Global Optimization with Native Space Semi-Norm Bounds

David Eriksson

Center for Applied Mathematics
Cornell University

dme65@cornell.edu

March 2, 2017

Joint work with David Bindel and Christine Shoemaker

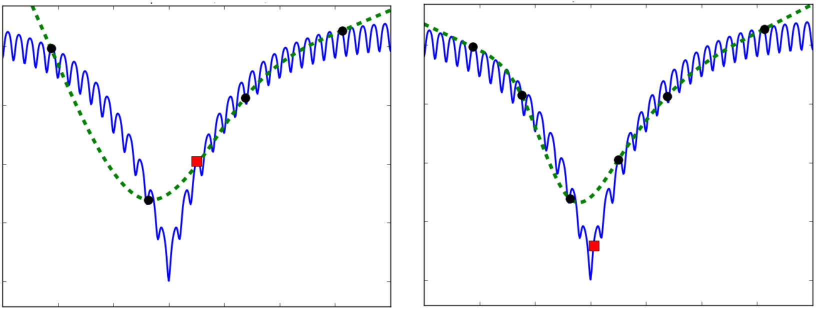
Background: Global optimization

- Global optimization problem (GOP)

$$\begin{array}{ll} \text{minimize} & f(x) \\ x \in \Omega & \end{array}$$

- $f : \Omega \to \mathbb{R}$ a continuous, deterministic, expensive black-box
- $\Omega \subset \mathbb{R}^d$ is compact (usually a hypercube)

Background: Surrogate optimization



- Use a surrogate $\hat{f}$ (- - -) to approximate f (———)
- Common surrogates: RBFs, Kriging, MARS, polynomials

Main idea: Sample, fit the surrogate $\hat{f}$, repeat

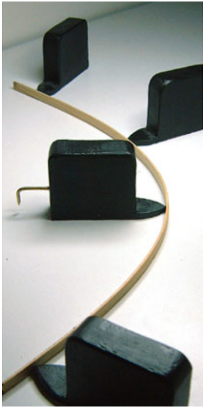
Theorem (Törn and Zilinskas)

Convergence of GOP for all $f \in \mathcal{C}(\Omega) \implies$ *dense sampling.*

Possible retorts:

- Give up on global convergence
  (Con: Can get arbitrarily bad answers in principle)
- Use methods that eventually sample densely
  (Con: Eventually, we all die)
- Assume a more regular class of functions
  (Our approach today)

Cubic splines and beam bending



- The bending energy for a beam is:

$$\Psi[u] = \frac{1}{2} \int_{\alpha}^{\beta} u''(x)^2 \, dx$$

- Natural spline minimizes this energy subject to interpolation

More splines

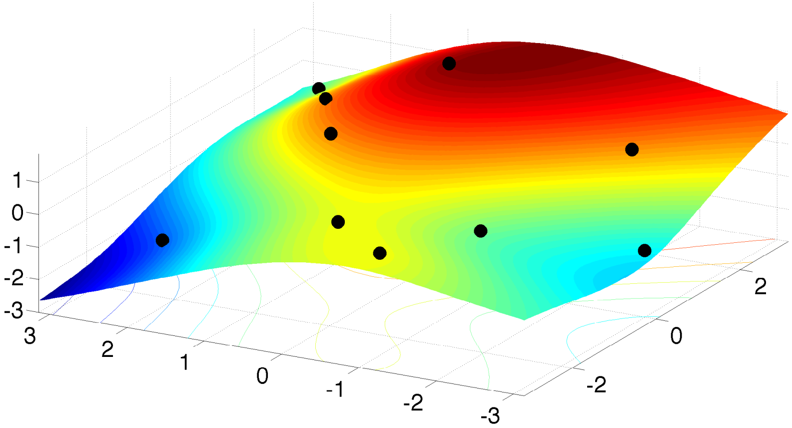
- A particular representation of a piece-wise cubic:

$$s(x) = c_0 + c_1 x + \sum_{j=1}^{n} \lambda_j |x - x_j|^3$$

- Make natural: Add $s(x_j) = f(x_j)$, $\sum_{j=1}^{n} \lambda_j = 0$, $\sum_{j=1}^{n} \lambda_j x_j = 0$
- Can write $\Psi[s] = \frac{1}{6}\lambda^T \Phi \lambda$ where $\Phi_{ij} = |x_i - x_j|^3$
- Want to minimize Ψ subject to $P^T \lambda = 0$ and interpolation
- The KKT conditions are:

$$\begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ c \end{bmatrix} = \begin{bmatrix} f_X \\ 0 \end{bmatrix}, \qquad \text{where } P^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$

Beyond beams bending



$$\Psi[u] = \frac{1}{2}\int_{\Omega} (\nabla^2 u)^2 \, d\Omega$$

Functional form of the interpolant:

$$s_{f,X}(x) = \sum_{j=1}^{n} \lambda_j \varphi(\|x - x_j\|) + p(x)$$

Interpolation constraints:

$$s(x_i) = f(x_i), \qquad i = 1, \ldots, n$$

Discrete orthogonality:

$$\sum_{j=1}^{n} \lambda_j q(x_j) = 0, \qquad \forall q \in \Pi_{k-1}^d$$

- $X = \{x_i\}_{i=1}^n$ pairwise distinct interpolation nodes
- $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is CPD of order k
- $p \in \Pi_{k-1}^d$ a polynomial in d dims of degree at most $k-1$

Conditional positive definite RBFs

φ is conditionally positive definite of order k if for all $X = \{x_1, \dots, x_n\}$ distinct and $\lambda \neq 0$ s.t.

$$\sum_{j=1}^{n} \lambda_j q(x_j) = 0, \qquad \forall q \in \Pi_{k-1}^d$$

we have that

$$\sum_{i,j} \lambda_i \lambda_j \varphi(\|x_i - x_j\|) > 0.$$

Popular RBF kernels

Name	$\varphi(x)$	Order	Example
Gaussian	$e^{-\epsilon^2 \|x\|^2}$	$k = 0$	
Inverse multiquadric	$\left(1 + \epsilon^2 \|x\|^2\right)^{\beta},\ \beta < 0$	$k = 0$	$\frac{1}{\sqrt{1+\epsilon^2 \|x\|^2}}$
Multiquadric	$(-1)^{\lceil \beta \rceil} \left(1 + \epsilon^2 \|x\|^2\right)^{\beta},\ 0 < \beta \notin \mathbb{N}$	$k = \lceil \beta \rceil$	$\sqrt{1 + \epsilon^2 \|x\|^2}$
Radial powers	$(-1)^{\lceil \beta/2 \rceil} \|x\|^{\beta},\ 0 < \beta \notin 2\,\mathbb{N}$	$k = \lceil \beta/2 \rceil$	$\|x\|^3$
Thin-plate spline	$(-1)^{\beta+1} \|x\|^{2\beta} \log(\|x\|),\ \beta \in \mathbb{N}$	$k = \beta + 1$	$\|x\|^2 \log(\|x\|)$

- Cubic RBF + linear tail is popular for surrogate optimization
- Gaussian RBF is popular in ML
- Choice of shape parameter $\epsilon > 0$ is critical

Native spaces and semi-inner products

- The RBF space $\mathcal{A}_{\varphi,k}$ is the space of functions of the form

$$s_{f,X}(x) = \sum_{j=1}^{n} \lambda_j \varphi(\|x - x_j\|) + p(x)$$

that satisfy

$$\sum_{j=1}^{n} \lambda_j q(x_j) = 0, \qquad \forall q \in \Pi_{k-1}^{d}.$$

- $\mathcal{A}_{\varphi,k}$ can be equipped with the semi-inner product

$$\langle s, u \rangle = (-1)^k \sum_{i=1}^{n(s)} \lambda_i u(x_i)$$

for s, $u \in \mathcal{A}_{\varphi,k}$.

Semi-norms and energy

- We can define a semi-norm on $\mathcal{A}_{\varphi,k}$ via

$$
\begin{aligned}
|s_{f,X}|^2 &:= \langle s_{f,X}, s_{f,X} \rangle \\
&= (-1)^k \sum_{i=1}^{n} \lambda_i s_{f,X}(x_i) \\
&= (-1)^k \sum_{i,j=1}^{n} \lambda_i \lambda_j \varphi(\|x_i - x_j\|) \\
&= (-1)^k \lambda^T \Phi \lambda.
\end{aligned}
$$

- Native space: Closure of splines under semi-norm
- Native space semi-norm:

$$
|f|_{\mathcal{N}_{\varphi,k}} = \sup_{X \subset \Omega,\, |X| < \infty} |s_{f,X}|
$$

For RBFs the KKT conditions of

$$\min_{x\in\Omega} \frac{1}{2}\lambda^T \Phi \lambda - \lambda^T f_X \text{ s.t. } P^T\lambda = 0$$

are

$$\begin{bmatrix} 0 & P^T \\ P & \Phi \end{bmatrix} \begin{bmatrix} c \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ f_X \end{bmatrix} \qquad (Aw = b)$$

where

- $\Phi_{ij} = \varphi(\|x_i - x_j\|)$
- $P_{ij} = \pi_j(x_i)$, and $\{\pi_j\}_{j=1}^m$ is a basis for $\mathcal{P}_{k-1}^d$

When is this well-posed?

- If $\mathrm{rank}(P) = m$
- $\deg(p) = k - 1$ is at least the order of the CPD kernel φ

Special cases of native spaces

- Native space for radial powers and thin-plate splines:

$$\mathsf{BL}_\ell(\mathbb{R}^d) = \{f \in \mathcal{C}(\mathbb{R}^d) : D^\alpha f \in L^2(\mathbb{R}^d),\ \ \forall |\alpha| = \ell,\ \alpha \in \mathbb{N}^d\}.$$

- Native space for Gaussians and (inverse) multiquadrics harder to characterize
    - These spaces are rather small
    - For the Gaussian, the Fourier transform of $f \in \mathcal{N}(\Omega)$ must decay faster than the Fourier transform of a Gaussian
    - These spaces are unlikely to contain functions in applications

Estimates for functions in the native space

- Generic error estimate:

$$|f(x) - s_{f,X}(x)| \le P_{X,\varphi}(x)\sqrt{|f|^2_{\mathcal{N}_{\varphi,k}} - |s_{f,X}|^2_{\mathcal{N}_{\varphi,k}}}$$

- Power function:

$$[P_{X,\varphi}(x)]^2 = \varphi(0) - v(x)^T A^{-1} v(x)$$

where

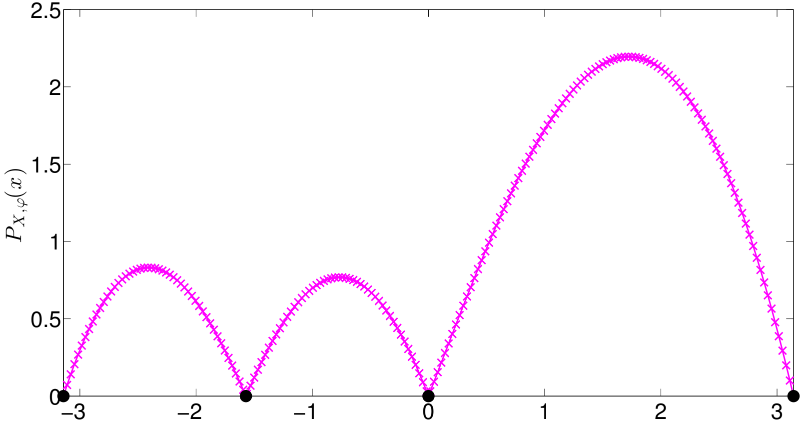
$$v(x) = [\pi_1(x), \ldots, \pi_m(x), \varphi(\|x - x_1\|), \ldots, \varphi(\|x - x_n\|)]^T.$$

- Can be seen as the Schur complement of the extended system:

$$\begin{bmatrix} A & v(x) \\ v(x)^T & \varphi(0) \end{bmatrix} \begin{bmatrix} w \\ \mu \end{bmatrix} = \begin{bmatrix} b \\ f(x) \end{bmatrix}$$

- $P_{X,\varphi}(x)$ tells us how stiff the surface is at a given point

Power function



- Power function for $X = [-\pi, -\pi/2, 0, \pi]$
- Cubic kernel + Linear tail

Lower bounds

- The error estimate gives a lower bound for $f(x)$:

$$f(x) \geq \ell_{f,X}(x) := s_{f,X}(x) - P_{X,\varphi}(x)\sqrt{|f|^2_{\mathcal{N}_{\varphi,k}} - |s_{f,X}|^2_{\mathcal{N}_{\varphi,k}}}$$

- Requires that we know $|f|_{\mathcal{N}_{\varphi,k}}$ or an upper bound
  - Use semi-norm of initial spline times a fudge factor?
- A natural thing to do is to minimize $\ell_{f,X}$
  - Potentially hard, since it can be multimodal
  - Evaluating $\ell_{f,X}$ cheap compared to f
  - Acceptable to brute-force

Algorithm

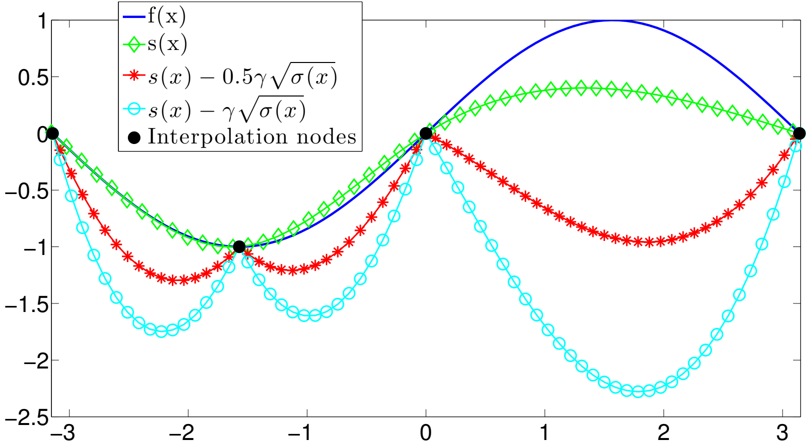
Algorithm 1: Optimization algorithm that minimizes the lower bound at each step

Tolerance ϵ
X_0 initial points
f_{X_0} initial function values
Build s_{f,X_0} from (X_0, f_{X_0})
$n \leftarrow 0$
while $\left|\min f_{X_n} - \min_{x\in\Omega} \ell_{f,X_n}(x)\right| > \epsilon$ **do**
    $y \leftarrow \arg\min_{x\in\Omega} \ell_{f,X_n}(x)$
    $X_{n+1} \leftarrow X_n \cup \{y\}$
    $f_{X_{n+1}} \leftarrow f_{X_n} \cup \{f(y)\}$
    Build $s_{f,X_{n+1}}$ from $(X_{n+1}, f_{X_{n+1}})$
    $n \leftarrow n+1$
end while

Gutmann + Vary the semi-norm budget

- Unlikely that all energy will be used for one point
- **Solution:** Vary the fraction of energy that is used in $\ell_{f,X}$
- Gutmann proposed sampling based on a target value
  - Samples where the least energy is needed to reach target value
  - This makes the surface less bumpy
  - Target values are cycled
- We can do similarly with the amount of energy that we use
- Energy is more natural than target values

Vary the semi-norm budget



- Exploration vs exploitation

Convergence rates and fill-in distance

- At the global minimum x^* :

$$\begin{aligned}|f(x^*) - \ell_{f,X_n}(x^*)| &= |f(x^*) - s_{f,X_n}(x^*) + \gamma P_{X_n,\varphi}(x^*)| \\ &\leq |f(x^*) - s_{f,X_n}(x^*)| + \gamma P_{X_n,\varphi}(x^*) \\ &\leq 2\gamma P_{X_n,\varphi}(x^*)\end{aligned}$$

- Convergence rates for the power function depends on the fill-in distance:

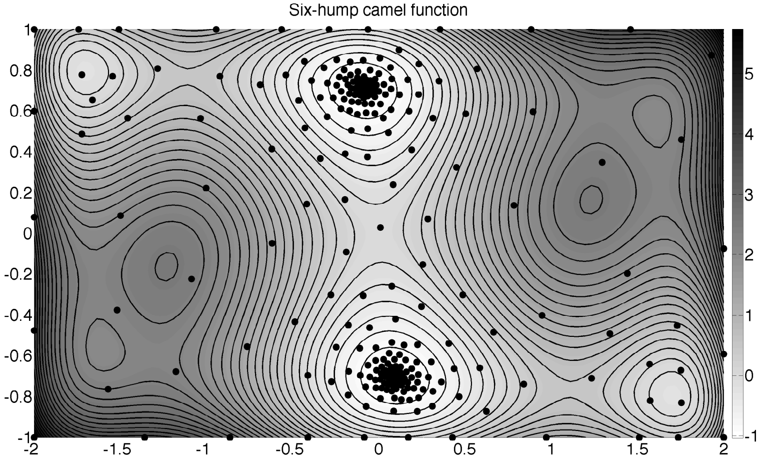
$$h_{X,\Omega} := \sup_{x\in\Omega} \min_{x_j \in X} \|x - x_j\|_2.$$

- Can be shown that:

$$|f(x) - s_{f,X_n}(x)| \leq C\sqrt{F(h_{X_n,\Omega})}\,|f|_{\mathcal{N}_{\varphi,k}}, \qquad \forall x \in \Omega,$$

- Problem: Our goal was to not sample densely, so $h_{X,\Omega}$ may be large
- ϵ-modification gives this rate, but this is an undesirable solution

Sampling pattern



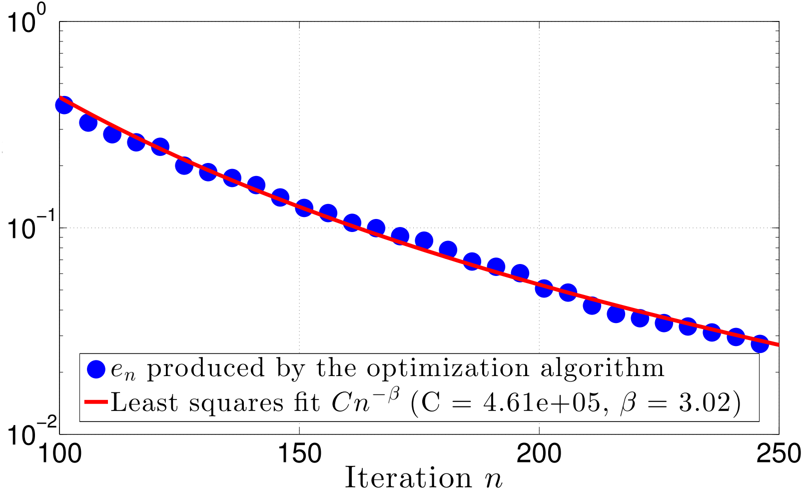


Figure: Convergence rates for the Camel function

- Looking for: $e_n = f(x^*) - \min_{x \in \Omega} \ell_{f,X_n}(x) = Cn^{-\beta}$
- $\beta = 3/4$ is expected from theory for cubic kernel + linear tail

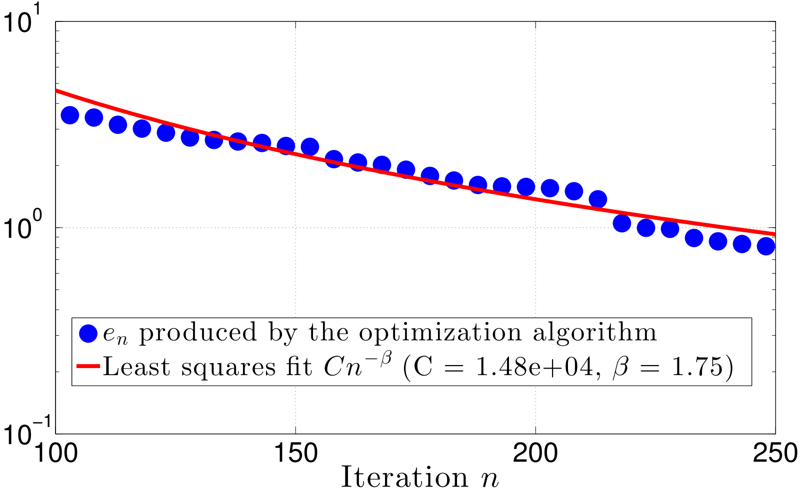


Figure: Convergence rates for Hartman3

- Looking for: $e_n = f(x^*) - \min_{x\in\Omega} \ell_{f,X_n}(x) = Cn^{-\beta}$
- $\beta = 1/2$ is expected from theory for cubic kernel + linear tail

Conclusions

Covered today:

- Connection between energy budgets and optimization
- Globally convergent algorithm that does not sample densely
- Numerical convergence rates agree with RBF theory
- Sampling patterns are beautiful

Next steps:

- Estimation of the semi-norm
- Deal with functions that are not in the native space
- The algorithm will be added to pySOT (github.com/dme65/pySOT)
- Use our algorithm on a real-world optimization problem

Thank you for your attention!