Modeling an Augmented Lagrangian for Improved Blackbox Constrained Optimization

Robert B. Gramacy* Genetha A. Gray† Sébastien Le Digabel‡
Herbert K.H. Lee§ Pritam Ranjan¶ Garth Wells‖ Stefan M. Wild**

March 20, 2014

Abstract

Constrained blackbox optimization is a difficult problem, with most approaches coming from the mathematical programming literature. The statistical literature is sparse, especially in addressing problems with nontrivial constraints. This situation is unfortunate because statistical methods have many attractive properties: global scope, handling noisy objectives, sensitivity analysis, and so forth. To narrow that gap, we propose a combination of response surface modeling, expected improvement, and the augmented Lagrangian numerical optimization framework. This hybrid approach allows the statistical model to think globally and the augmented Lagrangian to act locally. We focus on problems where the constraints are the primary bottleneck, requiring expensive simulation to evaluate and substantial modeling effort to map out. In that context, our hybridization presents a simple yet effective solution that allows existing objective-oriented statistical approaches, like those based on Gaussian process surrogates and expected improvement heuristics, to be applied to the constrained setting with minor modification. This work is motivated by a challenging, real-data benchmark problem from hydrology where, even with a simple linear objective function, learning a nontrivial valid region complicates the search for a global minimum.

Key words: surrogate model, emulator, Gaussian process, nonparametric regression and sequential design, expected improvement, additive penalty method

*Corresponding author: The University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago IL, 60605; rbgramacy@chicagobooth.edu

†Most of the work was done while at Sandia National Laboratories, Livermore, CA

‡GERAD and Département de mathématiques et génie industriel, École Polytechnique de Montréal, Montréal, QC H3C 3A7, Canada

§Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064

¶Department of Mathematics and Statistics, Acadia University, Wolfville, NS B4P 2R6, Canada

‖Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK

**Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

1 Introduction

The area of mathematical programming has produced efficient algorithms for nonlinear optimization, most of which have provable convergence properties. They include algorithms for optimizing under constraints and for handling so-called *blackbox* functions, where evaluation requires running an opaque computer code revealing little about the functional form of the objective and/or constraints. Many modern blackbox solvers converge without derivative information and require only weak regularity conditions. Since their search is focused locally, however, only local solutions are guaranteed.

Statistical approaches to blackbox optimization have the potential to offer more global scope. Methods based on Gaussian process (GP) emulation and expected improvement (EI, Jones et al., 1998) enjoy global convergence properties and compare favorably with classical alternatives when objective evaluations are expensive, simulated by (noisy) Monte Carlo (Picheny et al., 2013) or when there are many local optima. In more conventional contexts, however, nonstatistical approaches are usually preferred. Global search is slower than local search; hence, for easier problems, the statistical methods underperform. Additionally, statistical methods are more limited in their ability to handle constraints. Here, we explore a hybrid approach that pairs a global statistical perspective with a classical augmented Lagrangian localization technique for accommodating constraints.

We consider constrained optimization problems of the form

$$\min_{x} \left\{ f(x) : c(x) \leq 0, x \in \mathcal{B} \right\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ denotes a scalar-valued objective function, $c : \mathbb{R}^d \to \mathbb{R}^m$ denotes a vector[1] of constraint functions, and $\mathcal{B} \subset \mathbb{R}^d$ denotes a known, bounded, and convex region. Here we take $\mathcal{B} = \{x \in \mathbb{R}^d : l \leq x \leq u\}$ to be a hyperrectangle, but it could also include other constraints known in advance. Throughout, we will assume that a solution of (1) exists; in particular, this means that the feasible region $\{x \in \mathbb{R}^d : c(x) \leq 0\} \cap \mathcal{B}$ is nonempty. In (1), we note the clear distinction made between the known bound constraints $\mathcal{B}$ and the constraints $c_1(x), \ldots, c_m(x)$, whose functional forms may not be known.

The abstract problem in (1) is challenging when the constraints c are nonlinear, and even more difficult when evaluation of at least one of f and c requires blackbox simulation. In Section 2 we review local search algorithms from the numerical optimization literature that allow for blackbox f and c. The statistical literature, by contrast, has offered solutions only in certain contexts. For example, Schonlau et al. (1998) adapted EI for blackbox f and known c; Gramacy and Lee (2011) considered blackbox f and blackbox $c \in \{0, 1\}$; and Williams et al. (2010) considered blackbox f and c coupled by an integral operator. These methods work well in their chosen contexts but are limited in scope.

The current state of affairs is unfortunate because statistical methods have much to offer. Beyond searching more globally, they can offer robustness, natively facilitate uncertainty quantification, and enjoy a near monopoly in the noisy observation case (e.g., Taddy et al., 2009). In many real-world optimization problems, handling constraints presents the

[1] Vector inequalities are taken componentwise (i.e., for $a, b \in \mathbb{R}^d$, $a \leq b$ means $a_i \leq b_i$ for all $i = 1, \ldots, d$).

biggest challenge; many have a simple, known objective f (e.g., linear, such as total cost $f(x) = \sum_i x_i$) but multiple complicated, simulation-based constraints (e.g., indicating if expenditures so-allocated meet policy/physical requirements). And yet, to our knowledge, this important case is unexplored in the statistical literature. In Section 5 we present a hydrology problem meeting that description: despite having a simple linear objective function, learning a highly nonconvex feasible region complicates the search for a global minimum.

One way forward is to force the problem (1) into a more limited existing statistical framework. For example, *integrated expected conditional improvement* (Gramacy and Lee, 2011) could be applied by treating $c(x)$ as binary. However, that would require discarding information about the *distance* to the boundary separating feasible ("valid") and infeasible ("invalid") regions. Recognizing this situation, we develop a statistical approach based on the augmented Lagrangian (AL, e.g., Bertsekas, 1982), a tool from mathematical programming that converts a problem with general constraints into a sequence of unconstrained (or simply constrained) problems that explicitly leverages knowledge about distance to constraint boundaries. Under specific conditions we can derive closed-form expressions, like EI, to guide the optimization, and we explore Monte Carlo alternatives for other cases. The result is a scheme that compares favorably with modern alternatives in the mathematical programming literature, especially in the presence of several nonglobal minima.

Although the approach we advocate is general, for specificity in this paper we focus on blackbox optimization problems for which the objective f is known while the constraints c require simulation. This setting all but rules out statistical comparators whose emphasis is on modeling f and treat c as an inconvenience. Throughout, we note how our approach can be extended to unknown f by pairing it with standard surrogate modeling techniques.

The remainder of the paper is organized as follows. We first describe a toy problem that introduces the challenges in this area. Then, in Section 2, we review statistical optimization and introduce the AL framework for handling constraints. Section 3 contains the bulk of our methodological contribution, combining statistical surrogates with the AL. Section 4 describes implementation details and provides results for our toy example. Section 5 provides a similar comparison for a challenging real-data hydrology problem. We conclude in Section 6 with a discussion focused on the potential for further extensions and efficiency gains.

A toy problem. We begin by introducing a test problem of the form (1) that illustrates some of the challenges involved. It has a linear objective in two variables:

$$\min_x \left\{ x_1 + x_2 : c_1(x) \leq 0,\ c_2(x) \leq 0,\ x \in [0,1]^2 \right\}, \tag{2}$$

where the two nonlinear constraints are given by

$$c_1(x) = \frac{3}{2} - x_1 - 2x_2 - \frac{1}{2}\sin\left(2\pi(x_1^2 - 2x_2)\right), \quad c_2(x) = x_1^2 + x_2^2 - \frac{3}{2}.$$

Figure 1 shows the feasible region and the three local optima, with x^A being the unique global minimizer. We note that at each of these solutions, the second constraint is strictly satisfied and the first constraint holds with equality. For x^C, the lower bound on x_1 is also

$$\begin{aligned} x^A &\approx [0.1954,\ 0.4044], \\ f\left(x^A\right) &\approx 0.5998, \\ x^B &\approx [0.7197,\ 0.1411], \\ f\left(x^B\right) &\approx 0.8609, \\ x^C &= [0,\ 0.75], \\ f\left(x^C\right) &= 0.75, \end{aligned}$$

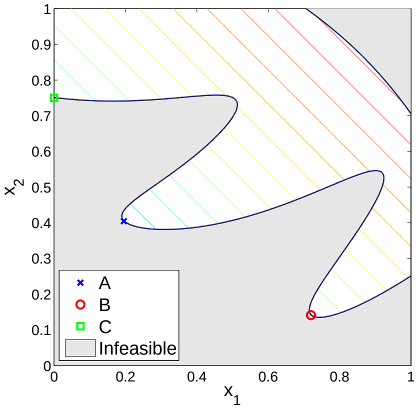


Figure 1: Toy problem (2) and its local minimizers; only x^A is a global minimizer.

binding because if this bound were not present, x^C would not be a local solution. The second constraint may seem uninteresting, but it reminds us that the solution may not be on every constraint boundary and thereby presents a challenge to methods designed to search that boundary in a blackbox setting. This toy problem has several characteristics in common with the real-data hydrology problem detailed in Section 5. Notably, the two problems both have a linear objective and highly nonlinear, nonconvex constraint boundaries.

2 Elements of hybrid optimization

Here we review the elements we propose hybridizing—statistical response surface models, expected improvement, and the augmented Lagrangian (AL)—in the context of derivative-free solvers. Implementations of the specific algorithms (particularly leveraging AL) that serve as our main comparators are detailed in Section 4.

2.1 Surrogate modeling framework for optimization

Examples of statistical models being used to guide optimization date back at least to Mockus et al. (1978) and Box and Draper (1987). Although the technique has evolved over the years, the basic idea is to train a flexible regression model f^n on input-output pairs $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ and use aspects of f^n to help choose $x^{(n+1)}$. One implementation involves searching the mean of a predictive surface $f^n(x)$ derived from f^n, which serves as a *surrogate* for the true $f(x)$, for global or local minima. Most modern practitioners would choose Gaussian process (GP) regression models as *emulators* f^n for a deterministic objective, f, since GPs produce highly accurate, conditionally normal predictive distributions and can interpolate the data if desired. For a review of GP regression for computer experiments, see Santner et al. (2003). For a more general overview of the so-called surrogate modeling framework for optimization, see Booker et al. (1999). This literature is focused primarily on the objective f, with far less attention to constraints (e.g., Audet et al., 2000; Sasena, 2002). Known constraints are usually accommodated by restricting search to the valid region.

2.2 Expected improvement

In initial usage, outlined above, the full statistical potential of f^n remained untapped: estimated uncertainties—a hallmark of any statistical endeavor—captured in the predictive distributions were not being used. Jones et al. (1998) changed this state of affairs by recognizing that the conditionally normal equations provided by a GP emulator $f^n(x)$, completely described by mean function $\mu^n(x)$ and variance function $\sigma^{2n}(x)$, could be used together to balance exploitation and exploration toward a more efficient global search scheme. They defined an improvement statistic $I(x) = \max\{0, f^n_{\min} - Y(x)\}$, where $f^n_{\min}$ is the minimum among the n y-values seen so far, and $Y(x) \sim f^n(x)$ is a random variable. The improvement assigns large values to inputs x, where $Y(x)$ is likely below $f^n_{\min}$. Jones et al. showed that the *expected improvement* (EI) could be calculated analytically in the Gaussian case:

$$\mathbb{E}\{I(x)\} = (f^n_{\min} - \mu^n(x))\Phi\left(\frac{f^n_{\min} - \mu^n(x)}{\sigma^n(x)}\right) + \sigma_n(x)\phi\left(\frac{f^n_{\min} - \mu^n(x)}{\sigma^n(x)}\right), \tag{3}$$

where Φ and ϕ are the standard normal cdf and pdf, respectively. The equation reveals a balance between exploitation ($\mu^n(x)$ under $f^n_{\min}$) and exploration ($\sigma^n(x)$). The expectation is essentially tallying the mass of the predictive distribution for $Y(x)$ that is under $f^n_{\min}$.

Leveraging the analytic EI, Jones et al. proposed the *efficient global optimization* (EGO) algorithm where each trial involved using a branch-and-bound scheme to search for the largest $\mathbb{E}\{I(x)\}$. In a later paper, Schonlau et al. (1998) provided an analytical form for a *generalized EI* based on a powered up improvement measure $I^g(x) = |f^n_{\min} - Y(x)|^g \mathbb{I}_{\{Y(x)<f^n_{\min}\}}$. Special cases of $\mathbb{E}\{I^g(x)\}$ for $g = 0, 1, 2$ lead to searches via $\Pr(Y(x) < f^n_{\min})$, the original EI criterion, and a hybrid criterion $\mathbb{E}\{I(x)\}^2 + \mathbb{V}\mathrm{ar}[I(x)]$, respectively.

Under relatively weak regularity conditions, search algorithms based on EI variations converge to the global optimum. EGO, which specifically pairs GP emulators with EI, can be seen as one example of a wider family of routines. For example, radial basis function emulators have been used with similar success in the context of local search (Wild and Shoemaker, 2013). Although weak from a technical viewpoint, the computer model regularities required are rarely reasonable in practice. They ignore potential feedback loops between surface fits, predictive distributions, improvement calculations, and search (see, e.g., Bull, 2011); in practice, these can pathologically slow convergence and/or lead to local rather than global solutions. Practitioners instead prefer hybrids between global EI and deliberately local search (e.g., Taddy et al., 2009; Gramacy and Le Digabel, 2011).

For a wider review of the literature, we recommend the tutorial by Brochu et al. (2010) and the Ph.D. thesis of Boyle (2007), both on "Bayesian optimization."[2] It is revealing that

[2] This is the terminology preferred by the machine learning community, combining EI-type blackbox optimization and reinforcement learning. The Bayesian perspective goes back to Mockus et al. (1978), but many of the techniques can be justified on less idealogical terms.

in a combined near-200 pages of text, the treatment of constraints is virtually neglected; this is clearly an underserved area.

2.3 Augmented Lagrangian framework

Augmented Lagrangian methods are a class of algorithms for constrained nonlinear optimization that enjoy favorable theoretical properties for finding local solutions from arbitrary starting points. The main device used by these methods is the augmented Lagrangian, which, for the inequality constrained problem (1), is given by

$$L_A(x;\lambda,\rho) = f(x) + \lambda^\top c(x) + \frac{1}{2\rho}\sum_{i=1}^{m} \max\left(0, c_i(x)\right)^2, \tag{4}$$

where $\rho > 0$ is a *penalty parameter* and $\lambda \in \mathbb{R}_+^m$ serves the role of *Lagrange multiplier.*

The first two terms in (4) correspond to the Lagrangian, which is the merit function that defines stationarity for constrained optimization problems. Without the second term, (4) reduces to an *additive penalty method* (APM) approach to constrained optimization. APM-based comparators are used as benchmarks for the hydrology problem in Section 5. Without considerable care in choosing the scale of penalization, however, APMs can introduce ill-conditioning in the resulting subproblems.

We focus on AL-based methods in which the original nonlinearly constrained problem is transformed into a sequence of nonlinear problems where only the bound constraints $\mathcal{B}$ are imposed. In particular, given the current values for the penalty parameter, ρ^{k-1}, and approximate Lagrange multipliers, λ^{k-1}, one approximately solves the subproblem

$$\min_x \left\{ L_A(x;\lambda^{k-1},\rho^{k-1}) : x \in \mathcal{B} \right\}. \tag{5}$$

Given a candidate solution x^k, the penalty parameter and approximate Lagrange multipliers are updated and the process repeats. Algorithm 1 gives a specific form of these updates. Functions f and c are evaluated only when solving (5), comprising the "inner loop" of the scheme. For additional details on AL-based methods, see, e.g., Nocedal and Wright (2006).

```
Require: $\lambda^0 \geq 0$, $\rho^0 > 0$
for $k = 1, 2, \ldots$ (i.e., each "outer" iteration) do
    Let $x^k$ (approximately) solve (5)
    Set $\lambda_i^k = \max\left(0, \lambda_i^{k-1} + \frac{1}{\rho^{k-1}} c_i(x^k)\right)$, $i = 1, \ldots, m$
    If $c(x^k) \leq 0$, set $\rho^k = \rho^{k-1}$; otherwise, set $\rho^k = \frac{1}{2}\rho^{k-1}$
end for
```

Algorithm 1: Basic augmented Lagrangian framework.

We note that termination conditions have not been explicitly provided in Algorithm 1. In our setting of blackbox optimization, termination is dictated primarily by a user's computational budget. Our empirical comparisons in Sections 4–5 involve tracking the best (valid) value of the objective over increasing budgets determined by the number of evaluations of the blackbox (i.e., the cumulative number of inner iterations). Outside that context, however, one could stop when all constraints are sufficiently satisfied and the (approximated) gradient of the Lagrangian is sufficiently small; for example, given thresholds $\eta_1, \eta_2 \geq 0$, one could stop when $\left\|\max\left\{c(x^k), 0\right\}\right\| \leq \eta_1$ and $\left\|\nabla f(x^k) + \sum_{i=1}^m \lambda_i^k \nabla c_i(x^k)\right\| \leq \eta_2$.

2.4 Derivative-free augmented Lagrangian methods

The inner loop of Algorithm 1 can accommodate a host of methods for solving the unconstrained (or simply constrained) subproblem (5). Solvers can leverage derivatives of the objective and/or constraint functions if they are available, or be *derivative-free* when they are not. We specifically focus on the derivative-free case because this subsumes blackbox optimization (see, e.g., Conn et al., 2009). In our comparisons in Sections 4–5 we consider two solvers for the inner loop. We now briefly introduce how these solvers can be situated within the AL framework; software/implementation details are deferred until later.

Direct Search: Loosely, direct search involves probing the objective at stencils centered on the current best input value. The outputs obtained on the stencil determine the placement and size of the next stencil. A modern overview of direct search methods can be found in Kolda et al. (2003). In particular, we consider the mesh adaptive direct search (MADS) algorithm (Audet and Dennis, 2006). MADS is a directional direct-search method that uses dense sets of directions and generates trial points on a spatial discretization called a mesh. The most important MADS parameters are the initial and minimal poll sizes, which define the limits for the *poll size parameter*, determining the stencil size, and the *maximum mesh index*, which limits poll size reductions after a failed iteration (when a stencil does not find an improved solution). In the context of Algorithm 1 it makes sense to allow the initial poll size parameter to take a software-recommended/default value but to set the maximum mesh index to $k-1$, prescribing a finer subproblem as outer iterations progress.

Model-based: These are closest in spirit to the statistical methods we propose. Model-based optimization employs local approximation models, typically based on local polynomials (e.g., Conn et al., 2009) or nonlinear kernels such as radial basis functions (e.g., Wild and Shoemaker, 2013), which are related to GPs. Here we consider the trust-region-based method that was previously used as an AL inner solver by Kannan and Wild (2012). This method builds quadratic approximation models $q^f, q^{c_1}, \ldots, q^{c_m}$ about the current iterate x^{k-1} by interpolating the values of $f, c_1, \ldots, c_m$ at design points. The AL subproblem (5) is then approximately solved by locally solving a sequence of quadratic problems of the form

$$\min_x \left\{ q^f(x) + \sum_{i=1}^m \lambda_i^{k-1} q^{c_i}(x) + \frac{1}{2\rho^{k-1}} \sum_{i=1}^m \left[\max\left(0, q^{c_i}(x)\right)^2\right]_Q \,:\, x \in \mathcal{B}^{k-1} \right\}, \tag{6}$$

where $[\cdot]_Q$ denotes a truncation to a quadratic form and $\mathcal{B}^{k-1}$ is a local neighborhood ("trust region") of x^{k-1}. The choice of quadratic models enables the efficient solution of (6).

3 Statistical surrogate additive penalty methods

The methods above are not designed for global optimization, and it is hard to predict which local minima they will ultimately converge to when several minima are present. Hybridizing with statistical surrogates offers the potential to improve this situation. Here we introduce the basic idea and explore variations. The simplest approach involves deploying a statistical surrogate directly on the AL (4), but this has obvious shortcomings. To circumvent these, we consider separately modeling the objective function f and each constraint function c_i. We then pursue options for using the surrogate to solve (5), either via the predictive mean or EI, which has an enlightening closed-form expression in a special case.

3.1 Surrogate modeling the augmented Lagrangian

Consider deploying GP regression-based emulation of the AL (4) in order to find x^k. In each iteration of the inner loop (step 2 of Algorithm 1), proceed as follows. Let n denote the total number of blackbox evaluations obtained throughout all previous "inner" and "outer" iterations, collected as $(x^{(1)}, f^{(1)}, c^{(1)}), \dots, (x^{(n)}, f^{(n)}, c^{(n)})$. Then form $y^{(i)} = L_A(x^{(i)}; \lambda^{k-1}, \rho^{k-1})$ via $f^{(i)}$ and $c^{(i)}$, and fit a GP emulator to the n pairs $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$. Optimization can be guided by minimizing $\mu^n(x)$ in order to find $x^{(n+1)}$ or via EI following Eq. (3) with $Y(x) \equiv Y_{\ell^n}(x) \sim \mathcal{N}(\mu^n(x), \sigma^{2n}(x))$. Approximate convergence can be determined by various simple heuristics, from the number of iterations passing without improvement, to monitoring the maximal EI (Gramacy and Polson, 2011) over the trials.

At first glance this represents an attractive option, being modular and facilitating a global-local tradeoff. It is modular in the sense that standard software can be used for emulation and EI. It is global because the GP emulates the entire data seen so far, and EI balances exploration and exploitation. It is local because, as the AL "outer" iterations progress, the (global) "inner" searches organically concentrate near valid regions.

Several drawbacks become apparent, however, upon considering the nature of the composite objective (4). For example, the $y^{(i)}$ values, in their relationship with the $x^{(i)}$s, are likely to exhibit behavior that requires nonstationary surrogate models, primarily because of the final squared term in the AL. Most out-of-the-box GP regression methods assume stationarity, which means that the functional structure (wiggliness, correlation, differentiability, noise) must be uniform throughout the input space. The squared term amplifies the effects of $c(x)$ away from the boundary with the valid region, so it is not consistent with the assumption of stationarity. A related challenge is the max in (4), which produces kinks near the boundary of the valid region, with the regime changing behavior across that boundary.

Modern GP methods accommodate nonstationarity (Schmidt and O'Hagan, 2003; Paciorek and Schervish, 2006) and even regime-changing behavior (Gramacy and Lee, 2008). To our knowledge, however, only the latter option is paired with public software. That method leverages treed partitioning, whose divide-and-conquer approach can accommodate limited differentiability and stationarity challenges, but only if regime changes are roughly axis-aligned. Partitioning, however, does not parsimoniously address effects amplified quadratically in space. In fact, no part of the above scheme, whether surrogate modeling (via GPs or otherwise) or EI-search, acknowledges the *known* quadratic relationship between objective (f) and constraints (c). By treating the entire apparatus as a blackbox, it discards potentially useful information. Moreover, when the objective portion (f) is completely known, as in our motivating example(s), the fitting method needlessly models a known quantity, which is inefficient (see, e.g., Kannan and Wild, 2012).

3.2 Separately modeling the pieces of the composite

Those shortcomings can be addressed by deploying surrogate models separately on the components of the AL, rather than wholly to the composite. With separate models, stationarity assumptions are less likely to be violated since modeling can commence on quantities prior to the problematic square and max operations. Separately estimated emulators, $f^n(x)$ for the objective and $c^n(x) = (c_1^n(x), \dots, c_m^n(x))$ for the constraints, can provide predictive distributions for $Y_{f^n}(x)$ and $Y_c^n(x) = (Y_{c_1}^n(x), \dots, Y_{c_m}^n(x))$, respectively. The n superscripts, which we drop below, serve here as a reminder that we propose to solve the "inner" AL subproblem (5) using all n data points seen so far. Samples from those distributions, obtained trivially via GP emulators, are easily converted into samples from the composite

$$Y(x) = Y_f(x) + \lambda^\top Y_c(x) + \frac{1}{2\rho}\sum_{i=1}^m \max(0, Y_{c_i}(x))^2, \tag{7}$$

serving as a surrogate for $L_A(x; \lambda, \rho)$. When the objective f is known, we can forgo calculating $f^n(x)$ and swap in a deterministic $f(x)$ for $Y_f(x)$.

As in Section 3.1, there are several ways to choose new trial points using the composite *distribution* of the random variable(s) in (7), for example, by searching the predictive mean or EI. We first consider the predictive mean approach and defer EI to Section 3.3. We have $\mathbb{E}\{Y(x)\} = \mathbb{E}\{Y_f(x)\} + \lambda^\top \mathbb{E}\{Y_c(x)\} + \frac{1}{2\rho}\sum_{i=i}^m \mathbb{E}\{\max(0, Y_{c_i}(x))^2\}$. The first two expectations are trivial under normal GP predictive equations, giving

$$\mathbb{E}\{Y(x)\} = \mu_f^n(x) + \lambda^\top \mu_c^n(x) + \frac{1}{2\rho}\sum_{i=1}^m \mathbb{E}\{\max(0, Y_{c_i}(x))^2\}, \tag{8}$$

via a vectorized $\mu_c^n = (\mu_{c_1}^n, \dots, \mu_{c_m}^n)^\top$. An expression for the final term, which involves $\mathbb{E}\{\max(0, Y_{c_i}(x))^2\}$, can be obtained by recognizing its argument as a powered improvement for $-Y_{c_i}(x)$ over zero, that is, $I_{-Y_{c_i}}^{(0)}(x) = \max\{0, 0 + Y_{c_i}(x)\}$. Since the power is 2, an expectation-variance relationship can be exploited to obtain

$$\begin{aligned}\mathbb{E}\{\max(0, Y_{c_i}(x))^2\} &= \mathbb{E}\{I_{-Y_{c_i}}(x)\}^2 + \mathbb{V}\mathrm{ar}[I_{-Y_{c_i}}(x)] \qquad (9)\\ &= \sigma_{c_i}^{2n}(x)\left[\left(-\frac{\mu_{c_i}^n(x)}{\sigma_{c_i}^n(x)} + 1\right)\Phi\left(-\frac{\mu_{c_i}^n(x)}{\sigma_{c_i}^n(x)}\right) - \frac{\mu_{c_i}^n(x)}{\sigma_{c_i}^n(x)}\phi\left(-\frac{\mu_{c_i}^n(x)}{\sigma_{c_i}^n(x)}\right)\right],\end{aligned}$$

by using a result from the generalized EI (Schonlau et al., 1998). Combining (8) and (9) completes the expression for $\mathbb{E}\{Y(x)\}$. When f is known, one simply replaces μ_f^n with f.

3.3 New expected improvement

The composite random variable $Y(x)$ in Eq. (7) does not have a form that readily suggests a familiar distribution, for any reasonable choice of f^n and c^n (e.g., under GP emulation), thwarting analytic calculation of EI. A numerical approximation is straightforward by Monte Carlo. Assuming normal predictive equations, we simply sample $y_f^{(t)}(x)$, $y_{c_1}^{(t)}(x), \dots, y_{c_m}^{(t)}(x)$ from $\mathcal{N}(\mu_f^n(x), \sigma_f^{2n}(x))$ and $\mathcal{N}(\mu_{c_i}^n, \sigma_{c_i}^{2n})$, respectively, and then average:

$$\begin{aligned}\mathbb{E}\{I_Y(x)\} &\approx \frac{1}{T}\sum_{t=1}^{T} \max(0, y_{\min}^n - y^{(t)}(x)) \qquad (10)\\ &= \frac{1}{T}\sum_{t=1}^{T} \max\left[0, y_{\min}^n - \left(y_f^{(t)}(x) + \lambda^\top y_c^{(t)}(x) + \frac{1}{2\rho}\sum_{i=1}^{m}\max(0, y_{c_i}^{(t)}(x))^2\right)\right].\end{aligned}$$

We find generally low Monte Carlo error, and hence very few samples (e.g., $T = 100$) suffice.

However, challenges arise in exploring the EI surface over $x \in \mathcal{X}$, since whole swaths of the input space emit numerically zero $\mathbb{E}\{I_Y(x)\}$. When f is known, whereby $Y_f(x) \equiv f(x)$, and when the outer loop is in later stages (large k), yielding smaller ρ^k, the portion of the input space yielding zero EI can become prohibitively large, complicating searches for improvement. The quadratic nature of the AL composite (7) causes Y to be bounded below for *any* Y_c-values under certain (λ, ρ), no matter how they are distributed.

To delve a little deeper, consider a single blackbox constraint $c(x)$, a known objective $f(x)$, and a slight twist on Eq. (7) obtained by removing the max. In this special case, one can derive an analytical expression for the EI under GP emulation of c. Let $I_Y = \max\{0, y_{\min} - Y\}$ be the improvement function for the composite Y, suppressing x to streamline notation. Calculating the EI involves the following integral, where $c(y)$ represents the density c^n of Y_c:

$$\mathbb{E}\{I_Y\} = \int_{-\infty}^{\infty} I_y c(y)\, dy = \int_\theta (y_{\min} - y)\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\, dy, \quad \theta = \{y : y < y_{\min}\}.$$

Substitution and integration by parts yield that when $\lambda^2 - 2(f - y_{\min})/\rho \geq 0$,

$$\begin{aligned}\mathbb{E}\{I_Y\} &= \left[y_{\min} - \left(\frac{\mu^2}{2\rho} + \lambda\mu + f\right) - \frac{\sigma^2}{2\rho}\right](\Phi(v_2) - \Phi(v_1)) \qquad (11)\\ &\quad + [\sigma\mu/\rho + \lambda\sigma](\phi(v_2) - \phi(v_1)] + \frac{\sigma^2}{2\rho}(v_2\phi(v_2) - v_1\phi(v_1)),\end{aligned}$$

$$\text{where} \quad v_1 = \frac{u_- - \mu}{\sigma}, \qquad v_2 = \frac{u_+ - \mu}{\sigma}, \qquad u_\pm = \frac{-\lambda \pm \sqrt{\lambda^2 - 2(f - y_{\min})/\rho}}{\rho^{-1}}.$$

Otherwise, when $\lambda^2 - 2(f - y_{\min})/\rho < 0$, we have that $\mathbb{E}\{I_Y\} = 0$. Rearranging gives $\rho\lambda < 2(f - y_{\min})$, which, as ρ is repeatedly halved, reveals a shrinking region of x values that lead to nonzero EI.

Besides pointing to analytically zero EI values under (7), the above discussion suggests two ideas. First, avoiding x values leading to $f(x) > y_{\min}$ will boost search efficiency by avoiding zero EI regions, an observation we exploit in Section 4.1. Second, dropping the max in (7) may lead to efficiency gains in two ways: from analytic rather than Monte Carlo evaluation, and via a more targeted search when f is a known monotone function, which is bounded below over the feasible region. In that case, a solution is known to lie on the boundary between valid and invalid regions. Dropping the max will submit large negative $c(x)$'s to a squared penalty, pushing search away from the interior of the valid region and towards the boundary.

While the single-constraint, known f formulation is too restrictive for most problems, some simple remedies are worthy of consideration. Extending to blackbox f, and modeling Y_f, is straightforward since $Y_f(x)$ features linearly in Eq. (7). Extending to multiple constraints is much more challenging. One option is to reduce a multiple constraint problem into a single one by estimating a single surrogate $c^n(x)$ for an aggregated constraint function, say, $\sum_i Y_{c_i}(x)$. Some care is required because summing positive and negative $Y_{c_i}(x)$ cancels valid and invalid values, potentially resulting in (extreme) information loss. A better approach would be to use $Y_c = \sum_i |Y_{c_i}(x)|$ or $Y_c = \sum_i \max(0, Y_{c_i}(x))$ even though that may result in challenging kinks to model. The former option, using absolute values, could lead to improvements when f is a known monotone function, exploiting that the solution is on the constraint boundary.[3] The advantage of modeling an aggregated final constraint term, before squaring, is that the analytic EI (11) can be used directly on the resulting $Y(x) = Y_f(x) + \lambda^\top Y_c(x) + \frac{1}{2\rho} Y_c(x)^2$. The disadvantage, beyond modeling challenges arising from kinks, is that information loss is always a concern when aggregating.

4 Implementation and illustration

In this section, we first dispense with the implementation details of our proposed methods (Section 3) and our comparators (Section 2.4). We then demonstrate how these methods fare on our motivating toy data (Section 1). All methods are initialized with $\lambda^0 = (0, \dots, 0)^\top$ and $\rho^0 = 1/2$. Throughout, we randomize over the initial x^0 by choosing it uniformly in $\mathcal{B}$.

4.1 Implementation for surrogate model-based comparators

Multiple variations were suggested in Section 3. We focus our comparative discussion here on those that performed best. To be clear, none of them performed poorly; but several are easy to criticize on an intuitive level, and those same methods are consistently dominated by their more sensible counterparts. In particular, we do not provide results for the simplistic approach of Section 3.1, which involved modeling a nonstationary composite AL response surface, since that method is dominated by the analog involving separated modeling of the constituent parts described in Section 3.2.

We entertain alternatives from Sections 3.2–3.3 that involve guiding the inner optimization with $\mathbb{E}\{Y\}$ and $\mathbb{E}\{I_Y\}$, following Eqs. (9–10), respectively. We note here that the results

[3] However, it may introduce complications when the constraint set includes ones that are not active/binding at the solution, which we discuss further in Section 6.

based on a Monte Carlo $\mathbb{E}\{Y\}$, via the analog of (10) without "$\max[0, y^n_{\min} -$", and the analytical alternative (9) are indistinguishable up to Monte Carlo error when randomizing over x^0. Taking inspiration from the analytic EI derivation for the special case in Section 3.3, we consider a variation on the numeric EI that discards the max term. We do not provide results based on the analytic expression (11), however, because doing so requires compromises on the modeling end, which lead to deterioration in performance. Therefore, in total we report results for four variations pairing one of $\mathbb{E}\{Y\}$ and $\mathbb{E}\{I_Y\}$ with the original AL (7) and a version obtained without the max, which are denoted by the acronyms EY, EI, EY-nomax, and EI-nomax, respectively.

Throughout we treat f as known and emulate each Y_{c_i} with separate GPs initialized with ten random input-output pairs from $\mathcal{B}$ (i.e., the outer loop of Algorithm 1 starts with $x^{(1:10)}$). For fast updates and MLE calculations we used `updateGP` and `mleGP` from the `laGP` package (Gramacy, 2013) for `R`. Each inner loop search in Algorithm 1 is based on a random set of 1,000 candidate x locations $\mathcal{X}^n$. We recognize that searching uniformly in the input space is inefficient when f is a known linear function. Instead, we consider random *objective improving candidates* (OICs) defined by $\mathcal{X} = \{x : f(x) < f^{n_*}_{\min}\}$, where $f^{n_*}_{\min}$ is the best value of the objective for the $n_* \leq n$ *valid* points found so far. If $n_* = 0$, then $f^{n_*}_{\min} = \infty$. A random set of candidates $\mathcal{X}^n$ is easy to populate by rejection sampling. A naïve sampler could have a high rejection rate if $\mathcal{X}$ is a very small fraction of the volume of $\mathcal{B}$; however, we find that even in that case the algorithm is very fast in execution.

A nice feature of OICs is that a fixed number $|\mathcal{X}^n|$ organically pack more densely as improved $f^{n_*}_{\min}$ are found. However, as progress slows in later iterations, the density will plateau, with two consequences: (1) impacting convergence diagnostics based on the candidates (like $\max \mathbb{E}\{I_Y\}$) and (2) causing the proportion of $\mathcal{X}^n$ whose EI is nonzero to dwindle. We address (1) by declaring approximate convergence, ending an inner loop search, if ten trials pass without improving $y^n_{\min}$. When $\mathbb{E}\{I_Y\}$ is guiding the search, earlier approximate convergence is declared when $\max_{x \in \mathcal{B}} \mathbb{E}\{I_Y(x)\} < \epsilon$, for some tolerance ϵ. Consequence (2) can be addressed by increasing $|\mathcal{X}^n|$ over time; however, we find it simpler to default to an $\mathbb{E}\{Y\}$-based search if less than, say, 5% of $\mathcal{X}^n$ gives nonzero improvement. This recognizes that the biggest gains to the exploratory features of EI are realized early in the search, when the risk of being trapped in an inferior local mode is greatest.

We close the description here by recognizing that while this explains some of the salient details of our implementation, many specifics have been omitted for space considerations. For full transparency please see the `optim.auglag` function and documentation in the `laGP` package. That routine implements all variations considered here.

4.2 Implementation for classical AL comparators

We now summarize the particular implementation details for our comparators.

Direct: For MADS we use the implementation in the `NOMAD` software (Le Digabel, 2011; Abramson et al., 2014). Beyond adaptations for the maximum mesh index in Section 2.4, software defaults are used throughout with the direction type set to "`OrthoMads n+1`" (Audet et al., 2012) and quadratic models (Conn and Le Digabel, 2013) disabled. `NOMAD` can handle constraints natively by using a progressive barrier approach (Audet and Dennis, 2009), which we include as a representative comparator from outside the APM framework.

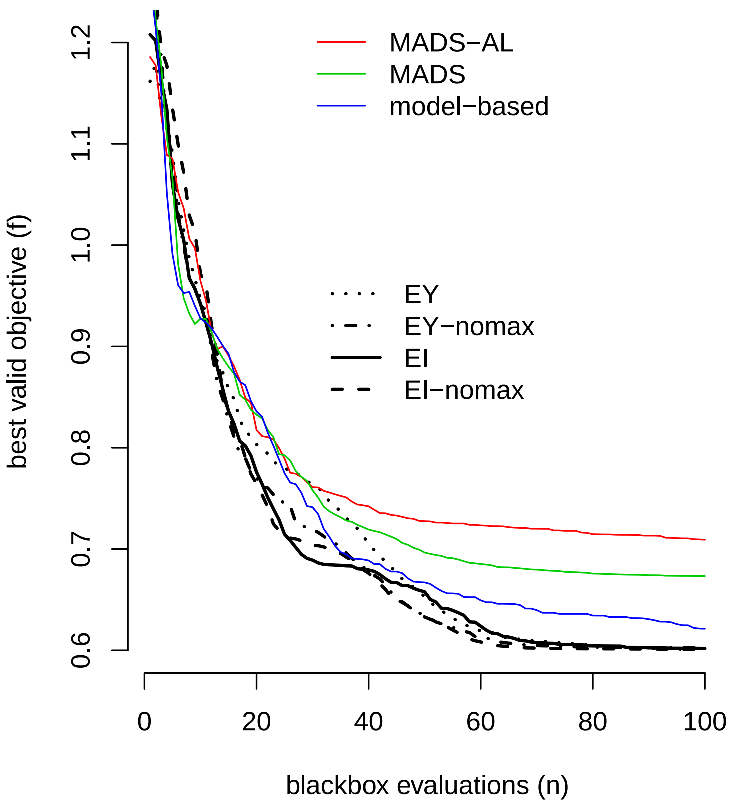
Model-based: We used the same code employed in Kannan and Wild (2012). A maximum of 50 blackbox evaluations were allotted to solving the subproblem (5), with early termination being declared if the norm of the gradient of the approximated AL [see (6)] was below 10^{-2}; for the toy problem this model gradient condition determined the inner termination, and for the motivating hydrology problem in Section 5 the budget determined the inner termination. The initial trust-region radius was taken to be $\Delta^0 = 0.2$ for the toy problem (2) and $\Delta^0 = 10,000$ for the hydrology problem. In order to remain consistent with Kannan and Wild (2012), a maximum of 5 outer iterations (see Algorithm 1) were performed. If there remained function/constraint evaluations in the overall budget, the method was rerun from a random starting point (without incorporating any of the history of previous run(s)).

4.3 Empirical results for the toy problem

Figure 2 summarizes the results of a Monte Carlo experiment for the toy problem described in Section 1. Each of 100 repetitions is initialized with a different random starting value in $\mathcal{B} = [0, 1]^2$. The graph in the figure records the average of the best valid value of the objective over the iterations. The plotted data coincide with the numbers shown in the middle section of the accompanying table for the 25th, 50th and final iteration. The other two sections show the 90% quantiles to give an indication of worst- and best-case behavior.

Since AL-based methods tend to focus just outside of active constraints, examining only strictly feasible points can lead to erroneous conclusions when comparing methods. Therefore, we have followed the convention in constrained optimization and tolerate a small degree of constraint violation when summarizing results for our classical comparators: we consider a point $x^{(j)}$ to be effectively valid if $\| \max(0, c(x^{(j)}))\|_\infty \leq 10^{-3}$.

Figure 2 indicates that all variations on our methods eventually outperform the classical comparators in terms of both average and worst-case behavior. All methods find the right global minima in five or more cases (5% results), but only the EI-based ones perform substantially better in the worst case (using the 95% numbers). In only one case out of 100 did EI not find the global minima, whereas 15% of the model-based runs failed to find it (these runs finding other local solutions instead). Except for a brief time near iteration $n = 50$, and ignoring the first 20 iterations where all methods are about equally good, EI-based comparators dominate EY analogues. There is a period ($n \approx 35$) where EI's average progress stalls temporarily. We observed that this usually marks a transitional period from exploratory to primarily exploitive behavior. Toward the end of the trials, the methods based on dropping the max from Eq. (7) win out. Ignoring regions of the space that give large negative values of the constraint seems to help once the methods have abandoned their more exploitative behavior. However, this comes at the cost of poorer performance earlier on.



n	25	50	100
95%			
EI	0.866	0.775	0.602
EI-nomax	0.906	0.770	0.603
EY	1.052	0.854	0.603
EY-nomax	1.042	0.796	0.603
MADS-AL	1.070	0.979	0.908
MADS	1.056	0.886	0.863
model	1.064	0.861	0.750
average			
EI	0.715	0.658	0.602
EI-nomax	0.715	0.633	0.601
EY	0.779	0.653	0.601
EY-nomax	0.743	0.634	0.603
MADS-AL	0.789	0.728	0.709
MADS	0.793	0.697	0.673
model	0.775	0.667	0.621
5%			
EI	0.610	0.602	0.600
EI-nomax	0.613	0.601	0.600
EY	0.607	0.601	0.600
EY-nomax	0.606	0.600	0.600
MADS-AL	0.600	0.600	0.600
MADS	0.608	0.600	0.599
model	0.600	0.599	0.599

Figure 2: Results for the motivating problem in Section 1 over 100 Monte Carlo repetitions with a random x^0. The plot tracks the average best valid value of the objective over 100 blackbox iterations; the table shows distributional information at iterations 25, 50, and 100.

5 Pump-and-treat hydrology problem

Worldwide, there are more than 10,000 contaminated land sites (Meer et al., 2008). Environmental cleanup at these sites has received increased attention over the past 20–30 years. Preventing the migration of contaminant plumes is vital to protect water supplies and prevent disease. One approach is pump-and-treat remediation, in which wells are strategically placed to pump out contaminated water, purify it, and inject the treated water back into the system. For some situations, pump-and-treat is an effective way of reducing high concentrations of contaminants and preventing their spread. One case study of such remediation is the 580-acre Lockwood Solvent Groundwater Plume Site, an EPA Superfund site located near Billings, Montana. As a result of industrial practices, the groundwater at this site is contaminated with volatile organic compounds that are hazardous to human health (United States Environmental Protection Agency, 2013). Figure 3 shows the location of the site and provides a simplified illustration of the two contaminant plumes that threaten the Yellowstone River. In order to prevent further expansion of these plumes, the placement of six pump-and-treat wells has been proposed, as shown in the figure.

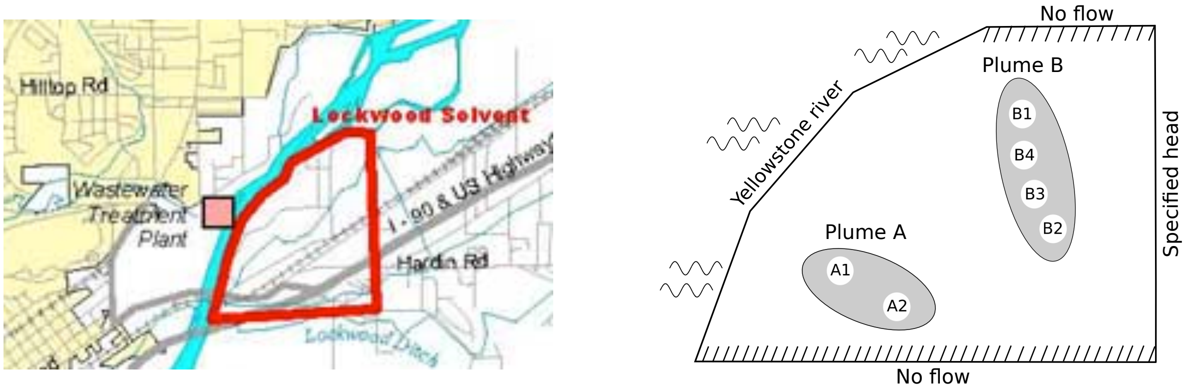


Figure 3: Lockwood site and its contaminant plumes. The map on the *left* identifies the Lockwood Solvent region and shows its proximity to the Yellowstone River and the city of Billings (image from the website of Agency for Toxic Substances & Disease Registry (2010)). The *right* panel illustrates the plume sites, its boundaries (including the Yellowstone river), and the proposed location of six remediation wells (A1, A2, B1, B2, B3, B4).

Mayer et al. (2002) posed the pump-and-treat problem as a constrained blackbox optimization problem, and Fowler et al. (2008) explored the applicability of a variety of derivative-free optimization approaches to solve this problem. For the version of the problem considered here, the pumping rates are varied in order to minimize the cost of operating the system subject to constraints on the contaminant staying within the plume boundaries. Letting x_j denote the pumping rate for well j, one obtains the constrained problem

$$\min_x \{f(x) = \sum_{j=1}^{6} x_j : c_1(x) \leq 0,\ c_2(x) \leq 0,\ x \in [0, 2 \cdot 10^4]^6\}.$$

The objective f is linear and describes the costs required to operate the wells. In the absence of the constraints c, the solution is at the origin and corresponds to no pumping and no remediation. The two constraints denote flow exiting the two contaminant plumes. An *analytic element method* groundwater model simulates the amount of contaminant exiting the boundaries and is treated as a blackbox (Matott et al., 2006). This model never returns negative values for the constraint, and this nonsmoothness—right at the constraint boundary—can present modeling challenges.

5.1 Some comparators

Matott et al. (2011) featured this example in a comparison of MATLAB and Python optimizers, treating constraints via APM. The results of this study are shown in the *left* panel of Figure 4 under a total budget of 1,000 evaluations. All comparators were initialized at

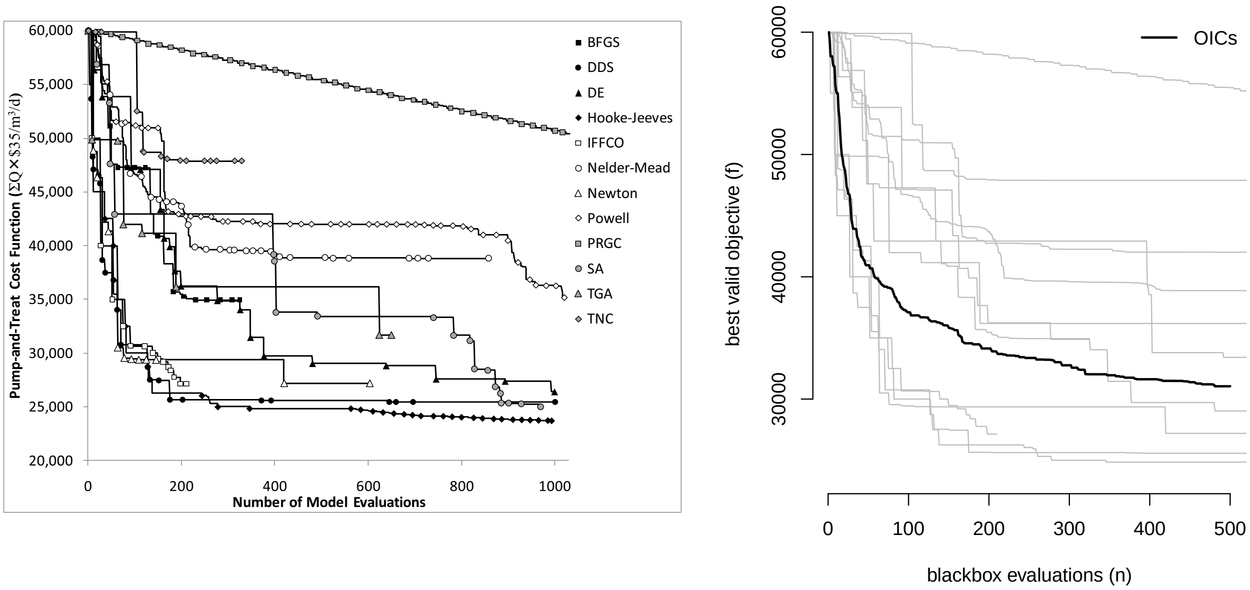


Figure 4: Progress of algorithms on the Lockwood problem; the vertical axes denote the value of the objective at the best valid iterate as a function of the number of optimization iterations. The *left* graph shows the results of algorithms compared by (Matott et al., 2011); the *right* one abstracts the *left* graph for further comparison (e.g., using OICs).

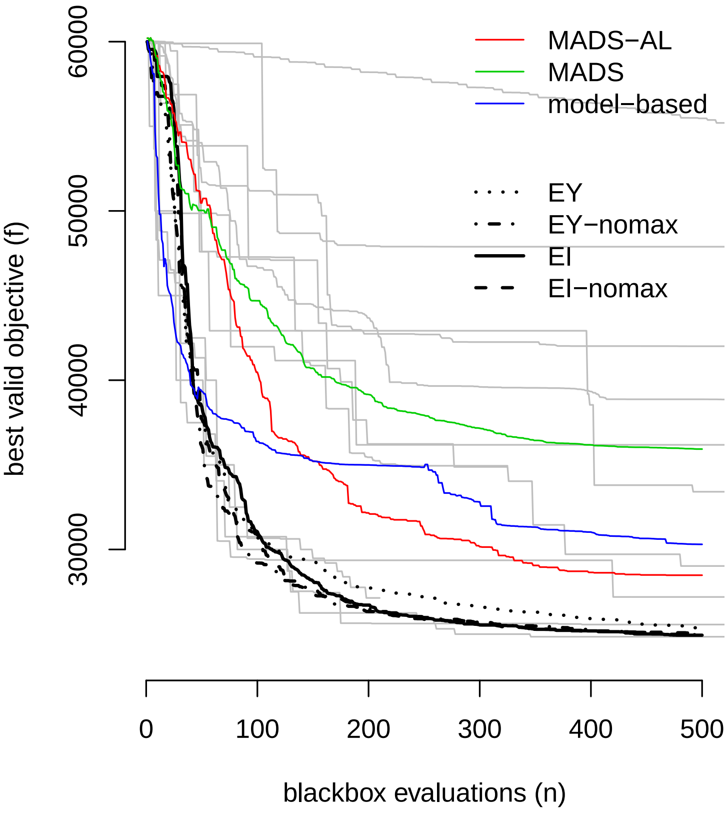
the valid input $x^0 = (10,000, \ldots, 10,000)^\top$. The `Hooke-Jeeves` algorithm, an unconstrained "direct search" method (Hooke and Jeeves, 1961), performed best. To abstract this set of results as a benchmark for our numerical studies, we superimpose our new results on the first 500 iterations of these trajectories. The *right* panel of Figure 4 shows an example, with a simple comparator overlaid based on stochastic search with OICs [Section 4.1].

One may find surprising the fact that simple stochastic search—based on sampling one OIC in each trial and updating the best valid value of the objective when a new one is found—performs well relative to much more thoughtful comparators. Since the method is stochastic, we are revealing its average behavior over thirty replicates. On average, it is competitive with the best group of methods for the first twenty-five iterations or so, suggesting that those methods, while implementing highly varied protocols, are not searching any better than randomly in early stages. We also observe that even after those early stages, OICs still outperform at least half the comparators for the remaining trials. Those methods are getting stuck in local minima, whereas OICs are shy of clumsily global. However pathologically slow a random search like this may be to converge, its success on this problem illustrates a clear benefit to exploration rather than exploitation in early stages.

5.2 Using augmented Lagrangians

Figure 5 shows the results of a Monte Carlo experiment set up like the one in Section 4.3. In this case each of thirty repetitions was initialized randomly with $x^0 \in \mathcal{B} = [0,\, 20,000]^6$.

The comparators from Section 5.1 are shown in gray; note, however, that these used a fixed starting location x^0, *not* a random one—that is, they were not included in the Monte Carlo. From the figure we can see that the relative ordering of our comparators is roughly the same as for the toy problem, except that results vary for the classical ones depending on the number of evaluations, n. The surrogate-model-based average and worst-case behaviors are better than those of the other AL comparators and are competitive with those of the best APMs from Matott et al. We note that many of the individual Monte Carlo runs of our EI- and EY- based methods outperformed all APM comparators, including `Hooke-Jeeves`.



n	100	200	500
	95%		
EI	37584	28698	25695
EI-nomax	43267	28875	25909
EY	36554	32770	27362
EY-nomax	36446	29690	26220
MADS-AL	60000	49020	32663
MADS	60000	60000	60000
model	60000	60000	35730
	average		
EI	31048	26714	24936
EI-nomax	30874	26333	25032
EY	30701	27721	25359
EY-nomax	29199	26493	24954
MADS-AL	40474	32162	28479
MADS	44694	39157	35931
model	36378	34994	30304
	5%		
EI	27054	25119	24196
EI-nomax	26506	24367	24226
EY	25677	24492	24100
EY-nomax	25185	24487	24128
MADS-AL	30776	26358	24102
MADS	30023	26591	23571
model	25912	25164	24939

Figure 5: Results for the Lockwood problem over 30 Monte Carlo repetitions with a random x^0. The plot tracks the average best valid value of the objective over blackbox iterations; the table shows more distributional information at iterations 25, 50, and 100.

In fact, one has reason to believe that the initializing value x^0 used by those methods is a tremendous help. For example, when running `MADS` (no AL) with that same value, it achieved the best result in our study, 23,026. That MADS' average behavior is much worse suggests extreme differential performance depending on the quality of initialization, particularly with regard to the validity of the initial value x^0. In fact, the 95% section of the table reveals that a substantial proportion (5-10%) of the repetitions resulted in no valid solution (up to the

10^{-3} tolerance) even after exhausting the full budget of $n = 500$ iterations.[4] Had the Matott et al. comparison been randomly initialized, we expect that the best comparators would similarly have fared worse. By contrast, in experiments with the surrogate-based methods using the same valid x^0 we found (not shown) no differences, up to Monte Carlo error, in the final solutions we obtained.

6 Discussion

We explored a hybridization of statistical global optimization with an amenable mathematical programming approach to accommodating constraints. In particular, we combine Gaussian process surrogate modeling and expected improvement methods from the design of computer experiments literature with an additive penalty method that has attractive convergence properties: the augmented Lagrangian. The main advantage of this pairing is that it reduces a constrained optimization into an unconstrained one, for which statistical methods are more mature. Statistical methods are not known for their rapid convergence to local optima, but they are more conservative than their mathematical programming analogues: in many cases offering better global solutions for similar computational effort (number of blackbox function evaluations).

This paper has demonstrated the clear potential of such an approach. We have extended the idea of EI to a composite objective arising from the AL and showed that the most sensible variations on such schemes consistently outperform similar methods leveraging a more traditional optimization framework whose focus is usually more local. Still, we see opportunities for further improvement. For example, we anticipate gains from a more aggressive hybridization that acknowledges that the statistical methods fail to "penetrate" into local troughs, particularly toward the end of a search. In the unconstrained context, Gray et al. (2007) and Taddy et al. (2009) have had success pairing EI with the `APPS` direct search method (Kolda et al., 2003). Gramacy and Le Digabel (2011) took a similar tack with MADS. Both setups port provable local convergence from the direct method to a more global search context by, in effect, letting the direct solver take over toward the end of the search in order to "drill down" to a final solution.

Other potential extensions involve improvements on the statistical modeling front. For example, our models for the constraints in Section 3.2 are explicitly independent for each c_i, $i = 1, \ldots, m$, leaving untapped potential to leverage cross correlations (e.g., Williams et al., 2010). Moreover, ideas from multiobjective optimization may prove helpful in our multiconstraint format. Treating them as we do in a quadratic composite (via the AL) represents one way forward; however, keeping them separated with Pareto-optimal-like strategies may prove advantageous as a way to reconcile competing constraint information. A good starting point from the statistical literature may be the work of Svenson and Santner (2012) or Picheny (2013), both of which consider EI-like methods.

There may be alternative ways to acknowledge—in the known monotone objective (f) case, as in both of our examples—that the solution lies on a constraint boundary. Our

[4] We put 60,000 in as a placeholder for these cases.

ideas for this case (e.g., dropping the max in the AL (7)) are attractive because they can be facilitated by a minor coding change, but they yield just modest improvements. It is also risky when the problem includes nonbinding constraints at the solution, by inappropriately inflating the importance of candidates well inside the valid region according to one constraint, but well outside for another. The slack variable approach of Kannan and Wild (2012) may present an attractive remedy for this case. It may also be more natural when the $c(x)$ returns only nonnegative values, as in our hydrology example. Alternatively, knowledge that the solution lies on a boundary could be exploited by letting a fitted classification surface explicitly sample there (e.g., Lee et al., 2010). Such an approach would benefit from further hybridization with an EI-like scheme so as not to focus on parts of the boundary that are not improving on the objective (Lindberg and Lee, 2015).

In closing, however, we remark that perhaps extra complication, which is what many of the above ideas entail, may not be pragmatic from an engineering perspective. The AL is a simple framework and its hybridization with GP models and EI is relatively straightforward, allowing existing statistical software to be leveraged directly (e.g., `laGP` was easy to augment to accommodate all the new methodology described here). This is attractive because, relative to the mathematical programming literature, statistical optimization has few constrained optimization methods readily deployable by practitioners. The statistical optimization literature is still in its infancy in the sense that bespoke implementation is required for most novel applications. By contrast, software packages such as `NOMAD` and `TAO` (Munson et al., 2012) generally work right out of the box. It is hard to imagine matching that engineering capability for difficult constrained optimization problems with statistical methodology if we insist on those methods being even more intricate than the current state of the art.

Acknowledgments

We would like to thank the American Institute of Mathematics for hosting them in a series of small group meetings. Lee was supported by National Science Foundation grant DMS-0906720. Wild was supported by the Applied Mathematics activity within the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under Contract No. DE-AC02-06CH11357. Le Digabel and Ranjan's research is supported by discovery grants from the Natural Sciences and Engineering Research Council of Canada. Thanks to David Lindberg for the Lockwood diagram in the right panel of Figure 3.

References

Abramson, M. A., Audet, C., Couture, G., Dennis, Jr, J. E., Le Digabel, S., and Tribes, C. (2014). "The NOMAD project." Software available at `http://www.gerad.ca/nomad`.

Agency for Toxic Substances & Disease Registry (2010). "Public Health Assessment of the Lockwood Solvent Groundwater Plume." `http://www.atsdr.cdc.gov/HAC/pha/pha.asp?docid=1228&pg=3`.

Audet, C. and Dennis, Jr, J. E. (2006). "Mesh Adaptive Direct Search Algorithms for Constrained Optimization." *SIAM J. on Optimization*, 17, 1, 188–217.

— (2009). "A Progressive Barrier for Derivative-Free Nonlinear Programming." *SIAM J. on Optimization*, 20, 1, 445–472.

Audet, C., Dennis Jr, J., Moore, D., Booker, A., and Frank, P. (2000). "Surrogate-Model-Based Method for Constrained Optimization." In *AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*.

Audet, C., Ianni, A., Le Digabel, S., and Tribes, C. (2012). "Reducing the Number of Function Evaluations in Mesh Adaptive Direct Search Algorithms." Tech. Rep. G-2012-43, Les Cahiers du GERAD. To appear in *SIAM J. on Optimization*.

Bertsekas (1982). *Constrained Optimization and Lagrange Multiplier Methods*. New York, NY: Academic Press.

Booker, A. J., Dennis Jr, J. E., Frank, P. D., Serafani, D. B., Torczon, V., and Trosset, M. W. (1999). "A Rigorous Framework for Optimisation of Expensive Functions by Surrogates." *Structural Optimization*, 17, 1–13.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model Building and Response Surfaces*. Oxford: Wiley.

Boyle, P. (2007). "Gaussian Processes for Regression and Optimization." Ph.D. thesis, Victoria University of Wellington.

Brochu, E., Cora, V. M., and de Freitas, N. (2010). "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning." Tech. rep., University of British Columbia. ArXiv:1012.2599v1.

Bull, A. D. (2011). "Convergence Rates of Efficient Global Optimization Algorithms." *J. of Machine Learning Research*, 12, 2879–2904.

Conn, A. R. and Le Digabel, S. (2013). "Use of Quadratic Models with Mesh-Adaptive Direct Search for Constrained Black Box Optimization." *Optimization Methods and Software*, 28, 1, 139–158.

Conn, A. R., Scheinberg, K., and Vicente, L. N. (2009). *Introduction to Derivative-Free Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Fowler, K. R., Reese, J. P., Kees, C. E., Dennis, J. E., Kelley, C. T., Miller, C. T., Audet, C., Booker, A. J., Couture, G., Darwin, R. W., Farthing, M. W., Finkel, D. E., Gablonsky, J. M., Gray, G. A., and Kolda, T. G. (2008). "A Comparison of Derivative-free Optimization Methods for Water Supply and Hydraulic Capture Community Problems." *Advances in Water Resources*, 31, 5, 743–757.

Gramacy, R. B. (2013). `laGP`*: Local Approximate Gaussian Process Regression*. R package version 1.0.

Gramacy, R. B. and Le Digabel, S. (2011). "The Mesh Adaptive Direct Search Algorithm with Treed Gaussian Process Surrogates." Tech. Rep. G-2011-37, Les Cahiers du GERAD.

Gramacy, R. B. and Lee, H. K. H. (2008). "Bayesian Treed Gaussian Process Models with an Application to Computer Modeling." *J. of the American Statistical Association*, 103, 1119–1130.

— (2011). "Optimization under Unknown Constraints." In *Bayesian Statistics 9*, eds. J. Bernardo, S. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 229–256. Oxford University Press.

Gramacy, R. B. and Polson, N. G. (2011). "Particle Learning of Gaussian Process Models for Sequential Design and Optimization." *J. of Computational and Graphical Statistics*, 20, 1, 102–118.

Gray, G. A., Martinez-Canales, M., Taddy, M., Lee, H. K. H., and Gramacy, R. B. (2007). "Enhancing Parallel Pattern Search Optimization with a Gaussian Process Oracle." In *Proceedings of the 14th NECDC*.

Hooke, R. and Jeeves, T. A. (1961). ""Direct Search" Solution of Numerical and Statistical Problems." *J. of the Association for Computing Machinery (ACM)*, 8, 2, 212–229.

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

Kannan, A. and Wild, S. M. (2012). "Benefits of Deeper Analysis in Simulation-based Groundwater Optimization Problems." In *Proceedings of the XIX International Conference on Computational Methods in Water Resources (CMWR 2012)*.

Kolda, T. G., Lewis, R. M., and Torczon, V. (2003). "Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods." *SIAM Review*, 45, 385–482.

Le Digabel, S. (2011). "Algorithm 909: NOMAD: Nonlinear Optimization with the MADS Algorithm." *ACM Trans. on Mathematical Software*, 37, 4, 44:1–44:15.

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.

Lindberg, D. and Lee, H. K. H. (2015). "Optimization Under Constraints by Applying an Asymmetric Entropy Measure." *J. of Computational and Graphical Statistics*. To appear.

Matott, L. S., Leung, K., and Sim, J. (2011). "Application of MATLAB and Python Optimizers to Two Case Studies Involving Groundwater Flow and Contaminant Transport Modeling." *Computers & Geosciences*, 37, 11, 1894–1899.

Matott, L. S., Rabideau, A. J., and Craig, J. R. (2006). "Pump-and-Treat Optimization Using Analytic Element Method Flow Models." *Advances in Water Resources*, 29, 5, 760–775.

Mayer, A. S., Kelley, C. T., and Miller, C. T. (2002). "Optimal Design for Problems Involving Flow and Transport Phenomena in Subsurface Systems." *Advances in Water Resources*, 25, 1233–1256.

Meer, J. T. M. T., Duijne, H. V., Nieuwenhuis, R., and Rijnaarts, H. H. M. (2008). "Prevention and Reduction of Pollution of Groundwater at Contaminated Megasites: Integrated Management Strategy, and Its Application on Megasite Cases." In *Groundwater Science and Policy: An International Overview*, ed. P. Quevauviller, 405–420. RSC Publishing.

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

Munson, T., Sarich, J., Wild, S. M., Benson, S., and Curfman McInnes, L. (2012). "TAO 2.0 Users Manual." Technical Memo. ANL/MCS-TM-322, Argonne National Laboratory.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. 2nd ed. Springer.

Paciorek, C. J. and Schervish, M. J. (2006). "Spatial Modelling Using a New Class of Nonstationary Covariance Functions." *Environmetrics*, 17, 5, 483–506.

Picheny, V. (2013). "Multiobjective Optimization Using Gaussian Process Emulators via Stepwise Uncertainty Reduction." `http://arxiv.org/abs/1310.0732`.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013). "Quantile-based Optimization of Noisy Computer Experiments with Tunable Precision." *Technometrics*, 55, 1, 2–13.

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

Sasena, M. J. (2002). "Flexibility and Efficiency Enhancement for Constrained Global Design Optimization with Kriging Approximations." Ph.D. thesis, University of Michigan.

Schmidt, A. M. and O'Hagan, A. (2003). "Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations." *J. of the Royal Statistical Society, Series B*, 65, 745–758.

Schonlau, M., Jones, D. R., and Welch, W. J. (1998). "Global Versus Local Search in Constrained Optimization of Computer Models." In *New Developments and Applications in Experimental Design*, vol. 34, 11–25. Institute of Mathematical Statistics.

Svenson, J. D. and Santner, T. J. (2012). "Multiobjective Optimization of Expensive Black-Box Functions via Expected Maximin Improvement." Tech. rep., Ohio State.

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

United States Environmental Protection Agency (2013). "Lockwood Solvent Groundwater Plume." `http://www2.epa.gov/region8/lockwood-solvent-ground-water-plume`.

Wild, S. M. and Shoemaker, C. A. (2013). "Global Convergence of Radial Basis Function Trust-Region Algorithms for Derivative-Free Optimization." *SIAM Review*, 55, 2, 349–371.

Williams, B. J., Santner, T. J., Notz, W. I., and Lehman, J. S. (2010). "Sequential Design of Computer Experiments for Constrained Optimization." In *Statistical Modeling and Regression Structures*, eds. T. Kneib and G. Tutz, 449–472. Springer-Verlag.