# A Stepwise uncertainty reduction approach to constrained global optimization

**Victor Picheny**
INRA, 31326 Castanet Tolosan, France, victor.picheny@toulouse.inra.fr

## Abstract

Using statistical emulators to guide sequential evaluations of complex computer experiments is now a well-established practice. When a model provides multiple outputs, a typical objective is to optimize one of the outputs with constraints (for instance, a threshold not to exceed) on the values of the other outputs. We propose here a new optimization strategy based on the stepwise uncertainty reduction paradigm, which offers an efficient trade-off between exploration and local search near the boundaries. The strategy is illustrated on numerical examples.

## 1 INTRODUCTION

We consider single-objective optimization problems, subject to non-linear constraints and box constraints:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ s.t. \quad & g^i(\mathbf{x}) \leq T_i, \quad i \in \{1, \dots, q\} \\ & \mathbf{x} \in \mathbb{X} \subset \mathbb{R}^d \end{aligned} \tag{1}$$

More specifically, we assume that $f(.)$ and the $g^i(.)$'s are outputs of a complex, expensive-to-evaluate computer model parameterized by a vector $x$ (if size $d \geq 1$), $\mathbb{X}$ being the parameters intervals of variation.

In order to cope with the limited number of model evaluations (due to their computational cost), a well-established practice consists of using statistical emulators, based on a small experimental set, to approximate the model outputs and guide a (parcimonious) sequential sampling strategy. In the celebrated article of Jones et al. (1998), a Gaussian process emulator is used and the experiments are chosen according to an associated infill criterion, the *Expected Improvement* (EI), that expresses a trade-off between exploration and intensification to achieve global optimization. Such strategy, called *Efficient Global Optimization* (EGO), can cope with constraints, as long as they are explicit or inexpensive to compute.

In the case considered here, constraints are evaluated together with the objective function. Finding efficient sampling strategies is complex, as one may want to avoid spending too many evaluations on the unfeasible regions, while exploring the regions close to the boundaries, where the optimum is likely to be. This problem has been addressed by several authors (Schonlau et al., 1998; Sasena et al., 2002; Parr et al., 2012; Gramacy & Lee, 2011), essentially by combining the Expected Improvement with feasibility indicators (expected feasibility, probability of feasibility, etc.).

We propose here to address this issue alternatively, first by providing a measure of the uncertainty on the minimizer location, in the spirit of Villemonteix et al. (2009), then by finding sequentially the measurement that achieves, in expectation, the maximum reduction of this uncertainty. This approach has the advantage of incorporating rigorously the constraints in the measurements decision, while reflecting more precisely the actual users' objective: finding the minimizer rather than the minimal value of the function.

We start by providing a rapid introduction to Gaussian-process-based emulation. We describe our strategy first in the unconstrained case, then we show how to incorporate the constraints. Finally, the method is illustrated and compared to alternatives on numerical examples.

## 2 GAUSSIAN PROCESS MODELING

We consider the standard Gaussian process model (Cressie, 1993; Rasmussen & Williams, 2006), where the output of interest $y$ is assumed to be one realiza-

Appearing in Proceedings of the $17^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

tion of a Gaussian process

$$Y(.) \sim \mathcal{GP}\left(\mathbf{h}(.)^T \boldsymbol{\beta}, k(.,.)\right) \tag{2}$$

where $\mathbf{h}(.)^T = (h_1(.), \dots, h_p(.))$ is a vector of trend functions, $\boldsymbol{\beta}$ a vector of (unknown) coefficients and $k(.,.)$ is a known covariance kernel.

Conditionally on the event

$$\mathcal{A}_n = \{Y(\mathbf{x}_1) = y_1, \dots, Y(\mathbf{x}_n) = y_n\},$$

we have the predictive distribution:

$$Y(.|\mathcal{A}_n) \sim \mathcal{GP}\left(m_n(.), c_n(.,.)\right), \tag{3}$$

with:

$$m_n(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1}(\mathbf{y}_n - \mathbf{H}_n \hat{\boldsymbol{\beta}}),$$

$$\begin{aligned} c_n(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}') \\ &+ \left(\mathbf{h}(\mathbf{x})^T - \mathbf{k}_n(\mathbf{x})^T \mathbf{K}_n^{-1} \mathbf{H}_n\right)^T \left(\mathbf{H}_n^T \mathbf{K}_n^{-1} \mathbf{H}_n\right)^{-1} \\ &\left(\mathbf{h}(\mathbf{x}')^T - \mathbf{k}_n(\mathbf{x}')^T \mathbf{K}_n^{-1} \mathbf{H}_n\right), \end{aligned}$$

where $\mathbf{y}_n = (y_1, \dots, y_n)^T$, $\mathbf{K}_n = (k(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$, $\mathbf{k}_n(\mathbf{x})^T = (k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n))$, $\mathbf{H}_n = \left(\mathbf{h}(\mathbf{x}_1)^T, \dots, \mathbf{h}(\mathbf{x}_n)^T\right)^T$, and $\hat{\boldsymbol{\beta}} = \left(\mathbf{H}_n^T \mathbf{K}_n^{-1} \mathbf{H}_n\right)^{-1} \mathbf{H}_n^T \mathbf{K}_n^{-1} \mathbf{y}_n$. In addition, the *prediction variance* is defined as $s_n^2(\mathbf{x}) = c_n(\mathbf{x}, \mathbf{x})$. Figure 1 shows an example of GP modeling based on seven observations.

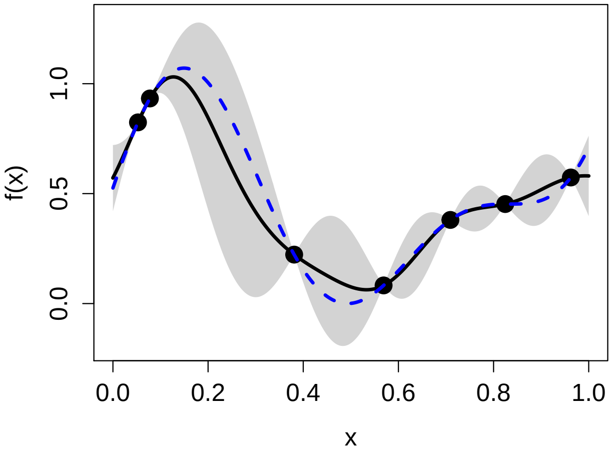

Figure 1: Gaussian process model based on seven observations. The dotted line is the actual function, the plain line is the predictive mean $m_n(\mathbf{x})$, and the grey area the 95% predictive confidence interval $m_n(\mathbf{x}) \pm 2s_n(\mathbf{x})$.

In the rest of the article, the subscript $n$ denotes the conditioning on $\mathcal{A}_n$, unless stated otherwise.

Classically, the covariance kernel depends on unknown parameters that are inferred from an initial set of responses, using maximum likelihood estimates for instance. Usually, the estimates are used as face value, but updated when new observations are added to the model. The reader can refer to Stein (1999) (chapter 6), Rasmussen & Williams (2006) (chapter 5) or Roustant et al. (2012) for detailed calculations and implementation issues.

When several outputs are predicted simultaneously, it is possible to take their dependency into account (Kennedy & O'Hagan, 2001; Craig et al., 2001). However, in this work we consider all the processes $F, G^1, \dots, G^q$ independent, hence modelled as above.

## 3 GP-BASED CONSTRAINED OPTIMIZATION

The EGO algorithm of Jones et al. (1998) consists of the sequential enrichment of the design of experiments by adding a measurement at a point that maximizes the expected improvement: $\mathbf{x}_{n+1} = \arg\max EI_n(\mathbf{x})$, where

$$\begin{aligned} EI_n(\mathbf{x}) &:= \mathbb{E}\left[I_n(\mathbf{x})\right] \\ &= \mathbb{E}\left[\max\left(0, \min_{i \in \{1,\dots,n\}}(y_i) - Y(\mathbf{x})\right) |\mathcal{A}_n\right]. \end{aligned}$$

To handle constraints, Schonlau et al. (1998) introduced a *feasability* function $F(\mathbf{x})$, that equals one when the constraints are satisfied and zero otherwise. The points are then taken where the maximum expected *feasible* improvement is achieved, that is:

$$\mathbb{E}\left[I_n(\mathbf{x}) \cap F_n(\mathbf{x})\right] := EI_n(\mathbf{x}) \prod_{i=1}^{q} \mathbb{P}\left(G^i(\mathbf{x}) \le T_i\right),$$

by independence of $F, G^1, \dots, G^q$. The expected feasability can be computed using the GP model, as:

$$\mathbb{P}\left(G^i(\mathbf{x}) \le T_i\right) = \Phi\left(\frac{T^i - m_n^i(\mathbf{x})}{s_n^i(\mathbf{x})}\right).$$

In Sasena et al. (2002), this approach is compared to the use of additive penalties but no significant difference is found. However, both approaches tend to avoid frontier regions (where the optimum is supposedly located), leading to poor convergence speed (Audet et al., 2000; Parr et al., 2010). Parr et al. (2012) use a multi-objective approach and sample points that realize optimal (in the Pareto sense) trade-offs between expected improvement and expected feasibility.

Gramacy & Lee (2011) use a GP classifier for the constraint and a so-called *Integrated Expected Conditional*

*Improvement* (IECI) sampling criterion. Contrarily to the EI-based approaches, this criterion accounts for the fact that an unfeasible design can also provide useful information, at it measures improvement over the entire design region. The constraints are handled using expected feasibility. Although theoretically appealing, this criterion was found only moderately efficient (see section 7.2. The new criterion we describe in the two following sections resembles the IECI in spirit, but considers improvement probabilities instead of expectations (section 4) while the constraints are handled by considering all the possible sampling scenarii (section 5).

# 4 UNCONSTRAINED OPTIMIZATION BY UNCERTAINTY REDUCTION

Box-constrained optimization using stepwise uncertainty reduction has been introduced in Picheny (2013). We recall here the main principles that are necessary for the constrained scenario.

The principle of stepwise uncertainty reduction (Villemonteix et al., 2009; Bect et al., 2012; Chevalier et al., 2012) is to define an uncertainty measure related to the desired objective, and choose sequentially the measurements that decrease best, in expectation, this uncertainty. Given an uncertainty measure $\Gamma_n$ (computed with the GP model conditioned by $\mathcal{A}_n$), we need to compute $\mathbb{E}_n\left[\Gamma_{n+1} | Y(\mathbf{x}_{n+1}) = Y_{n+1}\right]$, that is, the expectation of the new uncertainty measure knowing that $\mathbf{x}_{n+1}$ is added to the measurements. In the following, we abusively denote by $|Y_{n+1}$ such conditioning to lighten the notations. Conditionally on $\mathcal{A}_n$, $Y_{n+1}$ is random with mean $m_n(\mathbf{x}_{n+1})$ and variance $s_n^2(\mathbf{x}_{n+1})$. Updating the model (3) by adding $\mathbf{x}_{n+1}$ to the design while accounting for $Y_{n+1}$'s distribution allows us to calculate $\mathbb{E}_n\left[\Gamma_{n+1}|Y_{n+1}\right]$.

Coming back to optimization, consider that $n$ measurements have been performed, and the current best measurement is $f_n^{\min} = \min(f_1, \ldots, f_n)$. At any design $\mathbf{x}$, the probability $p_n(\mathbf{x}, f_n^{\min}) := \mathbb{P}_n\left(F(\mathbf{x}) \leq f_n^{\min}\right)$, often referred to as *probability of improvement* (Jones, 2001), is

$$p_n(\mathbf{x}, f_n^{\min}) = \Phi\left(\frac{f_n^{\min} - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right), \tag{4}$$

where $\Phi(.)$ is the cumulative distribution function (CDF) of the standard Gaussian distribution.

Now, we wish to define a measure of uncertainty we have about the location of the minimizer $\mathbf{x}^*$ of $f$. Integrating the probability of improvement over the design space, we obtain

$$\begin{aligned} ev_n &= \mathbb{E}_{\mathbb{X}}\left[\mathbb{P}_n\left(F(\mathbf{x}) \leq f_n^{\min}\right)\right] \\ &= \int_{\mathbb{X}} \Phi\left(\frac{f_n^{\min} - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right) d\mathbf{x}, \end{aligned} \tag{5}$$

which is the expected volume of the excursion set below $f_n^{\min}$. A large volume indicates that the optimum is not yet precisely located, as many designs are likely to be better than the current best one; on the contrary, a small volume indicates that very little can be gained by pursuing the optimization process. Hence, $ev_n$ defines a suitable uncertainty measure $\Gamma_n$ regarding optimization.

A critical advantage of this measure over possible alternatives is that its expected update ($\mathbb{E}_n\left[\Gamma_{n+1}|Y_{n+1}\right]$) is computationally tractable, as we show in the following. Inversely, the IAGO criterion of Villemonteix et al. (2009), based on the Shannon entropy of the minimizer, requires expensive conditional simulations; the IECI of Gramacy & Lee (2011), which is based on $\Gamma_n = \mathbb{E}_{\mathbb{X}}\left[EI_n(\mathbf{x})\right]$, uses only a simplification of $\mathbb{E}_n\left[\Gamma_{n+1}|Y_{n+1}\right]$.

Hypothesizing that a measurement $f_{n+1}$ is performed at a point $\mathbf{x}_{n+1}$, its benefit can be measured by the reduction of the expected volume of excursion set $\Delta = ev_n - ev_{n+1}$, with:

$$ev_{n+1}(\mathbf{x}_{n+1}) = \int_{\mathbb{X}} p_{n+1}\left(\mathbf{x}, \min\left(f_n^{\min}, f_{n+1}\right)\right) d\mathbf{x}.$$

Now, we want to obtain:

$$\begin{aligned} &EEV(\mathbf{x}_{n+1}) \\ &\quad= \mathbb{E}_n\left[EV_{n+1}|F_{n+1}\right] \\ &\quad= \textstyle\int_{\mathbb{X}} \mathbb{P}_n\left[F(\mathbf{x}) \leq \min\left(f_n^{\min}, F_{n+1}\right)|F_{n+1}\right] d\mathbf{x}. \end{aligned}$$

We note first that:

$$\begin{aligned} &\mathbb{P}_n\left[F(\mathbf{x}) \leq \min\left(f_n^{\min}, F_{n+1}\right)|F_{n+1}\right] \\ &= \mathbb{P}_n\left[F(\mathbf{x}) \leq F_{n+1} | F_{n+1} \leq f_n^{\min}\right] \mathbb{P}_n\left[F_{n+1} \leq f_n^{\min}\right] \\ &+ \mathbb{P}_n\left[F(\mathbf{x}) \leq f_n^{\min} | F_{n+1} \geq f_n^{\min}\right] \mathbb{P}_n\left[F_{n+1} \geq f_n^{\min}\right]. \end{aligned}$$

These two quantities are given by the following propositions (see Picheny (2013) for proofs):

$$\begin{aligned} &\mathbb{P}_n\left[F(\mathbf{x}) \leq a \middle| F_{n+1} \geq a\right] \mathbb{P}_n\left[F_{n+1} \geq a\right] \\ &\qquad = \mathbf{\Phi}_{-\rho}\left(-\bar{a}, \tilde{a}\right), \end{aligned} \tag{6}$$

$$\begin{aligned} &\mathbb{P}_n\left[F(\mathbf{x}) \leq F_{n+1} | F_{n+1} \leq a\right] \mathbb{P}_n\left[F_{n+1} \leq a\right] \\ &\qquad = \mathbf{\Phi}_{\nu}\left(\bar{a}, \eta\right), \end{aligned} \tag{7}$$

where $\mathbf{\Phi}_r$ is the Gaussian bivariate CDF with zero mean and covariance $\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$,

$$\rho = \frac{c_n(\mathbf{x}, \mathbf{x}_{n+1})}{s_n(\mathbf{x}_{n+1}) s_n(\mathbf{x})},$$

$$\begin{aligned}
\nu &= \frac{c_n(\mathbf{x},\mathbf{x}_{n+1}) - s_n^2(\mathbf{x}_{n+1})}{s_n(\mathbf{x}_{n+1})\sqrt{s_n^2(\mathbf{x}) + s_n^2(\mathbf{x}_{n+1}) - 2c_n(\mathbf{x},\mathbf{x}_{n+1})}},\\
\eta &= \frac{m_n(\mathbf{x}_{n+1}) - m_n(\mathbf{x})}{\sqrt{s_n^2(\mathbf{x}) + s_n^2(\mathbf{x}_{n+1}) - 2c_n(\mathbf{x},\mathbf{x}_{n+1})}},\\
\bar{a} &= \frac{a - m_n(\mathbf{x}_{n+1})}{s_n(\mathbf{x}_{n+1})} \text{ and } \widetilde{a} = \frac{a - m_n(\mathbf{x})}{s_n(\mathbf{x})}.
\end{aligned}$$

Hence, the expected new uncertainty measure is:

$$\begin{aligned}
EEV(\mathbf{x}_{n+1}) = \int_{\mathbb{X}} \Big[ & \mathbf{\Phi}_{\nu}\left(\overline{f}_n^{\min}, \eta\right) \\
& + \mathbf{\Phi}_{-\rho}\left(-\overline{f}_n^{\min}, \widetilde{f}_n^{\min}\right)\Big] d\mathbf{x}. \quad (8)
\end{aligned}$$

The one-step optimal strategy consists of adding the point that minimizes, in expectation, the uncertainty about the minimizer:

$$\mathbf{x}_{n+1} = \arg\min_{\mathbf{x}^+ \in \mathbb{X}} EEV(\mathbf{x}^+) \quad (9)$$

As in the EGO strategy (Jones et al., 1998), no closed-form expression exists for $\mathbf{x}_{n+1}$, and it must be found by running an inner optimization algorithm.

# 5 ACCOUNTING FOR CONSTRAINTS

For clarity purpose, we consider here a single constraint $g(\mathbf{x}) \leq T$; note that the reasoning and calculations transpose without difficulty to several constraints.

We assume that two GP models have been fit to the objective and the constraint. The models are denoted as $F(.)$ and $G(.)$ and their corresponding distributions (3) as well as other related quantities (4) are indexed by the superscripts $f$ and $g$, respectively.

**Volume of the admissible excursion sets** The overall idea is to define the volume of the excursion set below the current best point while accounting for constraints.

First, we need to restrict the current best solution $f_n^{\min}$ to the set of **admissible** points, that is:

$$f_n^{\min} = \min_{\text{s.t. } g_i \leq T} f_i.$$

If no existing point is feasible, we set $f_n^{\min} = +\infty$. Now, with an explicit admissible domain, say $C \in \mathbb{X}$, the uncertainty measure is simply the volume of "admissible excursion" below $f_n^{\min}$:

$$ev_n = \int_C \mathbb{P}_n\left(F(\mathbf{x}) \leq f_n^{\min}\right) d\mathbf{x}.$$

Here, the admissible domain is not known beforehand, but it can be inferred using the GP model of $g$. However, instead of defining an approximation for $C$, we use the probability that a point $\mathbf{x}$ is admissible, that is: $\mathbb{P}_n\left(G(\mathbf{x}) \leq T\right)$. By independence of $F$ and $G$, the probability that $\mathbf{x}$ is below $f_n^{\min}$ *and* admissible is $\mathbb{P}_n\left(F(\mathbf{x}) \leq f_n^{\min}\right) \times \mathbb{P}_n\left(G(\mathbf{x}) \leq T\right)$.

Hence, the expected volume of admissible excursion below the current minimum $f_n^{\min}$ is equal to:

$$\begin{aligned}
ev_n &= \mathbb{E}_{\mathbb{X}}\left[\mathbb{P}_n\left(F(\mathbf{x}) \leq f_n^{\min}\right)\mathbb{P}_n\left(G(\mathbf{x}) \leq T\right)\right]\\
&= \mathbb{E}_{\mathbb{X}}\left[p_n^f(\mathbf{x}, f_n^{\min}) p_n^g(\mathbf{x}, T)\right],
\end{aligned}$$

with $p_n$ as defined in (4).

**Volume update** As in the unconstrained case, the expectation of $EV_{n+1}$ defines the SUR criterion:

$$EEV(\mathbf{x}_{n+1}) = \mathbb{E}_n\left[EV_{n+1} | F_{n+1}, G_{n+1}\right].$$

When a new observation is performed at $\mathbf{x}_{n+1}$, both $F_{n+1}$ and $G_{n+1}$ are observed. Two cases are to be considered:

- $\mathbf{x}_{n+1}$ is admissible ($G_{n+1} \leq T$): the current minimum may or may not change, as: $f_{n+1}^{\min} = \min(f_n^{\min}, F(\mathbf{x}_{n+1}))$.
- $\mathbf{x}_{n+1}$ is not admissible ($G_{n+1} > T$): the current minimum remains unchanged : $f_{n+1}^{\min} = f_n^{\min}$.

We define the four following quantities, depending on the two cases:

$$\begin{aligned}
p_-^f(\mathbf{x}) &:= \mathbb{P}_n\Big[F(\mathbf{x}) \leq F_{n+1}^{\min}\Big| F_{n+1}, G_{n+1} \leq T\Big],\\
p_+^f(\mathbf{x}) &:= \mathbb{P}_n\Big[F(\mathbf{x}) \leq F_{n+1}^{\min}\Big| F_{n+1}, G_{n+1} > T\Big],\\
p_-^g(\mathbf{x}) &:= \mathbb{P}_n\Big[G(\mathbf{x}) \leq T\Big| G_{n+1} \leq T\Big]\mathbb{P}_n\left[G_{n+1} \leq T\right],\\
p_+^g(\mathbf{x}) &:= \mathbb{P}_n\Big[G(\mathbf{x}) \leq T\Big| G_{n+1} > T\Big]\mathbb{P}_n\left[G_{n+1} > T\right].
\end{aligned}$$

We have, by independence of $F$ and $G$ and law of total probability:

$$\mathbb{E}\left[P_{n+1}^f(\mathbf{x}) P_{n+1}^g(\mathbf{x})\right] = p_-^f(\mathbf{x}) p_-^g(\mathbf{x}) + p_+^f(\mathbf{x}) p_+^g(\mathbf{x}),$$

which will allow us to compute our criterion.

**$p_-^f(\mathbf{x})$ and $p_+^f(\mathbf{x})$ calculations:** If the new point is admissible, $p_-^f(\mathbf{x})$ writes as in the unconstrained case (Equation 8):

$$\begin{aligned}
p_-^f(\mathbf{x}) &= \mathbb{P}_n\left[F(\mathbf{x}) \leq \min(f_n^{\min}, F_{n+1}) | F_{n+1}\right]\\
&= \mathbf{\Phi}_{\nu}^f\left(\overline{f}_n^{\min}, \eta^f\right) + \mathbf{\Phi}_{-\rho}^f\left(-\overline{f}_n^{\min}, \widetilde{f}_n^{\min}\right),
\end{aligned}$$

with quantities as defined in (8).

Otherwise, the new point is not admissible, and we have:

$$p_+^f(\mathbf{x}) = \mathbb{P}_n\left[F(\mathbf{x}) \leq f_n^{\min} | F_{n+1}\right].$$

Using the following proposition (see Picheny (2013) for proofs):

$$\mathbb{P}_n\left[F(\mathbf{x}) \leq a \middle| F_{n+1}\right] = p_n\left(\mathbf{x}, a\right), \qquad (10)$$

we obtain:

$$p_+^f(\mathbf{x}) = p_n^f(\mathbf{x}, f_n^{\min}).$$

**$p_-^g(\mathbf{x})$ and $p_+^g(\mathbf{x})$ calculations:** Similarly to (6), we have (see Picheny (2013) for proofs):

$$\mathbb{P}_n\left[G(\mathbf{x}) \leq a \middle| G_{n+1} \leq a\right] \mathbb{P}_n\left[G_{n+1} \leq a\right] = \mathbf{\Phi}_\rho\left(\bar{a}, \tilde{a}\right),$$

hence:

$$p_-^g(\mathbf{x}) = \mathbf{\Phi}_\rho^g\left(\bar{T}, \tilde{T}\right),$$

with:

$$\bar{T} = \frac{T - m_n^g(\mathbf{x}_{n+1})}{s_n^g(\mathbf{x}_{n+1})} \text{ and } \tilde{T} = \frac{T - m_n^g(\mathbf{x})}{s_n^g(\mathbf{x})}.$$

$p_+^g(\mathbf{x})$ can be deduced from $p_-^g(\mathbf{x})$ by the law of total probability and the use of (10):

$$\begin{aligned} p_-^g(\mathbf{x}) + p_+^g(\mathbf{x}) &= \mathbb{P}_n\left(G(\mathbf{x}) \leq T | G_{n+1}\right) \\ &= p_n^g(\mathbf{x}, T) \end{aligned}$$

Finally, we obtain:

$$EEV(\mathbf{x}_{n+1}) = \int_{\mathbb{X}} \left[p_-^f(\mathbf{x}) p_-^g(\mathbf{x}) + p_+^f(\mathbf{x}) p_+^g(\mathbf{x})\right] d\mathbf{x},$$

which is equal, noting that $p_n(\mathbf{x}, a) = \Phi(\bar{a})$, to:

$$\begin{aligned} & EEV(\mathbf{x}_{n+1}) \\ = \; & \textstyle\int_{\mathbb{X}} \Big( \Big[\mathbf{\Phi}_\nu^f\left(\overline{f}_n^{\min}, \eta^f\right) + \mathbf{\Phi}_\rho^f\left(-\overline{f}_n^{\min}, \widetilde{f}_n^{\min}\right)\Big] \\ \times \; & \mathbf{\Phi}_\rho^g\left(\bar{T}, \tilde{T}\right) + \Phi(\bar{f}_n^{\min})\left[\Phi(\bar{T}) - \mathbf{\Phi}_\rho^g\left(\bar{T}, \tilde{T}\right)\right] \Big) d\mathbf{x}. \end{aligned}$$

# 6 COMMENTS

**Implementation** When $\mathbb{X}$ is not finite, $EEV(\mathbf{x}_{n+1})$ can only be computed approximately using numerical integration over $\mathbb{X}$, for instance using Gaussian quadrature or Monte-Carlo methods. This aspect can be limiting, in particular in high dimension where a large number of integration points would be required to achieve sufficient accuracy, making the procedure overly costly. Note however that the bivariate normal CDF can be computed very quickly using efficient programs, such as the R package `pbivnorm` (Kenkel, 2012).

At each step $n$, searching for the best candidate point (9) may be done using an internal optimization loop, which is typical in GP-based optimization (Jones, 2001). Global optimization algorithms may be used (e.g. population-based), as the criterion is likely to be a multimodal function.

As a stopping criterion for the algorithm, a natural choice is the difference $ev_n - EEV(\mathbf{x}_{n+1})$, that indicates the potential gain of another iteration.

**Convergence** By sequentially sampling at the point that provides the best reduction (in expectation) of the volume of the excursion set below the current minimum, we assume that this volume eventually tends to zero, meaning that the true optimum is observed and the conditional GP variance tends to zero. In practice, we observe (section 7.2) that our algorithm converges to the optimum with competitive speed compared to existing alternatives. However, as for now, we do not provide theoretical guarantees convergence.

In general, few results have been established on the convergence of GP-based optimization. Recently, Vazquez & Bect (2010) and Grünewälder et al. (2010), followed by Bull (2011), addressed the convergence of the EGO algorithm. Srinivas et al. (2012) provide regret bounds for the sequential maximization of the probability of improvement. These approaches seem difficult to transpose here, making the question of proof of convergence out of reach of the present work.

**Advantages and limits** The proposed criterion defines a trade-off betweeen exploration of the design space and sampling intensification in promising regions. The risk of sampling in unfeasible regions is here accounted for automatically, without resorting to heuristic strategies and without parameters to tune.

Besides limitations due to the numerical integration, the method proposed here relies on Gaussian process modeling, hence may be efficient only on certain classes of functions: continuous, of small to moderate dimension, etc. This issue might be overcome by applying our strategy with other metamodeling approaches, in the spirit of Chipman et al. (2012) for instance.

# 7 EXPERIMENTS

## 7.1 Illustration

First, we illustrate the strategy on a two-dimensional example. Both $f$ and $g$ are realizations of stationary GPs with Matérn covariance (Rasmussen & Williams,

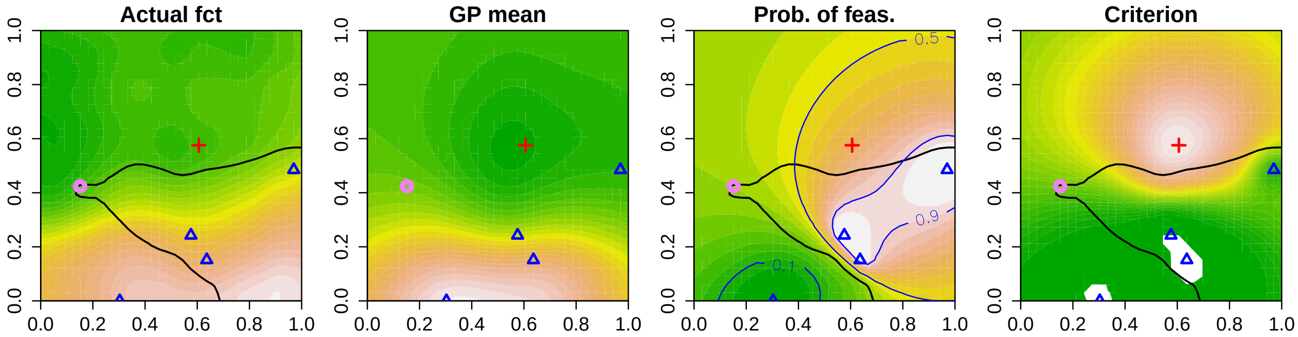

Figure 2: Left to right: actual function and feasibility domain (black curve); predictive mean of $F$; probability of feasability ($p_n(\mathbf{x}, T)$); SUR criterion ($EEV(\mathbf{x}_{n+1})$). Circles: actual optimum; triangles: initial set of observations; red crosses: new chosen observation.

2006, chap.4) with regularity parameter $\nu = 5/2$, indexed by a $34 \times 34$ regular grid on $[0, 1]^2$. For $F$, the covariance range is taken as $\theta_F = \sqrt{2}/4$ and for $G$ $\theta_G = \sqrt{2}/3$. We choose $T$ such that 30% of the design space is feasible. Both krigings are based on 6 points randomly chosen on the grid. Figure 2 shows the actual function and constraint limit, as well as the important initial quantities given by the GP models. The first added observation can be seen as an exploration step.

Figure 3 shows the experimental set after 10 iterations of the SUR algorithm. Only two observations are far from the actual feasibility boundary. Five observations form a cluster around the actual optimum. The actual feasible region is well identified.

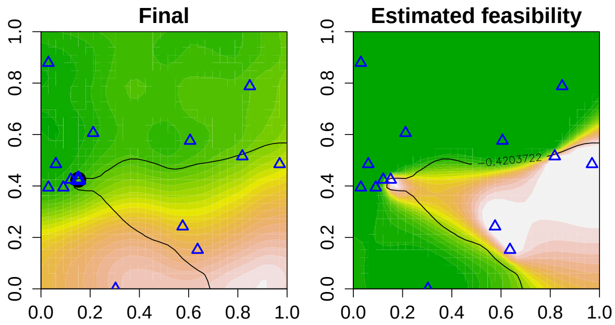

Figure 3: Observation set after 10 iterations and predicted feasibility.

### 7.2 Comparison to existing GP-based methods

Here, we compare our strategy to two strategies. First, the state-of-the-art EGO algorithm is considered, modified to account for constraints as in Schonlau et al. (1998). The second strategy is based on the IECI criterion (Integrated Expected Conditional Improvement), as proposed in Gramacy & Lee (2011).

The test functions $F$ and $G$ are realizations of Gaussian processes indexed by discrete sets. Considering discrete problems allows us to get rid of the integration and optimization issues, as the integrals over the set can be computed exactly (as the sum of the integrand over all elements) and the search for the maximizer of the criterion can be performed by evaluating it over all the elements. Using GP test functions allows us to vary space dimension and objective function activity easily. It also allows us to compare the strategies under ideal conditions, as there is no modeling error.

For both $F$ and $G$, the covariance functions are the Matérn ones with regularity parameter $\nu = 5/2$. The range for $F$ is taken as $\theta_F = \sqrt{d}/10$. For the constraint, we consider two cases: a moderately difficult constraint, with a low activity function modeled by a $G$ range of $\theta_G = \sqrt{d}/5$ and a threshold $T$ chosen such that 25% of the initial domain is feasible, and a difficult constraint with a high activity ($\theta_G = \sqrt{d}/10$) and a small feasible domain (10% of the initial domain).

We consider two-dimensional and four-dimensional cases. For 2D, the finite set is a regular grid of size $37 \times 37$; the initial experimental set consists of four points randomly chosen, and 36 experiments are added sequentially. For 4D, the finite set is chosen as 2,000 points taken from the Sobol sequence; the initial experimental set consists of 10 points randomly chosen, and 80 experiments are added sequentially. The results are given in terms of convergence of the objective function: $f_n^{\min} - f(\mathbf{x}^*)$ ($\mathbf{x}^*$ being the actual minimum) in Figures 4 and 5. The curves show the 10%, 50% and 90% percentiles over 100 repetitions.

On the 2D problems, for both setups, the SUR strategy clearly outperforms the two other methods. For the easy constraint case, the EI strategy works similarly in terms of mean performance, but is substantially poorer in terms of 90% percentile (which indicates that for several runs the strategy converges to

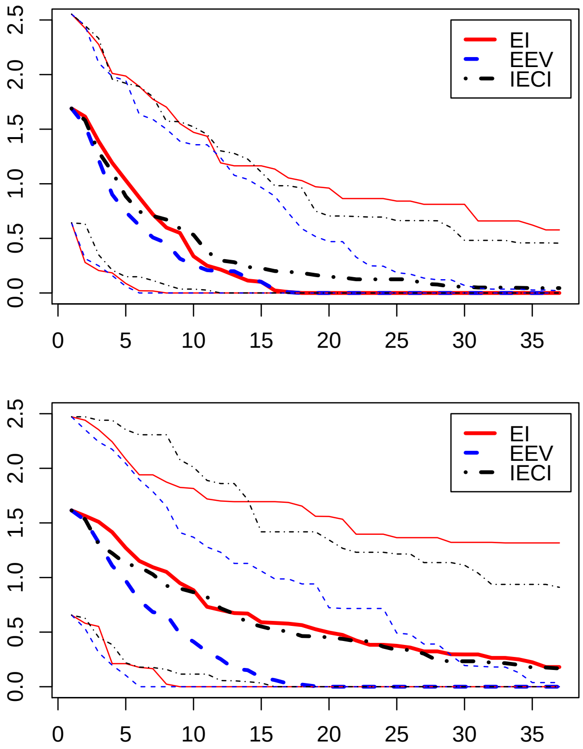

Figure 4: Performance of the different methods on the 2D problems. The curves show the 10%, 50% and 90% percentiles of $f_n^{\min} - f(\mathbf{x}^*)$ as a function of the iteration number over 50 runs. Top: easy constraint; bottom: difficult constraint.

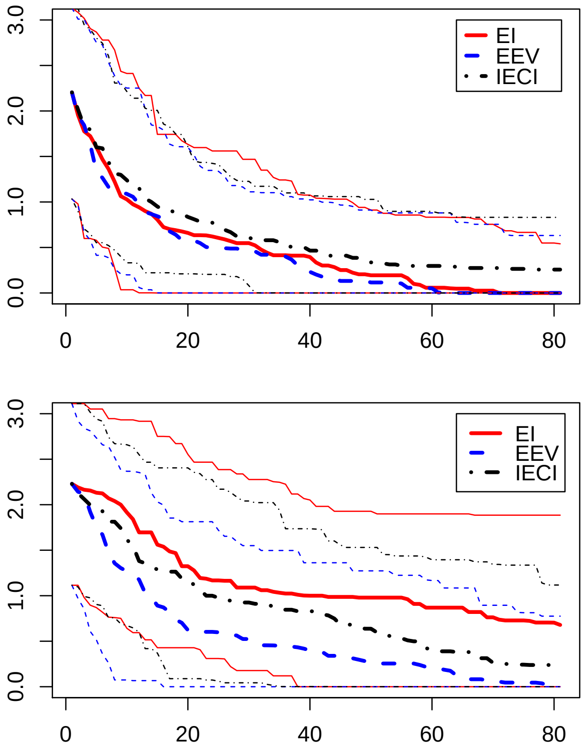

Figure 5: Performance of the different methods on the 4D problems. X-axis: iteration; y-axis: $f_n^{\min} - f(\mathbf{x}^*)$. Top: easy constraint; bottom: difficult constraint.

a local optimum). The IECI strategy converges more slowly. The difference is more clear for the difficult constraint, as the SUR strategy converges at the same rate as for the easy case, while the other methods are more affected.

On the 4D problems, for the simpler constraint, EI and EEVR have comparable performances, while IECI is slower. For the difficult constraint, EEVR works best and EI is the most penalized method.

### 7.3 Comparison to classical algorithms

Finally, we compare our approach to non-GP-based alternatives on a problem taken from Parr et al. (2012). The objective is taken as

$$f(x_1,x_2) = \left(x_2 - \frac{5.1x_1^2}{4\pi^2} + \frac{5x_1}{\pi} - 6\right)^2 + 10\left(\left(1-\frac{1}{8\pi}\right)\cos(x_1)+1\right) + \frac{5x_1+25}{15}$$

with $x_1 \in [-5,10]$ and $x_2 \in [0,15]$. The constraint is

$$g(x_1,x_2) = \Big(4-2.1x_1^2+\frac{1}{3}x_1^4)x_1^2 + x_1x_2 + (4x_2^2-4)x_2^2 + 3\sin(6(1-x_1)) + 3\sin(6*(1-x_2))\Big),$$

with $x_1 \in [-1,1]$ and $x_2 \in [-1,1]$. The constraint is satisfied if $g(x_1,x_2) \geq 6$. The objective function is relatively smooth, with two local minima and one global minimum, but the constraint is highly multimodal and the admissible space consists of three narrow regions, which makes this problem challenging. The optimization problem is represented in Figure 6.

Here, for our approach the initial design set consists of a 8-point Latin hypercube design and 22 points are added sequentially to the design. A Matérn covariance with $\nu = 5/2$ is chosen, with covariance parameters estimated using maximum likelihood using the `R` package `DiceKriging` (Roustant et al., 2012). As alternatives, we use the function `sumt` of the `R` package `clue` (Hornik, 2005), which solves constrained problems by the L-BFGS-B method with (adaptive) penalization, and the Matlab®routine `fmincon` (based on interior point method). As a measure of performance, we observe if the best design found is in one of the three feasible regions, or if no feasible point is found. Each

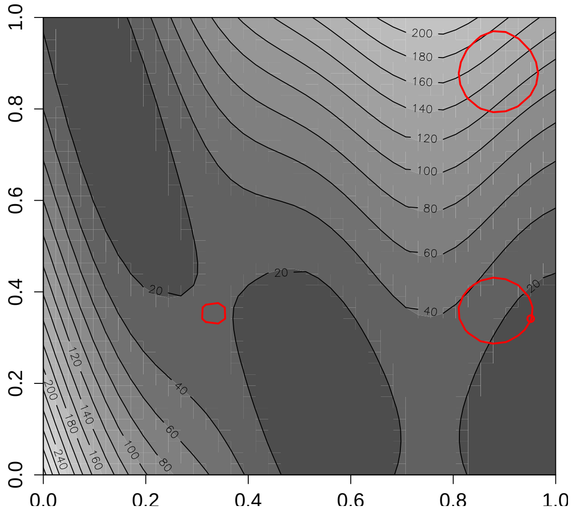

Figure 6: Objective function and contour lines (red) defining the admissible space.

method is run 100 times (changing the initial design for EEV and the starting point for `sumt` and `fmincon`) The resuls are reported in Table 1 in terms of percentages over 100 restarts for all the methods. In addition, Figure 7 shows the evolution of the percentages for the EEV criterion.

After 22 iterations of our approach, all the runs have converged to a feasible design, which is almost always in the global optimum region. Both non-GP based methods have a large percentage of failed runs (33% and 42%, respectively), and a majority of runs identify only a local optimum.

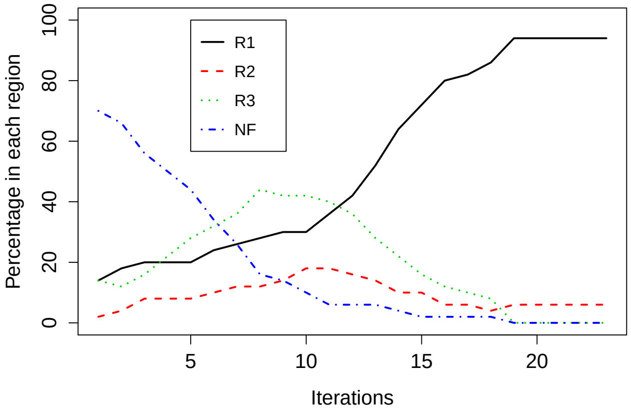

Figure 7: Evolution of the percentage of optima found in each region for the EEV criterion. R1 is the global optimum region, R2 and R3 are local optima regions and NF means non feasible.

| Method | R1 | R2 | R3 | Non feasible |
|---|---|---|---|---|
| EEV (12 ite.) | 42% | 16% | 36% | 6% |
| EEV (22 ite.) | 94% | 6% | 0% | 0% |
| `fmincon` | 17% | 34% | 16% | 42% |
| `sumt` | 15% | 33% | 19% | 33% |

Table 1: Location of the optimum found by the different methods. Region 1 (R1) corresponds to the global optimum, region 3 to the worst local optimum.

## 8 CONCLUSION

In this paper, we proposed a stepwise uncertainty reduction approach to address constrained optimization problems. We compared our approach to two GP-based alternatives using realizations of Gaussian processes as test problems, and to a classical non-GP approach. The proposed method compared favorably with alternatives, converging faster and more robustly to the optimum, in particular when the constraint function is complex.

The efficiency of the method lies in the ability of the GP model to approximate the objective and constraint functions. This implies some limits of the method in terms of problem dimensionality, functions regularity or minimal sample size. In addition, we stress here that the overall procedure (GP model fitting and update, iterative maximization of a criterion based on numerical integration) is computationally costly, making its use relevant for expensive computer models only.

Further developments may include studying of the method efficiency on actual engineering problems, accounting for correlations between objective and constraints, and addressing alternative optimization cases, such as boolean constraints and hidden constraints.

Finally, the issue of theoretical guarantees of our approach has been left aside here. Future work may address this important problem.

## References

Audet, C., Booker, A. J., Dennis Jr, J., Frank, P. D. & Moore, D. W. (2000). A surrogate-model-based method for constrained optimization. In *8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, September 6-8, Long Beach, CA*. AIAA.

Bect, J., Ginsbourger, D., Li, L., Picheny, V. & Vazquez, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing* **22**, 773–793.

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *The Journal of Machine Learning Research* **12**, 2879–2904.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V. & Richet, Y. (2012). Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. `http://hal.inria.fr/hal-00641108/en`.

Chipman, H., Ranjan, P. & Wang, W. (2012). Sequential design for computer experiments with a flexible bayesian additive model. *Canadian Journal of Statistics* **40**, 663–678.

Craig, P. S., Goldstein, M., Rougier, J. C. & Seheult, A. H. (2001). Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association* **96**, 717–729.

Cressie, N. (1993). *Statistics for Spatial Data, revised edition*, vol. 928. Wiley, New York.

Gramacy, L. & Lee, H. (2011). Optimization under unknown constraints. *Bayesian Statistics* **9**, 229.

Grünewälder, S., Audibert, J.-Y., Opper, M. & Shawe-Taylor, J. (2010). Regret bounds for gaussian process bandit problems. In *13th International Conference on Artificial Intelligence and Statistics (AISTATS), May 13-15, Italy.*

Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software* **14**.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization* **21**, 345–383.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization* **13**, 455–492.

Kenkel, B. (2012). *pbivnorm: Vectorized Bivariate Normal CDF.* R package version 0.5-1.

Kennedy, M. C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 425–464.

Parr, J., Holden, C. M., Forrester, A. I. & Keane, A. J. (2010). Review of efficient surrogate infill sampling criteria with constraint handling. In *2nd international conference on engineering optimization, September 6-9, Lisbon, Portugal.*

Parr, J., Keane, A., Forrester, A. & Holden, C. (2012). Infill sampling criteria for surrogate-based optimization with constraint handling. *Engineering Optimization* **44**, 1147–1166.

Picheny, V. (2013). Multiobjective optimization using gaussian process emulators via stepwise uncertainty reduction. *arXiv preprint arXiv:1310.0732* .

Rasmussen, C. & Williams, C. (2006). *Gaussian processes for machine learning.* MIT Press.

Roustant, O., Ginsbourger, D. & Deville, Y. (2012). Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software* **51**, 1–55.

Sasena, M. J., Papalambros, P. & Goovaerts, P. (2002). Exploration of metamodeling sampling criteria for constrained global optimization. *Engineering optimization* **34**, 263–278.

Schonlau, M., Welch, W. J. & Jones, D. R. (1998). Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series* , 11–25.

Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *Information Theory, IEEE Transactions on* **58**, 3250–3265.

Stein, M. (1999). *Interpolation of spatial data: some theory for kriging.* Springer Verlag.

Vazquez, E. & Bect, J. (2010). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference* **140**, 3088–3095.

Villemonteix, J., Vazquez, E. & Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization* **44**, 509–534.