Relaxed Gaussian process interpolation: a goal-oriented approach to Bayesian optimization

Sébastien Petit$^{1,2}$
Joint work with Julien Bect$^1$ and Emmanuel Vazquez$^1$

$^1$Université Paris-Saclay, CentraleSupélec, Laboratoire des Signaux et Systèmes
$^2$ Safran Aircraft Engines

GDR Mascot-Num, juin 2022
Outline of the presentation

1. Introduction

2. Building predictive distributions with GPs

3. Relaxed Gaussian processes (reGP)
   - Predictive distributions with interpolation relaxation
   - Application to Bayesian optimization
   - Convergence analysis of reGP

4. Conclusion
1 Introduction
Goal-oriented modeling

• Consider the task of building a prediction of a function

\[ f : X \subset \mathbb{R}^d \rightarrow \mathbb{R} \]

from evaluations at \( x_1, x_2, \ldots \) using a Gaussian process model.
Goal-oriented modeling

• Consider the task of building a prediction of a function

\[ f : X \subset \mathbb{R}^d \rightarrow \mathbb{R} \]

from evaluations at \( x_1, x_2, \ldots \) using a Gaussian process model.

• When such a prediction is used inside a Bayesian optimization algorithm, for example in a minimization problem, it is particularly important to get good predictive distributions on a range of function values corresponding to low values of the function.
The Steep function
The Steep function

Figure: A stationary GP for building predictive distributions
Our proposal: relaxed Gaussian process

Figure: Relax interpolation constraints above $t$!
Building predictive distributions with GPs
Building predictive distributions with GPs

Given data points \((x_i, f(x_i)), i = 1, \ldots, n\), a standard practice to get predictive distributions for \(f\) is the following procedure:

1. Choose a model \(\xi \sim \text{GP}(\mu_\theta, k_\theta)\), with \(\theta \in \Theta\).

2. Select \(\theta\) by maximum likelihood:
   
   \[ L(\theta; Z_n) = -\ln(p(Z_n|\theta)) \]
   
   with \(Z_n = (\xi(x_1), \ldots, \xi(x_n))^T\).

3. Compute the posterior distribution \(\xi|Z_n\).
Building predictive distributions with GPs

Given data points \((x_i, f(x_i)), i = 1, \ldots, n\), a standard practice to get predictive distributions for \(f\) is the following procedure:

1. choose a model \(\xi \sim \text{GP} (\mu_\theta, k_\theta)\), with \(\theta \in \Theta\)
Building predictive distributions with GPs

Given data points \((x_i, f(x_i)), i = 1, \ldots, n\), a standard practice to get predictive distributions for \(f\) is the following procedure:

1. choose a model \(\xi \sim \text{GP}(\mu_\theta, k_\theta)\), with \(\theta \in \Theta\)

2. select \(\theta\) by maximum likelihood

\[
\mathcal{L} (\theta; Z_n) = - \ln (p(Z_n | \theta))
\]

with \(Z_n = (\xi(x_1), \ldots, \xi(x_n))^T\)
Building predictive distributions with GPs

Given data points \((x_i, f(x_i)), i = 1, \ldots, n\), a standard practice to get predictive distributions for \(f\) is the following procedure:

1. choose a model \(\xi \sim \text{GP}(\mu_\theta, k_\theta)\), with \(\theta \in \Theta\)
2. select \(\theta\) by maximum likelihood

\[
\mathcal{L}(\theta; \underline{Z}_n) = -\ln(p(\underline{Z}_n | \theta))
\]

with \(\underline{Z}_n = (\xi(x_1), \ldots, \xi(x_n))^T\)

3. compute the posterior distribution \(\xi | \underline{Z}_n\)
3 Relaxed Gaussian processes (reGP)

Predictive distributions with interpolation relaxation

Application to Bayesian optimization

Convergence analysis of reGP
Objective

- Consider $n$ points $x_n = (x_1, \ldots, x_n)$ in $\mathbb{X}$ and let $z_n = (f(x_1), \ldots, f(x_n))$ be the vector of the corresponding values of $f$. 
Objective

- Consider $n$ points $x_n = (x_1, \ldots, x_n)$ in $\mathbb{X}$ and let $z_n = (f(x_1), \ldots, f(x_n))$ be the vector of the corresponding values of $f$.
- Consider a family $\xi \sim \text{GP}(\mu_\theta, k_\theta)$, with $\theta \in \Theta$. 
Objective

- Consider $n$ points $x_n = (x_1, \ldots, x_n)$ in $\mathbb{X}$ and let $z_n = (f(x_1), \ldots, f(x_n))$ be the vector of the corresponding values of $f$.
- Consider a family $\xi \sim \text{GP}(\mu_\theta, k_\theta)$, with $\theta \in \Theta$.
- Our objective is to obtain (good) predictive distributions of $f$ below a threshold $t^{(0)}$, on the range $Q = (-\infty, t^{(0)})$. We accept degraded predictions above $t^{(0)}$.

Goal-oriented modeling
Overview

Given $x_n, z_n, t^{(0)}$, and a parametrized GP model $\xi \sim \text{GP}(\mu_\theta, k_\theta)$:

- Select $t$ automatically above $t^{(0)}$
- Choose $\widehat{\theta}_n \in \Theta$ and modified observations $z_n^* \in \mathbb{R}^n$
- reGP: $\xi \mid Z_n = z_n^*$
reGP predictive distribution given $t$

- Suppose that $\xi \sim \text{GP} (0, k)$ with a **fixed $k$ for simplicity**. Let $\mu_n$ be the posterior mean of $\xi$. We have (Kimeldorf & Wahba 1970)

$$
\mu_n = \arg\min \left\{ h \in \mathcal{H}(X) \bigg| \begin{array}{l}
\| h \|_{\mathcal{H}(X)} \\
h(x_n) = z_n
\end{array} \right\}
$$

with $\mathcal{H}(X)$ the RKHS attached to $k$. 


reGP predictive distribution given $t$

- Suppose that $\xi \sim \text{GP}(0, k)$ with a fixed $k$ for simplicity. Let $\mu_n$ be the posterior mean of $\xi$. We have (Kimeldorf & Wahba 1970)

$$\mu_n = \arg\min_{h \in \mathcal{H}(X)} \|h\|_{\mathcal{H}(X)}.$$ 

with $\mathcal{H}(X)$ the RKHS attached to $k$.

- The core idea is to build a predictive distribution with a mean given by the relaxed interpolator:

$$\tilde{\mu}_n = \arg\min_{h \in \mathcal{H}(X)} \|h\|_{\mathcal{H}(X)},$$ 

with $x_n = (x_{n,0}, x_{n,1})$ and $z_n = (z_{n,0}, z_{n,1})$, such that $z_{n,0} < t$ and $z_{n,1} \geq t$ wlog.
**reGP predictive distribution given** $t$

Recall that $z_n = (z_{n,0}, z_{n,1})$, where $z_{n,0} < t$ and $z_{n,1} \geq t$.

**Definition**

The predictive distribution reGP is defined as the conditional distribution $P^n_t$ of $\xi$ given

$$
\begin{cases}
\xi(x_{n,0}) = z_{n,0} \\
\xi(x_{n,1}) = z_{n,1}^*
\end{cases}
$$

where $z_{n,1}^*$ is the solution of the extended negative log likelihood

$$
(\hat{\theta}_n, z_{n,1}^*) = \underset{\theta \in \Theta, z_{n,1} \geq t}{\text{argmin}} \mathcal{L}(\theta; z_{n,0}, z_{n,1}).
$$
reGP predictive distribution given \( t \)

Recall that \( z_n = (z_{n, 0}, z_{n, 1}) \), where \( z_{n, 0} < t \) and \( z_{n, 1} \geq t \).

**Definition**

The predictive distribution reGP is defined as the conditional distribution \( P^t_n \) of \( \xi \) given

\[
\begin{align*}
\xi(x_{n, 0}) &= z_{n, 0} \\
\xi(x_{n, 1}) &= z^*_n, 1
\end{align*}
\]

where \( z^*_n, 1 \) is the solution of the extended negative log likelihood

\[
\left( \hat{\theta}_n, z^*_n, 1 \right) = \operatorname{argmin}_{\theta \in \Theta, \ z_{n, 1} \geq t} \mathcal{L} \left( \theta; z_{n, 0}, z_{n, 1} \right).
\]

For a given \( \theta \), the mean of the reGP predictive distribution is given by the relaxed interpolator \( \tilde{\mu}_n \).
• In more details, let $\mu_\theta = (\mu_\theta(x_1), \ldots, \mu_\theta(x_n))^T$, $K_\theta = (k_\theta(x_i, x_j))_{i,j}$, and write $z_n = \left( z_{n,0}^T, z_{n,1}^T \right)^T$, then

$$L(\theta; z_{n,0}, z_{n,1}) \propto \log(\det(K_\theta)) + (z_n - \mu_\theta)^T K_\theta^{-1}(z_n - \mu_\theta) + \text{constant},$$

where the quadratic form is with respect to $(z_n, 1)$.

• The likelihood can be optimized jointly w.r.t. $\theta$ and $z_{n,1}$ with an L-BFGS-B algorithm (Byrd et al., 1995), for instance.
• In more details, let \( \mu_\theta = (\mu_\theta(x_1), \ldots, \mu_\theta(x_n))^T \),
\[ K_\theta = (k_\theta(x_i, x_j))_{i,j}, \]
and write \( z_n = (z_{n,0}^T, z_{n,1}^T)^T \), then
\[
\mathcal{L} \left( \theta; z_n, 0, z_n, 1 \right) \propto \log \left( \det \left( K_\theta \right) \right) + (z_n - \mu_\theta)^T K_\theta^{-1} (z_n - \mu_\theta) + \text{constant},
\]
quadratic form w.r.t. \((z_n, 1)\)

• The likelihood can be optimized jointly w.r.t. \( \theta \) and \( z_n, 1 \) with an L-BFGS-B algorithm (Byrd et al., 1995), for instance.

• Moreover, note that,
\[
\mathcal{L} \left( \theta; z_n, 0, z_n, 1 \right) = -\ln \left( p \left( z_n, 0 \mid \theta \right) \right) - \ln \left( p \left( z_n, 1 \mid \theta, z_n, 0 \right) \right),
\]
likelihood under \( t \) imputation term
Overview

Given $x_n$, $z_n$, $t^{(0)}$, and a parametrized GP model $\xi \sim \text{GP}(\mu_\theta, k_\theta)$:

- **Select $t$ automatically above $t^{(0)}$**
- **Choose $\hat{\theta}_n \in \Theta$ and modified observations $z_n^* \in \mathbb{R}^n$**
- **reGP: $\xi \mid Z_n = z_n^*$**
Choosing $t \geq t^{(0)}$

Which data are useful for prediction below $t^{(0)}$?
Choosing $t \geq t^{(0)}$

Which data are useful for prediction below $t^{(0)}$?

$\longrightarrow$ select a good value of $t$ between:

- $t = +\infty$ yielding a standard GP,
- $t = t^{(0)}$, when only the data within $Q = (-\infty, t^{(0)})$ help for prediction
Choosing $t \geq t^{(0)}$

Which data are useful for prediction below $t^{(0)}$?
→ select a good value of $t$ between:

- $t = +\infty$ yielding a standard GP,
- $t = t^{(0)}$, when only the data within $Q = (-\infty, t^{(0)})$ help for prediction

We consider a **finite set of candidates** $t^{(0)} < t^{(1)} < \cdots < t^{(K)} = +\infty$

$$Q = \begin{array}{cccccc}
\cdots \\
-\infty & t^{(0)} & t^{(1)} & t^{(2)} & \cdots & t^{(K)} = +\infty
\end{array}$$
A goal-oriented goodness-of-fit criterion?

• How to select $t$ among $t^{(0)} < t^{(1)} < \cdots < t^{(K)} = +\infty$?
A goal-oriented goodness-of-fit criterion?

- How to select \( t \) among \( t^{(0)} < t^{(1)} < \cdots < t^{(K)} = +\infty \)?

- Recall the standard leave-one-out cross-validation (Currin et al. 1988)

\[
\frac{1}{n} \sum_{i=1}^{n} S \left( P_{n,-i}; z_i \right)
\]

where

\( P_{n,-i} \) is the loo predictive distribution at \( x_i \)

and \( S \) is a scoring rule (see, e.g., Gneiting & Raftery 2007):

\( S(P; z) \) represents a loss when using the distribution \( P \) to predict \( z \)
A goal-oriented goodness-of-fit criterion?

• How to select t among $t^{(0)} < t^{(1)} < \cdots < t^{(K)} = +\infty$?

• Recall the standard leave-one-out cross-validation (Currin et al. 1988)

$$\frac{1}{n} \sum_{i=1}^{n} S(P_{n,-i}; z_i)$$

where

→ $P_{n,-i}$ is the loo predictive distribution at $x_i$

→ and $S$ is a scoring rule (see, e.g., Gneiting & Raftery 2007): $S(P; z)$ represents a loss when using the distribution $P$ to predict $z$

• A scoring rule focusing on $Q = (-\infty, t^{(0)})$?
Truncated continuous ranked probability score

Recall the continuous ranked probability score \((\text{Matheson and Winkler 1976; Hersbach 2000; Gneiting 2004})\):

\[
S^{\text{CRPS}}(P, z) = \int_{-\infty}^{+\infty} (F_P(u) - \mathbb{1}_{z \leq u})^2 \, d\mu,
\]

• If \(z < t(0)\) → the \(t\)CRPS asks that \(P \approx \delta_z\)
• If \(z \geq t(0)\) → the value of \(S_{t\text{CRPS}}(0)\) does not depend on the specific value \(z\), and decreases when \(P\) is concentrated above \(t(0)\)
• We give closed-form expressions for \(S_{t\text{CRPS}}(0)\) when \(P\) is Gaussian
Truncated continuous ranked probability score

Recall the continuous ranked probability score (Matheson and Winkler 1976; Hersbach 2000; Gneiting 2004):

$$S_{\text{CRPS}}(P, z) = \int_{-\infty}^{+\infty} (F_P(u) - 1_{z \leq u})^2 \, du,$$

We propose the truncated continuous ranked probability score:

$$S^{t_{\text{CRPS}}}_{t^{(0)}}(P; z) = \int_{-\infty}^{t^{(0)}} (F_P(u) - 1_{z \leq u})^2 \, du.$$

- If $z < t^{(0)} \rightarrow$, the tCRPS asks that $P \simeq \delta_z$
- If $z \geq t^{(0)} \rightarrow$ the value of $S^{t_{\text{CRPS}}}_{t^{(0)}}(P; z)$ does not depend on the specific value $z$, and decreases when $P$ is concentrated above $t^{(0)}$
- We give closed-form expressions for $S^{t_{\text{CRPS}}}_{t^{(0)}}(P; z)$ when $P$ is Gaussian
The cross-validation criterion to select $t$ using the tCRPS scoring rule is called the LOO-tCRPS
Application to Bayesian optimization
Efficient Global Optimization (EGO, Jones et al. 1998)

- Given $f : \mathbb{X} \rightarrow \mathbb{R}$, consider the minimization problem

$$m = \min_{x \in \mathbb{X}} f$$

- construct a sequence of evaluations points $x_1, x_2, \ldots$, such that for $n > 0$, $m_n = \min f(x_1), \ldots, f(x_n)$ is close to $m$
Efficient Global Optimization (EGO, Jones et al. 1998)

- Given $f : \mathbb{X} \rightarrow \mathbb{R}$, consider the minimization problem

$$m = \min_{x \in \mathbb{X}} f$$

- construct a sequence of evaluations points $x_1, x_2, \ldots$, such that for $n > 0$, $m_n = \min f(x_1), \ldots, f(x_n)$ is close to $m$

- Standard Bayesian approach $\Rightarrow$ build predictive distributions for $f$ using a GP $\xi$, and use the strategy

$$x_{n+1} = \arg\max_{x \in \mathbb{X}} \rho_n(x)$$

where $\rho_n$ is the expected improvement (Mockus $\sim$ 1970s)

$$\rho_n(x) = \mathbb{E}((m_n - m_{n+1})_+ \mid \underline{Z}_n = \underline{z}_n)$$

with $\underline{Z}_n = (\xi(x_1), \ldots, \xi(x_n))$
Efficient global optimization with relaxation (EGO-R)

At each step $n$, construct a reGP model

$$t_n^{(0)} \rightarrow t_n \rightarrow \hat{\theta}_n; Z^*_n \rightarrow \text{reGP}$$

and choose $x_{n+1}$ using the expected improvement

$$\rho_n(x) = \mathbb{E} \left( (m_n - m_{n+1})_+ \mid Z_n = Z^*_n \right),$$

sequentially under the reGP distribution.
Efficient global optimization with relaxation (EGO-R)

At each step $n$, construct a reGP model

and choose $x_{n+1}$ using the expected improvement

$$
\rho_n(x) = \mathbb{E} ((m_n - m_{n+1})_+ \mid Z_n = Z_n^*)
$$

sequentially under the reGP distribution.

Several strategies for determining a validation threshold $t_n^{(0)}$ may be considered (e.g., set a fixed threshold $t_n^{(0)} = t^{(0)}$ from an initial DoE)
EGO-R on a “difficult” function: Goldstein-Price

Figure: average number of iterations to reach a target on the horizontal axis (top) and fraction of repetitions reaching the target.
Relaxation on Goldstein-Price

Figure: relaxed observations \((z_n, 0, z_n^*, 1)\) versus \(z_n\).
EGO-R on an “easy” function: log of Golstein-Price

Figure: average number of iterations to reach a target (top) and fraction of repetitions reaching the target (bottom)
Convergence analysis of reGP
Setup

• Consider $\xi \sim \text{GP}(0, k)$, where $k$ is a fixed covariance with finite smoothness (e.g., a Matérn covariance function with regularity $0 < \nu < \infty$).

• Denote by $\mathcal{H}(X)$ the reproducing kernel Hilbert space (RKHS) attached to $k$. 
Convergence of EGO-R

Prop.

Let $f \in \mathcal{H}(X)$ and consider the EGO-R algorithm for building

$(x_n)_{n \geq 1} \in X^\mathbb{N}$, with

$$t_n^{(0)} > \min(f(x_1), \ldots, f(x_n)),$$

\forall n \geq 1.

Then, the sequence $(x_n)_{n \geq 1}$ is dense in $X$. 
Known results about GPs

- Denote by $\mu_{n,f}$ the posterior distribution of $\xi$ given the data
- If $(x_n)_{n \geq 1}$ is dense, then $\mu_{n,f} \to f$ in $\mathcal{H}(X)$
Known results about GPs

- Denote by $\mu_{n,f}$ the posterior distribution of $\xi$ given the data.
- If $(x_n)_{n \geq 1}$ is dense, then $\mu_{n,f} \to f$ in $H(X)$.
- If $k$ has smoothness $\nu > 0$ and $X$ is “nice”, then

$$
\|f - \mu_{n,f}\|_{L^\infty(X)} \lesssim h_n^\nu \|f\|_{H(X)}, \quad h_n = \sup_{x \in X} \min_{1 \leq i \leq n} \|x - x_i\|
$$

(Arcangeli et al. 2007)
Convergence of reGP

Let \( t \) be a fixed relaxation threshold and \( \mathcal{H}(t, f) \) be the space of functions \( g \in \mathcal{H}(X) \) such that, for all \( x \in X \)

\[
\begin{cases}
  g(x) \geq t & \text{if } f(x) \geq t, \\
  g(x) = f(x) & \text{otherwise}.
\end{cases}
\]

(3)

Figure: \( \mathcal{H}(t, f) \)
Convergence of reGP

Let $\tilde{\mu}_{n,f} : X \to \mathbb{R}$ be the mean of the reGP predictive distribution.

Prop.

Assume $(x_n)_{n \geq 1}$ is dense. Then, the sequence $(\tilde{\mu}_{n,f})_{n \geq 1}$ converges to the unique minimum norm element $s_{t,X}$ of $\mathcal{H}(t,f)$.

Figure: $\mathcal{H}(t,f)$
Convergence analysis of reGP

Let $X_0 = \{ x \in X, f(x) < t \}$. For every $g \in \mathcal{H}(t, f)$, standard GP interpolation on $g$ yields

$$\| f - \mu_{n,g} \|_{L^\infty(X_0)} \lesssim h_n^\nu \| g \|_{\mathcal{H}(X)}.$$
Convergence analysis of reGP

Let $X_0 = \{ x \in X, f(x) < t \}$. For every $g \in \mathcal{H}(t, f)$, standard GP interpolation on $g$ yields

$$\|f - \mu_{n,g}\|_{L^\infty(X_0)} \lesssim h_n^\nu \|g\|_{\mathcal{H}(X)}.$$

Observe that $s_{t,X}$ optimizes the bound.
Convergence analysis of reGP

Let $X_0 = \{ x \in X, f(x) < t \}$. For every $g \in \mathcal{H}(t, f)$, standard GP interpolation on $g$ yields

$$\| f - \mu_{n,g} \|_{L^\infty(X_0)} \lesssim h_n^\nu \| g \|_{\mathcal{H}(X)}.$$

Observe that $s_{t,X}$ optimizes the bound.

**Prop.**

For $n \geq 1$ and a “nice” $B \subset X_0$:

$$\| f - \tilde{\mu}_{n,f} \|_{L^\infty(B)} \lesssim h_n^\nu \| s_{t,X} \|_{\mathcal{H}(X)}.$$
Convergence analysis of reGP

Let $X_0 = \{ x \in X, f(x) < t \}$. For every $g \in \mathcal{H}(t, f)$, standard GP interpolation on $g$ yields

$$\| f - \mu_{n,g} \|_{L^\infty(X_0)} \lesssim h_n \| g \|_{\mathcal{H}(X)}.$$

Observe that $s_{t,X}$ optimizes the bound.

Prop.

For $n \geq 1$ and a “nice” $B \subset X_0$:

$$\| f - \tilde{\mu}_{n,f} \|_{L^\infty(B)} \lesssim h_n \| s_{t,X} \|_{\mathcal{H}(X)}.$$

We also have a weaker kind of guarantee outside $X_0$.
Convergence analysis of reGP

What about $\|f\|_{\mathcal{H}(X)}/\|s_{t,X}\|_{\mathcal{H}(X)}$?
Convergence analysis of reGP

What about $\|f\|_{\mathcal{H}(X)}/\|s_{t,X}\|_{\mathcal{H}(X)}$?

**Prop.**

If $\max f > t$ and $k$ has finite smoothness $0 < \nu < \infty$, then

$$\sup_{g \in \mathcal{H}(t,f)} \|g\|_{\mathcal{H}(X)} = \infty.$$
Conclusion & perspectives

- **reGP →** a goal-oriented GP-based model to good obtain predictive distributions on a range of function values (if we accept degraded predictions outside the range of interest)
- Easy to use
- Low / moderate algorithmic complexity
- EGO-R seems preferable to EGO and can sometimes achieve significant convergence acceleration
- Main future work: extend reGP to the case of regression