Title
Modeling an Augmented Lagrangian for Blackbox Constrained Optimization

Permalink
https://escholarship.org/uc/item/60z1260d

Journal
Technometrics, 58(1)

ISSN
0040-1706

Authors
Gramacy, Robert B
Gray, Genetha A
Le Digabel, Sébastien
et al.

Publication Date
2016-01-02

DOI
10.1080/00401706.2015.1014065

Peer reviewed

Modeling an Augmented Lagrangian for Blackbox Constrained Optimization

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

March 4, 2015

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 optimization approaches for blackbox problems 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) surrogates 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. Until very recently the statistical literature has, by contrast, only offered solutions tailored to 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. In many real-world optimization problems, handling constraints presents the biggest challenge; many have

[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 $i = 1, \ldots, d$).

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 an existing statistical framework. For example, one could treat $c(x)$ as binary (Gramacy and Lee, 2011), ignoring information about the *distance* to the boundary separating feasible ("valid") and infeasible ("invalid") regions. To more fully utilize all available information, we propose 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.

For the unconstrained problem(s) we develop novel surrogate modeling and EI techniques tailored to the form of the AL, similar to the way that Parr et al. (2012) deploy EI on a *single* conversion from constrained to unconstrained problems. Borrowing the AL setup, and utilizing an appropriate sequence of unconstrained problems, weakens the burden on the statistical optimizer—deployed in this context as a subroutine—and leverages convergence guarantees from the mathematical programming literature. Under specific conditions we can derive closed-form expressions, like EI, to guide the optimization subproblems, and we explore numerical/Monte Carlo alternatives for other cases. Importantly, our use of Monte Carlo is quite unlike optimization by stochastic search, e.g., simulated annealing (SA, Kirkpatrick et al., 1983). SA, and other methods utilizing inhomogeneous Markov chains, offer global convergence guarantees asymptotically. However, in our experience and indeed in our empirical studies herein, such schemes are less than ideal when expensive blackbox evaluation severely limits the number of simulations that can be performed.

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 that emphasize 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 synthetic 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 extension.

A toy problem. Consider the following test problem of the form (1) with a (known) linear objective in two variables, $f(x) = x_1 + x_2$, bounding box $\mathcal{B} = [0, 1]^2$, and two (blackbox)

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

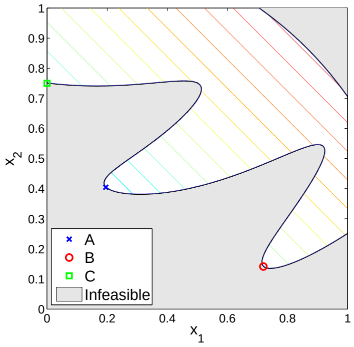


Figure 1: Toy problem and its local minimizers; only x^A is a global minimizer.

nonlinear constraints 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 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 hybridize: response surface models, expected improvement, and the augmented Lagrangian. Implementations details are deferred to Section 4.

2.1 Surrogate modeling framework for optimization

Examples of statistical models guiding optimization date back at least to Mockus et al. (1978). The technique has since evolved, but the basic idea still involves training a flexible regression model f^n on input-output pairs $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ and using aspects of f^n to help choose $x^{(n+1)}$. One option is to search the mean of a predictive surface $f^n(x)$ derived from f^n, serving as a *surrogate* for the true $f(x)$, for minima. Gaussian process (GP) regression models provide attractive f^n for a deterministic objective, f, since GPs produce highly accurate, conditionally normal predictive distributions and can interpolate if desired (see, e.g., Santner et al., 2003). This so-called surrogate modeling framework for optimization is focused primarily on the objective f. Extensions have been made to accommodate constraints (e.g., Audet et al., 2000), often 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 surrogate $f^n(x)$, 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:

$$\mathbb{E}\{I(x)\} = (f^n_{\min} - \mu^n(x))\Phi\left(\frac{f^n_{\min} - \mu^n(x)}{\sigma^n(x)}\right) + \sigma^n(x)\phi\left(\frac{f^n_{\min} - \mu^n(x)}{\sigma^n(x)}\right), \qquad (2)$$

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 ($\sigma^n(x)$).

Leveraging (2), Jones et al. proposed an *efficient global optimization* (EGO) scheme using branch-and-bound to maximize $\mathbb{E}\{I(x)\}$. In a later paper, Schonlau et al. (1998) provided an analytical form for a *generalized EI* based on a powered up improvement $I_g(x) = |f^n_{\min} - Y(x)|^g \mathbb{I}_{\{Y(x) < f^n_{\min}\}}$. 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, and the hybrid $\mathbb{E}\{I(x)\}^2 + \mathbb{V}\mathrm{ar}[I(x)]$, respectively.

Under weak regularity conditions, search algorithms based on EI converge to the global optimum. EGO, which specifically pairs GP surrogates with EI, can be seen as one example of a wider family of routines. For example, radial basis function surrogates 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; in practice, these can pathologically slow convergence and/or lead to local rather than global solutions. Practitioners instead increasingly prefer hybrids between global EI and local search (e.g., Gramacy and Le Digabel, 2011).

2.3 Augmented Lagrangian framework

Augmented Lagrangian methods (see, e.g., Nocedal and Wright, 2006) 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

$$L_A(x;\lambda,\rho) = f(x) + \lambda^\top c(x) + \frac{1}{2\rho}\sum_{j=1}^{m} \max\left(0, c_j(x)\right)^2 , \tag{3}$$

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

The first two terms in (3) correspond to the Lagrangian, which is the merit function that defines stationarity for constrained optimization problems. Without the second term, (3) reduces to an *additive penalty method* (APM) approach to constrained optimization. Unless considerable care is taken 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

$$\min_x \left\{ L_A(x;\lambda^{k-1},\rho^{k-1}) : x \in \mathcal{B} \right\}. \tag{4}$$

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 (4), comprising the "inner loop" [step 2].

```
Require: $\lambda^0 \geq 0$, $\rho^0 > 0$
1: for $k = 1, 2, \ldots$ (i.e., each "outer" iteration) do
2:     Let $x^k$ (approximately) solve (4)
3:     Set $\lambda_j^k = \max\left(0, \lambda_j^{k-1} + \frac{1}{\rho^{k-1}} c_j(x^k)\right)$, $j = 1, \ldots, m$
4:     If $c(x^k) \leq 0$, set $\rho^k = \rho^{k-1}$; otherwise, set $\rho^k = \frac{1}{2}\rho^{k-1}$
5: 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 the outer loop 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 $\left\|\nabla f(x^k) + \sum_{j=1}^{m} \lambda_i^k \nabla c_j(x^k)\right\| \leq \eta_2$. Determining convergence within the inner loop [step 2], is solver dependent; common solvers in our motivating context are discussed below. The theory for global convergence of the overall AL scheme is forgiving about the criteria used to end each local search.

2.4 Derivative-free augmented Lagrangian methods

The inner loop [step 2] of Algorithm 1 can accommodate a host of methods for solving the simply constrained subproblem (4). Solvers can leverage derivatives of the objective and/or constraint functions when available, or be *derivative-free* otherwise. 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 benchmark solvers for the inner loop. We now briefly introduce how these solvers can be situated within the AL framework; software/implementation details and convergence detection are provided in the supplementary material, which also contains the details of three additional comparators that do not leverage the AL.

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. 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 interpolating quadratic approximation models to the objective and constraints. The AL subproblem (4) is then approximately solved by locally solving a sequence of quadratics.

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. The simplest approach involves deploying a statistical surrogate directly on the AL (3), but this has drawbacks. To circumvent these, we consider separately modeling the objective f and each constraint c_j. We then pursue options for using the surrogate to solve (4), 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 surrogate modeling of the AL (3) 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 $(x^{(1)}, f^{(1)}, c^{(1)}), \ldots, (x^{(n)}, f^{(n)}, c^{(n)})$. Then form $y^{(i)} = L_A(x^{(i)}; \lambda^{k-1}, \rho^{k-1})$ via $f^{(i)}$ and $c^{(i)}$, and fit a GP surrogate 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. (2) 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 actual 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 surrogate modeling and EI. It is global because the GP is trained on 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 (3). 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, amplifying the effects of $c(x)$ away from the boundary with the valid region. Most out-of-the-box GP regression methods assume stationarity, and will therefore be a poor match. A related challenge is the max in (3), which produces kinks near the boundary of the valid region, with the regime changing behavior across that boundary.

Modern GP methods accommodate nonstationarity (e.g., Schmidt and O'Hagan, 2003) 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 (unless it accommodates a known mean) 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. To clarify, here we take "separate" to mean *independent* as that simplifies many matters, although extensions to correlated models may yield improvements. Under independence, surrogates $f^n(x)$ for the objective and $c^n(x) = (c_1^n(x), \dots, c_m^n(x))$ for the constraints provide 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 (4) using all n data points seen so far. Samples from those distributions, obtained trivially via GP surrogates, are easily converted into samples from the composite, serving as a surrogate for $L_A(x; \lambda, \rho)$:

$$Y(x) = Y_f(x) + \lambda^\top Y_c(x) + \frac{1}{2\rho} \sum_{j=1}^m \max(0, Y_{c_j}(x))^2. \tag{5}$$

When 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 (5), 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 $\mathbb{E}\{Y(x)\} = \mathbb{E}\{Y_f(x)\} + \lambda^\top \mathbb{E}\{Y_c(x)\} + \frac{1}{2\rho}\sum_{i=i}^m \mathbb{E}\{\max(0, Y_{c_j}(x))^2\}$. The first two expectations are trivial under normal GP predictive equations, giving

$$\mathbb{E}\{Y(x)\} = \mu_f^n(x) + \lambda^\top \mu_c^n(x) + \frac{1}{2\rho} \sum_{j=1}^m \mathbb{E}\{\max(0, Y_{c_j}(x))^2\}, \tag{6}$$

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

$$\begin{aligned} \mathbb{E}\{\max(0, Y_{c_j}(x))^2\} &= \mathbb{E}\{I_{-Y_{c_j}}(x)\}^2 + \mathbb{V}\mathrm{ar}[I_{-Y_{c_j}}(x)] \qquad (7) \\ &= \sigma_{c_j}^{2n}(x) \left[\left(1 + \left(\frac{\mu_{c_j}^n(x)}{\sigma_{c_j}^n(x)} \right)^2 \right) \Phi\left(\frac{\mu_{c_j}^n(x)}{\sigma_{c_j}^n(x)} \right) + \frac{\mu_{c_j}^n(x)}{\sigma_{c_j}^n(x)} \phi\left(\frac{\mu_{c_j}^n(x)}{\sigma_{c_j}^n(x)} \right) \right], \end{aligned}$$

by using a result from the generalized EI (Schonlau et al., 1998). Combining (6) and (7) completes the expression for $\mathbb{E}\{Y(x)\}$. When f is known, simply replace μ_f^n with f.

3.3 New expected improvement

The composite random variable $Y(x)$ in Eq. (5) does not have a form that readily suggests a familiar distribution, for any reasonable choice of f^n and c^n (e.g., under a GP model), complicating analytic calculation of EI. A numerical approximation is straightforward by Monte Carlo. Assuming normal predictive equations, 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 $\mathcal{N}(\mu_{c_j}^n, \sigma_{c_j}^{2n})$, respectively, and then average:

$$\begin{aligned} \mathbb{E}\{I_Y(x)\} &\approx \frac{1}{T}\sum_{t=1}^{T} \max(0, y_{\min}^n - y^{(t)}(x)) \qquad (8)\\ &= \frac{1}{T}\sum_{t=1}^{T} \max\left[0, y_{\min}^n - \left(y_f^{(t)}(x) + \lambda^\top y_c^{(t)}(x) + \frac{1}{2\rho}\sum_{j=1}^{m}\max(0, y_{c_j}^{(t)}(x))^2\right)\right], \end{aligned}$$

where $y_{\min}^n = \min_{i=1,\ldots,n}\{f^{(i)} + \lambda^\top c^{(i)} + \frac{1}{2\rho}\sum_{j=1}^m \max(0, c_j^{(i)})^2\}$ is the best value of the AL observed so far, given the current λ and ρ values. 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 (5) 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. (5) where a new composite $\tilde{Y}$ is defined by removing the max. In this special case, one can derive an analytical expression for the EI under GP model for c. Let $I_{\tilde{Y}} = \max\{0, \tilde{y}_{\min} - \tilde{Y}\}$ be the improvement function for the new composite $\tilde{Y}$, suppressing x to streamline notation. Calculating the EI involves the following integral, where $c(y_c)$ represents the density c^n of Y_c, which for a GP is specified by some μ and σ^2:

$$\mathbb{E}\{I_{\tilde{Y}}\} = \int_{-\infty}^{\infty} I_{\tilde{y}} c(y_c)\, dy_c = \int_\theta (\tilde{y}_{\min} - \tilde{y}) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_c-\mu)^2}\, dy_c, \quad \theta = \{y_c : \tilde{y} < \tilde{y}_{\min}\}.$$

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

$$\begin{aligned} \mathbb{E}\{I_Y\} &= \left[\tilde{y}_{\min} - \left(\frac{\mu^2}{2\rho} + \lambda\mu + f\right) - \frac{\sigma^2}{2\rho}\right](\Phi(v_2) - \Phi(v_1)) \qquad (9)\\ &\quad + [\sigma\mu/\rho + \lambda\sigma](\phi(v_2) - \phi(v_1)] + \frac{\sigma^2}{2\rho}(v_2\phi(v_2) - v_1\phi(v_1)),\\ \text{where} \quad v_1 &= \frac{u_- - \mu}{\sigma}, \qquad v_2 = \frac{u_+ - \mu}{\sigma}, \qquad u_\pm = \frac{-\lambda \pm \sqrt{\lambda^2 - 2(f - \tilde{y}_{\min})/\rho}}{\rho^{-1}}. \end{aligned}$$

Otherwise, when $\lambda^2 - 2(f - \tilde{y}_{\min})/\rho < 0$, we have that $\mathbb{E}\{I_{\tilde{Y}}\} = 0$. Rearranging gives $\rho\lambda^2 < 2(f - \tilde{y}_{\min})$, which, as ρ is repeatedly halved, reveals a shrinking region non-zero EI.

Besides pointing to analytically zero EI values under (5), 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. Second, dropping the max in (5) 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 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, especially when Monte Carlo is not feasible. Extending to blackbox f, and modeling Y_f, is straightforward since $Y_f(x)$ features linearly in Eq. (5). 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, $\sum_i Y_{c_j}(x)$. Some care is required because summing can result in information loss which may be extreme if positive and negative $Y_{c_j}(x)$ cancel valid and invalid values. A better approach would be to use $Y_c = \sum_i |Y_{c_j}(x)|$ or $Y_c = \sum_i \max(0, Y_{c_j}(x))$ 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. However, it may introduce complications when the constraint set includes constraints that are not active/binding at the solution, which we discuss further in Section 6.

4 Implementation and illustration

In this section, we first provide implementation details for our proposed methods (Section 3), and then demonstrate how they fare on our motivating toy data (Section 1). Implementation details for our comparators (Section 2.4) are provided as supplementary material. which includes details for three further comparators used in our study: a MADS modification which hands constraints natively, simulated annealing, and an "asymmetric entropy" heuristic (Lindberg and Lee, 2015). 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 easily criticized, 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, requiring modeling a nonstationary composite AL, since that method is dominated by separated modeling of the constituent parts, as 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. (7–8), respectively. We note here that the results based on a Monte Carlo $\mathbb{E}\{Y\}$, via the analog of (8) without "$\max[0, y^n_{\min} -$", and the analytical alternative (7) 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 entertain a variation on the numeric EI that discards the max term. We do not provide results based on the analytic expression (9), however, because doing so requires modeling compromises that hamper effective search. In total we report results for four variations pairing one of $\mathbb{E}\{Y\}$ and $\mathbb{E}\{I_Y\}$ with the original AL (5) 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 model each Y_{c_j} with separate GPs initialized with ten random (space filling) 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—when designs are augmented as Algorithm 1 progresses—we used `updateGP` and `mleGP`, from the `laGP` package (Gramacy, 2014) for `R`. Each inner loop search in Algorithm 1 is based on a random set of 1,000 candidate x locations $\mathcal{X}^n$. Spacing candidates uniformly in $\mathcal{B}$ 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 even when $\mathcal{X}$ is small relative to $\mathcal{B}$.

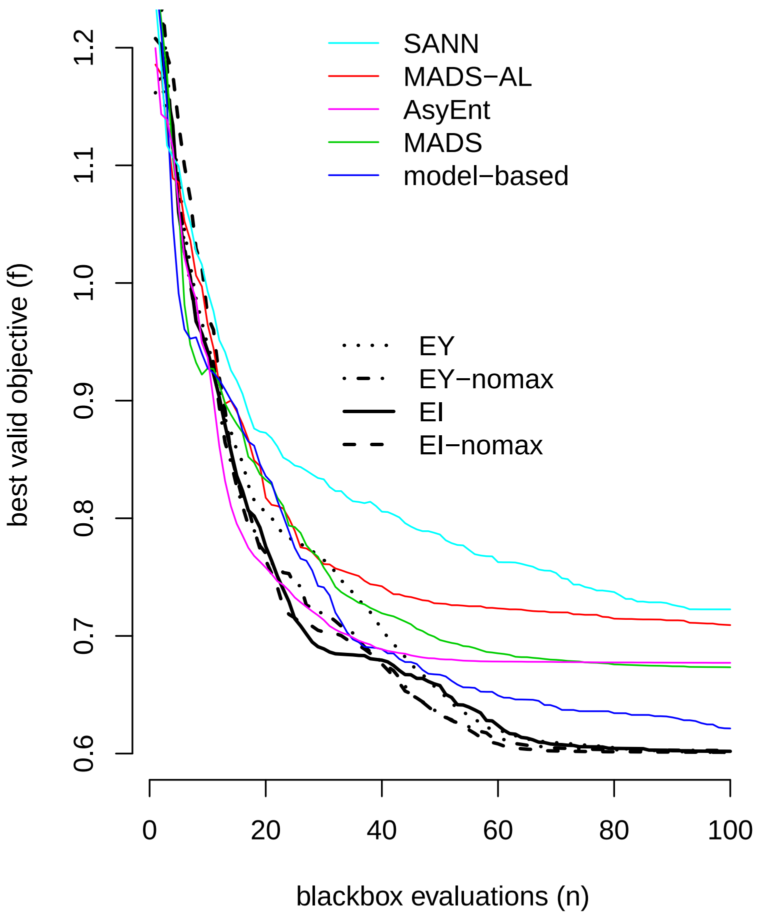
A nice feature of OICs is that a fixed number $|\mathcal{X}^n|$ organically pack more densely as improved $f^{n_*}_{\min}$ are found. However, as progress slows in later iterations, the density will plateau, with two consequences: (1) impacting convergence diagnostics based on the candidates (like $\max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

While the above explains some of the salient details of our implementation, many specifics have been omitted for space considerations. For full transparency please see `optim.auglag` in the `laGP` package, implementing all variations considered here. Gramacy (2014), the package vignette, provides a worked-code illustration for the toy problem considered below.

4.2 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 randomly 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.

Figure 2 indicates that all variations on our methods eventually outperform all 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



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
SANN	1.013	0.940	0.855
MADS-AL	1.070	0.979	0.908
AsyEnt	0.825	0.761	0.758
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
SANN	0.837	0.771	0.716
MADS-AL	0.789	0.728	0.709
AsyEnt	0.733	0.680	0.677
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
SANN	0.648	0.630	0.612
MADS-AL	0.600	0.600	0.600
AsyEnt	0.610	0.601	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.

first 20 iterations for which all methods perform similarly, 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 primarily exploratory to primarily exploitive behavior. Toward the end of the trials, the methods based on dropping the max from Eq. (5) win out. Ignoring regions of the space giving large negative values of $c(x)$ seems to help in these latter stages, but it can hinder performance earlier on.

SA fares worst in this example; however, it does better in our real-data one below. Although we summarize results for the first 100 evaluations, we ran the SA comparator to convergence and report here that each repetition did eventually converge to the global minimum. Reaching convergence, however, took an average of 8,240 evaluations. Compared with our EI-based results, this represents an enormous expense, which foreshadows further discussion in Section 6 about why stochastic search may be less than ideal for optimization of expensive computer simulation experiments. The asymmetric entropy method ("AsyEnt") performs well once active search begins after ten space-filling evaluations. Subsequently its progress plateaus indicating a struggle to consistently hone-in on solutions near the boundary.

5 Pump-and-treat hydrology problem

We turn now to our motivating example. 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 to prevent contaminant spread. A case study of one such remediation is the 580-acre Lockwood Solvent Groundwater Plume Site, an EPA Superfund site located near Billings, Montana, where industrial practices have led to groundwater contamination (United States Environmental Protection Agency, 2013). Figure 3 shows the location of the site and provides a simplified illustration of the contaminant plumes that threaten the Yellowstone River. To prevent further expansion of these plumes, the placement of six pump-and-treat wells has been proposed, as shown in the figure.

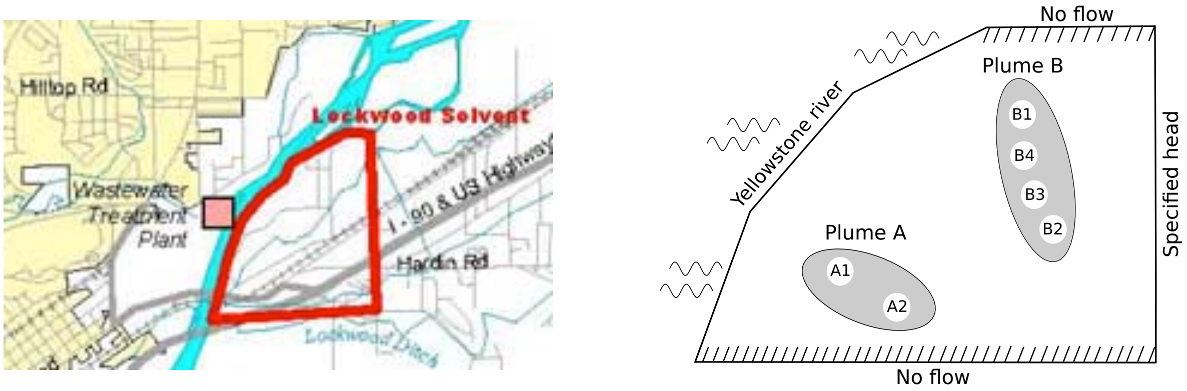


Figure 3: Lockwood site and its contaminant plumes. The map on the *left* identifies the Lockwood Solvent region and shows its proximity to the Yellowstone River and the city of Billings (image from the website of Agency for Toxic Substances & Disease Registry (2010)). The *right* panel illustrates the 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. 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

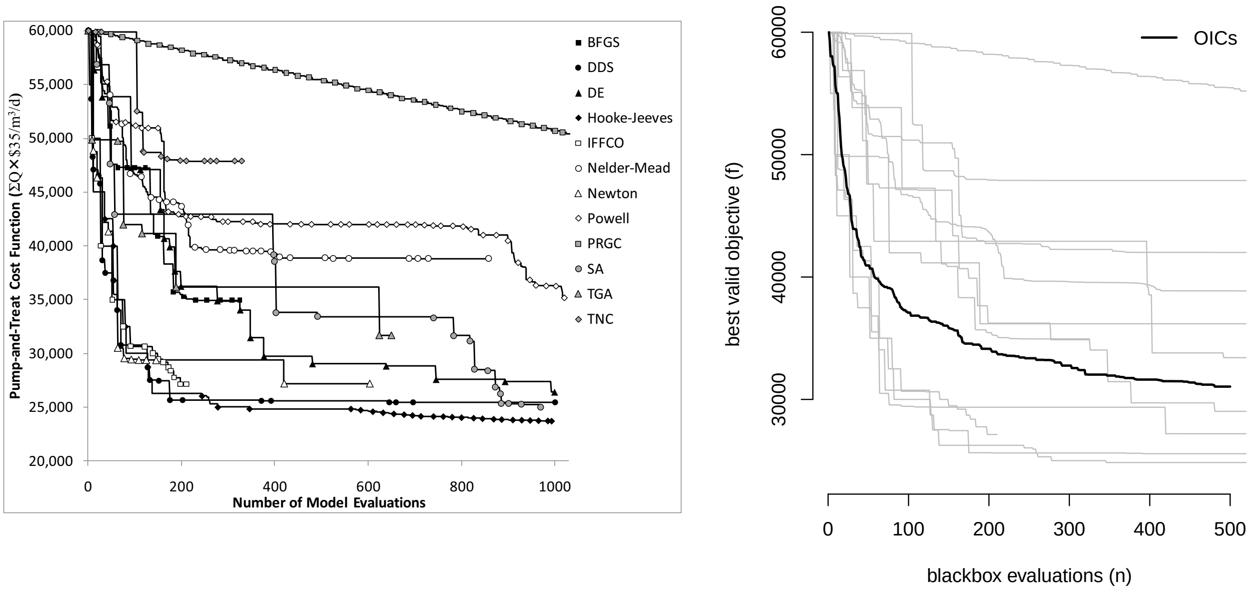


Figure 4: Progress of algorithms on the Lockwood problem; the vertical axes denote the value of the objective at the best valid iterate as a function of the number of optimization iterations. The *left* graph shows the results of algorithms compared by (Matott et al., 2011); the *right* one abstracts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

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 the valid input $x^0 = (10,000, \ldots, 10,000)^\top$. To abstract these trajectories as a benchmark we shall superimpose our new results on an image of the first 500 iterations.

The *right* panel of Figure 4 shows an example, with a simple comparator overlaid based on stochastic search with OICs (Section 4.1). It may be surprising 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. That average 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. 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 over exploitation in early stages.

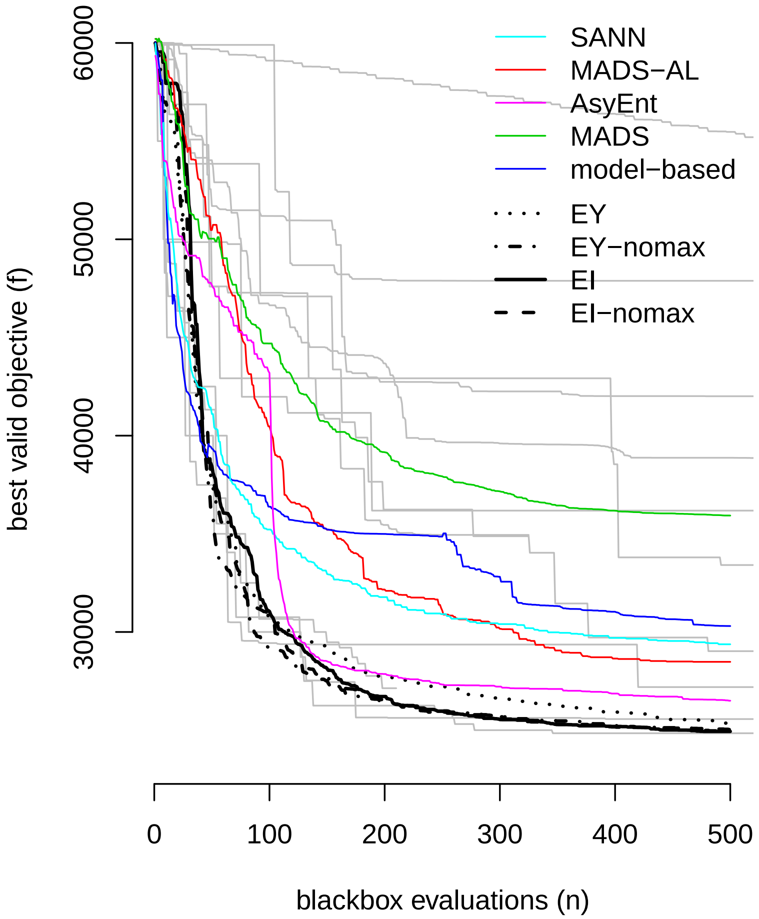
5.2 Using augmented Lagrangians

Figure 5 shows the results of a Monte Carlo experiment set up like the one in Section 4.2. In this case each of thirty repetitions was initialized randomly with $x^0 \in \mathcal{B} = [0, 20000]^6$. Note that comparators from Section 5.1 (in gray) used a single fixed starting location x^0, *not* a random one. From the figure we can see that the relative ordering of our proposed hybrid surrogate/AL comparators is roughly the same as for the toy problem. 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. Many of the individual Monte Carlo runs of our EI- and EY-based methods outperformed all APM comparators.

But some individual APM runs outperform our average results. We 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. Moreover, the 95% section of the table reveals that a substantial proportion (5-10%) of the repetitions resulted in no valid solution even after exhausting the full budget of $n = 500$ iterations.[2] 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.

The SA comparator performed rather better in this experiment compared our synthetic example in Section 4.2, out-performing all but one of the classical AL methods. Due to the expense of the blackbox simulations (and the likelihood that convergence could require thousands of iterations) we did not attempt to run SA to convergence to see if it is ultimately competitive with our EI-based methods. The story is similar for "AsyEnt". After the one hundred space-filling candidates, progress is rapid—beating out our classical comparators—but it quickly plateaus, never quite matching our surrogate-model-based AL ones. Interestingly, SA and AsyEnt have the worst best-case (5% and 500 iterations) results.

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



n	100	200	500
95%			
EI	37584	28698	25695
EI-nomax	43267	28875	25909
EY	36554	32770	27362
EY-nomax	36446	29690	26220
SANN	43928	35456	30920
MADS-AL	60000	49020	32663
AsyEnt	49030	29079	27445
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
SANN	35219	31777	29375
MADS-AL	40474	32162	28479
AsyEnt	43194	27860	26500
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
SANN	28766	27238	26824
MADS-AL	30776	26358	24102
AsyEnt	37483	26681	25377
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 100, 200, and 500.

6 Discussion

We explored a hybridization of statistical global optimization with an amenable mathematical programming approach to accommodating constraints. In particular, we combined 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 with fewer blackbox evaluations.

We 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, there is plenty of room for 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 Gramacy and Le Digabel (2011) have had success pairing EI with modern direct search optimizers. 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 were independent for each c_j, $j = 1, \dots, m$, leaving untapped potential to leverage cross correlations (e.g., Williams et al., 2010). 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; keeping them separated with Pareto-optimal-like strategies may be another promising option (see, e.g., Svenson and Santner, 2012; Picheny, 2013, 2014).

Better analytics offer the potential for further improvement. We resorted to Monte Carlo (MC) for two aspects of our calculations, however it is important to clarify we are not advocating a stochastic search. An analytic EI calculation, or a branch-and-bound-like scheme for optimizing it at each inner-loop search step, would eliminate MC errors and yield an entirely deterministic scheme. Sometimes a bit of stochasticity is welcome in statistical design applications, of which blackbox optimization is an example. However, there are clearly limits to the usefulness of random search in that setting, especially when simulations for the objective or constraint(s) are expensive. This is borne out in our empirical work where a simulated annealing (SA) implementation demonstrates lukewarm results under tight computational budgets. SA offers nice technical guarantees for global solutions but, as the R documentation for `optim`'s `method="SANN"` cautions, it "depends critically on the settings of the control parameters. It is not a general-purpose method but can be very useful in getting to a good value on a very rough surface."

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 ideas for this case (e.g., dropping the max in the AL (5)) 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, especially when $c(x)$ returns only nonnegative values like in our hydrology example, as might explicit learning of classification boundaries to guide sampling (e.g., Lee et al., 2011; Chevalier et al., 2014; 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 the new methods 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 like NOMAD, implementing MADS (see supplementary material) generally work 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 us in a series of small group meetings. Many thanks to an Associate Editor and Referee for thoughtful comments throughout the review process. 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

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.

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

Chevalier, C., Picheny, V., and Ginsbourger, D. (2014). "The KrigInv package: An efficient and user-friendly R implementation of Kriging-based inversion algorithms." *Computational Statistics and Data Analysis*, 71, 1021–1034.

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

Gramacy, R. 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. To appear in *Pacific Journal of Optimization*.

Gramacy, R. B. (2014). `laGP`*: Local Approximate Gaussian Process Regression*. R package version 1.1-3.

Gramacy, R. (2014). "`laGP`: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R." Tech. rep., The University of Chicago. Vignette for the `laGP` package.

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*.

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)*.

Kirkpatrick, S., Gelatt, C. D., and Vecci, M. P. (1983). "Optimization by simulated annealing." *Science*, 220, 671–680.

Lee, H. K. H., Gramacy, R. B., Linkletter, C., and Gray, G. A. (2011). "Optimization Subject to Hidden Constraints via Statistical Emulation." *Pacific J. of Optimization*, 7, 3, 467–478.

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.

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

Parr, J., Keane, A., Forrester, A., and Holden, C. (2012). "Infill sampling criteria for surrogate-based optimization with constraint handling." *Engineering Optimization*, 44, 1147–1166.

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

— (2014). "A Stepwise uncertainty reduction approach to constrained global optimization." In *Proceedings of the 7th International Conference on Artificial Intelligence and Statistics*, vol. JMPR W&CP 33, 787–795.

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.

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.

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.

Supplementary Materials

SM§1 Implementation details for comparators

Classical AL 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., 2014) 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 (4), with early termination being declared if the norm of the gradient of the quadratically approximated AL was below 10^{-2}; for the toy problem this model gradient condition determined inner-loop termination, and for the motivating hydrology problem in Section 5 the budget was The initial trust-region radius was taken to be $\Delta^0 = 0.2$ for the toy problem 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)).

A note on relaxing convergence: AL-based methods from the mathematical programming literature tend to focus just outside of active constraints, so examining only strictly feasible points may not lead to a fair comparison. Therefore, we follow 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 $\| \max(0, c(x^{(j)}))\|_\infty \leq 10^{-3}$. Similar concessions are not required for our EI-based methods, or other comparators. Only strictly valid results with $\| \max(0, c(x^{(j)}))\|_\infty \leq 0$ are reported for those cases.

Other comparators

To broaden the exercises we consider two further comparators: SA, as representative of the stochastic optimization literature; asymmetric entropy boundary exploration method of Lindberg and Lee (2015) as an alternative class of methods discussed further in Section 6.

For SA we use the `method="SANN"` option in the `optim` function for `R`, which is modeled after the Belisle (1992) base specification. Our usage leverages default settings throughout. To ensure that the default settings, particularly for the random walk proposal mechanism, are appropriate for our examples, we pre-scale the input bounding box $\mathcal{B}$ to lie in the unit cube. In our comparisons, we do not penalize SA by counting proposals lying outside $\mathcal{B}$ as blackbox evaluations against the total budget. To accommodate blackbox constraints we deploy an APM approach and sum a suitably scaled objective and (absolute value of) constraint(s) so that the relative weightings in the composite are about equal for each component. This is meant to show SA in an idealized setting, as generally such scales are unknown in advance.

The method of Lindberg and Lee (2015) uses a GP surrogate for the objective function and a classification GP for the constraint, extending Gramacy and Lee (2011). However, rather than mixing the processes to form an integrated improvement objective, it hybridizes entropy search (Gramacy and Polson, 2011) to focus near the classification boundary (of the valid region)—with ordinary EI—to bias the search away from the boundary and into the interior of the valid region. The search is started with a small space-filling design (10 points for the toy problem, 100 for the hydrology problem), and then additional points are chosen by maximizing the product of expected improvement and the fifth power of asymmetric entropy with a mode of 2/3, following recommendations from Lindberg and Lee.

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 `https://www.gerad.ca/nomad`.

— (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., Digabel, S. L., and Tribes, C. (2014). "Reducing the Number of Function Evaluations in Mesh Adaptive Direct Search Algorithms." *SIAM Journal on Optimization*, 24, 2, 621–642.

Belisle, C. (1992). "Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$." *Journal of Applied Probability*, 29, 885–895.

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.

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