Bayesian optimization

Dirk Husmeier
School of Mathematics & Statistics
University of Glasgow
How can we make parameter estimation of complex systems computationally viable?
How can we make parameter estimation of complex systems computationally viable?

A Review of Bayesian Optimization
Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams and Nando de Freitas

Illustrations from Carl Rasmussen et al., Neurocomputing 80, 2012
Unknown true log likelihood
Unknown true log likelihood

Not viable due to excessive computational costs
Unknown true log likelihood
Unknown true log likelihood
$\mathbb{E}[\ell(\theta)]$
\( J(\theta) := \mathbb{E}[\ell(\theta)] + K \sqrt{\text{Var}[\ell(\theta)]} \)
Maximum of \[ J(\theta) := \mathbb{E}[\ell(\theta)] + K \sqrt{\text{Var}[\ell(\theta)]} \]

\[ \mathbb{E}[\ell(\theta)] \]
Maximum of

$$J(\theta) := \mathbb{E}[\ell(\theta)] + K\sqrt{\text{Var}[\ell(\theta)]}$$

Trade-off between exploitation and exploration
Re-evaluation at the maximum
Re-evaluation at the maximum
Re-evaluation at the maximum
Maximum
Re-evaluation at the maximum
Continue until ....
For **minimization** (e.g. residual sum of squares) instead of **maximization** (e.g. maximum likelihood) it work the same, so let’s illustrate it again ...

The following illustrations are taken from a lecture given by Marc Deisenroth
Bayesian Optimization

Objective (Bayesian Optimization)
Minimize an objective function $g$, which is very expensive to evaluate.

Key Idea:
1. Build a model $\tilde{g}$ of the true objective function $g$
2. Find $\theta^* \in \arg \min_{\theta} \tilde{g}(\theta)$
3. Evaluate true objective $g$ at $\theta^*$
4. Update model $\tilde{g}$
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \ E[\tilde{g}(\theta)] - 2\sqrt{\text{V}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]} \]
Upper-Confidence-Bound (UCB) criterion to select next point

$$\theta^* \in \text{arg min}_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]}$$
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \ E[\tilde{g}(\theta)] - 2 \sqrt{V[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[
\theta^* \in \arg \min_{\theta} \ E[\tilde{g}(\theta)] - 2\sqrt{V[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB) criterion to select next point**

\[
\theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[ \theta^* \in \arg\min_\theta \quad E[\tilde{g}(\theta)] - 2\sqrt{V[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

$$\theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]}$$
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB) criterion to select next point**

\[ \theta^* \in \arg \min_\theta \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[
\theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{V}[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB) criterion to select next point**

\[
\theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[
\theta^* \in \arg \min_{\theta} \ E[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[
\theta^* \in \arg \min_\theta \quad E[\tilde{g}(\theta)] - 2\sqrt{V[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[ \theta^* \in \operatorname{arg\,min}_\theta \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[
\theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB)** criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\mathbb{V}[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- **Upper-Confidence-Bound (UCB) criterion to select next point**

\[
\theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2\sqrt{\text{Var}[\tilde{g}(\theta)]}
\]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[ \theta^* \in \arg \min_{\theta} E[\tilde{g}(\theta)] - 2\sqrt{V[\tilde{g}(\theta)]} \]
Bayesian Optimization: Illustration

- Upper-Confidence-Bound (UCB) criterion to select next point

\[ \theta^* \in \arg \min_{\theta} \quad \mathbb{E}[\tilde{g}(\theta)] - 2 \sqrt{\text{V}[\tilde{g}(\theta)]} \]

- Global minimum found after 10 function evaluations
Are there better methods than the Upper Confidence Bound (UCB) criterion?
Are there better methods than the Upper Confidence Bound (UCB) criterion?

Improvement-based methods

Information-based methods
Are there better methods than the Upper Confidence Bound (UCB) criterion?

Improvement-based methods

Information-based methods
We introduce the following random variable to indicate the *improvement* over the incumbent minimum:

\[ I(x) = \max\{f_{\min} - f(x), 0\} \]
The *Probability of Improvement* (PI) is the probability of the event \( \{ I(x) > 0 \} \)

\[
\text{PI}(x) = \mathbb{P}\{ I(x) > 0 \}
\]
The *Probability of Improvement* (PI) is the probability of the event \( \{ I(x) > 0 \} \) or, equivalently, \( \{ f(x) < f_{\text{min}} \} \):

\[
\text{PI}(x) = \mathbb{P}\{ I(x) > 0 \}
= \mathbb{E}\{ f(x) < f_{\text{min}} \}
= \mathbb{P}\{ f(x) < f_{\text{min}} \}
\]
The Probability of Improvement (PI) is the probability of the event \( \{I(x) > 0\} \) or, equivalently, \( \{f(x) < f_{\text{min}}\} \):

\[
PI(x) = \mathbb{P}\{I(x) > 0\} = \mathbb{E}1\{f(x) < f_{\text{min}}\} = \mathbb{P}\{f(x) < f_{\text{min}}\} = \mathbb{P}\left\{ \frac{f(x) - \hat{f}(x)}{s(x)} < \frac{f_{\text{min}} - \hat{f}(x)}{s(x)} \right\}
\]
The Probability of Improvement (PI) is the probability of the event \( \{ I(x) > 0 \} \) or, equivalently, \( \{ f(x) < f_{\text{min}} \} \):

\[
\text{PI}(x) = \mathbb{P}\{ I(x) > 0 \} \\
= \mathbb{E}1 \{ f(x) < f_{\text{min}} \} \\
= \mathbb{P}\{ f(x) < f_{\text{min}} \} \\
= \mathbb{P}\left\{ \frac{f(x) - \hat{f}(x)}{s(x)} < \frac{f_{\text{min}} - \hat{f}(x)}{s(x)} \right\} \\
= \int_{-\infty}^{\frac{f_{\text{min}} - \hat{f}(x)}{s(x)}} \phi(z) \, dz
\]

Normal pdf
The Probability of Improvement (PI) is the probability of the event \( \{ I(x) > 0 \} \) or, equivalently, \( \{ f(x) < f_{\min} \} \):

\[
\text{PI}(x) = \mathbb{P}\{I(x) > 0\} \\
= \mathbb{P}\{f(x) < f_{\min}\} \\
= \mathbb{P}\left\{ \frac{f(x) - \hat{f}(x)}{s(x)} < \frac{f_{\min} - \hat{f}(x)}{s(x)} \right\} \\
= \int_{-\infty}^{\frac{f_{\min} - \hat{f}(x)}{s(x)}} \phi(z) \, dz \\
= \Phi\left( \frac{f_{\min} - \hat{f}(x)}{s(x)} \right)
\]

Normal pdf

Normal cdf
The **Probability of Improvement (PI)** is the probability of the event \( \{ I(x) > 0 \} \) or, equivalently, \( \{ f(x) < f_{\text{min}} \} \):

\[
\text{PI}(x) = \mathbb{P}\{ I(x) > 0 \} = \mathbb{E} \{ f(x) < f_{\text{min}} \} = \mathbb{P}\{ f(x) < f_{\text{min}} \}
\]
The *Probability of Improvement* (PI) is the probability of the event $\{I(x) > 0\}$ or, equivalently, $\{f(x) < f_{\text{min}}\}$:

$$
\text{PI}(x) = \mathbb{P}\{I(x) > 0\} \\
= \mathbb{E}\{f(x) < f_{\text{min}}\} \\
= \mathbb{P}\{f(x) < f_{\text{min}}\}
$$
The **Probability of Improvement (PI)** is the probability of the event \( \{ I(\mathbf{x}) > 0 \} \) or, equivalently, \( \{ f(\mathbf{x}) < f_{\text{min}} \} \):

\[
\text{PI}(\mathbf{x}) = \mathbb{P}\{ I(\mathbf{x}) > 0 \} = \mathbb{E}\{ f(\mathbf{x}) < f_{\text{min}} \} = \mathbb{P}\{ f(\mathbf{x}) < f_{\text{min}} \}
\]

Better alternative: expectation of the quantity itself rather than the indicator function!
The **Probability of Improvement (PI)** is the probability of the event \( \{ I(x) > 0 \} \) or, equivalently, \( \{ f(x) < f_{\min} \} \):

\[
\text{PI}(x) = \mathbb{P}\{ I(x) > 0 \} = \mathbb{E} \mathbf{1}_{\{ f(x) < f_{\min} \}} = \mathbb{P}\{ f(x) < f_{\min} \}
\]

The expected value of the random variable \( I(x) \) is the **Expected Improvement (EI)** acquisition function:

\[
\text{EI}(x) = \mathbb{E}[I(x)]
\]
The **Probability of Improvement (PI)** is the probability of the event \( \{I(x) > 0\} \) or, equivalently, \( \{f(x) < f_{\text{min}}\} \):

\[
\text{PI}(x) = \mathbb{P}\{I(x) > 0\} \\
= \mathbb{E}[I(x) < f_{\text{min}}] \\
= \mathbb{P}\{f(x) < f_{\text{min}}\}
\]

The expected value of the random variable \( I(x) \) is the **Expected Improvement (EI)** acquisition function:

\[
\text{EI}(x) = \mathbb{E}[I(x)]
\]

The derivation of this expression is slightly more involved ...
Let $\phi(z) = (\sqrt{2\pi})^{-1} \exp(-z^2/2)$ be the standard Gaussian pdf.
Let $\phi(z) = (\sqrt{2\pi})^{-1} \exp(-z^2/2)$ be the standard Gaussian pdf. Then,

$$\phi'(z) = \frac{d}{dz} \phi(z)$$
Let $\phi(z) = (\sqrt{2\pi})^{-1} \exp(-z^2/2)$ be the standard Gaussian pdf. Then,

$$\phi'(z) = \frac{d}{dz} \phi(z) = \phi(z) \times \left(-\frac{1}{2} \times 2z\right)$$
Let $\phi(z) = \left(\sqrt{2\pi}\right)^{-1} \exp(-z^2/2)$ be the standard Gaussian pdf. Then,

$$\phi'(z) = \frac{d}{dz} \phi(z) = \phi(z) \times \left( -\frac{1}{2} \times 2z \right) = -z\phi(z)$$
$EI(x) = \mathbb{E}[I(x)]$

Define $u = \left\{ f_{\text{min}} - \hat{f}(x) \right\} / s(x)$
\[ EI(x) = \mathbb{E}[I(x)] \]
\[ = \mathbb{E}[\max\{f_{\text{min}} - f(x), 0\}] \]

Define \( u = \frac{f_{\text{min}} - \hat{f}(x)}{s(x)} \)
\[ EI(x) = \mathbb{E}[I(x)] \]
\[ = \mathbb{E}[\max\{f_{\text{min}} - f(x), 0\}] \]
\[ = \mathbb{E} \left[ \{f_{\text{min}} - f(x)\} 1\{f(x) < f_{\text{min}}\} \right] \]

Define \( u = \{f_{\text{min}} - \hat{f}(x)\}/s(x) \)
\[ \text{EI}(x) = \mathbb{E}[I(x)] \]
\[ = \mathbb{E}[\max\{f_{\text{min}} - f(x), 0\}] \]
\[ = \mathbb{E} \left[ (f_{\text{min}} - f(x)) \mathbf{1}\{f(x) < f_{\text{min}}\} \right] \]
\[ = \int_{-\infty}^{\infty} (f_{\text{min}} - y) \mathbf{1}\{y < f_{\text{min}}\} \phi(y \mid \hat{f}(x), s^2(x)) \, dy \]

Define \( u = \frac{f_{\text{min}} - \hat{f}(x)}{s(x)} \)

where \( \phi(x \mid \mu, \sigma^2) \) and \( \Phi(x \mid \mu, \sigma^2) \) represent the pdf and cdf of a \( \mathcal{N}(\mu, \sigma^2) \) distribution evaluated at \( x \) respectively. When \( \mu = 0 \) and \( \sigma^2 = 1 \) we will simply write \( \phi(x) \) and \( \Phi(x) \) for brevity.
$EI(\mathbf{x}) = \mathbb{E}[I(\mathbf{x})]$

$$= \mathbb{E}[\max\{f_{\text{min}} - f(\mathbf{x}), 0\}]$$

$$= \mathbb{E} \left[ \{f_{\text{min}} - f(\mathbf{x})\} 1\{f(\mathbf{x}) < f_{\text{min}}\} \right]$$

$$= \int_{-\infty}^{\infty} \{f_{\text{min}} - y\} 1\{y < f_{\text{min}}\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy$$

$$= \int_{-\infty}^{f_{\text{min}}} \{f_{\text{min}} - y\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy$$

Define $u = \{f_{\text{min}} - \hat{f}(\mathbf{x})\} / s(\mathbf{x})$

where $\phi(x \mid \mu, \sigma^2)$ and $\Phi(x \mid \mu, \sigma^2)$ represent the pdf and cdf of a $\mathcal{N}(\mu, \sigma^2)$ distribution evaluated at $x$ respectively. When $\mu = 0$ and $\sigma^2 = 1$ we will simply write $\phi(x)$ and $\Phi(x)$ for brevity.
Define $u = \{f_{\text{min}} - \hat{f}(\mathbf{x})\}/s(\mathbf{x})$

$$EI(\mathbf{x}) = \mathbb{E}[I(\mathbf{x})]$$
$$= \mathbb{E}\left[\max\{f_{\text{min}} - f(\mathbf{x}), 0\}\right]$$
$$= \mathbb{E}\left[\{f_{\text{min}} - f(\mathbf{x})\}1\{f(\mathbf{x}) < f_{\text{min}}\}\right]$$
$$= \int_{-\infty}^{\infty} \{f_{\text{min}} - y\}1\{y < f_{\text{min}}\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy$$
$$= \int_{-\infty}^{f_{\text{min}}} \{f_{\text{min}} - y\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy$$
$$= \int_{-\infty}^{u} \{f_{\text{min}} - (\hat{f}(\mathbf{x}) + s(\mathbf{x})z)\} \phi(z) \, dz$$

where $\phi(x \mid \mu, \sigma^2)$ and $\Phi(x \mid \mu, \sigma^2)$ represent the pdf and cdf of a $\mathcal{N}(\mu, \sigma^2)$ distribution evaluated at $x$ respectively. When $\mu = 0$ and $\sigma^2 = 1$ we will simply write $\phi(x)$ and $\Phi(x)$ for brevity.
Define \( u = \{ f_{\text{min}} - \hat{f}(\mathbf{x}) \} / s(\mathbf{x}) \)

\[
\text{EI}(\mathbf{x}) = \mathbb{E}[I(\mathbf{x})] \\
= \mathbb{E}[\max\{f_{\text{min}} - f(\mathbf{x}), 0\}] \\
= \mathbb{E} \left[ \{ f_{\text{min}} - f(\mathbf{x}) \} \mathbf{1}_{\{f(\mathbf{x}) < f_{\text{min}}\}} \right] \\
= \int_{-\infty}^{\infty} \{ f_{\text{min}} - y \} \mathbf{1}_{\{y < f_{\text{min}}\}} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy \\
= \int_{-\infty}^{f_{\text{min}}} \{ f_{\text{min}} - y \} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy \\
= \int_{-\infty}^{u} \{ f_{\text{min}} - (\hat{f}(\mathbf{x}) + s(\mathbf{x})z) \} \phi(z) \, dz \\
= \int_{-\infty}^{u} \{ f_{\text{min}} - \hat{f}(\mathbf{x}) - s(\mathbf{x})z \} \phi(z) \, dz
\]

where \( \phi(x \mid \mu, \sigma^2) \) and \( \Phi(x \mid \mu, \sigma^2) \) represent the pdf and cdf of a \( \mathcal{N}(\mu, \sigma^2) \) distribution evaluated at \( x \) respectively. When \( \mu = 0 \) and \( \sigma^2 = 1 \) we will simply write \( \phi(x) \) and \( \Phi(x) \) for brevity.
\[ EI(x) = \mathbb{E}[I(x)] \]
\[ = \mathbb{E}[\max\{f_{\min} - f(x), 0\}] \]
\[ = \mathbb{E}\left[ \{f_{\min} - f(x)\}1\{f(x) < f_{\min}\} \right] \]
\[ = \int_{-\infty}^{\infty} \{f_{\min} - y\}1\{y < f_{\min}\} \phi(y | \hat{f}(x), s^2(x)) \, dy \]
\[ = \int_{-\infty}^{f_{\min}} \{f_{\min} - y\} \phi(y | \hat{f}(x), s^2(x)) \, dy \]
\[ = \int_{-\infty}^{u} \{f_{\min} - (\hat{f}(x) + s(x)z)\} \phi(z) \, dz \]
\[ = \int_{-\infty}^{u} \{f_{\min} - \hat{f}(x) - s(x)z\} \phi(z) \, dz \]
\[ = \{f_{\min} - \hat{f}(x)\} \int_{-\infty}^{u} \phi(z) \, dz - s(x) \int_{-\infty}^{u} z \phi(z) \, dz \]

Define \( u = \{f_{\min} - \hat{f}(x)\} \) \( / s(x) \)
\[ \text{EI}(\mathbf{x}) = \mathbb{E}[I(\mathbf{x})] \]
\[ = \mathbb{E}[\max\{f_{\min} - f(\mathbf{x}), 0\}] \]
\[ = \mathbb{E} \left[ \{f_{\min} - f(\mathbf{x})\} \mathbb{1}\{f(\mathbf{x}) < f_{\min}\} \right] \]
\[ = \int_{-\infty}^{\infty} \{f_{\min} - y\} \mathbb{1}\{y < f_{\min}\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy \]
\[ = \int_{-\infty}^{f_{\min}} \{f_{\min} - y\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy \]
\[ = \int_{-\infty}^{u} \{f_{\min} - (\hat{f}(\mathbf{x}) + s(\mathbf{x})z)\} \phi(z) \, dz \]
\[ = \int_{-\infty}^{u} \{f_{\min} - \hat{f}(\mathbf{x}) - s(\mathbf{x})z\} \phi(z) \, dz \]
\[ = \{f_{\min} - \hat{f}(\mathbf{x})\} \int_{-\infty}^{u} \phi(z) \, dz - s(\mathbf{x}) \int_{-\infty}^{u} z \phi(z) \, dz \]

**Remember:**
\[ \phi'(z) = \frac{d}{dz} \phi(z) = \phi(z) \times \left( -\frac{1}{2} \times 2z \right) = -z \phi(z) \]
Define $u = \{f_{\text{min}} - \hat{f}(x)\}/s(x)$

$$EI(x) = \mathbb{E}[I(x)]$$

$$= \mathbb{E}[\max\{f_{\text{min}} - f(x), 0\}]$$

$$= \mathbb{E} \left[ \{f_{\text{min}} - f(x)\} 1\{f(x) < f_{\text{min}}\} \right]$$

$$= \int_{-\infty}^{\infty} \{f_{\text{min}} - y\} 1\{y < f_{\text{min}}\} \phi(y \mid \hat{f}(x), s^2(x)) \, dy$$

$$= \int_{-\infty}^{f_{\text{min}}} \{f_{\text{min}} - y\} \phi(y \mid \hat{f}(x), s^2(x)) \, dy$$

$$= \int_{-\infty}^{u} \{f_{\text{min}} - (\hat{f}(x) + s(x)z)\} \phi(z) \, dz$$

$$= \int_{-\infty}^{u} \{f_{\text{min}} - \hat{f}(x) - s(x)z\} \phi(z) \, dz$$

$$= \{f_{\text{min}} - \hat{f}(x)\} \int_{-\infty}^{u} \phi(z) \, dz - s(x) \int_{-\infty}^{u} z\phi(z) \, dz$$

$$= \{f_{\text{min}} - \hat{f}(x)\} \Phi(u) + s(x)\phi(u)$$
\[ EI(\mathbf{x}) = \mathbb{E}[I(\mathbf{x})] \]
\[ = \mathbb{E}[\max\{f_{\min} - f(\mathbf{x}), 0\}] \]
\[ = \mathbb{E}\left[\{f_{\min} - f(\mathbf{x})\}1\{f(\mathbf{x}) < f_{\min}\}\right] \]
\[ = \int_{-\infty}^{\infty} \{f_{\min} - y\}1\{y < f_{\min}\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy \]
\[ = \int_{-\infty}^{f_{\min}} \{f_{\min} - y\} \phi(y \mid \hat{f}(\mathbf{x}), s^2(\mathbf{x})) \, dy \]
\[ = \int_{-\infty}^{u} \{f_{\min} - (\hat{f}(\mathbf{x}) + s(\mathbf{x})z)\} \phi(z) \, dz \]
\[ = \int_{-\infty}^{u} \{f_{\min} - \hat{f}(\mathbf{x}) - s(\mathbf{x})z\} \phi(z) \, dz \]
\[ = \{f_{\min} - \hat{f}(\mathbf{x})\} \int_{-\infty}^{u} \phi(z) \, dz - s(\mathbf{x}) \int_{-\infty}^{u} z \phi(z) \, dz \]
\[ = \{f_{\min} - \hat{f}(\mathbf{x})\} \Phi(u) + s(\mathbf{x}) \phi(u) \]
\[ = s(\mathbf{x}) \{u \Phi(u) + \phi(u)\}. \]
$\text{EI}(x) = \mathbb{E}\{I(x)\}$

$= \{f_{\text{min}} - \hat{f}(x)\} \Phi(u) + s(x) \phi(u)$

The EI is made up of two terms.
\[ EI(\mathbf{x}) = \mathbb{E}\{ I(\mathbf{x}) \} \]

\[ = \left\{ f_{\min} - \hat{f}(\mathbf{x}) \right\} \Phi(u) + s(\mathbf{x}) \phi(u) \]

The EI is made up of two terms. The first term is increased by decreasing the predictive mean \( \hat{f}(\mathbf{x}) \),
\[ \text{EI}(x) = \mathbb{E}\{I(x)\} \]
\[ = \{f_{\text{min}} - \hat{f}(x)\} \Phi(u) + s(x) \phi(u) \]

The EI is made up of two terms. The first term is increased by decreasing the predictive mean \( \hat{f}(x) \), the second term is increased by increasing the predictive uncertainty \( s(x) \).
\[ \text{EI}(x) = \mathbb{E}\{I(x)\} \]
\[ = \left\{ f_{\text{min}} - \hat{f}(x) \right\} \Phi(u) + s(x) \phi(u) \]

The EI is made up of two terms. The first term is increased by decreasing the predictive mean \( \hat{f}(x) \), the second term is increased by increasing the predictive uncertainty \( s(x) \). This shows how EI balances exploitation and exploration.
Illustration: Study by Umberto Noè
Algorithm  Bayesian optimization.

1: Inputs:
   Initial design $\mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}}$
   Budget of $n_{\text{max}}$ function evaluations

2: for $n = n_{\text{init}}$ to $n_{\text{max}} - 1$ do

3: Update the GP: $f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x'))$

4: Compute the acquisition function $a_n(x)$

5: Auxiliary optimization: $x_{\text{next}} = \arg\max_{x \in X} a_n(x)$

6: Query $f$ at $x_{\text{next}}$ to obtain $y_{\text{next}}$

7: Augment data: $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}\}$

8: end for
Algorithm  Bayesian optimization.

1: **Inputs:**
   
   Initial design $D_{n_{\text{init}}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}}$
   
   Budget of $n_{\text{max}}$ function evaluations

2: for $n = n_{\text{init}}$ to $n_{\text{max}} - 1$ do

3:    Update the GP: $f(x) \mid D_n \sim \text{GP}(\hat{f}(x), s(x, x'))$

4:    Compute the acquisition function $a_n(x)$

5:    Auxiliary optimization: $x_{\text{next}} = \arg\max_{x \in \mathcal{X}} a_n(x)$

6:    Query $f$ at $x_{\text{next}}$ to obtain $y_{\text{next}}$

7:    Augment data: $D_{n+1} = D_n \cup \{x_{\text{next}}, y_{\text{next}}\}$

8: end for
EI\toptimum\ -\ Emulator's\ posterior\ mean

Training\ dataset\ at\ start

\[ \ell, \sigma_{\downarrow f} \] = [0.2534, 310.2622]
Algorithm Bayesian optimization.

1: Inputs:
   Initial design $\mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}}$
   Budget of $n_{\text{max}}$ function evaluations
2: for $n = n_{\text{init}}$ to $n_{\text{max}} - 1$ do
3:   Update the GP: $f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x'))$
4:   **Compute the acquisition function $a_n(x)$**
5:   Auxiliary optimization: $x_{\text{next}} = \arg\max_{x \in \mathbb{X}} a_n(x)$
6:   Query $f$ at $x_{\text{next}}$ to obtain $y_{\text{next}}$
7:   Augment data: $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}\}$
8: end for
GP mean, ± 2 SD and training data
Algorithm  Bayesian optimization.

1: Inputs:
   Initial design $\mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}}$
   Budget of $n_{\text{max}}$ function evaluations

2: for $n = n_{\text{init}}$ to $n_{\text{max}} - 1$ do

3:   Update the GP: $f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x'))$

4:   Compute the acquisition function $a_n(x)$

5:   **Auxiliary optimization:** $x_{\text{next}} = \arg\max_{x \in X} a_n(x)$

6: Query $f$ at $x_{\text{next}}$ to obtain $y_{\text{next}}$

7: Augment data: $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}\}$

8: end for
GP mean, ± 2 SD and training data

output, $y$

input, $x$

$\ell_\sigma \downarrow f = [0.2534, 310.2622]$
Algorithm  Bayesian optimization.

1: Inputs:
   Initial design \( \mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}} \)
   Budget of \( n_{\text{max}} \) function evaluations

2: for \( n = n_{\text{init}} \) to \( n_{\text{max}} - 1 \) do

3:   Update the GP: \( f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x')) \)

4:   Compute the acquisition function \( a_n(x) \)

5:   Auxiliary optimization: \( x_{\text{next}} = \arg\max_{x \in \mathcal{X}} a_n(x) \)

6:   Query \( f \) at \( x_{\text{next}} \) to obtain \( y_{\text{next}} \)

7:   Augment data: \( \mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}\} \)

8: end for
GP mean, ± 2 SD and training data

output, y

input, x

GP mean, ± 2 SD and training data

output, y

input, x
Algorithm Bayesian optimization.

1: Inputs:
   Initial design $\mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}}$
   Budget of $n_{\text{max}}$ function evaluations
2: for $n = n_{\text{init}}$ to $n_{\text{max}} - 1$ do
3:   Update the GP: $f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x'))$
4:   Compute the acquisition function $a_n(x)$
5:   Auxiliary optimization: $x_{\text{next}} = \arg \max_{x \in \mathcal{X}} a_n(x)$
6:   Query $f$ at $x_{\text{next}}$ to obtain $y_{\text{next}}$
7:   Augment data: $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}\}$
8: end for

Iterate until convergence
Iteraton 1

GP mean, ± 2 SD and training data

output, y

input, x

E(x)

El optimum location
  - Emulator’s posterior mean
  • Training dataset at start
Iteration 2

GP mean, ± 2 SD and training data

- optimum location
- previous evaluation
Iteration 3

GP mean, ± 2 SD and training data

output, y

input, x

E(x)
Iteration 4

GP mean, ± 2 SD and training data
Iteration 5

GP mean, ± 2 SD and training data
Iteration 6

GP mean, ± 2 SD and training data

output, y

1

0.1

0.2

0.3

0.4

0.5

0.6

0.7

0.8

0.9

1

input, x

0

-2

-4

4

3

2

1

0

-1

-2

-3

-4

0.0418

0.8395
Iteration 7

GP mean, ± 2 SD and training data
Iteration 8

GP mean, ± 2 SD and training data
Iteration 9

GP mean, ± 2 SD and training data

output, y

input, x

\[ [\ell, \sigma \downarrow f] = [0.0389, 0.8909] \]
Iteration 10

GP mean, ± 2 SD and training data

- Output, y
- Input, x

$\ell, \sigma_{\downarrow f} = [0.0394, 0.8692]$
Iteration 11

GP mean, $\pm$ 2 SD and training data

---

Output, $y$

Input, $x$

$\ell, \sigma_{\downarrow f} =$ [0.0402, 0.8650]
Iteration 12

GP mean, ± 2 SD and training data

output, y

input, x

$\sigma_{\downarrow f} = [0.0408, 0.8704]$
Are there better methods than the Upper Confidence Bound (UCB) criterion?

Improvement-based methods

Information-based methods
**Improvement-based methods**
favour points that are likely to improve on an incumbent target. This approach is intrinsically myopic.
Improvement-based methods favour points that are likely to improve on an incumbent target. This approach is intrinsically myopic.

Information-based methods consider the distribution $p(x_{\text{global}} \mid D_n)$

This distribution is implicitly induced by the posterior over objective functions.
Improvement-based methods favour points that are likely to improve on an incumbent target. This approach is intrinsically myopic.

Information-based methods consider the distribution \( p(x_{\text{global}} \mid D_n) \) and query the point that leads to the largest reduction in uncertainty about the location \( x_{\text{global}} \)
Improvement-based methods favour points that are likely to improve on an incumbent target. This approach is intrinsically myopic.

Information-based methods consider the distribution $p(x_{\text{global}} \mid D_n)$ and query the point that leads to the largest reduction in uncertainty about the location $x_{\text{global}}$.

\[ \text{ES}(x) = H[x_{\text{global}} \mid D_n] - \mathbb{E}\{H[x_{\text{global}} \mid D_n, x, y]\} \]

where the expectation is taken with respect to $p(y \mid D_n, x)$.
where the expectation is taken with respect to

$$p(y \mid \mathcal{D}_n, x)$$

Analytically intractable, numerical approximations are needed, like
- discretization of the space in which $x$ is defined
- Monte Carlo sampling

Several publications in the literature trying different approximations; see Shahriari et al. for a review.
Entropy Search

$$ES(x) = H[\mathbf{x}_{\text{global}} \mid \mathcal{D}_n]$$
$$- \mathbb{E}\left\{H[\mathbf{x}_{\text{global}} \mid \mathcal{D}_n, \mathbf{x}, y]\right\}$$

where the expectation is taken with respect to

$$p(y \mid \mathcal{D}_n, \mathbf{x})$$

Predictive Entropy Search

$$PES(x) = H[y \mid \mathcal{D}_n, \mathbf{x}]$$
$$- \mathbb{E}\left\{H[y \mid \mathcal{D}_n, \mathbf{x}, \mathbf{x}_{\text{global}}]\right\}$$

where the expectation is taken with respect to

$$p(\mathbf{x}_{\text{global}} \mid \mathcal{D}_n)$$
Entropy Search

\[ ES(x) = H[x_{\text{global}} \mid D_n] \]
\[ - \mathbb{E}\left\{ H[x_{\text{global}} \mid D_n, x, y]\right\} \]

where the expectation is taken with respect to \( p(y \mid D_n, x) \)

Predictive Entropy Search

\[ PES(x) = H[y \mid D_n, x] \]
\[ - \mathbb{E}\left\{ H[y \mid D_n, x, x_{\text{global}}]\right\} \]

where the expectation is taken with respect to \( p(x_{\text{global}} \mid D_n) \)
Entropy Search

\[
ES(x) = H[x_{\text{global}} \mid D_n] \\
- \mathbb{E}\{H[x_{\text{global}} \mid D_n, x, y]\}
\]

where the expectation is taken with respect to

\[
p(y \mid D_n, x)
\]

Predictive Entropy Search

\[
PES(x) = H[y \mid D_n, x] \\
- \mathbb{E}\{H[y \mid D_n, x, x_{\text{global}}]\}
\]

where the expectation is taken with respect to

\[
p(x_{\text{global}} \mid D_n)
\]

Mathematically equivalent, but different numerical approximations, which can be more efficient for PES.
Optimistic
LCB: Lower confidence bound

Improvement-based
PI: Probability of Improvement
EI: Expected Improvement

Information-based
ES: Entropy search
PES: Predictive entropy search
Optimistic
LCB: Lower confidence bound

Improvement-based
PI: Probability of Improvement
EI: Expected improvement
Scaled EI: new method
developed in our group

Information-based
ES: Entropy search
PES: Predictive entropy search
Umberto Noé
\[ \text{ScaledEI}(x) = \mathbb{E}[I(x)] / \{ \mathbb{V}[I(x)] \}^{1/2} \]

\[
\mathbb{V}[I(x)] = \mathbb{E}[I^2(x)] - \{ \mathbb{E}[I(x)] \}^2
\]

\[
= \mathbb{E}[\max\{f_{\min} - f(x), 0\}^2] - \{\text{EI}(x)\}^2
\]

\[
= \int_{-\infty}^{f_{\min}} \{f_{\min} - y\}^2 \phi(y | \hat{f}(x), s^2(x)) \, dy - \{\text{EI}(x)\}^2
\]

\[
= \int_{-\infty}^{u} \{f_{\min} - \hat{f}(x) - s(x)z\}^2 \phi(z) \, dz - \{\text{EI}(x)\}^2
\]

\[
= \int_{-\infty}^{u} \{[f_{\min} - \hat{f}(x)]^2 + z^2 s^2(x)
\]

\[
- 2zs(x)[f_{\min} - \hat{f}(x)]\phi(z) \, dz - \{\text{EI}(x)\}^2
\]

\[
= \{f_{\min} - \hat{f}(x)\}^2 \int_{-\infty}^{u} \phi(z) \, dz
\]

\[
+ s^2(x) \int_{-\infty}^{u} z^2 \phi(z) \, dz
\]

\[
- 2s(x)\{f_{\min} - \hat{f}(x)\} \int_{-\infty}^{u} z \phi(z) \, dz - \{\text{EI}(x)\}^2
\]

\[
= \{f_{\min} - \hat{f}(x)\}^2 \Phi(u) + 2s(x)\{f_{\min} - \hat{f}(x)\} \phi(u)
\]

\[
+ s^2(x) \int_{-\infty}^{u} (z^2 - 1) \phi(z) \, dz
\]

\[
+ s^2(x) \int_{-\infty}^{u} \phi(z) \, dz - \{\text{EI}(x)\}^2
\]

\[
= \{f_{\min} - \hat{f}(x)\}^2 \Phi(u) + 2s(x)\{f_{\min} - \hat{f}(x)\} \phi(u)
\]

\[
- s^2(x)u\phi(u) + s^2(x)\Phi(u) - \{\text{EI}(x)\}^2
\]

\[
= \{[f_{\min} - \hat{f}(x)]^2 + s^2(x)\} \Phi(u)
\]

\[
+ s(x)\{f_{\min} - \hat{f}(x)\} \phi(u) - \{\text{EI}(x)\}^2
\]

\[
= s^2(x)\{(u^2 + 1) \Phi(u) + u \phi(u)\} - \{\text{EI}(x)\}^2.
\]
Optimistic

- LCB: Lower confidence bound

Improvement-based

- PI: Probability of Improvement
- EI: Expected improvement
- Scaled EI

Information-based

- MES: Maximum value entropy search
  (Wang & Jegelka, 2017)

Naive

- MN: Negative GP predictive mean
- RND: Random search
Table 1. Key characteristics of the test functions.

<table>
<thead>
<tr>
<th>Test function</th>
<th>Abbreviation</th>
<th>Number of dimensions</th>
<th>Number of local minima</th>
<th>Number of global minima</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine Sine</td>
<td>CSF</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Rosenbrock</td>
<td>ROS</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Branin RCOS</td>
<td>BRA</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Goldstein and Price</td>
<td>GPR</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Six-Hump Camel</td>
<td>CAM</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Two-Dimensional Shubert</td>
<td>SHU</td>
<td>2</td>
<td>760</td>
<td>18</td>
</tr>
<tr>
<td>Hartman 3</td>
<td>HM3</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Shekel 5</td>
<td>SH5</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Shekel 7</td>
<td>SH7</td>
<td>4</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Shekel 10</td>
<td>SH10</td>
<td>4</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Hartman 6</td>
<td>HM6</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Rastrigin</td>
<td>RAS</td>
<td>10</td>
<td>11^{10}</td>
<td>1</td>
</tr>
<tr>
<td>Test function</td>
<td>RND</td>
<td>MN</td>
<td>LCB</td>
<td>PI</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>CSF</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ROS</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BRA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GPR</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CAM</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SHU</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HM3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SH5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SH7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SH10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HM6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAS</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Same</strong></td>
<td>0%</td>
<td>83%</td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Better</strong></td>
<td>100%</td>
<td>17%</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Worse</strong></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 2. Paired t-test of equality in the mean log10 distance for each acquisition function. Codes: 0 indicates a non-significant difference and 1 (-1) indicates that ScaledEI performed better (worse), i.e. it has a significantly lower (higher) average distance.
Umberto Noé et al.
Proc. CIBB 2017
Computational Intelligence Methods for Bioinformatics and Biostatistics
Elastic arteries

Systole

↓ Systolic/pulse pressure
Elastic arteries

Systole

Systolic/pulse pressure

Diastolic flow

Diastole
Pressure time courses
Flow and pressure time courses

(a) Flow rate ($q$) vs. time ($t$) for hypoxic and control conditions.
(b) Pressure ($p$) vs. time ($t$) for hypoxic and control conditions.

- Hypoxic
- Control

Legend:
- Black: Data
- Gray: Simulation
\[ \frac{\partial q}{\partial t} + \frac{\partial}{\partial x} \left( \frac{q^2}{A} \right) + \frac{A}{\rho} \frac{\partial p}{\partial x} = -\frac{2\pi \nu r}{\delta} \frac{q}{A} \]

Pressure and flow rate
\[
\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} \left( \frac{q^2}{A} \right) + \frac{A}{\rho} \frac{\partial p}{\partial x} = -\frac{2\pi \nu r}{\delta} \frac{q}{A}
\]

Time and axial coordinate

\[
\frac{\partial A}{\partial t} + \frac{\partial q}{\partial x} = 0
\]
\[
\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} \left( \frac{q^2}{A} \right) + \frac{A}{\rho} \frac{\partial p}{\partial x} = -\frac{2\pi \nu r}{\delta} q
\]

\[
\frac{\partial A}{\partial t} + \frac{\partial q}{\partial x} = 0
\]
\[
\begin{align*}
\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} \left( \frac{q^2}{A} \right) + \frac{A}{\rho} \frac{\partial p}{\partial x} &= -\frac{2\pi \nu r}{\delta} \frac{q}{A} \\
\frac{\partial A}{\partial t} + \frac{\partial q}{\partial x} &= 0
\end{align*}
\]
Data: Pressure and flow time series at different locations in different blood vessels
The equations get more complicated when ...
The equations get more complicated when including the small-vessel fluid dynamics.
Numerical simulation of blood flow and pressure drop in the pulmonary arterial and venous circulation

M. Umar Qureshi · Gareth D. A. Vaughan · Christopher Sainsbury · Martin Johnson · Charles S. Peskin · Mette S. Olufsen · N. A. Hill
Three critical unknown parameters

\[ \xi \] Indicator of vascular rarefaction

\[ f_L \] Stiffness of the large vessels

\[ f_S \] Stiffness of the small vessels
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.
Prior valid parameter domains

\[2.33 \leq \xi \leq 3\]

\[f_s \in [2.66 \times 10^4, 1.0666 \times 10^5]\]

\[f_L \in [1.33 \times 10^5, 5.33 \times 10^5]\]
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.

Parameter space

We don’t know in advance where the boundaries are …
Problem:
Certain parameter configurations violate the physical assumptions of the model and lead to a crash of the program.

We don’t know in advance where the boundaries are …
… so we have to learn them!

Parameter space
IDEA:
Combine a statistical emulator ...

... with a statistical classifier

Parameter space
Data

\((x_i, y_i, h_i)\)
Data

\((x_i, y_i, h_i)\)

Input vector
(model parameters)
Data

\((x_i, y_i, \ h_i)\)

Input vector (model parameters)

Output value (objective function)
Data

\((x_i, y_i, h_i)\)

Output value (objective function)

Input vector (model parameters)

\(h_i \in \{-1, 1\}\)

Crash indicators
We combine:

a GP model of the objective function, using the \((x_i, y_i)\) pairs;

a GP model of the failures, using the \((x_i, h_i)\) pairs.
Combine Bayesian optimization with a Gaussian Process classifier

\[ a(x) = \text{EI}(x) \Pr(C(x)) \]

- Standard acquisition function
- New acquisition function
- Indicator for constraint satisfaction
Algorithm 2 Bayesian optimization with hidden constraints.

1: **Inputs:**
   Initial design and corresponding failure labels: $\mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i, h_i)\}_{i=1}^{n_{\text{init}}}$
   Budget of $n_{\text{max}}$ function evaluations

2: **for** $n = n_{\text{init}}$ **to** $n_{\text{max}} - 1$ **do**
   3: Update the objective GP: $f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x'))$
   4: Update the failure GP: $h(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{h}(x), s_h(x, x'))$
   5: Compute the acquisition function: $a^*_n(x) = a_n(x) \times \Phi(0 \mid \hat{h}(x), s^2_h(x))$
   6: Solve the auxiliary optimization problem: $x_{\text{next}} = \arg\max_{x \in \mathcal{X}} a^*_n(x)$
   7: Query $f$ at $x_{\text{next}}$ to obtain $y_{\text{next}}$ and $h_{\text{next}}$
   8: Augment data: $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}, h_{\text{next}}\}$

9: **end for**

10: **Return:**
    Estimated minimum: $f_{\text{min}} = \min(y_1, \ldots, y_{n_{\text{max}}})$
    Estimated point of minimum: $x_{\text{min}} = \arg\min(y_1, \ldots, y_{n_{\text{max}}})$
Algorithm 2 Bayesian optimization with hidden constraints.

1: Inputs:
   Initial design and corresponding failure labels: \( \mathcal{D}_{n_{\text{init}}} = \{(x_i, y_i, h_i)\}_{i=1}^{n_{\text{init}}} \)
   Budget of \( n_{\text{max}} \) function evaluations

2: for \( n = n_{\text{init}} \) to \( n_{\text{max}} - 1 \) do
3:   Update the objective GP: \( f(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{f}(x), s(x, x')) \)
4:   Update the failure GP: \( h(x) \mid \mathcal{D}_n \sim \text{GP}(\hat{h}(x), s_h(x, x')) \)
5:   Compute the acquisition function: \( a_n^*(x) = a_n(x) \times \Phi(0 \mid \hat{h}(x), s^2_h(x)) \)
6:   Solve the auxiliary optimization problem: \( x_{\text{next}} = \arg\max_{x \in X} a_n^*(x) \)
7:   Query \( f \) at \( x_{\text{next}} \) to obtain \( y_{\text{next}} \) and \( h_{\text{next}} \)
8:   Augment data: \( \mathcal{D}_{n+1} = \mathcal{D}_n \cup \{x_{\text{next}}, y_{\text{next}}, h_{\text{next}}\} \)
9: end for

10: Return:
    Estimated minimum: \( f_{\text{min}} = \min(y_1, \ldots, y_{n_{\text{max}}}) \)
    Estimated point of minimum: \( x_{\text{min}} = \arg\min(y_1, \ldots, y_{n_{\text{max}}}) \)
Start with one parameter

\[ r_p^\xi = r_{d_1}^\xi + r_{d_2}^\xi \]
\[ r_p^\xi = r_{d_1}^\xi + r_{d_2}^\xi \]
Parameter to be inferred

\[ r_p = r_d + r_s \]
Parameter to be inferred

\[ r_p = r_1 + r_2 \]
Parameter to be inferred

\[ r_p^\xi = r_{d_1}^\xi + r_{d_2}^\xi, \quad 2.33 \leq \xi \leq 3.0 \]

Prior validity domain
RSS difference between measured and modelled pressure/flow time courses
Emulated RSS vs. Exponent

Expected Improvement vs. Exponent

Probability of success

\[ EI_{HCW}(x) = EI(x) \times P(h(x) = \text{no model failure}) \]
Probability of avoiding a crash
Table: The PDE parameters underlying the simulated data (Truth) and the estimated parameters (Estimate). Mean and standard error over the 15 design instantiations.

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>Estimate (n = 500)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>$f_L$</td>
<td>$2.6 \times 10^5$</td>
<td>$2.6005 \times 10^5$</td>
</tr>
<tr>
<td>$f_S$</td>
<td>50000</td>
<td>50003</td>
</tr>
<tr>
<td>$\xi$</td>
<td>2.76</td>
<td>2.7603</td>
</tr>
</tbody>
</table>
The PDE parameters underlying the simulated data (Truth) and the estimated parameters (Estimate). Mean and standard error over the 15 design instantiations.

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>Estimate ($n = 500$)</th>
<th></th>
<th>Estimate ($n = 100$)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Std. Err.</td>
<td>Mean</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>$f_L$</td>
<td>$2.6 \times 10^5$</td>
<td>$2.6005 \times 10^5$</td>
<td>189.31</td>
<td>$2.599 \times 10^5$</td>
<td>286.55</td>
</tr>
<tr>
<td>$f_S$</td>
<td>50000</td>
<td>50003</td>
<td>35.01</td>
<td>50038</td>
<td>55.96</td>
</tr>
<tr>
<td>$\xi$</td>
<td>2.76</td>
<td>2.7603</td>
<td>0.0002</td>
<td>2.76</td>
<td>0.0004</td>
</tr>
</tbody>
</table>
Comparison: global optimization with genetic algorithms: 3 days

Table: The PDE parameters underlying the simulated data (Truth) and the estimated parameters (Estimate). Mean and standard error over the 15 design instantiations.

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>Estimate (n = 500)</th>
<th>Estimate (n = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Err.</td>
<td>Mean</td>
</tr>
<tr>
<td>$f_L$</td>
<td>$2.6 \times 10^5$</td>
<td>$2.6005 \times 10^5$</td>
<td>189.31</td>
</tr>
<tr>
<td>$f_S$</td>
<td>500000</td>
<td>50003</td>
<td>35.01</td>
</tr>
<tr>
<td>$\xi$</td>
<td>2.76</td>
<td>2.7603</td>
<td>0.0002</td>
</tr>
</tbody>
</table>