Surrogate Optimization: A Brief Overview

David Bindel

5 August 2019

Department of Computer Science
Cornell University

Logistical notes

- Hard stop by 3:55 PM EDT for kid pickup!
- It is possible that I will not finish
- Slides under “talks” on my web page: `http://www.cs.cornell.edu/~bindel`

Basic setup

Single objective optimization:

$$\text{minimize } f(x) \text{ s.t. } x \in \Omega \subset \mathbb{R}^d$$

Assume

- Ω compact (and simple — e.g. a box)
- f is expensive to evaluate (and maybe noisy)
- We think the true f has some smoothness

Later: constraints, non-smoothness, multi-objective, etc

Common local strategy

Starting from initial guess $x_0 \in \Omega$:

- Build model $s(x) \approx f(x)$ near x_k
    - Linear (Taylor): Gradient descent and co
    - Linear (interp): Nelder-Mead, COBYLA, …
    - Quadratic (Taylor): Scaled gradient, (quasi-)Newton, …
    - Quadratic (interp): NEWUOA, UOBYQA, …
- Follow the model downhill to x_{k+1}
    - For Newton: minimize $s(x)$
    - Limit step based on model trustworthiness
    - Ex: line search, trust region
- Possibly project back to remain within Ω
    - More complex: Penalties, barriers, active sets, etc

Convergence of local methods

General: Convergence to stationary point (usually local min)

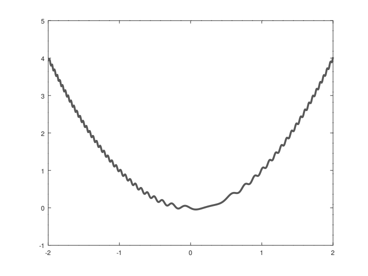
- From close enough initial guess
- Asymptotic convergence rates via model quality
- Globalize convergence via line search or trust region
    - Means “converge to a stationary point if reasonable”
    - Does *not* mean convergence to global minimizer
- May not need too much accuracy in early steps

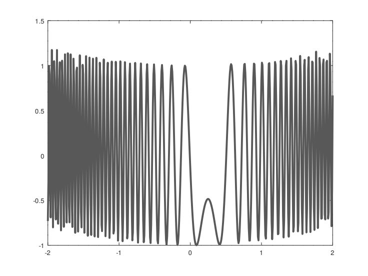
Comments:

- Convergence to global min is rare without more structure
- Can combine with other strategies (e.g. continuation)

From local to global

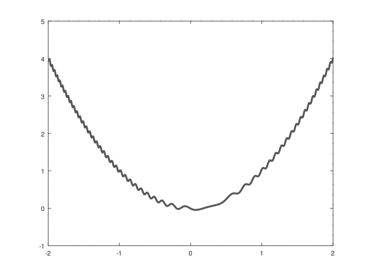
Can *exploit* model for local min; global requires we *explore*.

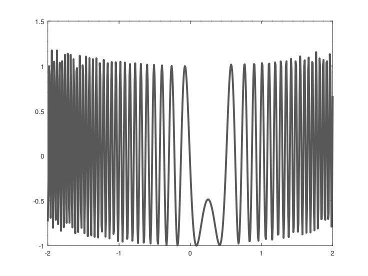




Torn and Zilinskas (1987): For *general* continuous *f*, convergence to *global* minimum $\implies$ dense sampling

From local to global



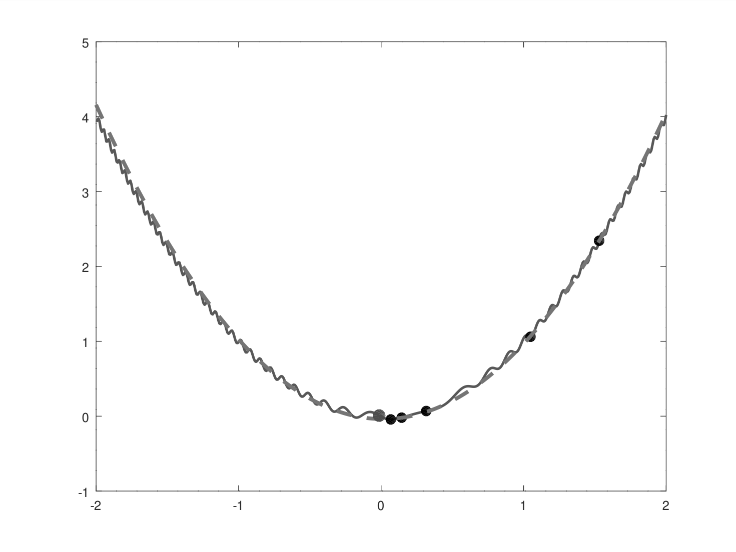


Both problems have local minima that cause issues:

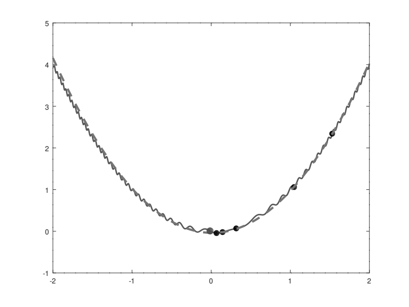
- Left: Good smooth approximations available
- Right: “Glassy” case — little obvious global guidance

These are not equally difficult (especially in high dimensions).
Some evidence that we may be (partly) nice.

From local to global: response surfaces



Response surface



Example approach (two-stage):

- Measure $f(x)$ on a sample set (experimental design)
- Fit a *surrogate* or *response surface* by least squares
- Minimize approximating function

Quality depends on model complexity and noise level

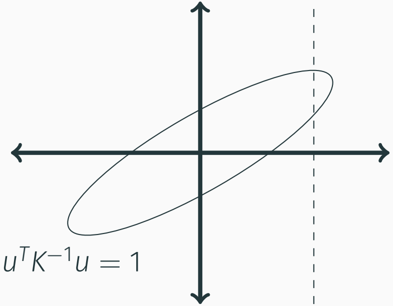
Variations on a theme

Basic idea: Replace expensive f by cheaper $\hat{f}$ (using data)

- Type of surrogate
  - Non-interpolatory (e.g. poly regression, smoothing splines)
  - Interpolatory (e.g. kernel interpolation approaches)
- Surrogate and hyperparameter selection
  - Noise parameters, length scales, etc
- Adaptivity of surrogate
  - Two-stage: Mostly fix surrogate after initial fit
  - One-stage: Continuously update surrogate
- Balancing exploration and exploitation
  - Bayesian interpretations (many types)
  - Frequentist / approximation theoretic
  - Candidate point framework

Will focus on Bayesian for the rest of today.

Being Bayesian



Let $U = (U_1, U_2) \sim N(0, K)$. Given $U_1 = u_1$, what is U_2?

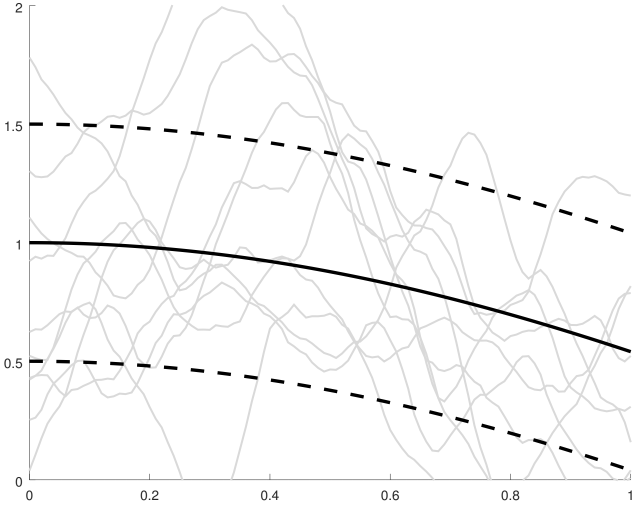
Posterior distribution: $(U_2 | U_1 = u_1) \sim N(w, S)$ where

$$w = K_{21} K_{11}^{-1} u_1$$

$$S = K_{22} - K_{21} K_{11}^{-1} K_{12}$$

How does this generalize to function approximation?

Basic ingredient: Gaussian Processes (GPs)



Basic ingredient: Gaussian Processes (GPs)

Our favorite continuous distributions over

$$\begin{array}{lll} \mathbb{R}: & \text{Normal}(\mu, \sigma^2), & \mu, \sigma^2 \in \mathbb{R} \\ \mathbb{R}^n: & \text{Normal}(\mu, C), & \mu \in \mathbb{R}^n, C \in \mathbb{R}^{n \times n}, C > 0 \\ \mathbb{R}^d \to \mathbb{R}: & \text{GP}(\mu, k), & \mu : \mathbb{R}^d \to \mathbb{R}, k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \end{array}$$

More technically, define GPs by looking at finite sets of points:

$$\forall X = (x_1, \ldots, x_n), x_i \in \mathbb{R}^d,$$

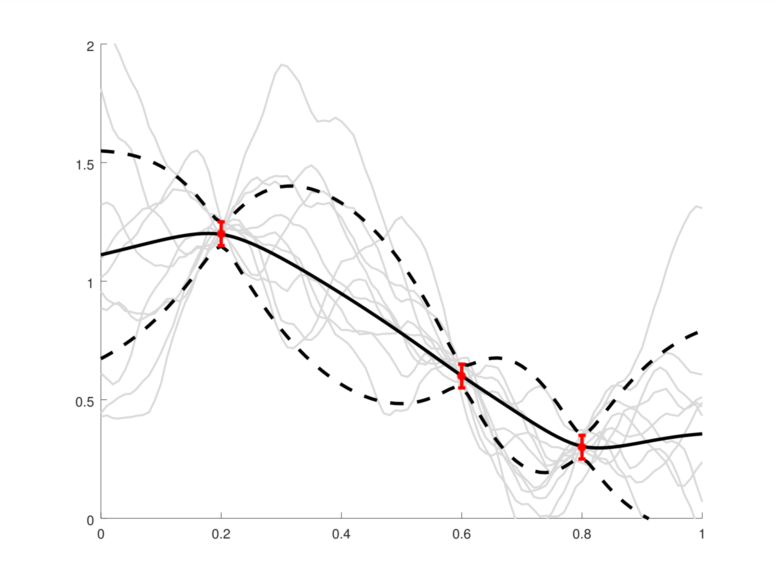
$$\text{have } f_X \sim N(\mu_X, K_{XX}), \text{ where}$$

$$f_X \in \mathbb{R}^n, \quad (f_X)_i \equiv f(x_i)$$

$$\mu_X \in \mathbb{R}^n, \quad (\mu_X)_i \equiv \mu(x_i)$$

$$K_{XX} \in \mathbb{R}^{n \times n}, \quad (K_{XX})_{ij} \equiv k(x_i, x_j)$$

Being Bayesian with GPs



Being Bayesian with GPs

Now consider prior of $f \sim \mathrm{GP}(\mu, k)$, noisy measurements

$$f_X \sim y + \epsilon, \quad \epsilon \sim N(0, W), \qquad \text{typically } W = \sigma^2 I$$

Posterior is $f \sim \mathrm{GP}(\mu', k')$ with

$$\begin{aligned} \mu'(x) &= \mu(x) + K_{xX} c & \tilde{K} &= K_{XX} + W \\ k'(x, x') &= K_{xx'} - K_{xX} \tilde{K}^{-1} K_{Xx'} & c &= \tilde{K}^{-1}(y - \mu_X) \end{aligned}$$

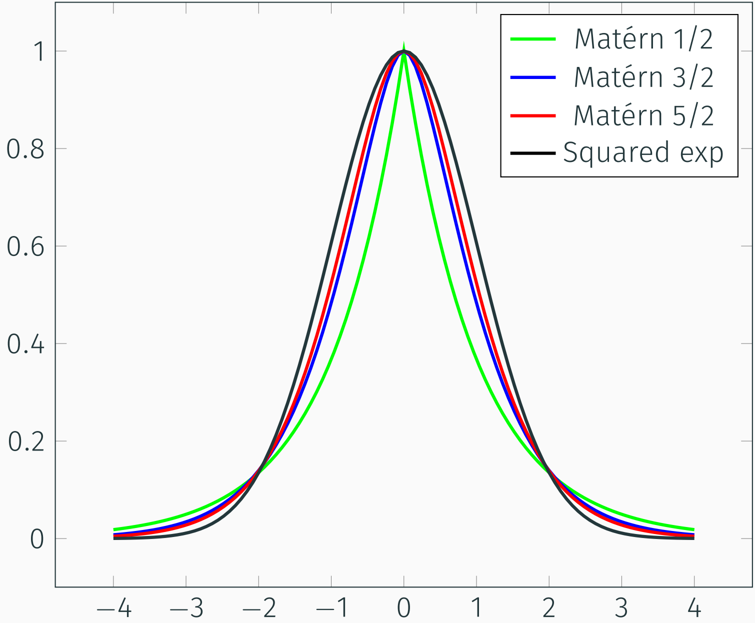
The expensive bit (for large n): solves with $\tilde{K}$.

Incorporating assumptions

Key places to inject assumptions on f:

- Kernel choice
  - Standard choices (Matérn, polyharmonic) are *universal*
  - … but better choices require less training data
  - Typically choose based on some belief about smoothness
- Mean field
  - Standard choices are constant or linear
  - Can bring in other shapes if more is known
- Can also include *covariates* — another time

Common Kernels



Hyper-parameter selection

Simplest approach is maximum likelihood estimation

$$\ell(\theta|y) = \log p(y|\theta) = \log\left[\frac{1}{\sqrt{2\pi\tilde{K}}}\exp\left(-\frac{1}{2}(y-\mu_X)^T\tilde{K}^{1}(y-\mu_X)\right)\right]$$

Decompose into data fidelity and model complexity terms

$$\begin{aligned}\ell(\theta|y) &= \ell_y(\theta|y) + \ell_K(\theta|y) - \frac{n}{2}\log(2\pi)\\ \ell_y(\theta|y) &= -\frac{1}{2}(y-\mu_X)^T\tilde{K}^{-1}(y-\mu_X)\\ \ell_K(\theta|y) &= -\frac{1}{2}\log\det\tilde{K}\end{aligned}$$

Alternatives like MAP and GCV involve similar computations.

Managing mean fields

Previously assumed fixed mean; can also choose

$$\mu(x) = \sum_j d_j \pi_j(x)$$

where $\{\pi_j\}$ usually span a low-degree polynomial space. Then

$$\begin{bmatrix} \tilde{K} & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

Posterior mean as before, variance only slightly complicated.

Initial experimental design

- Want well-posedness for determining the mean field
- Independent random points are usually to clustered
- Typical choices:
    - Latin hypercube (and symmetric)
    - Two-factorial designs (corners of hypercube)
- Start with small, “good enough” design, then adapt

The problem with naive adaptive sampling

Naive adaptive sampling:

- Fit GP surrogate
- Find where expected value is smallest
- Sample there and repeat

Good idea, but not guaranteed to find a stationary point!
Basic problem: Not enough exploration.

Adaptive sampling and acquisition functions

After initial experiment, adaptive phase

- Choose next point(s) by optimizing an *acquisition function*
- Typical choice: Expected Improvement (EI)

$$EI(x) = \mathbb{E}\left[\min_{i \leq k} f(x_i) - f(x)\right]_+$$

    - Basis of Efficient Global Optimization (EGO)
    - Greedy — only looks one step ahead
    - Known not to explore enough in many cases
    - But can evaluate in closed form
- Other common choices
    - Probability of improvement (PI)
    - Knowledge gradient (KG)
    - Upper/Lower Confidence Bound (UCB/LCB)

Summary: Basic Bayesian Optimization

- Choose appropriate prior (kernel and mean)
- Sample function using some experimental design
- Fit surrogate (and hyperparameters)
- While there is budget
  - Aux problem: optimize acquisition function
  - Evaluate function and update surrogate

Final notes

- No free lunch: results depend heavily on assumptions
  - Weak assumptions $\implies$ lots of exploration
  - Too strong $\implies$ may explore too little
  - "Black box" is (as usual) a misnomer
- Best for less than 20 dimensions
  - OK with *effective* low-dimensional structure
  - Can somewhat incorporate this into kernel
- There are useful surrogate methods other than BO!

Selected references

- Frazier (2018): Tutorial on Bayesian Optimization
- Peherstorfer, Willcox, Gunzburger (2018): Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization
- Jones (2001): Taxonomy of Global Optimization Methods Based on Response Surfaces