Les Cahiers du GERAD ISSN: 0711–2440

Modeling an Augmented Lagrangian for Improved Blackbox Constrained Optimization

R.B. Gramacy, G.A. Gray,
S. Le Digabel, H.K.H. Lee,
P. Ranjan, G. Weels, S.M. Wild

G–2014–17

March 2014

Les textes publiés dans la série des rapports de recherche HEC n'engagent que la responsabilité de leurs auteurs. La publication de ces rapports de recherche bénéficie d'une subvention du Fonds de recherche du Québec – Nature et technologies.

Modeling an Augmented Lagrangian for Improved Blackbox Constrained Optimization

Robert B. Gramacy

The University of Chicago Booth School of Business
Chicago, IL 60605, USA

rbgramacy@chicagobooth.edu

Genetha A. Gray

Most of the work was done while at
Sandia National Labs, Livermore, CA, USA

Sébastien Le Digabel

GERAD & Department of mathematics and industrial engineering
Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7

sebastien.le.digabel@gerad.ca

Herbert K.H. Lee

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

Pritam Ranjan

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

Garth Wells

Department of Engineering, University of Cambridge
Cambridge CB2 1PZ, UK

Stefan M. Wild

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

March 2014

Les Cahiers du GERAD

G–2014–17

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, Gaussian process, nonparametric regression and classification, sequential design, expected improvement.

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

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 offer more global scope compared to non-statistical approaches. Methods based on expected improvement (EI, Jones et al., 1998) enjoy global convergence properties and compare favorably to alternatives when objective evaluations are expensive, are simulated by Monte Carlo (Picheny et al., 2013), or are otherwise observed with noise, or when there are many local optima. In more conventional contexts, however, non-statistical approaches are usually preferred. We attribute this primarily to two factors: rates of convergence and treatment of constraints. Global search is slower than local search; so, for easier problems, the statistical methods underperform. Additionally, statistical methods are more limited in their ability to handle general constraints in optimization. Here we explore a hybrid approach that pairs a global statistical perspective (Gaussian process emulation) with a classical technique for accommodating constraints (the augmented Lagrangian approach).

We consider constrained nonlinear 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.

This abstract problem in (1) is a challenging problem for general nonlinear constraints c, and even more difficult when evaluation of at least one of f and c requires running an expensive, 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 only offered solutions in certain contexts. For example, Schonlau et al. (1998) adapted EI for unknown f and known c; Gramacy and Lee (2011) considered unknown f and unknown $c \in \{0, 1\}$; and Williams et al. (2010) considered f and c coupled by an integral operator. These methods work well in their chosen contexts, but are rather limited in scope.

The current state of affairs is unfortunate as we believe that statistical methods have much to offer. In addition to searching more globally, statistical methods can offer robustness (Taddy et al., 2009), natively facilitate uncertainty quantification, and (like statistical models for other engineering contexts) enjoy a near monopoly in the noisy observation case. Moreover, in many real-world optimization problems it is the handling of the constraints that is challenging and expensive; 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 this context, we consider in Section 5 a hydrology problem; even with a simple linear objective function, learning a highly nonconvex feasible region complicates the search for a global minimum.

One way forward is to force our abstract problem (1) into a more limited, but already developed, statistical framework. For example, the *integrated expected conditional improvement* (IECI) method of Gramacy and Lee (2009) could be applied by treating $c(x)$ as binary. However, this approach discards useful information about the *distance* to the boundary separating feasible ("valid") and infeasible ("invalid") regions. The key contribution of this paper is to develop a statistical approach based on the augmented Lagrangian (e.g.,

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

Bertsekas, 1982) used in the mathematical programming literature. Augmented Lagrangian methods convert a problem with general constraints into a sequence of unconstrained (or simply constrained) problems and fully leverage knowledge on how far a point is from satisfying a constraint. 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 end result is a scheme that compares favorably to modern alternatives in the mathematical programming literature, especially in the presence of several nonglobal minima.

Although the approach we advocate is quite general, for specificity in this paper, we focus on blackbox optimization problems for which the objective f is known while the constraints c are unknown (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 the advocated approach can be extended to unknown f by pairing it with standard surrogate modeling techniques.

The remainder of the paper is organized as follows. Section 1 describes a model problem that introduces the challenges in this area and makes connections to a motivating real-data problem from hydrology, which is presented in detail in Section 5. Section 2 reviews statistical optimization, particularly in the unconstrained context, and introduces the augmented Lagrangian (AL) framework for constrained optimization. Section 3 contains the essence of our methodological contribution, combining statistical surrogates with the AL. Section 4 describes implementation details and provides results on our model example. Section 5 then provides a similar comparison for our motivating 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 that illustrates some of the challenges of solving problems of the form (1). The problem consists of 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 binding, if this bound were not present, x^C would not be a local solution. The second constraint may seem uninteresting here, but it reminds us that the solution may not be on every constraint boundary, and it is important to accomodate this type of constraint when the constraints are unknown. This problem shares several characteristics of our pump-and-treat hydrology problem detailed in Section 5, notably 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) and its use with 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 where statistical models are used to guide an optimization date back at least to Mockus et al. (1978) and Box and Draper (1987). Although this technique has evolved over the years, the basic idea involves training a flexible regression model f^n on input-output pairs $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ and letting aspects of the predictive distribution $f^n(x)$ choose $x^{(n+1)}$ according to some criteria. One option is to search the

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

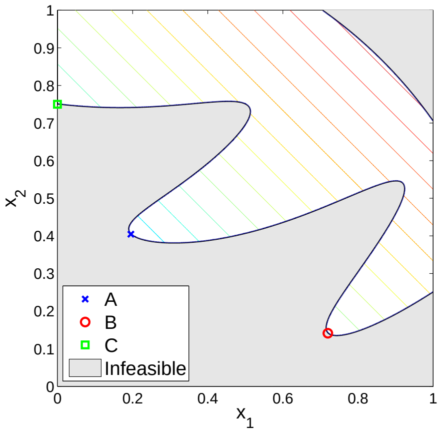


Figure 1: The test problem (2) and its local minimizers; only x^A is a global minimizer.

predictive mean surface derived from $f^n(x)$, 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 these models 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 towards 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 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 based on the probability $P(Y(x) < f^n_{\min})$, 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 similar 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, in a combined near-200 pages of text, the treatment of constraints is virtually neglected; this is clearly an underserved area. Schonlau et al. (1998) were the first to consider bounds on known constraint functions in the EI framework and Williams et al. (2010) considered an unknown constraint situation having a particular integral relationship with the objective. Gramacy and Lee (2011) provided a generic method for a single constraint coded in binary; as mentioned in Section 1, arbitrary constraints can be coerced to fit this model, but not without potentially substantial information loss.

2.3 Augmented Lagrangian framework

Augmented Lagrangian (AL) 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 only evaluated when solving (5), comprising the "inner loop" of the scheme. For additional details on AL-based methods, see, e.g., Nocedal and Wright (2006).

We note that termination conditions have not been explicitly provided in Algorithm 1. In our setting, termination is primarily dictated 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 $\eta_1, \eta_2 \geq 0$, stop when

$$\left\| \max\left\{ c(x^k), 0 \right\} \right\| \leq \eta_1 \qquad \text{and} \qquad \left\| \nabla f(x^k) + \sum_{i=1}^{m} \lambda_i^k \nabla c_i(x^k) \right\| \leq \eta_2.$$

[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.

```
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: A basic augmented Lagrangian framework.

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 as this subsumes blackbox optimization (see, e.g., Conn et al., 2009). In our comparisons in Sections 4–5 we consider two of solver 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 around 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. (2003a). In particular, we consider the mesh adaptive 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 to which local minima they will ultimately converge 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 equation (3) with $Y(x) \equiv Y_{\ell^n}(x) \sim \mathcal{N}(\mu^n(x), \sigma^{2n}(x))$. Approximate convergence can be determined by any number of 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 both 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. Finally, it is local because, as the AL "outer" iterations progress, the (global) "inner" searches organically concentrate near valid regions.

Several drawbacks, however, become apparent upon considering the nature of the composite objective (4). For example, the $y^{(i)}$ values are bound to exhibit behavior in their relationship with the $x^{(i)}$s 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, and/or 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.

There are modern GP methods that 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 $\mu_c^n = (\mu_{c_1}^n, \ldots, \mu_{c_m}^n)^\top$. 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, i.e., $I_{-Y_{c_i}}^{(0)}(x) = \max\{0, 0 + Y_{c_i}(x)\}$. Since the power is two, an expectation-variance relationship can be exploited to obtain

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

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, simply replace 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, simply sample $y_f^{(t)}(x)$, $y_{c_1}^{(t)}(x), \ldots, y_{c_m}^{(t)}(x)$ 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:

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

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

However, challenges arise in exploring the EI surface over $x \in \mathcal{X}$, as 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 yields that when $\lambda^2 - 2(f - y_{\min})/\rho \geq 0$,

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

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

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 increased, 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 things. 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, we recognize that 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 relatively 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 needed here because summing positive and negative $Y_{c_i}(x)$ cancels valid and invalid values, potentially resulting in (extreme) information loss. Instead it would be better to model $Y_c = \sum_i |Y_{c_i}(x)|$ or $Y_c = \sum_i \max(0, Y_{c_i}(x))$ even though that may result in challenging kinks to model with GPs, say. The former option, using absolute values, is only sensible when the constraint set is reduced to only include those that are active/binding at the solution (as is the special case considered above).[3] The advantage of modeling an aggregated final constraint term, pre-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

We first dispense with implementation details for our proposed methods (see Section 3) and our comparators (see Section 2.4). We then demonstrate how these methods fare on our motivating toy data from Section 1. All methods are initialized with $\lambda^0 = (0, \dots, 0)^\top$ and $\rho = 1/2$. 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 in our exercises, 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 non-stationary composite AL response surface, as that method is dominated by the analog involving separated modeling of the constituent parts described in Section 3.2.

We entertain alternatives from Section 3.2 that involve guiding the inner optimization with $\mathbb{E}\{Y\}$ and $\mathbb{E}\{I_Y\}$, following Eqs. (9) and (10), respectively. We note here that the results 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, which is appropriate for our monotone (linear) objective functions, f. However, we do not provide results based on the analytic expression (11) because that requires compromises on the modeling end, which leads 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, EI-nomax, respectively.

[3] We note in Section 6 that this is also the case when the inequality constraints are transformed into equality constraints $c(x) + s = 0$ by inclusion of slack variables $s \geq 0$.

Throughout we treat f as known and emulate each Y_{c_i} with separate GP response surface models 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 1000 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 convergence slows in later iterations the density will plateau, which will have 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 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 $\mathbb{E}\{Y\}$-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 are 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

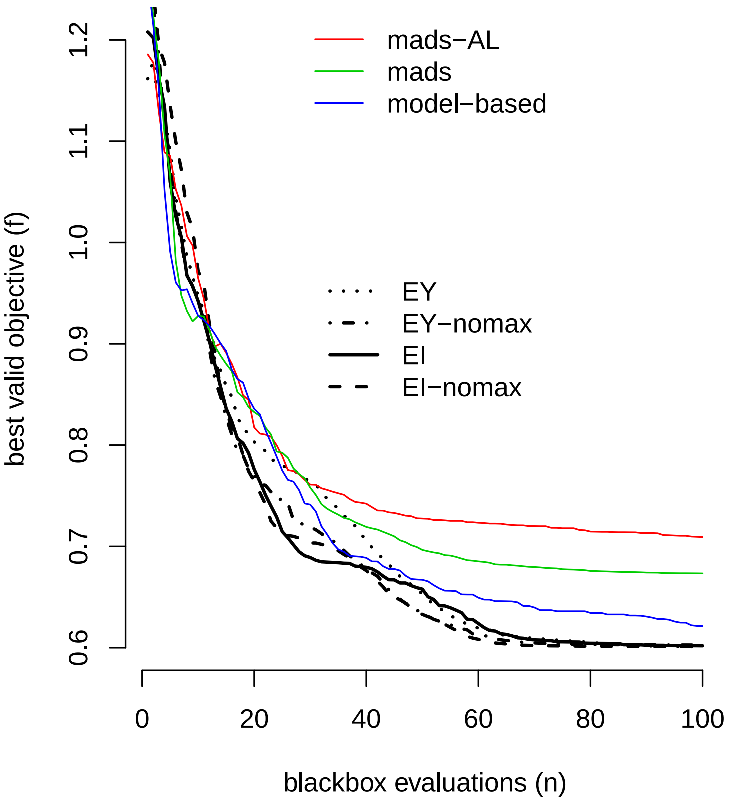
We now summarize the particular implementation details for our comparators.

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 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, 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 on toy data

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. Those numbers coincide ones 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.



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 on our toy motivating problem from Section 1 over 100 Monte Carlo repetitions with a random starting value. The plot tracks the average best valid value of the objective over blackbox iterations, and the table shows more detailed distributional information (95%, average 5%) at iterations 25, 50, and 100.

Since augmented Lagrangian-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. In particular, in these results 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 comparators both in terms of 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. 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 between $n = 25$ and $n = 50$ where EI's average progress stalls temporarily. We observed that this usually marks a transitional period from exploratory and to primarily exploitive behavior. Finally, towards the end of the trials the methods based on dropping the max from Eq. (7) win out. Ignoring regions of the space which 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.

5 Pump-and-treat hydrology problem

Worldwide, there are more than 10,000 contaminated land sites (Meer et al., 2008), and environmental clean-up at these sites has received increased attention over the last 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 the contaminated water, purify it, and inject it 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 the pump-and treat approach to remediation is the 580-acre Lockwood Solvent Groundwater Plume Site, an EPA Superfund site located near Billings, Montana. Due to 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 cartoon illustration of the two contaminated 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.

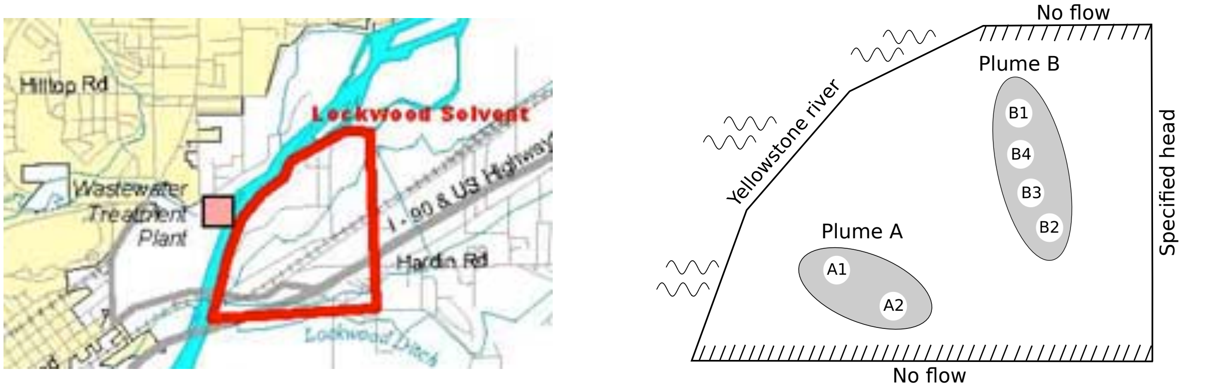


Figure 3: Map of the Lockwood site and illustration of the contaminant plumes. The map on the *left* identifies the Lockwood Solvent region (outlined in bold red) and shows its proximity to the Yellowstone River and the city of Billings. The image was borrowed from the website of Agency for Toxic Substances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

In Mayer et al. (2002), the pump-and-treat problem was posed as a blackbox optimization problem with constraints, and in Fowler et al. (2008) the authors explore the applicability of a wide variety of derivative-free optimization approaches to solve this problem. When formulated as an optimization problem, the number of wells, their locations, and their pumping rates are varied in order to minimize the cost of operating the system. Constraints are included to assure that the plumes are indeed contained. For the Lockwood site problem, the number of wells and their locations are specified and only the pumping rates are varied. This results in the following constrained optimization problem: $\min_{0 \leq x_j \leq 2\times 10^4} \; f(x) = \sum_{j=1}^{6} x_j$, subject to $c_{1,2}(x) \leq 0$, where x_j is the pumping rate for well j. The objective f is linear and describes the total pumping required to run the system. In the absence of constraints, the solution is at the origin and corresponds to no pumping and no remediation. Therefore, two quantified constraints on the amount of contaminant exiting the system are imposed. Constraint c_1 calculates how much of the plume enters the river, whereas c_2 accounts for flow at the southern and eastern boundaries, and computing both involves a computationally expensive simulation. A so-called *analytic element method* (AEM) groundwater model simulates the amount of contaminant exiting the boundaries each pumping scenario (Matott et al., 2006). The simulator thus returns $c(x) \geq 0$, with $c(x) = 0$ indicating satisfaction, producing a kink that 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 one-thousand evaluations. All comparators were initialized at $x_0 = (10000, \dots, 10000)^\top$, a valid input. The `Hooke-Jeeves` algorithm performed best, which is an early example of a "direct search" method (Hooke and Jeeves, 1961). To abstract this suite of results as a benchmark for our numerical studies, we will superimpose

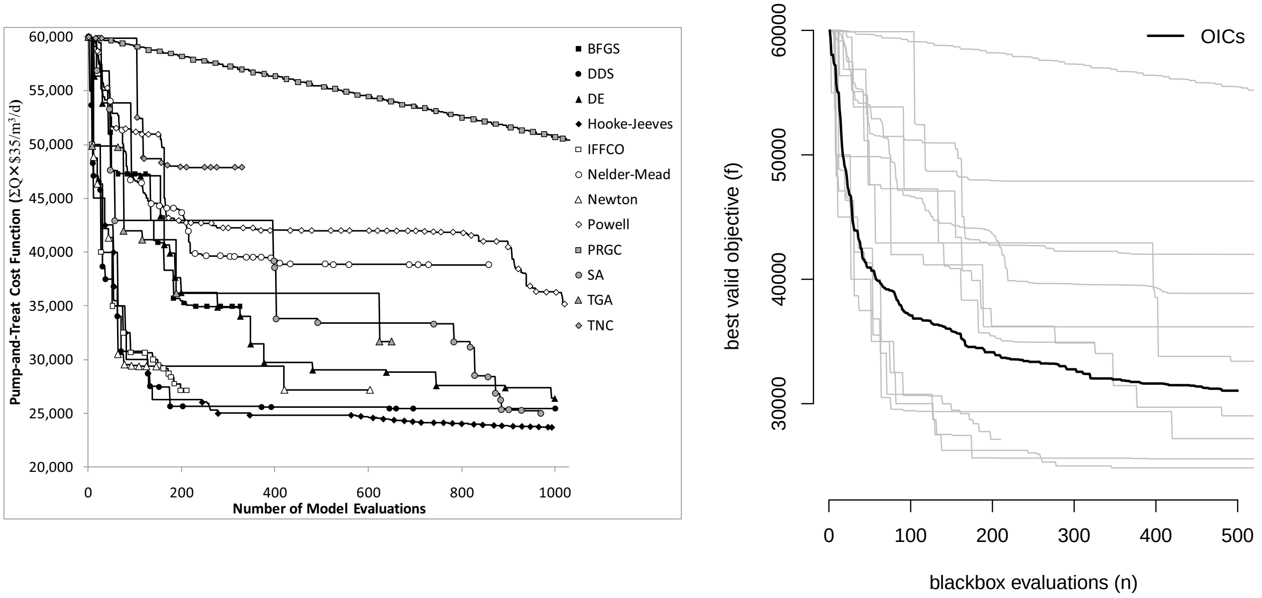


Figure 4: Progress of optimization algorithms on the Lockwood pump-and-treat problem. The *left* graph shows the results of the comparison study of the performance of optimization algorithms completed by (Matott et al., 2011); the *right* one abstracts away the *left* graph for further comparison, for example using OICs (shown). On both graphs, the x-axis counts the number of optimization iterations and the y-axis shows the value of the objective function at the best iterate that satisfies both constraints.

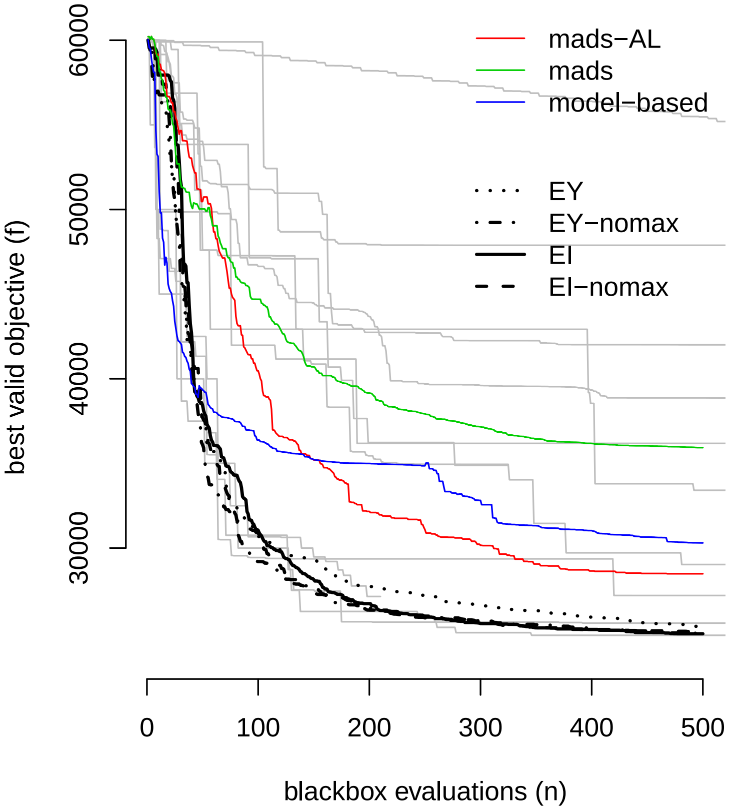
our *new* results on an analogue obtained after striping away identifying information and focusing on the first 500 iterations. The *right* panel of Figure 4 shows an example, with a simple comparator overlayed based on stochastic search with OICs [Section 4.1].

Indeed, that simple stochastic search is surprisingly good relative much more thoughtful comparators. To clarify, the search is based on sampling one OIC in each trial, and updating the best valid value of the objective when a new one is found. Since it is a stochastic method, we are revealing its average behavior over thirty replicates. On average, it is competitive with very 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. Also observe that even after those early stages, OICs still outperform at least half of the alternatives for the remainder of the 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 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 just like the one from Section 4.3. In this case each of the thirty repetitions was initialized randomly with $x^0 \in \mathcal{B} = [0, 20000]^6$. The comparators from Section 5.1 are shown in gray, however note these used a fixed starting location x_0, *not* a random one—i.e., they were not included in the Monte Carlo. From the figure we can see that the relative orderings of our comparators is roughly the same as for the toy problem, except results vary for the classical ones depending on then number of evaluations, n. The surrogate model-based average and worst case behavior is better than the other AL comparators, and competitive with the best APMs from Matott et al. It is worth noting that many of the individual Monte Carlo runs of our EI and EY- based methods outperformed all Matott et al. APM comparators, including `Hooke-Jeeves`.

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,



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 on the hydrology problem over 30 Monte Carlo repetitions with a random starting value. The plot tracks the average best valid value of the objective over blackbox iterations, and the table shows more detailed distributional information (95%, average 5%) at iterations 25, 50, and 100.

23,026. That MADS's average behavior is much worse suggests extreme differential performance depending on the quality of initialization, particularly with regards 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 1e-3 tolerance) even after as many as $n = 100$ iterations.[4] We can guess, then, that had the Matott et al. comparison been randomly initialized, the best comparators would similarly have fared much worse. By contrast, in experiments with the surrogate-based methods using the same helpful 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 processes (GP) surrogate modeling and expected improvement (EI) methods from the design of computer experiments literature with an additive penalty method (APM) that has attractive convergence properties: the augmented Lagrangian (AL). The main advantage of this pairing is that it reduces a constrained optimization into an unconstrained one, for which statistical methods are rather 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).

[4] We put 60000 in as a place holder for that case.

This paper has demonstrated clear potential for 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 in nature. Still, we see potential for further improvement. For example, we anticipate gains from a more aggressive hybridization which acknowledges that the statistical methods fail to "penetrate" into local troughs, particularly towards the end of a search. In the unconstrained context, Gray et al. (2007); Taddy et al. (2009) have had success with pairing EI with the APPS direct search method (Kolda et al., 2003b). Gramacy and Le Digabel (2011) took a similar tack with MADS. Both setups port provable local convergence from the direct method over to a more global search context by, in effect, letting the direct solver take over towards the end of the search process 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, $i = 1, \dots, m$, leaving untapped potential to leverage cross correlations (e.g., Williams et al., 2010). Moreover, ideas from multi-objective optimization may prove helpful in our multi-constraint 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 Svenson and Santner (2012) or Picheny (2013) who both 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 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 gives just modest improvements. It may be advantageous to exploit that knowledge more aggressively by, say, estimating a classification surface to explicitly direct sampling near boundaries (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 which are not improving on the objective (Lindberg and Lee, 2015).

However in closing 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 be leveraged directly—e.g., laGP was easy to augment to accommodate all of the new methodology described herein. That is attractive because, relative to the mathematical programming literature, statistical optimization has very few "horses in the race" when it comes to 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 application. Software packages like NOMAD and APPS generally work, by contrast, right out of the box. It is hard to imagine matching that engineering capability for hard constrained optimization problems with statistical methodology if we insist on those methods being even more intricate than the current state of the art.

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 Journal on Optimization*, 17, 1, 188–217.

— (2009). "A Progressive Barrier for Derivative-Free Nonlinear Programming." *SIAM Journal 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." *Les Cahiers du GERAD* G–2012–43, Tech. Rep., HEC Montréal. To appear in *SIAM Journal 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 Tutotial on Bayresian 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." *Journal 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*. MPS/SIAM Series on Optimization. Philadelphia, PA, USA: 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." *Les Cahiers du GERAD* G–2011–37, Tech. Rep., HEC Montréal.

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

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

— (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." *Journal 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." *Journal 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." *Journal 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. (2003a). "Optimization by direct search: new perspectives on some classical and modern methods." *SIAM Review*, 45, 385–482.

— (2003b). "Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods." *SIAM Review*, 45, 3, 385–482.

Le Digabel, S. (2011). "Algorithm 909: NOMAD: Nonlinear Optimization with the MADS algorithm." *ACM Transactions 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." *Journal 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 *In groundwater science and policy: an international overview*, ed. P. Quevauviller, 405–420. Cambridge: 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.

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." *ArXiv e-prints*.

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." *Journal 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*, no. 34 in IMS Lecture Notes - Monograph Series, 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., The Ohio State University.

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.