Iterated Brownian Bridge Kernels for Sequential Kriging Optimization

Greg Fasshauer\textsuperscript{1} and Mike McCourt\textsuperscript{2}

\textsuperscript{1}Colorado School of Mines, \textsuperscript{2}SigOpt

SIAM CSE 2017
Advances in Reproducing Kernel Hilbert Spaces Applied to Bayesian Optimization
Atlanta
March 2, 2017
Sequential Kriging Optimization (SKO)

Goal: Unconstrained minimization of a (costly) black-box function $f$.

Our Approach: Use kernel-based/kriging surrogate model for $f$. 

**SKO algorithm**:

1. Create (inexpensive) kriging prediction $s$ using $N$ function values $y = (y_1, \ldots, y_N)^T$ at locations $X = \{x_1, \ldots, x_N\}$.

2. Analyze the surrogate $s$ to determine the “best” location $x_{N+1}$ for the optimization of $f$.

3. Update $X = X \cup \{x_{N+1}\}$, $y = (y, y_{N+1})$ and $s$. Return to Step 2.

We judge “best” in terms of expected improvement \cite{JSW98, SWN03}, 

$$
\text{EI}(x) = \mathbb{E}[y^* - \text{Y}_x]
= \int_{-\infty}^{\infty} (y^* - y) \text{d}F_{\text{Y}_x}(y), 
$$

where $F_{\text{Y}_x}$ is the distribution associated with $y$-values and $y^*$ is the value upon which to improve (which should be $F_{\text{Y}_x}$-measurable).
Sequential Kriging Optimization (SKO)

**Goal:** Unconstrained minimization of a (costly) black-box function \( f \).

**Our Approach:** Use kernel-based/kriging surrogate model for \( f \).

**SKO algorithm:**

1. Create (inexpensive) kriging prediction \( s \) using \( N \) function values 
   \[ y = (y_1 \cdots y_N)^T \] at locations \( \mathcal{X} = \{x_1, \ldots, x_N\} \).
2. Analyze the surrogate \( s \) to determine the “best” location \( x_{N+1} \) for 
   the optimization of \( f \).
3. Update \( \mathcal{X} = \mathcal{X} \cup \{x_{N+1}\}, \ y = \begin{pmatrix} y \\ y_{N+1} \end{pmatrix} \) and \( s \). Return to Step 2.

We judge “best” in terms of expected improvement [JSW98, SWN03],

\[ EI(x) = E[(y^* - Yx) +] = \int_{-\infty}^{\infty} (y^* - y) + dF_{Yx}(y) \]

where \( F_{Yx}(y) \) is the distribution associated with \( y \)-values and 
\( y^* \) is the 
value upon which to improve (which should be \( F_{Yx} \)-measurable).

Greg Fasshauer
Sequential Kriging Optimization (SKO)

**Goal:** Unconstrained minimization of a (costly) black-box function $f$.

**Our Approach:** Use kernel-based/kriging surrogate model for $f$.

**SKO algorithm:**

1. Create (inexpensive) kriging prediction $s$ using $N$ function values $y = (y_1 \cdots y_N)^T$ at locations $X = \{x_1, \ldots, x_N\}$.
2. Analyze the surrogate $s$ to determine the “best” location $x_{N+1}$ for the optimization of $f$.
3. Update $X = X \cup \{x_{N+1}\}$, $y = \begin{pmatrix} y \\ y_{N+1} \end{pmatrix}$ and $s$. Return to Step 2.

We judge “best” in terms of expected improvement [JSW98, SWN03],

$$EI(x) = \mathbb{E}[(y^* - Y_x)_+] = \int_{-\infty}^{\infty} (y^* - y)_+ dF_{Y_x}(y),$$

where $F_{Y_x}$ is the distribution associated with $y$-values and $y^*$ is the value upon which to improve (which should be $F_{Y_x}$-measurable).
The Kriging Model

Given a Gaussian (zero-mean) random field $Y$ with covariance kernel $K$ and observations

$$Y = (Y_{x_1}, \ldots, Y_{x_N})^T,$$

$Y_{x_j}$ zero-mean random variables, the (simple) kriging predictor is of the form

$$\hat{Y}_x = \sum_{j=1}^{N} w_j(x) Y_{x_j} = w(x)^T Y,$$

$\hat{Y}_x$: zero-mean random variable,

$w(\cdot) = (w_1(\cdot), \ldots, w_N(\cdot))^T$: vector of weight functions.

“Optimal” weights $\hat{w}_j(\cdot)$ will minimize the MSE of the predictor, i.e.,

$$\text{MSE}(\hat{Y}_x) = \mathbb{E} \left[ \left( Y_x - w(x)^T Y \right)^2 \right].$$
Using the covariance kernel \( K \) of \( Y \), i.e., \( K(x, z) = \mathbb{E}[Y_x Y_z] \), we have

\[
\text{MSE}(\hat{Y}_x) = \mathbb{E} \left[ \left( Y_x - w(x)^T Y \right)^2 \right]
= \mathbb{E}[Y_x Y_x] - 2\mathbb{E}[Y_x w(x)^T Y] + \mathbb{E}[w(x)^T Y Y^T w(x)]
= K(x, x) - 2w(x)^T (k(x)) + w(x)^T K w(x).
\]
Using the covariance kernel $K$ of $Y$, i.e., $K(x, z) = \mathbb{E}[Y_x Y_z]$, we have

$$\text{MSE}(\hat{Y}_x) = \mathbb{E} \left[ \left( Y_x - w(x)^T Y \right)^2 \right]$$

$$= \mathbb{E}[Y_x Y_x] - 2\mathbb{E}[Y_x w(x)^T Y] + \mathbb{E}[w(x)^T Y Y^T w(x)]$$

$$= K(x, x) - 2w(x)^T (k(x)) + w(x)^T K w(x).$$

Differentiation and equating to 0 yields the optimum weight vector

$$\hat{w}(x) = K^{-1} k(x),$$
Using the covariance kernel $K$ of $Y$, i.e., $K(x, z) = \mathbb{E}[Y_x Y_z]$, we have

$$
\text{MSE}(\hat{Y}_x) = \mathbb{E} \left[ \left( Y_x - w(x)^T Y \right)^2 \right]
$$

$$
= \mathbb{E}[Y_x Y_x] - 2\mathbb{E}[Y_x w(x)^T Y] + \mathbb{E}[w(x)^T Y Y^T w(x)]
$$

$$
= K(x, x) - 2w(x)^T k(x) + w(x)^T K w(x).
$$

Differentiation and equating to 0 yields the optimum weight vector

$$
\hat{w}(x) = K^{-1} k(x),
$$

so that the (simple) kriging predictor

$$
\hat{Y}_x = k(x)^T K^{-1} Y
$$

is the best linear unbiased predictor.
Using the covariance kernel $K$ of $Y$, i.e., $K(x, z) = \mathbb{E}[Y_x Y_z]$, we have

$$\text{MSE}(\hat{Y}_x) = \mathbb{E} \left[ (Y_x - w(x)^T Y)^2 \right]$$

$$= \mathbb{E}[Y_x Y_x] - 2\mathbb{E}[Y_x w(x)^T Y] + \mathbb{E}[w(x)^T Y Y^T w(x)]$$

$$= K(x, x) - 2w(x)^T (k(x)) + w(x)^T K w(x).$$

Differentiation and equating to 0 yields the optimum weight vector

$$w(x)^* = K^{-1} k(x),$$

so that the (simple) kriging predictor

$$\hat{Y}_x = k(x)^T K^{-1} Y$$

is the best linear unbiased predictor. The surrogate $s$ corresponds to the kriging prediction (specific realization)

$$\hat{y}_x = k(x)^T K^{-1} y.$$
Expected Improvement

The kriging model provides a normal density $p(y|x, Y_x)$ for the EI computation:

$$EI(x) = \int_{-\infty}^{\infty} (y^* - y) dF_{Y_x}(y)$$

$$= \int_{-\infty}^{y^*} (y^* - \mu) dF_{Y_x}(y) - \int_{-\infty}^{y^*} (y - \mu) dF_{Y_x}(y)$$

$$= (y^* - \mu) F_{Y_x}(y^*) - \int_{-\infty}^{y^*} (y - \mu) \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left( - \frac{(y - \mu)^2}{2\sigma^2} \right) dy$$

$$= (y^* - \mu) F_{Y_x}(y^*) - \frac{1}{2} \frac{1}{\sqrt{2\pi}\sigma^2} \int_{\infty}^{(y^* - \mu)^2} \exp \left( - \frac{u}{2\sigma^2} \right) du$$

$$= (y^* - \mu) F_{Y_x}(y^*) + \sigma^2 F'_{Y_x}(y^*) .$$

Mean of $Y_x$: $\mu = \mathbb{E}[Y_x] = \hat{y}_x = k(x)^T K^{-1} y$

Variance of $Y_x$: $\sigma^2 = \mathbb{E}[(Y_x - \hat{y}_x)^2] = K(x, x) - k(x)^T K^{-1} k(x)$
Taking a Deterministic Kernel-based Perspective

Given data \( \{\mathcal{X}, \mathbf{y}\} \), write the kernel interpolant (surrogate) as

\[
s(\mathbf{x}) = \sum_{j=1}^{N} c_j K(\mathbf{x}, \mathbf{x}_j) = \mathbf{k}(\mathbf{x})^T \mathbf{c}, \quad \mathbf{x} \in \Omega \subseteq \mathbb{R}^d
\]

with \( K : \Omega \times \Omega \rightarrow \mathbb{R} \) a positive definite reproducing kernel.
Taking a Deterministic Kernel-based Perspective

Given data \( \{ \mathbf{x}, \mathbf{y} \} \), write the kernel interpolant (surrogate) as

\[
s(\mathbf{x}) = \sum_{j=1}^{N} c_j K(\mathbf{x}, \mathbf{x}_j) = \mathbf{k}(\mathbf{x})^T \mathbf{c}, \quad \mathbf{x} \in \Omega \subseteq \mathbb{R}^d
\]

with \( K : \Omega \times \Omega \rightarrow \mathbb{R} \) a positive definite reproducing kernel.

To find \( c_j \) solve the interpolation equations

\[
s(\mathbf{x}_i) = y_i, \quad i = 1, \ldots, N,
\]

which leads to a linear system \( K \mathbf{c} = \mathbf{y} \) with system matrix

\[
K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \ldots, N,
\]
Taking a Deterministic Kernel-based Perspective

Given data \( \{ \mathcal{X}, \mathbf{y} \} \), write the kernel interpolant (surrogate) as

\[
s(\mathbf{x}) = \sum_{j=1}^{N} c_j K(\mathbf{x}, \mathbf{x}_j) = \mathbf{k}(\mathbf{x})^T \mathbf{c}, \quad \mathbf{x} \in \Omega \subseteq \mathbb{R}^d
\]

with \( K : \Omega \times \Omega \rightarrow \mathbb{R} \) a positive definite reproducing kernel. To find \( c_j \) solve the interpolation equations

\[
s(\mathbf{x}_i) = y_i, \quad i = 1, \ldots, N,
\]

which leads to a linear system \( \mathbf{K} \mathbf{c} = \mathbf{y} \) with system matrix

\[
K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \ldots, N,
\]

so that – as above – the surrogate is given by

\[
s(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{y}.
\]
The kernel interpolant is more accurate (as measured by the pointwise error) than any other linear combination of the data very similar to BLUP. In particular, for any $f \in \mathcal{H}_K(\Omega)$

$$|f(x) - s(x)| \leq P_{K,x}(x)\|f\|_{\mathcal{H}_K(\Omega)}$$

with power function $P_{K,x}(x) = \sqrt{K(x,x) - k(x)^TK^{-1}k(x)}$.

The kernel interpolant provides the best approximation to $f$ from $\mathcal{H}_K(\Omega)$ in the $\| \cdot \|_{\mathcal{H}_K(\Omega)}$ norm.

The kernel interpolant is the minimum $\mathcal{H}_K(\Omega)$-norm interpolant.
Matérn Kernels

Popular in spatial statistics and approximation theory

\[ \kappa(\varepsilon r) = \frac{K_{d/2-\beta}(\varepsilon r)}{(\varepsilon r)^{d/2-\beta}}, \quad \beta > \frac{d}{2} \]

\[ K_\nu: \text{ modified Bessel functions of the second kind of order } \nu \]
Matérn Kernels

Popular in spatial statistics and approximation theory

$$\kappa(\varepsilon r) = \frac{K_{d/2 - \beta}(\varepsilon r)}{\varepsilon r^{d/2 - \beta}}, \quad \beta > \frac{d}{2}$$

$K_\nu$: modified Bessel functions of the second kind of order $\nu$

$$\kappa(\varepsilon r) = (1 + \varepsilon r)e^{-\varepsilon r}, \quad \beta = \frac{5}{2}, \ C^2$$

$$\kappa(\varepsilon r) = (1 + \varepsilon r + \frac{1}{3}(\varepsilon r)^2)e^{-\varepsilon r}, \quad \beta = \frac{7}{2}, \ C^4$$

$$\kappa(\varepsilon r) = (1 + \varepsilon r + \frac{2}{5}(\varepsilon r)^2 + \frac{1}{15}(\varepsilon r)^3)e^{-\varepsilon r}, \quad \beta = \frac{9}{2}, \ C^6$$
A Series Approach to Positive Definite Kernels

Every positive definite kernel $K$ has a Hilbert–Schmidt (or Mercer) series expansion:

$$K(x, z) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(x) \varphi_n(z), \quad x, z \in \Omega \subseteq \mathbb{R}^d,$$

where $(\lambda_n, \varphi_n)$ are $\rho$-orthonormal eigenpairs of a Hilbert–Schmidt integral operator $\mathcal{K} : L_2(\Omega, \rho) \rightarrow L_2(\Omega, \rho)$, i.e.,

$$\mathcal{K} \varphi_n = \lambda_n \varphi_n$$

$$\iff \int_{\Omega} K(x, z) \varphi_n(z) \rho(z) dz = \lambda_n \varphi_n(x), \quad x \in \Omega, \ n = 1, 2, \ldots.$$
Simple Example: Brownian Bridge Kernel

Defined as

\[ K(x, z) = \min(x, z) - xz = \begin{cases} 
    x(1 - z), & 0 \leq x \leq z \leq 1, \\
    z(1 - x), & 0 \leq z \leq x \leq 1.
\end{cases} \]

One can verify that the eigenvalues and normalized eigenfunctions

\[ \lambda_n = (n\pi)^{-2}, \quad \varphi_n(x) = \sqrt{2} \sin(n\pi x), \quad n = 1, 2, \ldots, \]

satisfy \[ K \varphi_n = \lambda_n \varphi_n \text{ (with } \rho(z) \equiv 1), \text{ i.e.,} \]

\[ \int_0^1 (\min(x, z) - xz) \sin(n\pi z) dz = (n\pi)^{-2} \sin(n\pi x), \quad n = 1, 2, \ldots. \]
Simple Example: Brownian Bridge Kernel

Defined as

\[
K(x, z) = \min(x, z) - xz = \begin{cases} 
  x(1 - z), & 0 \leq x \leq z \leq 1, \\
  z(1 - x), & 0 \leq z \leq x \leq 1.
\end{cases}
\]

One can verify that the eigenvalues and normalized eigenfunctions

\[
\lambda_n = (n\pi)^{-2}, \quad \varphi_n(x) = \sqrt{2} \sin(n\pi x), \quad n = 1, 2, \ldots,
\]

satisfy \(K\varphi_n = \lambda_n\varphi_n\) (with \(\rho(z) \equiv 1\)), i.e.,

\[
\int_0^1 (\min(x, z) - xz) \sin(n\pi z) dz = (n\pi)^{-2} \sin(n\pi x), \quad n = 1, 2, \ldots.
\]

The Hilbert-Schmidt expansion is a generalized Fourier series

\[
K(x, z) = \min(x, z) - xz = \sum_{n=1}^{\infty} \frac{2}{n^2\pi^2} \sin(n\pi x) \sin(n\pi z).
\]
Iterated Brownian Bridge Kernels

Green’s kernels of 1-D iterated modified Helmholtz equation

\[
\left( -\frac{d^2}{dx^2} + \varepsilon^2 I \right)^\beta K(x, z) = \delta(x - z), \quad x, z \in [0, 1], \quad \beta \in \mathbb{N}, \quad \varepsilon \geq 0,
\]

with boundary conditions

\[
\frac{d^{2\nu}}{dx^{2\nu}} K(0, z) = \frac{d^{2\nu}}{dx^{2\nu}} K(1, z) = 0, \quad \nu = 0, \ldots, \beta - 1.
\]
Iterated Brownian Bridge Kernels

Green’s kernels of 1-D iterated modified Helmholtz equation

\[
\left(-\frac{d^2}{dx^2} + \varepsilon^2 I\right)^\beta K(x, z) = \delta(x - z), \quad x, z \in [0, 1], \; \beta \in \mathbb{N}, \; \varepsilon \geq 0,
\]

with boundary conditions

\[
\frac{d^{2\nu}}{dx^{2\nu}} K(0, z) = \frac{d^{2\nu}}{dx^{2\nu}} K(1, z) = 0, \quad \nu = 0, \ldots, \beta - 1.
\]

The Hilbert–Schmidt expansion for these kernels is

\[
K(x, z) = \sum_{n=1}^{\infty} \frac{2}{(n^2\pi^2 + \varepsilon^2)^\beta} \sin(n\pi x) \sin(n\pi z),
\]

i.e., the eigenvalues and eigenfunctions are

\[
\lambda_n = \frac{1}{(n^2\pi^2 + \varepsilon^2)^\beta}, \quad \varphi_n(x) = \sqrt{2} \sin(n\pi x).
\]
## Properties of Kernels

<table>
<thead>
<tr>
<th></th>
<th>Matérn</th>
<th>IBB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive definite</td>
<td>on $\mathbb{R}^d$ (when $\beta &gt; \frac{d}{2}$) via Bochner–Schoenberg theory, i.e., radial Fourier transform $F_d \kappa(\omega) = \left(1 + \omega^2\right)^{-\beta} &gt; 0$</td>
<td>on $[0, 1]$ via Hilbert–Schmidt/Mercer expansion</td>
</tr>
<tr>
<td>RKHS</td>
<td>classical Sobolev spaces $H^\beta(\mathbb{R}^d)$</td>
<td>subspaces of $H^\beta([0, 1])$ with appropriate BCs</td>
</tr>
<tr>
<td>Fundamental solution of</td>
<td>$d$-dimensional iterated modified Helmholtz operator, i.e., $D = (-\nabla^2 + \varepsilon^2 I)^\beta$</td>
<td>$1$-dimensional iterated modified Helmholtz operator, i.e., $D = (-\frac{d^2}{dx^2} + \varepsilon^2 I)^\beta$</td>
</tr>
<tr>
<td>Parameters</td>
<td>kernel scale and smoothness</td>
<td>kernel scale and smoothness</td>
</tr>
<tr>
<td>$\varepsilon$ and $\beta$</td>
<td>specify</td>
<td></td>
</tr>
<tr>
<td>$\varepsilon \to 0$ limit of interpolant</td>
<td>polyharmonic spline interpolant</td>
<td>interpolating piecewise polynomial spline of degree $2\beta - 1$ with special BCs</td>
</tr>
</tbody>
</table>

Greg Fasshauer

IBB Kernels in SKO
Two-Parameter IBB Kernels

\[ \beta = 1, \varepsilon = 0 \]

\[ \beta = 1, \varepsilon = 10 \]

Smoothness increases with \( \beta \)
Localization increases with \( \varepsilon \)
Two-Parameter IBB Kernels

$\beta = 2, \varepsilon = 0$

$\beta = 2, \varepsilon = 10$

Smoothness increases with $\beta$

Localization increases with $\varepsilon$
Two-Parameter IBB Kernels

$\beta = 3, \varepsilon = 0$

$\beta = 3, \varepsilon = 10$

**Smoothness** increases with $\beta$

**Localization** increases with $\varepsilon$
Two-Parameter IBB Kernels

\[ \beta = 7, \varepsilon = 0 \]

\[ \beta = 7, \varepsilon = 10 \]

Smoothness increases with \( \beta \)
Localization increases with \( \varepsilon \)
Two-Parameter IBB Kernels

\[ \beta = 20, \varepsilon = 0 \]

\[ \beta = 20, \varepsilon = 50 \]

Smoothness increases with \( \beta \)
Localization increases with \( \varepsilon \)
Comparison: $C^2$ Matérn (dashed) vs. IBB (solid) Kernel

$\beta = 2, \varepsilon = 30$

$\beta = 2, \varepsilon = 50$

Smoothness increases with $\beta$
Localization increases with $\varepsilon$
Comparison: $C^6$ Matérn (dashed) vs. IBB (solid) Kernel

$\beta = 4, \varepsilon = 30$

$\beta = 4, \varepsilon = 50$

Smoothness increases with $\beta$
Localization increases with $\varepsilon$
Numerical Experiments

**Motivation:** Standard SKO tends to do lots of boundary sampling.

**Remedy:** Enforce boundary conditions via IBB kernels.

Reasoning behind our approach

- IBB kernels yield **lower variance at the boundary**.
- Therefore the EI criterion should cause IBB kernels to implicitly prefer searching away from the boundary.

We compare SKO with a **fixed maximum number of points** for

- $C^4$ full-space radial Matérn kernels to
- tensor product of $C^4$ ($\beta = 3$) IBB kernels.

We include a **linear mean** and fit using generalized least squares:

$$s(x) = k(x)^T K^{-1} (y - Pb) + p(x)^T b,$$
$$b = (P^T K^{-1} P)^{-1} P^T K^{-1} y$$
Comparing Kernels in 2D SKO ($f$ satisfies BCs)

The full space Matérn does well, but spends many points sampling around the boundary.

Radial Matérn kernel ($\varepsilon = 5$)
Comparing Kernels in 2D SKO ($f$ satisfies BCs)

The IBB kernel takes advantage of the fact that $f$ satisfies the BCs, resulting in tighter sampling near the minimum.

Iterated Brownian bridge kernel ($\varepsilon = 5$)
Comparing Kernels in 2D SKO (linear BC violation)

$f$ has multiple local minima, and is composed of a BC satisfying function plus a linear term. Many samples around the boundary.

Radial Matérn kernel ($\varepsilon = 8$)
Comparing Kernels in 2D SKO (linear BC violation)

Sampling more concentrated near minimum and about half the variance.

Iterated Brownian bridge kernel ($\varepsilon = 8$)
Comparing Kernels in 2D SKO (no BC structure)

The full-space kernel produces an overall more accurate approximation, but at the cost of sampling near the boundary.

Radial Matérn kernel ($\varepsilon = 6$)
Comparing Kernels in 2D SKO (no BC structure)

Even a function with no BC structure can still be effectively optimized so long as the minimum is not near the boundary.

Iterated Brownian bridge kernel ($\varepsilon = 6$)
Comparing Kernels in 3D SKO ($f$ satisfies BCs)

Similar behavior occurs in higher dimensions ...

Radial Matérn kernel ($\epsilon = 5$)
Comparing Kernels in 3D SKO ($f$ satisfies BCs)

though the results are harder to visualize.

Iterated Brownian Bridge kernel ($\varepsilon = 5$)
Comparing Kernels in 3D SKO ($f$ satisfies BCs)

The impact of boundary sampling is more pronounced in higher dimensions and here the 60 sampled points are clearly distributed differently based on the kernel.
Summary

- Introduced an effective variant of SKO for functions whose minimum occurs away from the boundary.
- Accomplished by using kernels with built-in BCs (such as IBB kernels)
- Future work
  - Identifying and dealing with problems whose minimum does occur near the boundary.
- Other applications of IBB kernels
  - numerical solution of PDEs via (space-time) collocation

MATLAB code available at
http://math.iit.edu/~mccomic/gaussqr
References I


Comparing Kernels in 2D SKO (solution near/on boundary)

Problems may arise when using IIB kernels if the minimum lies on the boundary.

- This movie shows a parametrized version of the first test function, where the minimum approaches the boundary.
- The full space Matérn kernel performs consistently.
- The IBB kernel is unable to construct a surrogate that wants to sample towards the boundary.

Adaptation

Try to identify this condition and modify the polynomial mean to produce a surrogate that is willing to search a boundary when compelled to do so.