the problem), and so the estimated values are overly influenced by the prior mean. Thus we only display parts of the regions in the first two plots to better show the estimated probabilities. Sampled points which violated the constraints are shown with asterisks. One can see that the largest probabilities of constraint violations occurred for large values of the ESI parameter, for jointly small values of the individual and child individual parameters, and for values of the public and child public parameters which are in the corners of the space.
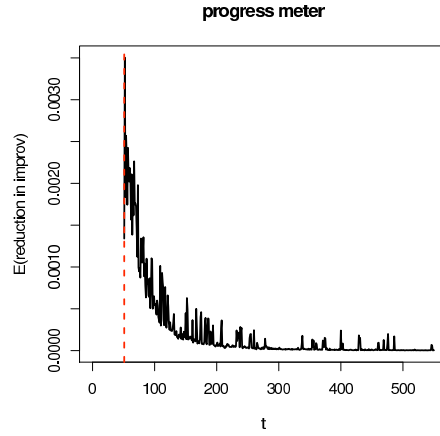
Figure 8: Progress meter (7) for the health policy optimization.

Figure 8 shows the progress meter (7) over the 500 optimization rounds which can be used as a heuristic check of convergence. As in previous examples, the noisiness in the meter is due to the LHD predictive grid of 100 candidates at which

the IECI is evaluated in each round. After about 250 samples the IECI seems to have "bottomed-out". However, further progress can be made to reduce the frequency and magnitude of the "up-spikes" in the remaining optimization rounds, and thereby obtain higher confidence that the constrained global minimum has been obtained.

6. DISCUSSION

We have introduced a statistical approach to optimization under unknown constraints by an integrated conditional expected improvement (IECI) statistic. The idea is consider how the improvement at reference locations (y) conditional on candidates (x) may be used to augment a design. Without considering constraints, the resulting statistic is a less greedy—aggregated—version of the standard expected improvement (EI) statistic. Another way to obtain a less greedy EI is to raise the improvement to a power g (Schonlau et al., 1998). The IECI approach, by contrast, does not require such a tuning parameter. In the presence of unknown constraints, IECI allows us to coherently consider how design candidates adjust the improvement at reference locations believed to satisfy the constraint. Our method was illustrated on two synthetic examples and a motivating problem from health care policy.

We envisage many ways that our methodology may be extended and improved. Understanding of convergence of statistical optimization algorithms is scant at best, and IECI is no exception. While we provide a sensible heuristic that seems to work well in our examples, much remains to be done in this area. It may also be sensible to model the constraint as a function of the inputs (x) and the real-valued response ($Z(x)$). An example of where this would be handy is when $C = \{x : Z(x) < k\}$, for some constant c. Our dual-GP modeling framework may easily be extended to allow uncertainty in Z (real-valued) responses to filter through, as predictors, into the surrogate model for the classification labels. A more difficult extension to involves accommodating *hidden constraints* (Lee et al., 2010): where evaluation of the real-valued response fails, e.g., due to a lack of convergence in a simulation. Finally, it may be worthwhile to consider surrogate models beyond GPs. Dynamic trees for regression and classification show considerable promise (Taddy et al., 2009a).

REFERENCES

Abrahamsen, P. (1997). "A Review of Gaussian Random Fields and Correlation Functions." Tech. Rep. 917, Norwegian Computing Center, Box 114 Blindern, N-0314 Oslo, Norway.

Carvalho, C., Johannes, M., Lopes, H., and Polson, N. (2008). "Particle Learning and Smoothing." Discussion Paper 2008-32, Duke University Dept. of Statistical Science.

Girosi, F., Cordova, A., Eibner, C., Gresenz, C. R., Keeler, E., Ringel, J., Sullivan, J., Bertko, J., Buntin, M. B., and Vardavas, R. (2009). "Overview of the COMPARE Microsimulation Model." Tech. Rep. WR-650, RAND.

Gramacy, R. and Polson, N. (2010). "Particle learning of Gaussian process models for sequential design and optimization." Tech. Rep. arXiv:0909.5262, University of Cambridge.

Gramacy, R. B. (2005). "Bayesian Treed Gaussian Process Models." Ph.D. thesis, University of California, Santa Cruz.

Gramacy, R. B. and Lee, H. K. H. (2008). "Bayesian treed Gaussian process models with an application to computer modeling." *Journal of the American Statistical Association*, 103, 1119–1130.

— (2009). "Adaptive Design and Analysis of Supercomputer Experiment." *Technometrics*, 51, 2, 130–145.

Gramacy, R. B. and Taddy, M. A. (2010). "Categorical Inputs, Sensitivity Analysis, Optimization and Importance Tempering with tgp Version 2, an R Package for Treed Gaussian Process Models." *Journal of Statistical Software*, 33, 6, 1–48.

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

Kennedy, M. and O'Hagan, A. (2001). "Bayesian Calibration of Computer Models (with discussion)." *Journal of the Royal Statistical Society, Series B*, 63, 425–464.

Lee, H., Gramacy, R., Linkletter, C., and Gray, G. (2010). "Optimization Subject to Hidden Constraints via Statistical Emulation." Tech. Rep. UCSC-SOE-10-10, University of California, Santa Cruz, Department of Applied Mathematics and Statistics.

Neal, R. M. (1998). "Regression and classification using Gaussian process priors (with discussion)." In *Bayesian Statistics 6*, ed. e. a. J. M. Bernardo, 476–501. Oxford University Press.

O'Hagan, A., Kennedy, M. C., and Oakley, J. E. (1999). "Uncertainty Analysis and Other Inference Tools for Complex Computer Codes." In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. Dawid, and A. Smith, 503–524. Oxford University Press.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). "Design and Analysis of Computer Experiments." *Statistical Science*, 4, 409–435.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York, NY: Springer-Verlag.

Schonlau, M., Jones, D., and Welch, W. (1998). "Global versus local search in constrained optimization of computer models." In *New Developments and applications in experimental design*, no. 34 in IMS Lecture Notes - Monograph Series, 11–25. Institute of Mathematical Statistics.

Seo, S., Wallat, M., Graepel, T., and Obermayer, K. (2000). "Gaussian Process Regression: Active Data Selection and Test Point Rejection." In *Proceedings of the International Joint Conference on Neural Networks*, vol. III, 241–246. IEEE.

Taddy, M., Gramacy, R., and Polson, N. (2009a). "Dynamic trees for learning and design." Tech. Rep. arXiv:0912.1636, University of Chicago.

Taddy, M., Lee, H. K. H., Gray, G. A., and Griffin, J. D. (2009b). "Bayesian Guided Pattern Search for Robust Local Optimization." *Technometrics*, 51, 389–401.