Modeling an augmented Lagrangian for blackbox constrained optimization

Robert B. Gramacy
The University of Chicago Booth School of Business
bobby.gramacy.com

Joint with Genetha A. Gray, Sébastien Le Digabel, Herbert K.H. Lee, Pritam Ranjan, Garth Wells, Stefan Wild

CASD/George Mason University — Oct 2015

Blackbox constrained optimization

Consider constrained optimization problems of the form

$$\min_{x} \{f(x) : c(x) \leq 0, x \in \mathcal{B}\}, \quad \text{where}$$

- $f : \mathbb{R}^d \to \mathbb{R}$ is a scalar-valued objective function
- $c : \mathbb{R}^d \to \mathbb{R}^m$ is a vector of constraint functions
- $\mathcal{B} \subset \mathbb{R}^d$ is known, bounded, and convex

This is a challenging problem when c are non-linear, and when evaluation of f and/or c requires expensive (blackbox) simulation.

Here is a toy problem to fix ideas.

- A linear objective in two variables:

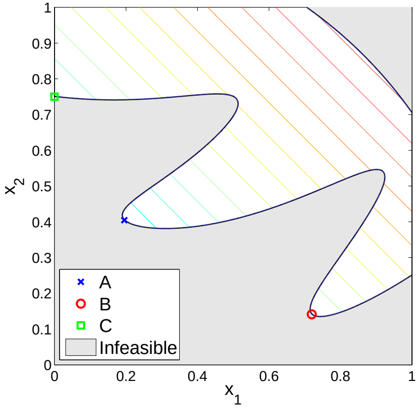
$$\min_x \left\{ x_1 + x_2 : c_1(x) \leq 0,\ c_2(x) \leq 0,\ x \in [0,1]^2 \right\}$$

- where two non-linear constraints are given by

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

Even when treating $f(x) = x_1 + x_2$ as known, this is a hard problem when $c(x)$ is treated as a blackbox.

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



- $c_2(x)$ may seem uninteresting, but it reminds us that solutions may not exist on every boundary

Solvers

Mathematical programming has efficient algorithms for non-linear (blackbox) optimization (under constraints) with

- provable local convergence properties,
- lots of polished open source software

Statistical approaches e.g., EI (Jones et al., 1998)

- enjoy global convergence properties,
- excel when simulation is expensive, noisy, non-convex

... but offer limited support for constraints.
(Schonlau et al., 1998; G & Lee, 2011; Williams et al., 2010)

A hybrid proposal

Combine (global) statistical objective-only optimization tools

a) response surface modeling/emulation: training a flexible model f^n on $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ to guide choosing $x^{(n+1)}$ (e.g., Mockus, et al., 1978, Booker et al., 1999)

b) expected improvement (EI) via Gaussian process (GP) emulation (Jones, et al., 1998)

... with a tool from mathematical programming

c) augmented Lagrangian (AL): converting a problem with general constraints into a sequence of simply constrained ones (e.g., Bertsekas, 1982)

Gaussian process (GP) surrogate/regression models make popular emulators.

As predictors, they are

- rarely beaten in out-of-sample tests,
- have appropriate coverage, and can interpolate

Using data $D = (X, Y)$, where X is an $n \times p$ design matrix, the $n \times 1$ response vector Y has MVN likelihood:

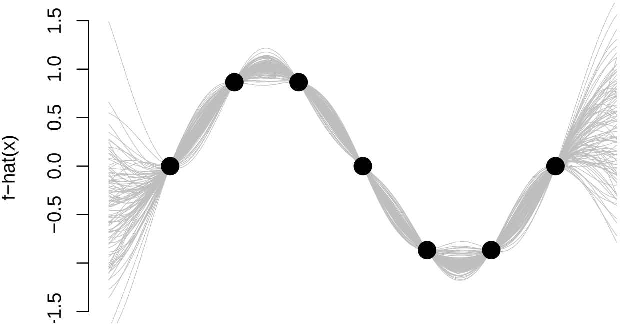
$$Y \sim \mathcal{N}_n(0, \tau^2 K), \quad \text{where} \quad K_{ij} = K(x_i, x_j)$$

often with prior $\pi(\tau^2) \propto \tau^{-2}$ (Berger et al., 2001)

The predictive equations have

$$\text{mean} \quad \mu^n(x|D,K) = k^\top(x)K^{-1}Y,$$
$$\text{and scale} \quad \sigma^{2n}(x|D,K) = \frac{\psi[K(x,x) - k^\top(x)K^{-1}k(x)]}{n},$$

where $k^\top(x)$ is the n-vector whose i^{th} component is $K(x, x_i)$.



Expected Improvement

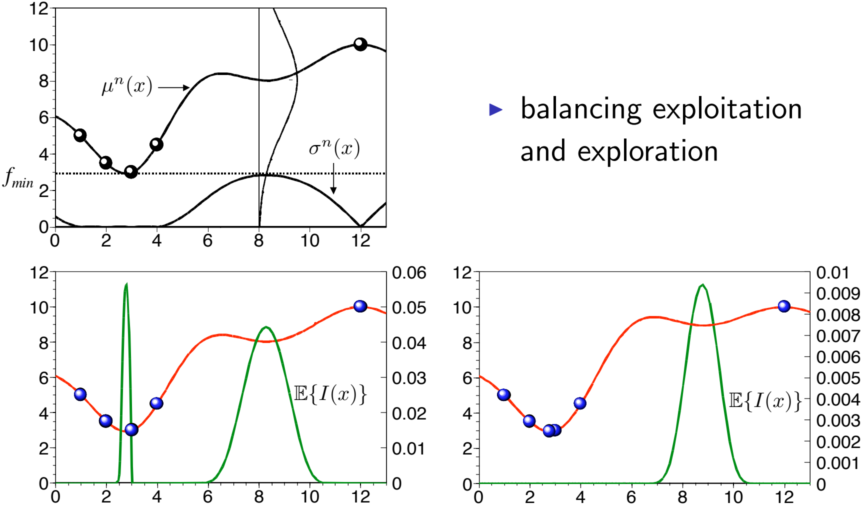
Suppose the predicting equations from f^n are conditionally normal, i.e., from a GP: $Y(x) \sim \mathcal{N}(\mu^n(x), \sigma^{2n}(x))$

Define the improvement as

$$I(x) = \max\{0, f^n_{\min} - Y(x)\}$$

Then, its expectation (EI) has a closed form expression:

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



- balancing exploitation and exploration

(Jones, et al., 1998)

Augmented Lagrangian

AL methods for constrained nonlinear optimization have favorable theoretical properties for finding local solutions.

The main tool is the AL:

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

- $\rho > 0$ is a penalty parameter
- $\lambda \in \mathbb{R}_+^m$ serves as a Lagrange multiplier; omitting this term leads to a so-called additive penalty method (APM)

AL-based methods transform a constrained problem into a sequence of simply constrained problems.

Given $(\rho^{k-1}, \lambda^{k-1})$,

1. approximately solve the subproblem

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

2. update:
   - $\lambda_j^k = \max\left(0, \lambda_j^{k-1} + \frac{1}{\rho^{k-1}} c_j(x^k)\right)$, $j = 1, \ldots, m$
   - If $c(x^k) \leq 0$, set $\rho^k = \rho^{k-1}$; otherwise, set $\rho^k = \frac{1}{2}\rho^{k-1}$

... then repeat, incrementing k.

- Functions f and c are only evaluated when solving the subproblem(s), comprising an "inner loop".

Statistical surrogate AL

AL methods are not designed for global optimization.

- Convergence results have a certain robustness,
- but only local solutions are guaranteed.

Hybridizing with surrogate models offers a potential remedy.

- Focus is on finding x^k in the “inner loop”,
- using evaluations $(x^{(1)}, f^{(1)}, c^{(1)}), \dots, (x^{(n)}, f^{(n)}, c^{(n)})$ collected over all “inner” and “outer” loops $\ell = 1, \dots, k-1$.

There are several options for how exactly to proceed.

One option is easy to rule out.

Let $y^{(i)} = L_A(x^{(i)}; \lambda^{k-1}, \rho^{k-1})$ via $f^{(i)}$ and $c^{(i)}$. I.e.,

$$y^{(i)} = f(x^{(i)}) + (\lambda^{k-1})^\top c(x^{(i)}) + \frac{1}{2\rho^{k-1}} \sum_{j=1}^{m} \max\left(0, c_j(x^{(i)})\right)^2$$

- fit a GP emulator f^n to the n pairs $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$
- guide "inner loop" search by the predictive mean or EI

Benefits include:

- modular
- facilitates global–local tradeoff

But modeling this x–y relationship presents serious challenges.

$$y^{(i)} = f(x^{(i)}) + (\lambda^{k-1})^\top c(x^{(i)}) + \frac{1}{2\rho^{k-1}} \sum_{j=1}^{m} \max\left(0, c_j(x^{(i)})\right)^2$$

Inherently nonstationarity.

- square amplifies and max creates kinks

Fails to exploit known structure.

- a quadratic form

Needlessly models a (potentially) known quantity.

- many interesting problems have linear f

Separated modeling

Shortcomings can be addressed by separately/independently modeling each component of the AL.

- f^n emitting $Y_{f^n}(x)$
- $c^n = (c_1^n, \ldots, c_m^n)$ emitting $Y_c^n(x) = (Y_{c_1}^n(x), \ldots, Y_{c_m}^n(x))$

The distribution of the composite random variable

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

can serve as a surrogate for $L_A(x; \lambda, \rho)$.

- simplifications when f is known

The composite posterior mean is available in closed form, e.g., under GP priors.

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

A result from generalized EI (Schonlau et al., 1998) gives

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

Expected improvement for AL

The simplest way to evaluate the EI is via Monte Carlo:

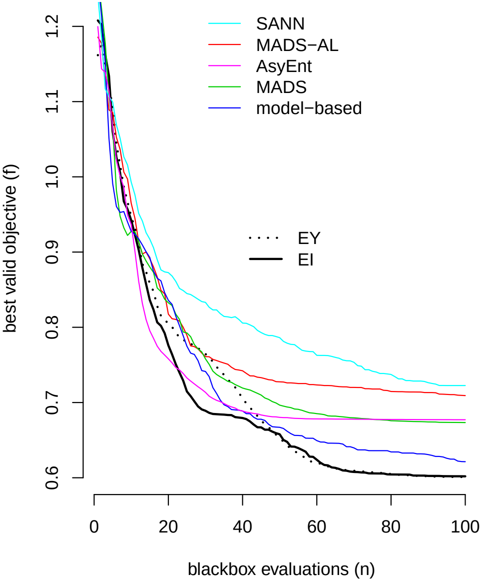
- take 100 samples $Y_f^{(i)}(x)$ and $Y_c^{(i)}(x)$
- then $EI(x) \approx \frac{1}{100} \sum_{i=1}^{100} \max\{0, y_{\min}^n - Y^{(i)}(x)\}$

The "max" in the AL makes analytic calculation intractible.

But you can remove the "max" by introducing slack variables

- turning inequality into equality constraints
- and making the AL composite $Y(x)$ a simple quadratic.
- The EI then becomes a one-dimensional integral of non-central chi-squared quantities.

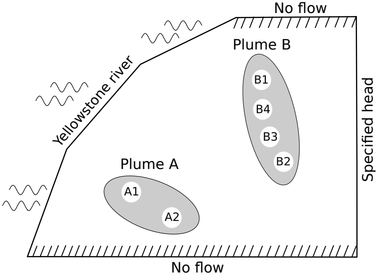
Results on toy data



n	25	50	100
95%			
EI	0.866	0.775	0.602
EY	1.052	0.854	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
5%			
EI	0.610	0.602	0.600
EY	0.607	0.601	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

Benchmark problem

Two contaminant plumes threaten a valuable water source: the Yellowstone River.



To prevent further expansion of these plumes, six pump-and-treat wells have been proposed.

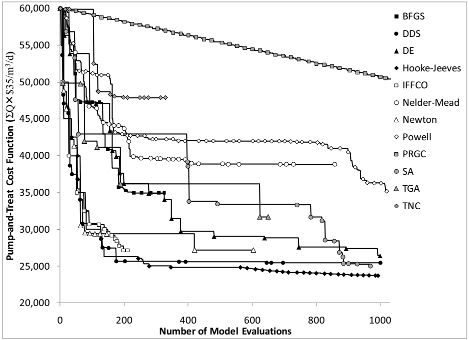
Mayer et al. (2002) first posed the pump-and-treat problem as a constrained blackbox optimization.

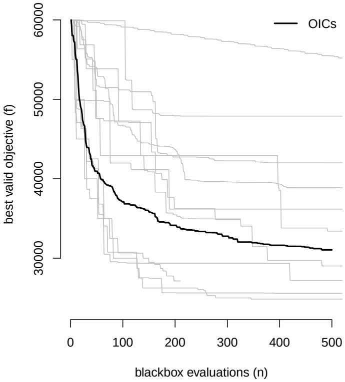
If x_j denotes the pumping rate for well j, then

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

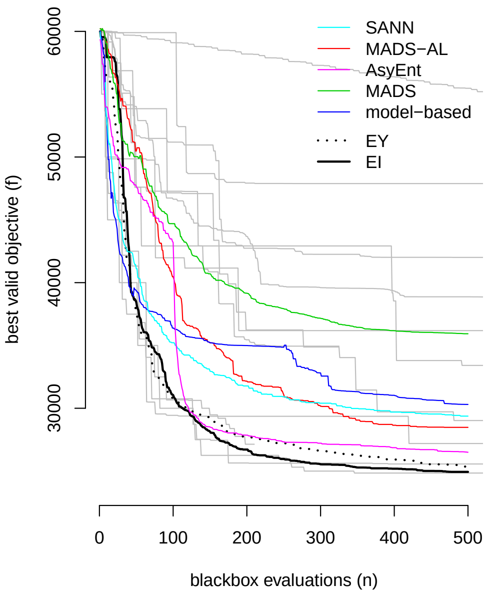
- f is linear, describing costs to operate the wells
- c_1 and c_2 denote plume flow exiting the boundary: simulated via an analytic element method groundwater model

Matott et al. (2011) compared MATLAB and Python optimizers, treating constraints via APM.





- initialized at $x_j^0 = 10^4$



n	100	200	500
	95%		
EI	37584	28698	25695
EY	36554	32770	27362
SANN	43928	35456	30920
MADS-AL	60000	49020	32663
AsyEnt	49030	29079	27445
MADS	60000	60000	60000
model	60000	60000	35730
	5%		
EI	27054	25119	24196
EY	25677	24492	24100
SANN	28766	27238	26824
MADS-AL	30776	26358	24102
AsyEnt	37483	26681	25377
MADS	30023	26591	23571
model	25912	25164	24939

Summarizing

Nontrivial multiple blackbox constraints present serious challenges to optimization

- even when the objective is simple/known.

The augmented Lagrangian method from mathematical programming is a nice framework for handling constraints

- but only local convergence is guaranteed.

Statistical surrogate modeling and expected improvement nicely hybridize with the AL:

- implementation is straightforward (see `laGP` on CRAN).