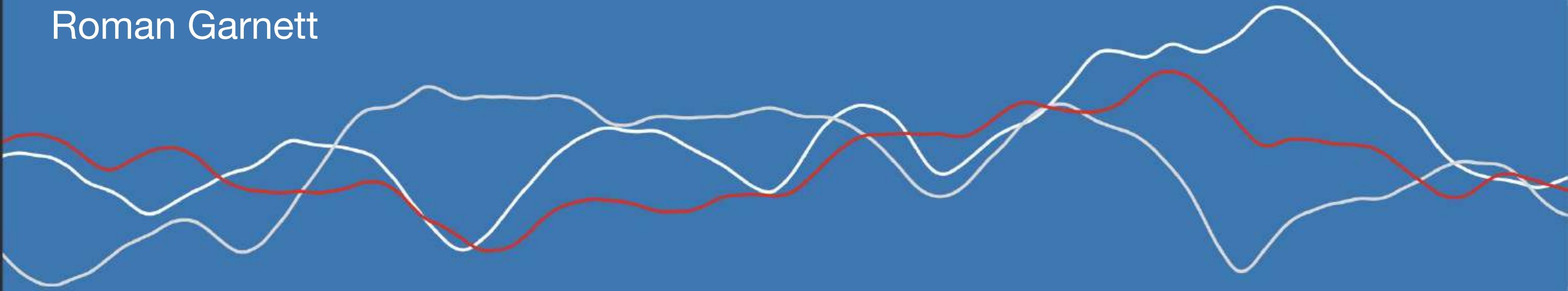
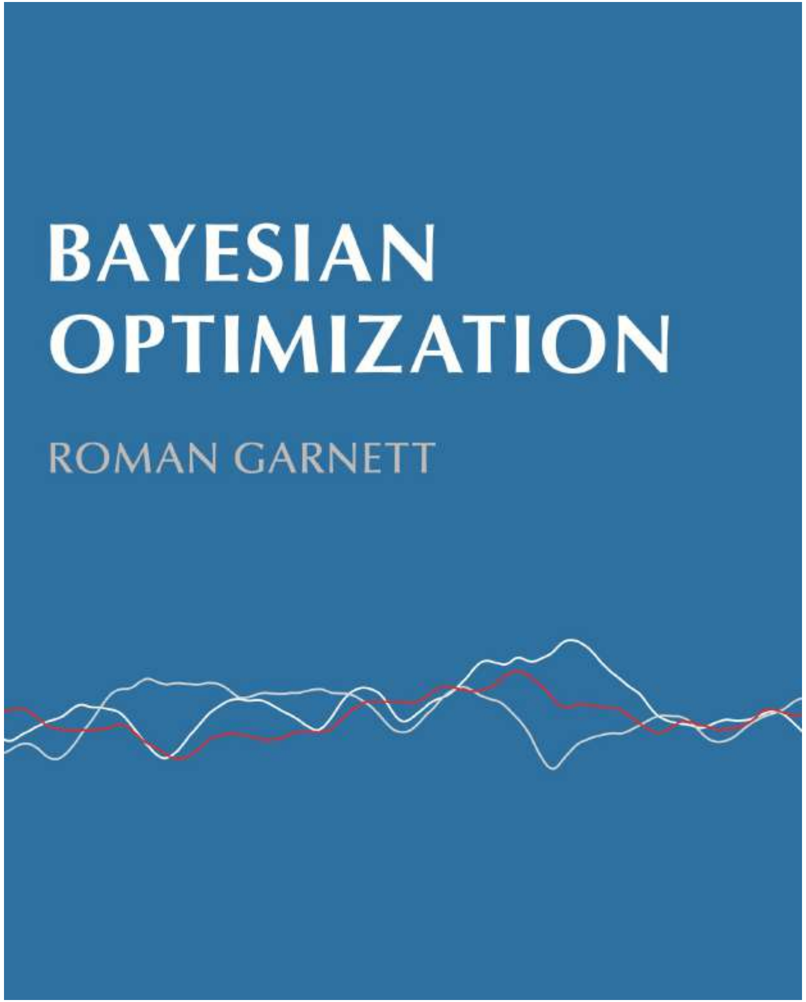
Some random things I learned writing the BayesOpt book

Roman Garnett





bayesoptbook.com

(Semi-) Joking advice: Don’t write a book…

Book timeline...
(4 authors, January 2013)

Bayesian Optimization book

N

Nando de Freitas <nando@cs.ubc.ca>
to me, Michael, Frank, Nando

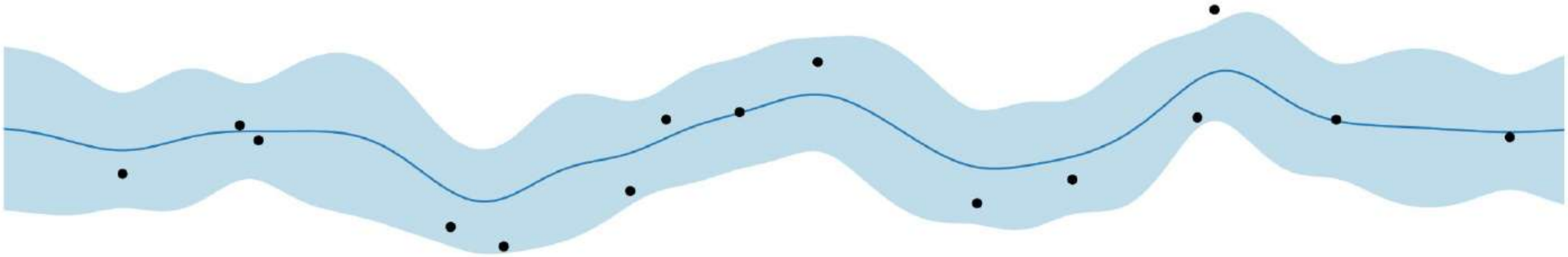
OK guys. I think it's time for us to do this seriously.

Expected Improvement with Noise

Simple Recipe for Bayesian Experimental Design

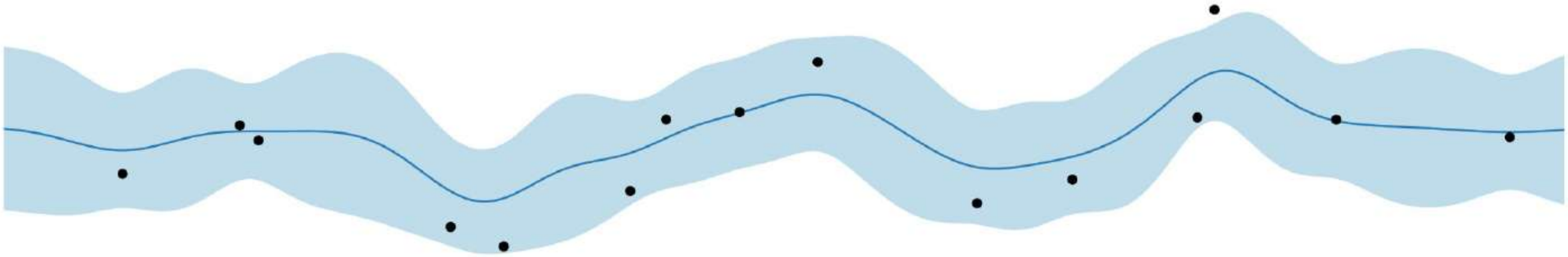
Step 1: build a model of (noisy) observations (x, y)

- latent function model, $p(f)$ — e.g., GP
- observation model, $p(y \mid x, \phi)$, $\phi = f(x)$ — e.g., Gaussian noise



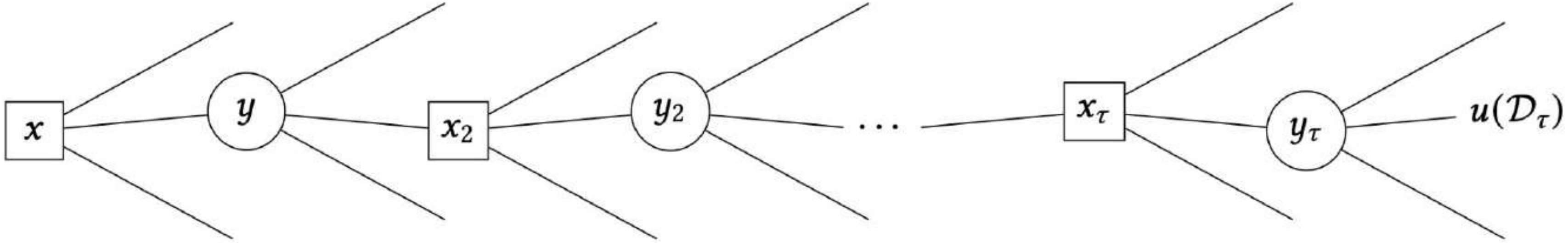
Simple Recipe for Bayesian Experimental Design

Step 2: choose a utility function $u(D)$, $D = \{(x, y)\}$



Simple Recipe for Bayesian Experimental Design

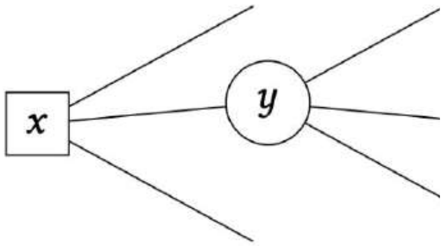
Step 3: give up on the optimal policy



Simple Recipe for Bayesian Experimental Design

Step 4: derive a policy via one-step lookahead (greedily maximize one-step expected gain in utility *D* → *D’*)

$$\alpha(x;\mathcal{D}) = \mathbb{E}\big[u(\mathcal{D}') \mid x, \mathcal{D}\big] - u(\mathcal{D})$$



(...nothing to see here...)

$u(\mathcal{D}_\tau)$

Simple Recipe for Bayesian Experimental Design

Step 4: derive a policy via one-step lookahead

$$\alpha(x;\mathcal{D}) = \mathbb{E}\big[u(\mathcal{D}') \mid x, \mathcal{D}\big] - u(\mathcal{D})$$

(wrt noisy observation *y*! consequence: in general, penalizes high noise)

Prevalent in BayesOpt!

Utility	**Policy**
simple reward	expected improvement
global simple reward	knowledge gradient
information gain	mutual information (aka entropy search)

Noiseless expected improvement

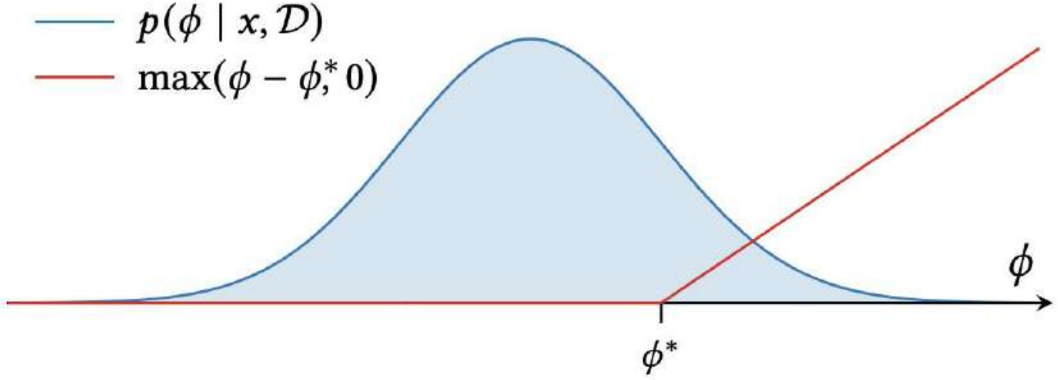
utility (best seen value):

$u(D) = \phi^* = \max \mathbf{f}$

marginal gain:

$\max(\phi - \phi^*, 0)$

expected utility easy to compute, has nice properties, etc.



Noiseless expected improvement

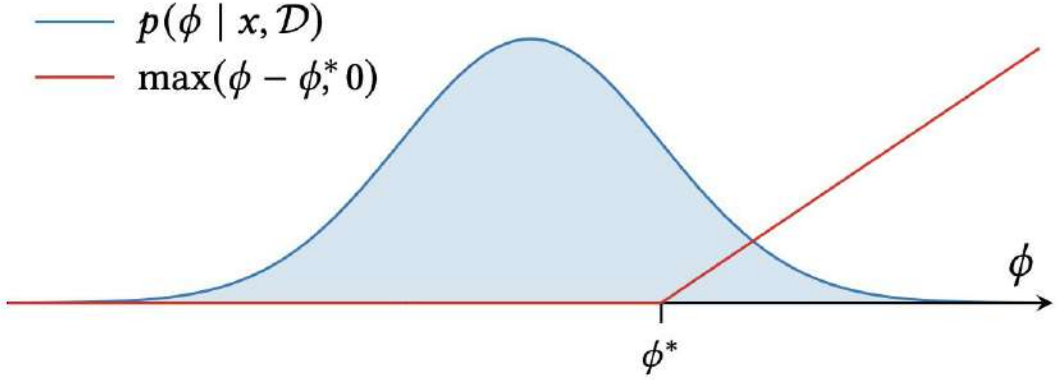
utility (best seen value):

$$u(D) = \phi^* = \max \mathbf{f}$$

marginal gain:

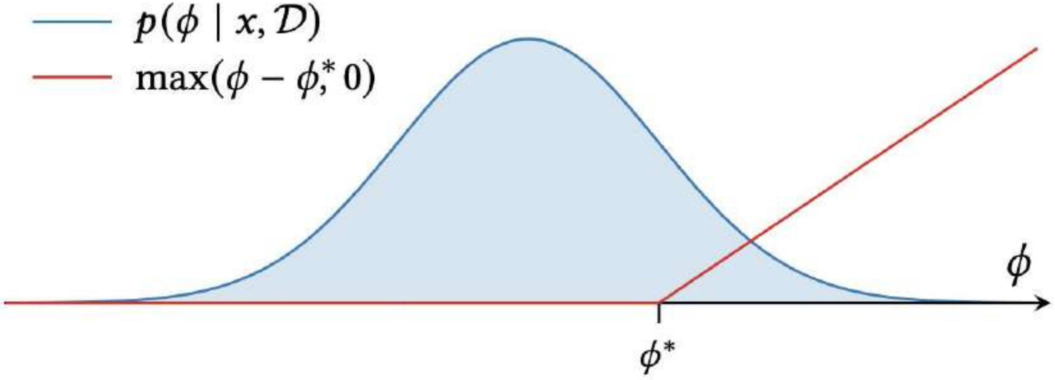
$$\max(\phi - \phi^*, 0)$$

expected utility easy to compute, has nice properties, etc.



$$\alpha_{\mathrm{EI}}(x; \mathcal{D}) = (\mu - \phi^*)\,\Phi\!\left(\frac{\mu - \phi^*}{\sigma}\right) + \sigma\phi\!\left(\frac{\mu - \phi^*}{\sigma}\right)$$

Noiseless expected improvement



expected utility easy to compute, has nice properties, etc.

$$\alpha_{\mathrm{EI}}(x; \mathcal{D}) = (\mu - \phi^*)\,\Phi\left(\frac{\mu - \phi^*}{\sigma}\right) + \sigma\phi\left(\frac{\mu - \phi^*}{\sigma}\right)$$

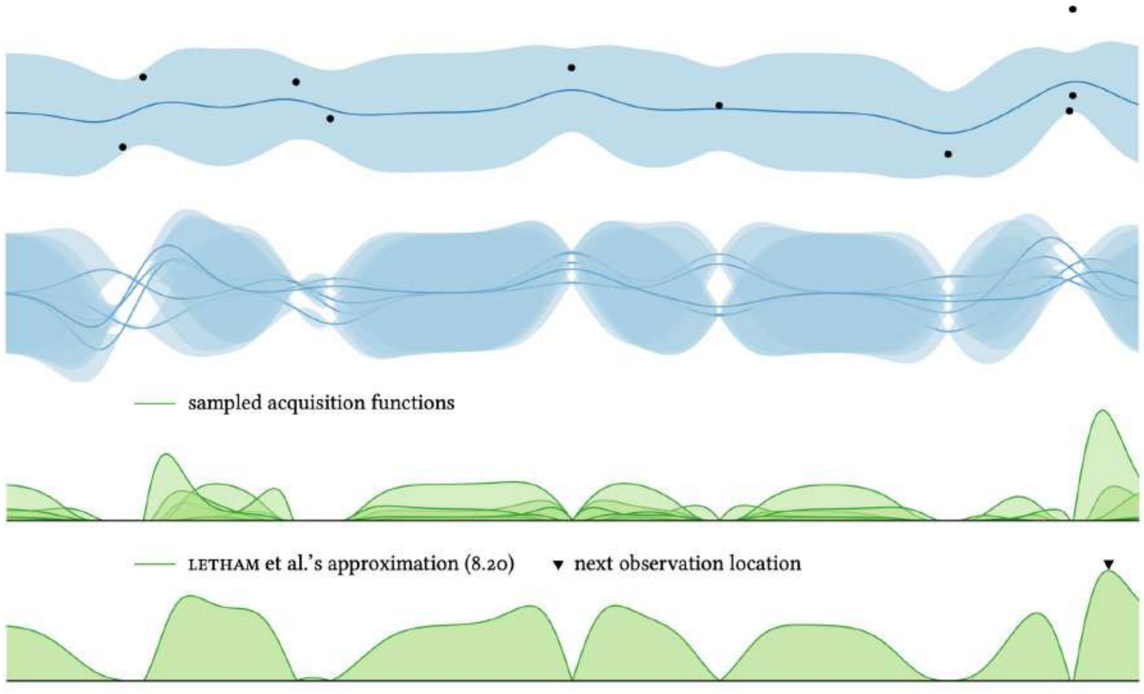
very tempting to start here and try to “fix” this!

“Fixing” the expected utility

- plug-in estimators: use noiseless EI with “guess” of max **f**
- expectation of EI

  with respect to **f**

  (Letham, et al. 2019)



Let’s start with utility!

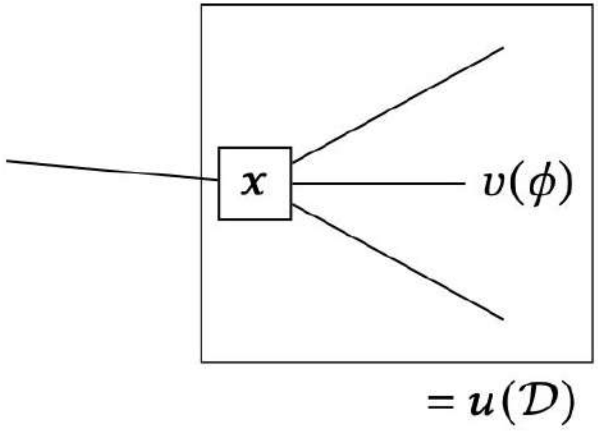
idea: consider gathering data to support a recommendation *after* optimization

action space: visited locations **x**

utility: risk-neutral

optimal recommendation:

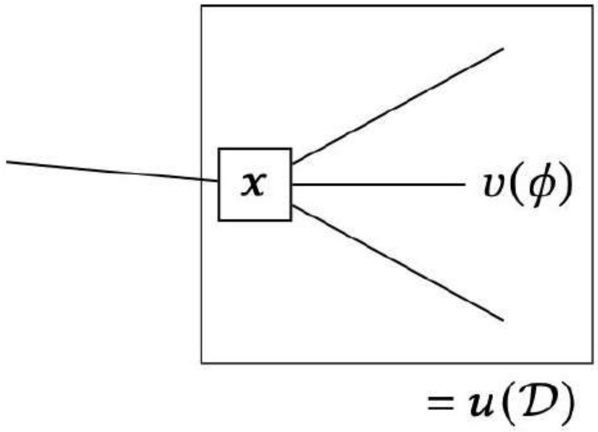
maximum of posterior mean on **x** = $u(D)$



The noisy setting: Utility

maximum of posterior mean on **x** = $u(D)$

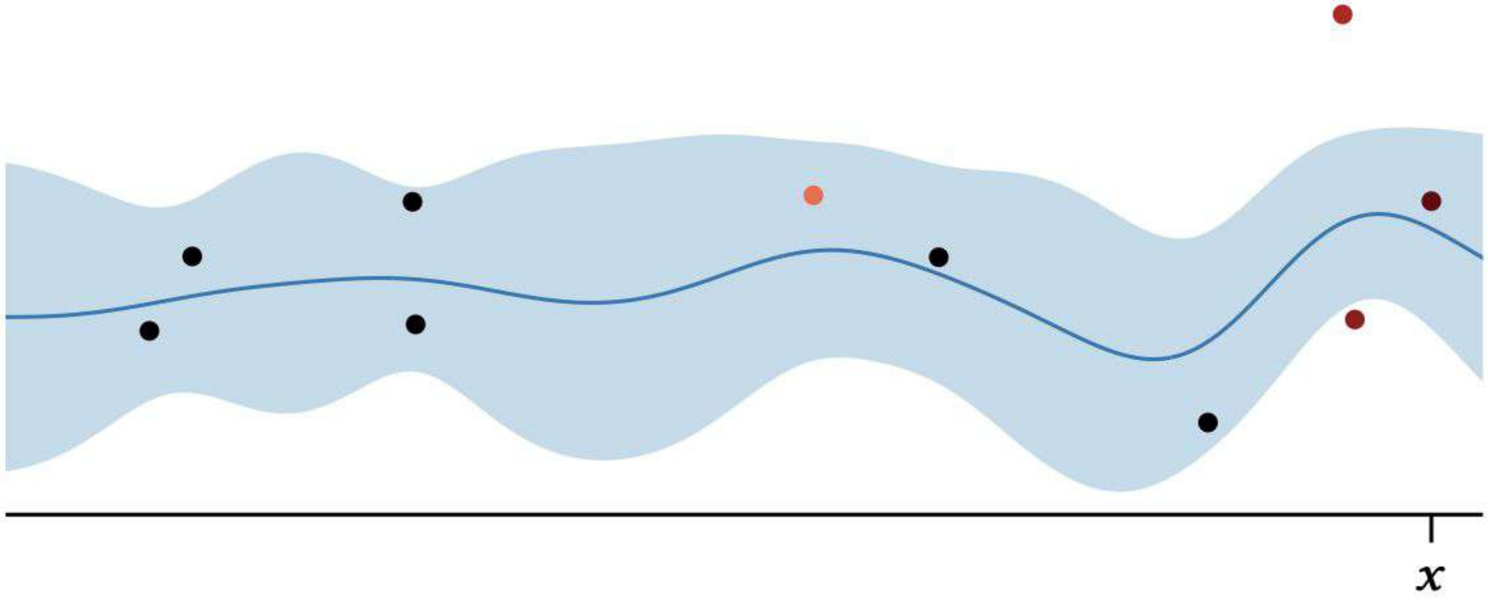
- compatible with noiseless EI!
- compatible with knowledge gradient! (just a different action space)



The difficulty

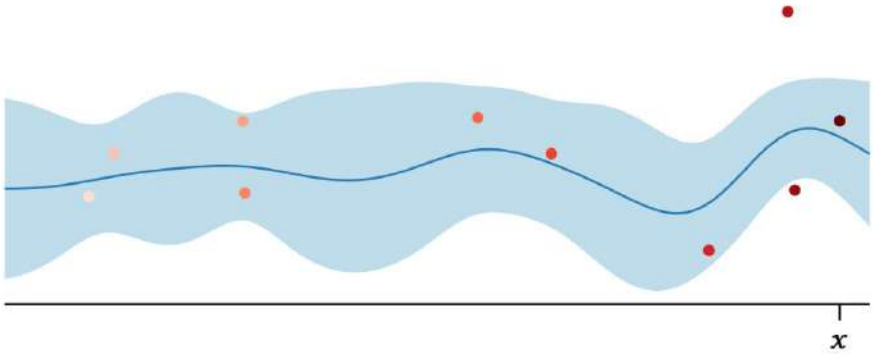
maximum of posterior mean can be anywhere!

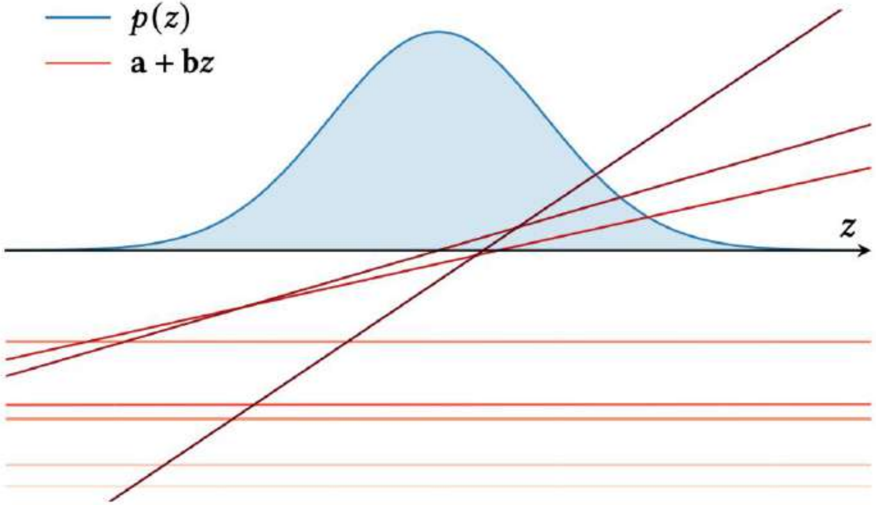
local reasoning of just $f(x)$, y not enough!



The fix (Frazier, et al. 2009)

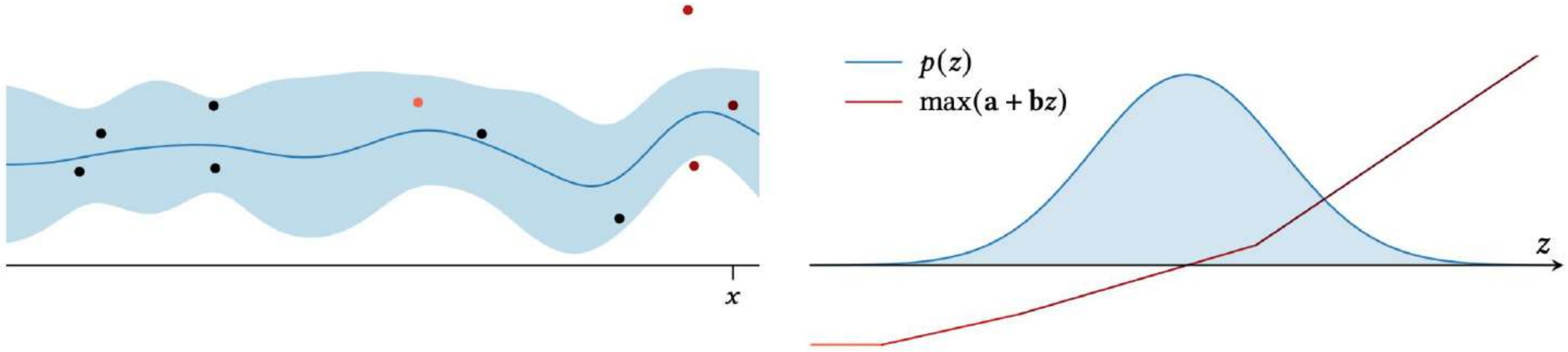
posterior mean update is linear in observed value





The fix (Frazier, et al. 2009)

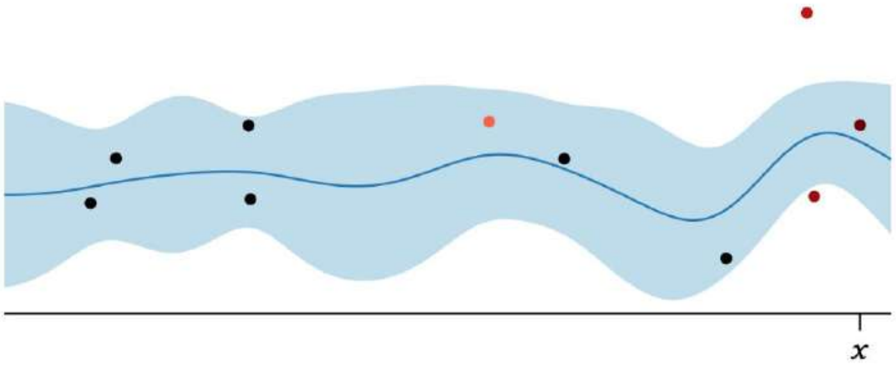
can compute piecewise linear update to max in $O(n^2 \log n)$

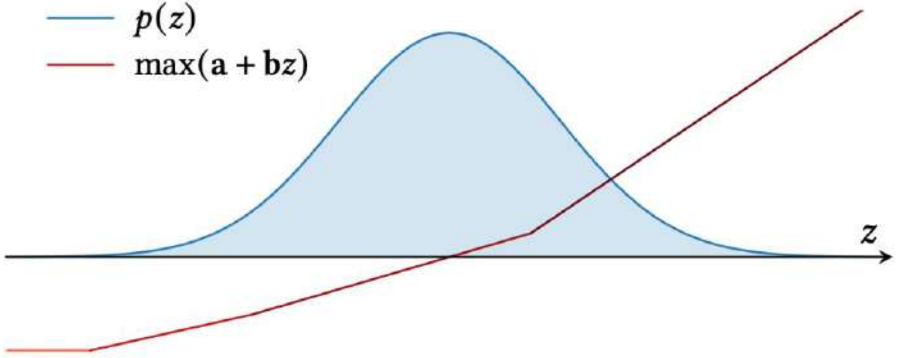


The fix (Frazier, et al. 2009)

sums of standard normal CDFs, PDFs as before

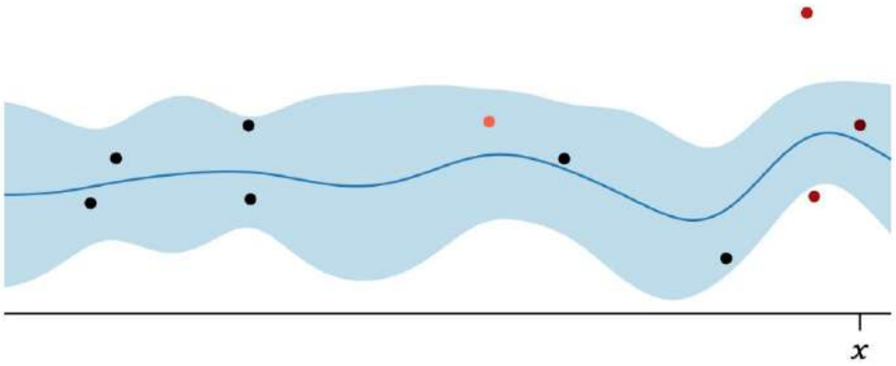
$$\sum_i a_i\left[\Phi(c_{i+1}) - \Phi(c_i)\right] + b_i\left[\phi(c_i) - \phi(c_{i+1})\right]$$

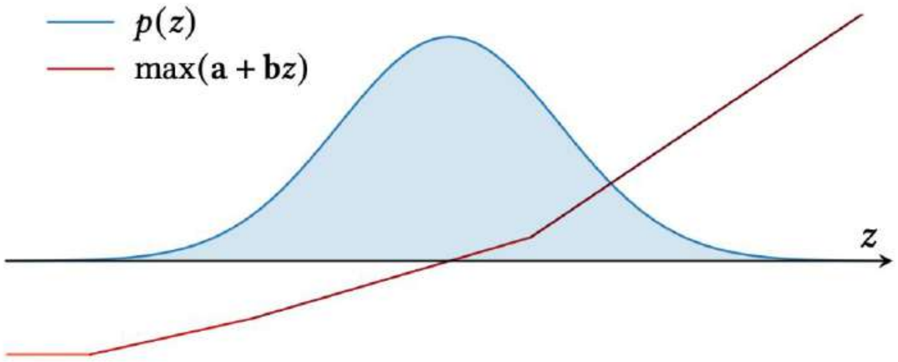




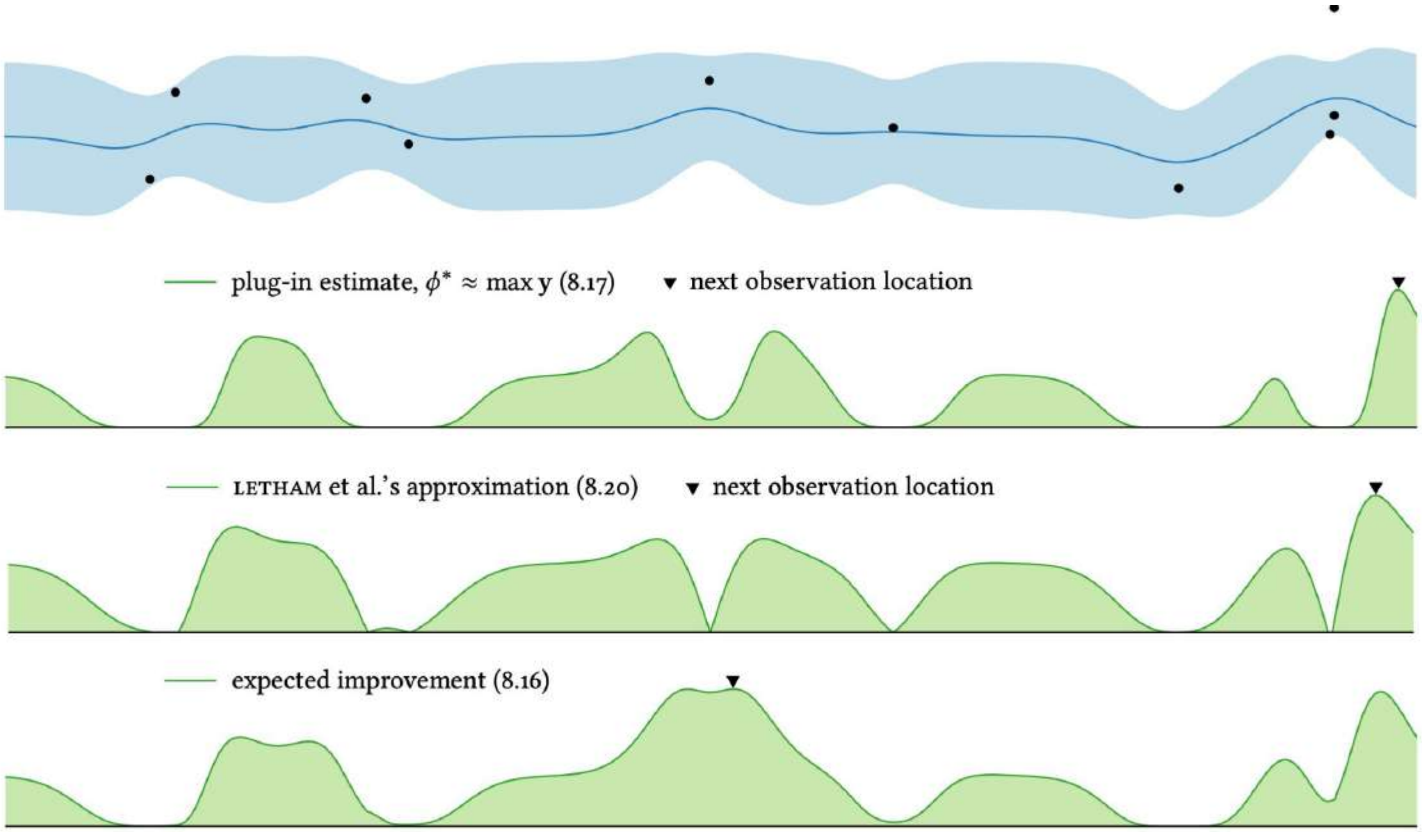
The result

- handles hetereoskedastic noise automatically / correctly
- handles correlations in / global nature of posterior mean
- noiseless EI special case
- closed form



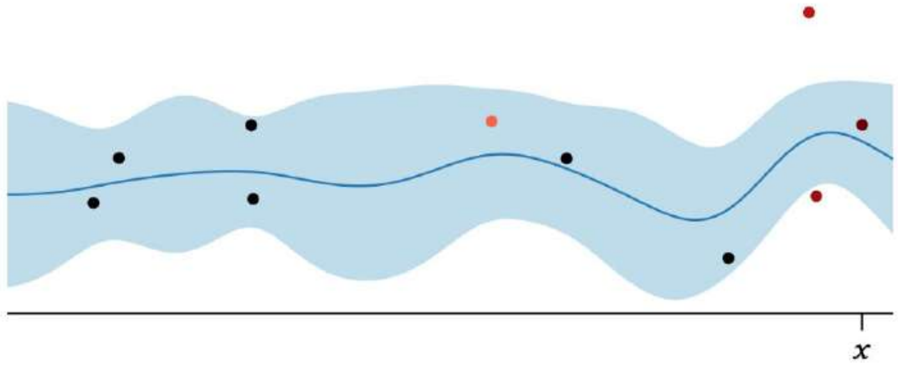


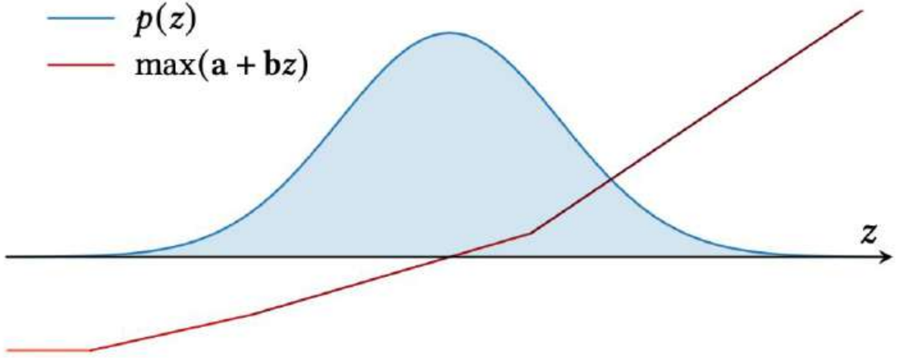
Alternative approaches



Why?

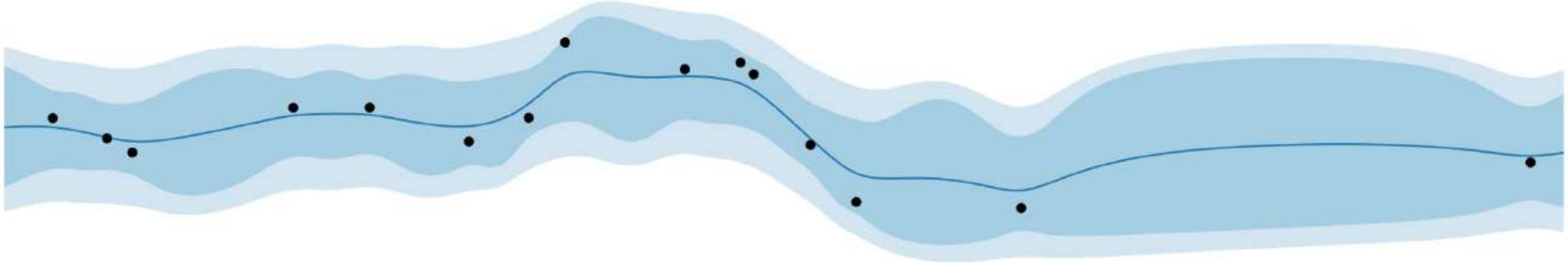
- ignores correlations in posterior mean update
- assumption of exact observations in expectation does not match true observation model
- (but honestly this is all fine for highish SNR)

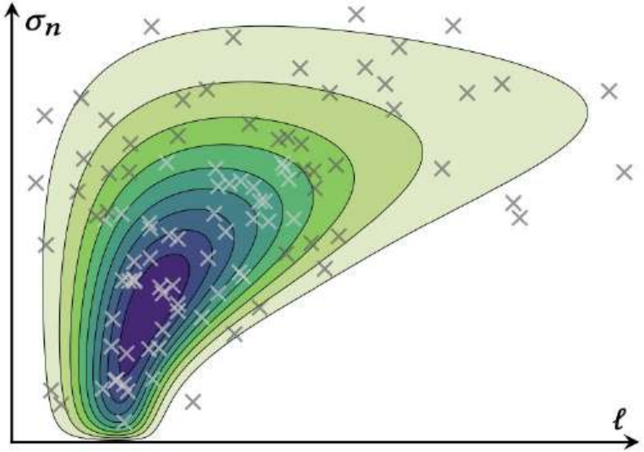


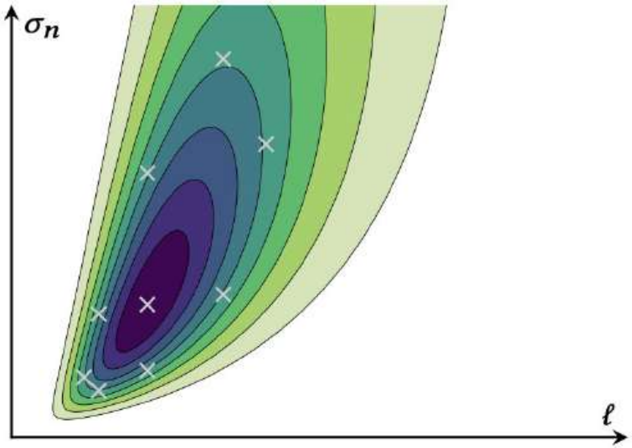


Marginalizing Hyperparameters in Policy

Marginalizing hyperparameters







Standard approach

Let utility $u(D; \theta)$ depend on θ and integrate the hyperprameter-conditional acquisition function against the hyperparameter posterior

$$\int \alpha(x; \mathcal{D}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}$$

Standard approach

Let utility $u(D; \theta)$ depend on θ and integrate the hyperprameter-conditional acquisition function against the hyperparameter posterior

$$\int \alpha(x; \mathcal{D}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}$$

blind to uncertainty in θ!

Standard approach

Let utility $u(D; \theta)$ depend on θ and integrate the hyperprameter-conditional acquisition function against the hyperparameter posterior

$$\int \alpha(x; \mathcal{D}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}$$

blind to uncertainty in θ!

Alternative approach

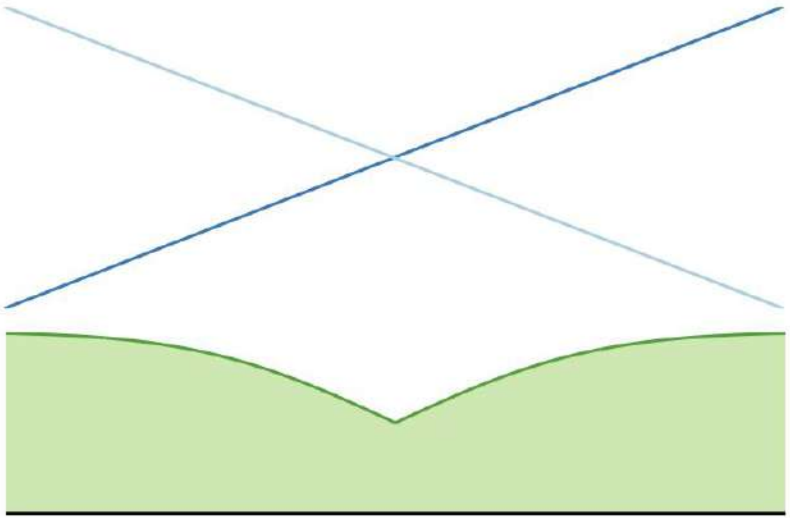
Define utility with respect to marginal model from the beginning!

E.g., for EI or KG, use θ marginal posterior mean (for a terminal recommendation we'd be marginalizing θ, right?)

$$\int \mu_{\mathcal{D}}(x;\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta},$$

Example

- function is $f(x) = x$ or $f(x) = -x$
- knowledge gradient
- for standard approach, acquisition function is flat! (maximum of θ-conditional posterior mean always equal)
- for alternative approach, get sensible answers (prefer sampling on boundary)



History of BayesOpt

Who first proposed the following policies?

probability of improvement?

expected improvement?

upper confidence bound?

knowledge gradient?

What I thought...

probability of improvement?	Harold Kushner, 1964
expected improvement?	Jonas Mockus, 1972
upper confidence bound?	Cox and John, 1998
knowledge gradient?	Frazier, et al., 2009

I was wrong!

probability of improvement?	Harold Kushner, 1964
expected improvement?	~~Jonas Mockus, 1972~~
upper confidence bound?	~~Cox and John, 1998~~
knowledge gradient?	~~Frazier, et al. 2009~~

Okay we can agree on this right? (1964)

H. J. KUSHNER
RIAS, Inc.,
Baltimore, Md.

A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise[1]

A versatile and practical method of searching a parameter space is presented. Theoretical and experimental results illustrate the usefulness of the method for such problems as the experimental optimization of the performance of a system with a very general multipeak performance function when the only available information is noise-distributed samples of the function. At present, its usefulness is restricted to optimization with respect to one system parameter. The observations are taken sequentially; but, as opposed to the gradient method, the observation may be located anywhere on the parameter interval. A sequence of estimates of the location of the curve maximum is generated. The location of the next observation may be interpreted as the location of the most likely competitor (with the current best estimate) for the location of the curve maximum. A Brownian motion stochastic process is selected as a model for the unknown function, and the observations are interpreted with respect to the model. The model gives the results a simple intuitive interpretation and allows the use of simple but efficient sampling procedures. The resulting process possesses some powerful convergence properties in the presence of noise; it is nonparametric and, despite its generality, is efficient in the use of observations. The approach seems quite promising as a solution to many of the problems of experimental system optimization.

Surprise twist! (Kushner, 1962)

JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 5, 150-167 (1962)

A Versatile Stochastic Model of a Function of Unknown and Time Varying Form

HAROLD J. KUSHNER

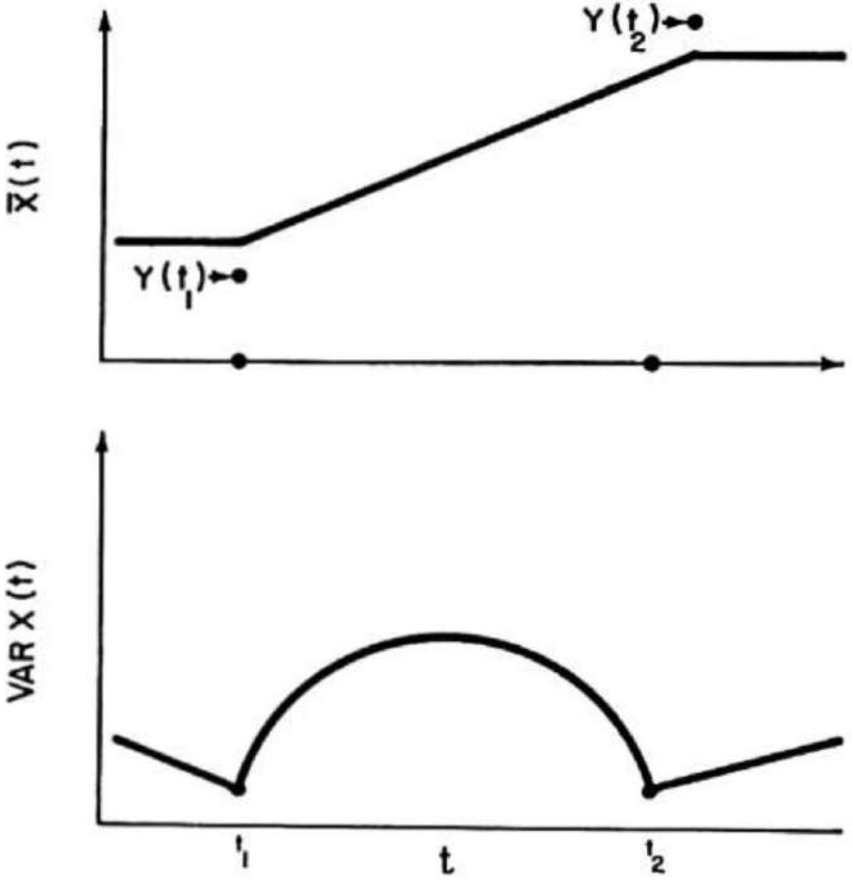
Massachusetts Institute of Technology,
Lincoln Laboratories, Lexington 73, Massachusetts

Submitted by Lotfi Zadeh

Properties of a random walk model of an unknown function are studied. The model is suitable for use in the following (among others) problem. Given a system with a performance function of unknown, time varying, and possibly multipeak form (with respect to a single system parameter), and given that the only information available are noise perturbed samples of the function at selected parameter settings, then determine the successive parameter settings such that the sum of the values of the observations is maximum. An attempt to avoid the optimal search problem through the use of several intuitively reasonable heuristics is presented.

Objective Model (Kushner, 1962)

- Wiener process prior
- additive Gaussian noise



Policy desiderata (Kushner, 1962)

- sample densely

1. As N (the total number of observations) tends to infinity, every region of greater than zero size is sampled at least once.

2. For large N, the initial observations will tend to be information gathering (or play the long shot) and be taken near the point of maximum curve variance.

3. The final observations are taken at points where the expected "pay off" (in whatever sense the observations pay off) will be maximum.

Policy desiderata (Kushner, 1962)

- sample densely
- explore more at the beginning of search

1. As N (the total number of observations) tends to infinity, every region of greater than zero size is sampled at least once.
2. For large N, the initial observations will tend to be information gathering (or play the long shot) and be taken near the point of maximum curve variance.
3. The final observations are taken at points where the expected "pay off" (in whatever sense the observations pay off) will be maximum.

Policy desiderata (Kushner, 1962)

- sample densely
- explore more at the beginning of search
- exploit more at end of search

1. As N (the total number of observations) tends to infinity, every region of greater than zero size is sampled at least once.

2. For large N, the initial observations will tend to be information gathering (or play the long shot) and be taken near the point of maximum curve variance.

3. The final observations are taken at points where the expected "pay off" (in whatever sense the observations pay off) will be maximum.

Policies (Kushner, 1962)

Policy B: probability of improvement (will see again)

B. Sample at the t point ($\hat{t}$) at which ($\epsilon = \epsilon(N, n)$ is a positive sequence)

$$P(X_t \geq \bar{X}^* + \epsilon) = 1 - \Phi\left(\frac{\bar{X}_t + \epsilon}{\sqrt{\mathrm{Var}\, X_t}}\right) \tag{3.2}$$

is maximum.

Policies (Kushner, 1962)

Policy A: upper confidence bound!

A. The location of every observation is selected on the basis of a balance between properties 2 and 3. The simplest such balance is a linear weighing. We select the point at which

$$\sqrt{\operatorname{Var} X_t} + f(N, n)\,(\bar{X}_t - \bar{X}^*) \qquad (3.1)$$

is maximum.

As far as I can tell...

upper confidence bound?	Harold Kushner, 1962
probability of improvement?	Harold Kushner, ~~1964~~ 1962
expected improvement?	
knowledge gradient?	

Further Development (Kushner, 1964)

- same model
- probability of improvement
- (what happened to UCB?)

H. J. KUSHNER
RIAS, Inc.,
Baltimore, Md.

A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise[1]

A versatile and practical method of searching a parameter space is presented. Theoretical and experimental results illustrate the usefulness of the method for such problems as the experimental optimization of the performance of a system with a very general multipeak performance function when the only available information is noise-distributed samples of the function. At present, its usefulness is restricted to optimization with respect to one system parameter. The observations are taken sequentially; but, as opposed to the gradient method, the observation may be located anywhere on the parameter interval. A sequence of estimates of the location of the curve maximum is generated. The location of the next observation may be interpreted as the location of the most likely competitor (with the current best estimate) for the location of the curve maximum. A Brownian motion stochastic process is selected as a model for the unknown function, and the observations are interpreted with respect to the model. The model gives the results a simple intuitive interpretation and allows the use of simple but efficient sampling procedures. The resulting process possesses some powerful convergence properties in the presence of noise; it is nonparametric and, despite its generality, is efficient in the use of observations. The approach seems quite promising as a solution to many of the problems of experimental system optimization.

Very thoughtful! (Kushner, 1964)

- very practical
- computational notes
- careful scheduling of improvement thresholds

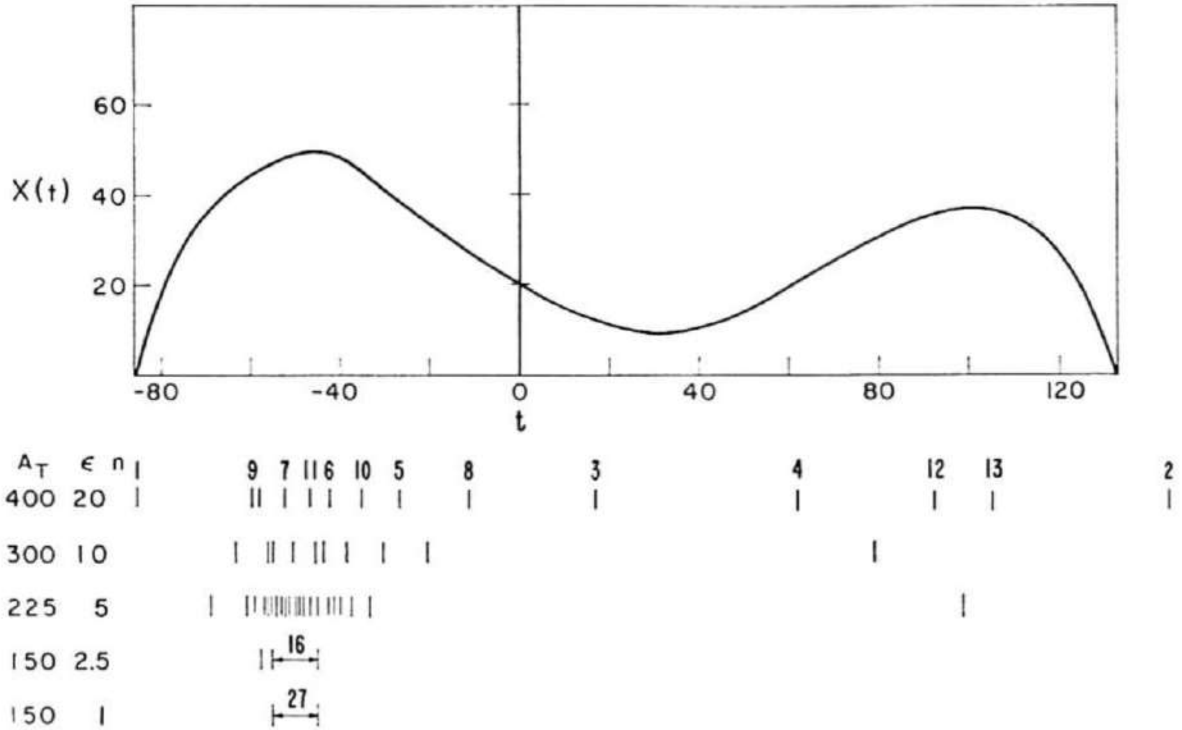


Fig. 4 Experimental results with no observation noise; locations of observations

Very thoughtful! (Kushner, 1964)

- very practical
- computational notes
- scheduling of improvement thresholds
- handling noise

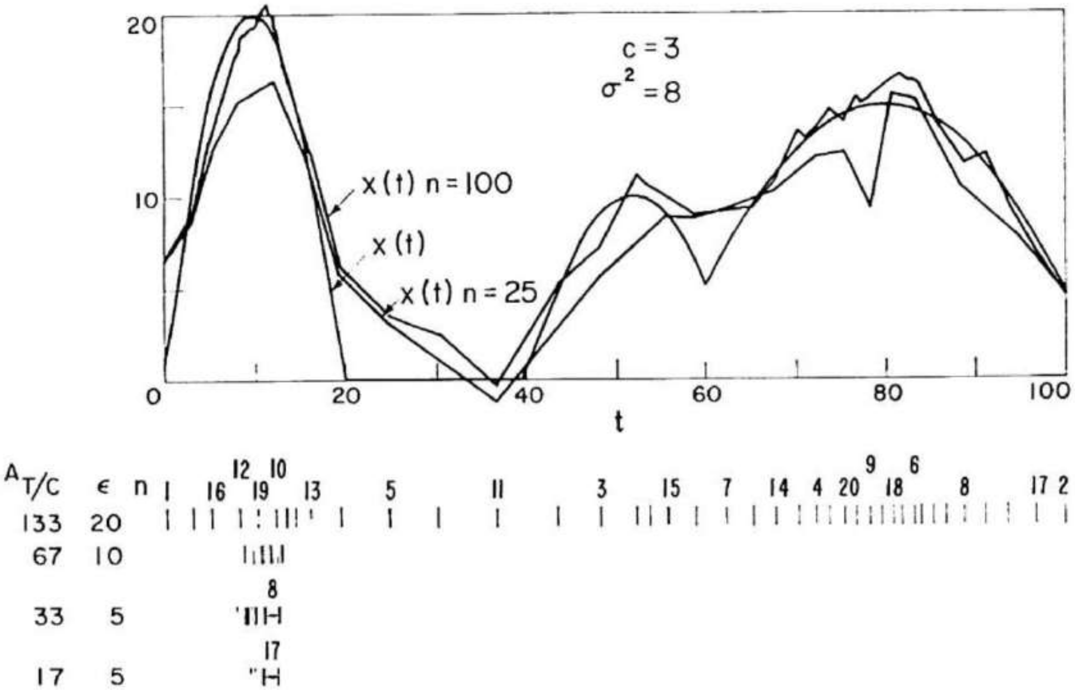
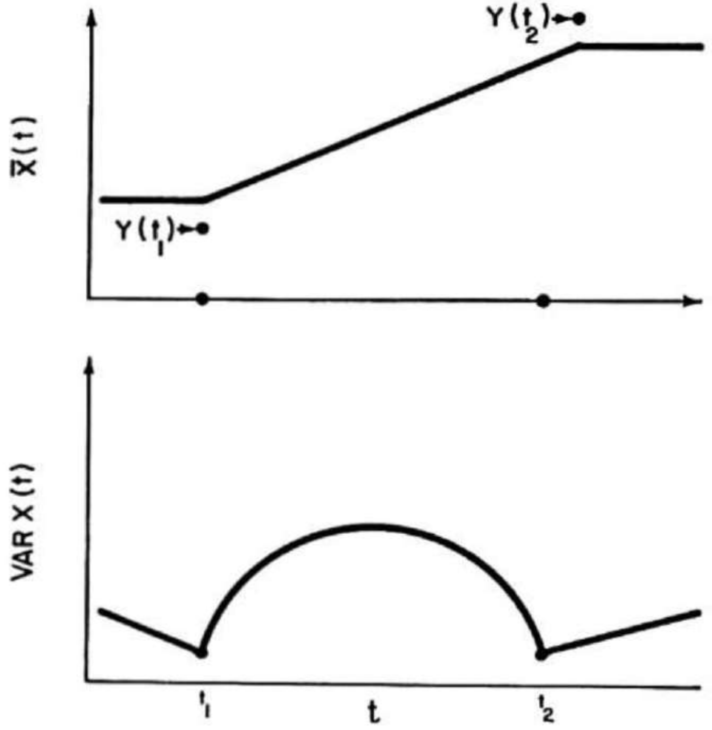
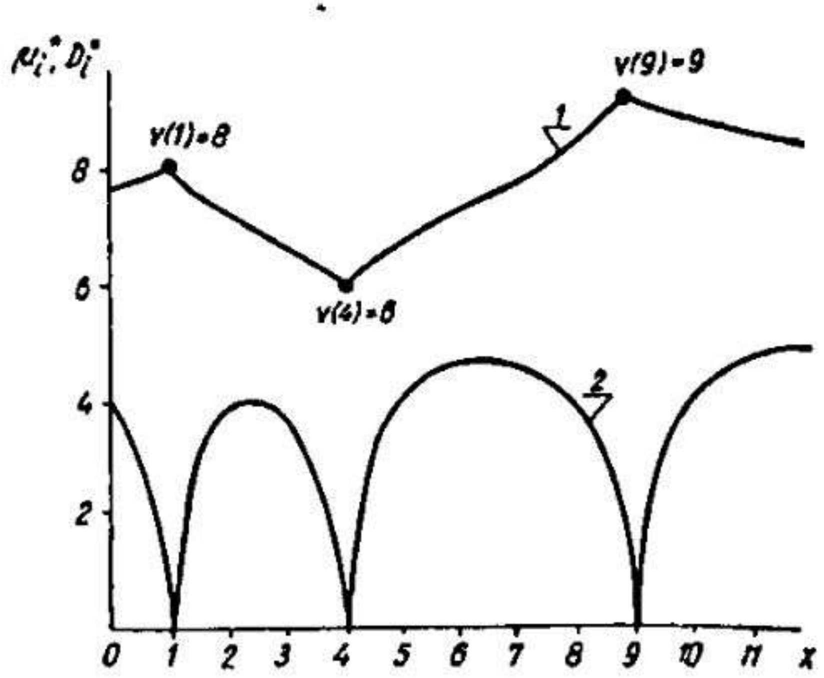


Fig. 7 Experimental results with observation noise

Aside: Gauss-Markov models

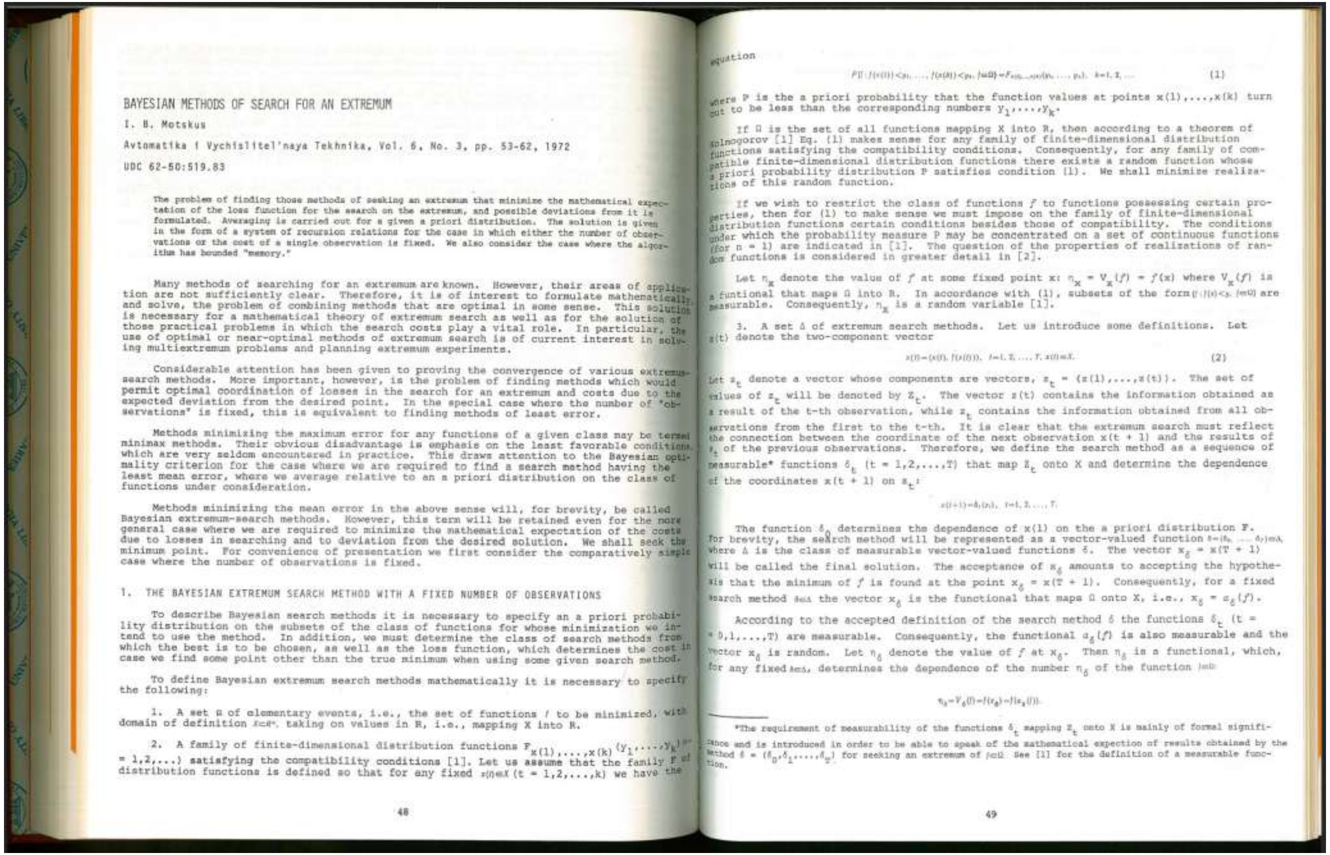


Wiener process (Kushner)



OU process (Šaltyanis)

Okay but this is EI right? (Mockus 1972)



BAYESIAN METHODS OF SEARCH FOR AN EXTREMUM

I. B. Motskus

Avtomatika i Vychislitel'naya Tekhnika, Vol. 6, No. 3, pp. 53-62, 1972

UDC 62-50:519.83

The problem of finding those methods of seeking an extremum that minimize the mathematical expectation of the loss function for the search on the extremum, and possible deviations from it is formulated. Averaging is carried out for a given a priori distribution. The solution is given in the form of a system of recursion relations for the case in which either the number of observations or the cost of a single observation is fixed. We also consider the case where the algorithm has bounded "memory."

Many methods of searching for an extremum are known. However, their areas of application are not sufficiently clear. Therefore, it is of interest to formulate mathematically and solve, the problem of combining methods that are optimal in some sense. This solution is necessary for a mathematical theory of extremum search as well as for the solution of those practical problems in which the search costs play a vital role. In particular, the use of optimal or near-optimal methods of extremum search is of current interest in solving multiextremum problems and planning extremum experiments.

Considerable attention has been given to proving the convergence of various extremum-search methods. More important, however, is the problem of finding methods which would permit optimal coordination of losses in the search for an extremum and costs due to the expected deviation from the desired point. In the special case where the number of "observations" is fixed, this is equivalent to finding methods of least error.

Methods minimizing the maximum error for any functions of a given class may be termed minimax methods. Their obvious disadvantage is emphasis on the least favorable conditions, which are very seldom encountered in practice. This draws attention to the Bayesian optimality criterion for the case where we are required to find a search method having the least mean error, where we average relative to an a priori distribution on the class of functions under consideration.

Methods minimizing the mean error in the above sense will, for brevity, be called Bayesian extremum-search methods. However, this term will be retained even for the more general case where we are required to minimize the mathematical expectation of the costs due to losses in searching and to deviation from the desired solution. We shall seek the minimum point. For convenience of presentation we first consider the comparatively simple case where the number of observations is fixed.

1. THE BAYESIAN EXTREMUM SEARCH METHOD WITH A FIXED NUMBER OF OBSERVATIONS

To describe Bayesian search methods it is necessary to specify an a priori probability distribution on the subsets of the class of functions for whose minimization we intend to use the method. In addition, we must determine the class of search methods from which the best is to be chosen, as well as the loss function, which determines the cost in case we find some point other than the true minimum when using some given search method.

To define Bayesian extremum search methods mathematically it is necessary to specify the following:

1. A set Ω of elementary events, i.e., the set of functions f to be minimized, with domain of definition $X \subset R^n$, taking on values in R, i.e., mapping X into R.

2. A family of finite-dimensional distribution functions $F_{x(1),\dots,x(k)}(y_1,\dots,y_k)$ ($k = 1,2,\dots$) satisfying the compatibility conditions [1]. Let us assume that the family F of distribution functions is defined so that for any fixed $x(t) \in X$ ($t = 1,2,\dots,k$) we have the

equation

$$P\{f: f(x(1))<y_1, \dots, f(x(k))<y_k, f\in\Omega\} = F_{x(1),\dots,x(k)}(y_1,\dots,y_k), \quad k=1,2,\dots \quad (1)$$

where P is the a priori probability that the function values at points $x(1),\dots,x(k)$ turn out to be less than the corresponding numbers $y_1,\dots,y_k$.

If Ω is the set of all functions mapping X into R, then according to a theorem of Kolmogorov [1] Eq. (1) makes sense for any family of finite-dimensional distribution functions satisfying the compatibility conditions. Consequently, for any family of compatible finite-dimensional distribution functions there exists a random function whose a priori probability distribution P satisfies condition (1). We shall minimize realizations of this random function.

If we wish to restrict the class of functions f to functions possessing certain properties, then for (1) to make sense we must impose on the family of finite-dimensional distribution functions certain conditions besides those of compatibility. The conditions under which the probability measure P may be concentrated on a set of continuous functions (for n = 1) are indicated in [1]. The question of the properties of realizations of random functions is considered in greater detail in [2].

Let η_x denote the value of f at some fixed point x: $\eta_x = V_x(f) = f(x)$ where $V_x(f)$ is a funtional that maps Ω into R. In accordance with (1), subsets of the form $\{f: f(x)<y, f\in\Omega\}$ are measurable. Consequently, η_x is a random variable [1].

3. A set Δ of extremum search methods. Let us introduce some definitions. Let $z(t)$ denote the two-component vector

$$z(t) = (x(t), f(x(t))), \quad t=1,2,\dots,T, \ x(t)\in X. \quad (2)$$

Let z_t denote a vector whose components are vectors, $z_t = (z(1),\dots,z(t))$. The set of values of z_t will be denoted by Z_t. The vector $z(t)$ contains the information obtained as a result of the t-th observation, while z_t contains the information obtained from all observations from the first to the t-th. It is clear that the extremum search must reflect the connection between the coordinate of the next observation $x(t+1)$ and the results of z_t of the previous observations. Therefore, we define the search method as a sequence of measurable* functions δ_t ($t = 1,2,\dots,T$) that map Z_t onto X and determine the dependence of the coordinates $x(t+1)$ on z_t:

$$x(t+1) = \delta_t(z_t), \quad t=1,2,\dots,T.$$

The function δ_0 determines the dependence of $x(1)$ on the a priori distribution F. For brevity, the search method will be represented as a vector-valued function $\delta = (\delta_0, \dots, \delta_T) \in \Delta$, where Δ is the class of measurable vector-valued functions δ. The vector $x_\delta = x(T+1)$ will be called the final solution. The acceptance of x_δ amounts to accepting the hypothesis that the minimum of f is found at the point $x_\delta = x(T+1)$. Consequently, for a fixed search method $\delta\in\Delta$ the vector x_δ is the functional that maps Ω onto X, i.e., $x_\delta = a_\delta(f)$.

According to the accepted definition of the search method δ the functions δ_t ($t = 0,1,\dots,T$) are measurable. Consequently, the functional $a_\delta(f)$ is also measurable and the vector x_δ is random. Let η_δ denote the value of f at x_δ. Then η_δ is a functional, which, for any fixed $\delta\in\Delta$, determines the dependence of the number η_δ of the function $f\in\Omega$

$$\eta_\delta = V_\delta(f) = f(x_\delta) = f(a_\delta(f)).$$

*The requirement of measurability of the functions δ_t mapping Z_t onto X is mainly of formal significance and is introduced in order to be able to speak of the mathematical expection of results obtained by the method $\delta = (\delta_0,\delta_1,\dots,\delta_T)$ for seeking an extremum of $f\in\Omega$. See [1] for the definition of a measurable function.

Nope, knowledge gradient! lol (Mockus 1972)

One of the simplifications for the solution of the equations (2) is "one-stage" method [1] [3] when at each stage it is assumed that the following observation is the last one. In such a case the sequence of observations is defined by the equations

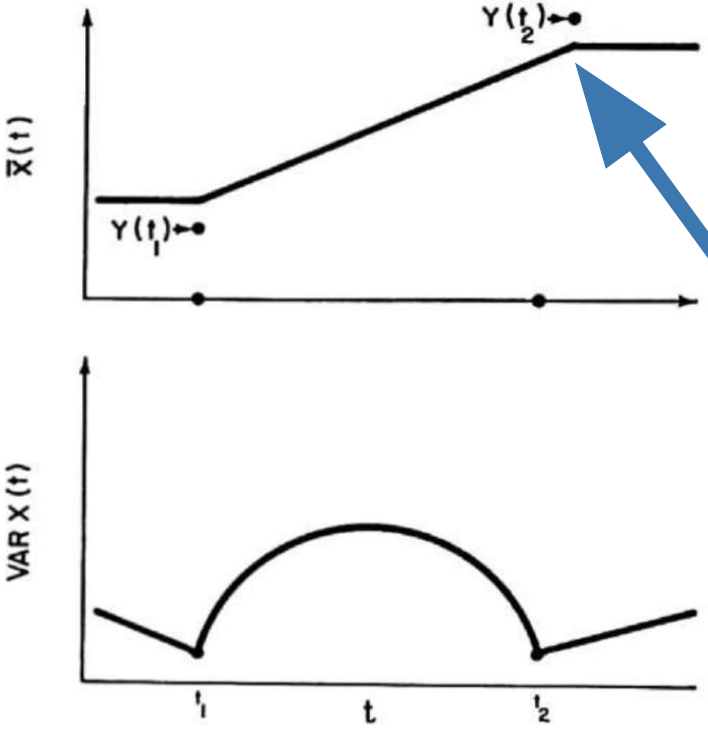
$$E\{u(z_n, f(x_{n+1}), x_{n+1}) | z_n\} = \min_{x \in A} E\{u(z_n, f(x), x) | z_n\}$$

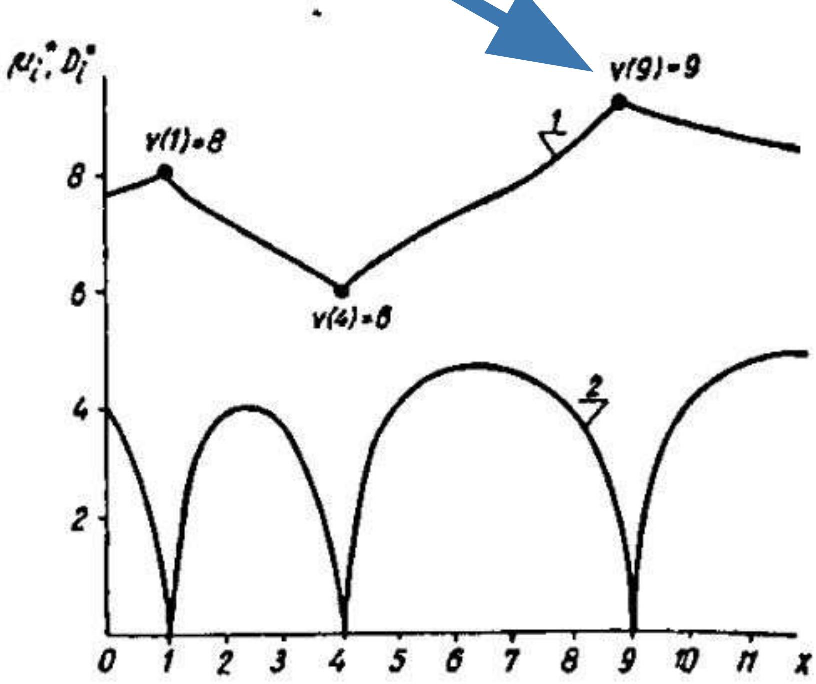
where

$$u(z_{n+1}) = \min_{x \in A} E\{f(x) | z_{n+1}\}, \quad n = 0, \ldots, N.$$

The one-stage Bayesian method converges to the minimum of any continuous function under the conditions of theorem 1.

Equivalent to EI for G-M





Maximum of posterior mean occurs at observation location...

As far as I can tell...

upper confidence bound?	Harold Kushner, 1962
probability of improvement?	Harold Kushner, ~~1964~~ 1962
expected improvement?	
knowledge gradient?	Jonas Mockus, 1972

What about EI? (Šaltyanis, 1971)

ONE METHOD OF MULTIEXTREMUM OPTIMIZATION

V. R. Shaltyanis

Avtomatika i Vychislitel'naya Tekhnika, Vol. 5, No. 3, pp. 33-38, 1971

UDC 62-505

A nonlocal optimization method is proposed which utilizes all the information on the results of tests. The assumptions made lead to an algorithm which is optimum on average for one optimization step. Results of experimental investigations of the algorithm are given.

2. Choice of the loss function. Henceforth we will consider search for the minimum value of the target function, our assumption being that the treatment of the maximization problem will be similar. The smallest known value of the target function will be denoted by $\omega_p = \min_{j=\overline{1,p}} \omega_j$. The effectiveness of the effectiveness of the $(p + 1)$-th trial will be measured by the difference $\Delta w_{p+1} = w_p - w_{p+1}$, while the average effectiveness will be measured by the mathematical expectation $M[\Delta\omega_{p+1}]$.

Expected Improvement (Šaltyanis, 1971)

- OU process prior on objective function
- experiments in up to 32 dimensions!
- very familiar comparison to random search...

Fig. 3. The quantity w_p as a function of the number of tests p: 1) Monte Carlo method; 2) proposed method.

As far as I can tell...

upper confidence bound?	Harold Kushner, 1962
probability of improvement?	Harold Kushner, ~~1964~~ 1962
expected improvement?	Šaltyanis, 1971
knowledge gradient?	Jonas Mockus, 1972

Thank you!