A GRADUAL SENSITIVITY-BASED KERNEL TO IMPROVE BAYESIAN OPTIMIZATION

Lise Kastner, Bertrand Cuissart, Jean-Luc Lamotte

University of Normandy, GREYC UNICAEN, CODAG, 14000 Caen, FRANCE

lise.kastner@unicaen.fr, bertrand.cuissart@unicaen.fr, jean-luc.lamotte@unicaen.fr

GREYC Electronics and Computer Science Laboratory

Context

GOAL → identify settings that lead to productive outcomes, in a limited number of trials: not possible to test many settings (cost/time)

METHOD → integration of Sensitivity analysis into Bayesian Optimization

1. Bayesian optimization (BO)

A solution is to use **Bayesian optimization**:

1. Model a **Gaussian Process** (GP) on the observations, calculating the **mean** μ and **standard deviation** σ, for each setting e

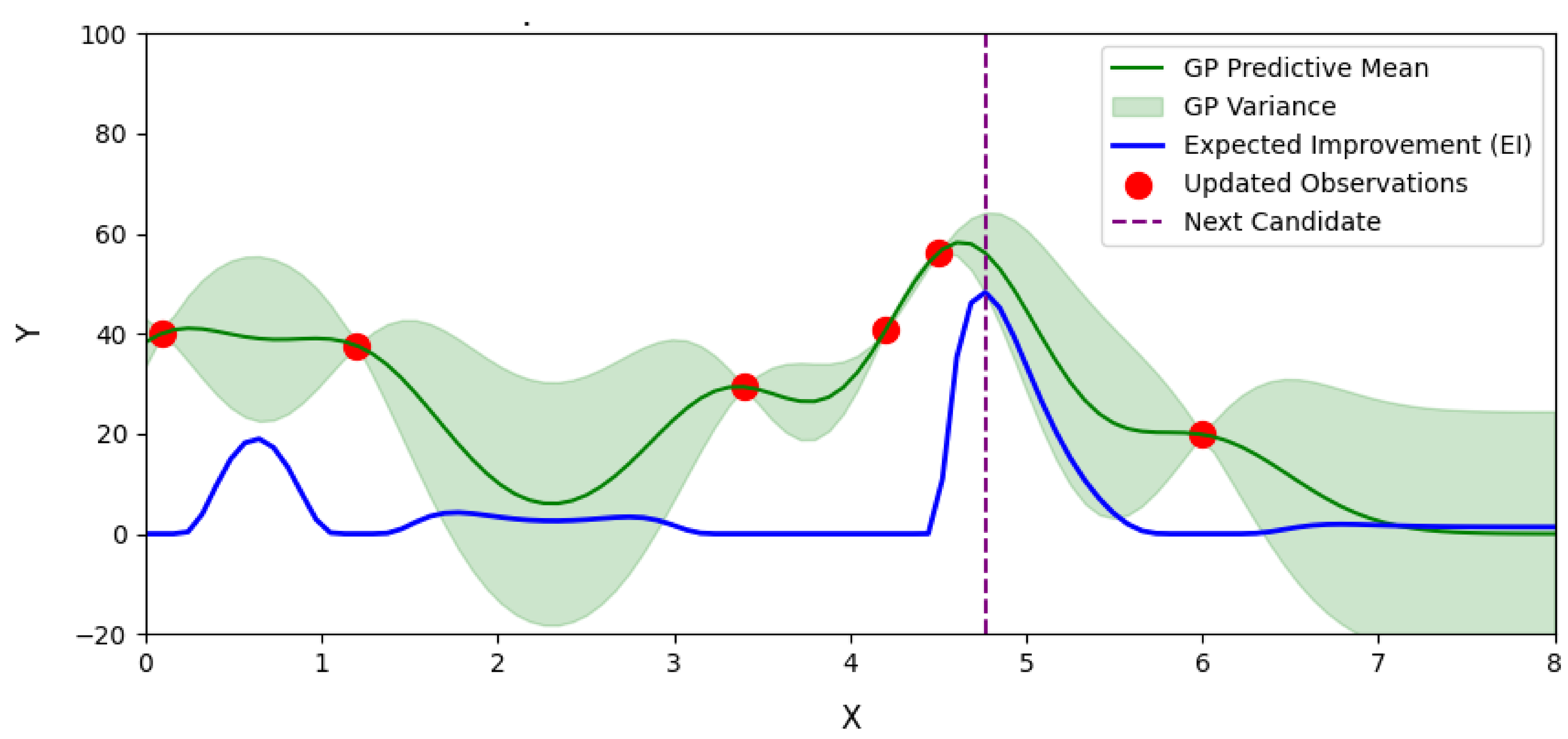
2. The GP is defined by a **kernel covariance** function $k(e, e')$, which is a similarity distance between 2 settings e and e'
   — **Automatic Relevance Determination kernel** :

$$k_{RBF_{ARD}}(e, e') = \exp\left(-\frac{1}{2}\sum_{k=1}^{m}\frac{\alpha \cdot (e_k - e'_k)^2}{l_k^2}\right)$$

   where :
   — $\alpha = 1$ for the baseline kernel
   — l_k is the length scale associated to the X_k parameter,
   — m is the total number of parameters.

3. The next setting is determined by maximizing an **acquisition function** based on 2 strategies :
   — **Exploration** : minimize the predictive variance
   — **Exploitation** : maximize the predictive mean

   The **Expected improvement** function combines the 2 strategies.



⇒ BO is effective but there is a need to accelerate the active learning process due to resource constraints.

2. Sensitivity analysis

Sensitivity indices: quantifies input parameters influence on the output

The **HSIC** indices (*Hilbert-Schmidt Independance Criterion*) are used, a **kernel-based method** suited for problems with **limited observations**.

The HSIC index of X_i on Y measures the **dependence between** $P_Y P_{X_k}$ **and** P_{Y,X_k}, **the marginal and joint distributions** of X_k and Y.

$$HSIC(X_k, Y)_{\mathcal{F},\mathcal{G}} = \| \mu(P_Y P_{X_k}) - \mu(P_{Y,X_k}) \|$$

⇒ A greater difference between the joint and marginal distributions implies a stronger influence of X_k on Y.

3. GSBK : Gradual Sensitivity-Based Kernel

Automatic Relevance Determination kernel :

1. **Baseline**: $\alpha = 1$
2. **Sensitivity-based kernel**: $\alpha = S_k$ where S_k is the HSIC index of X_k.
   ⇒ The more sensitive a parameter is, the more it contributes to the kernel.
3. **Gradual Sensitivity-based Kernel (GSBK)**:

$$\alpha = exp(d \cdot S_k \cdot C_k)$$

   where :
   — d is the dimension
   — C_k is a **coefficient of stability** where for each variable X_k, S_{k-n} is the list of the n last estimated sensitivity indices of X_k.

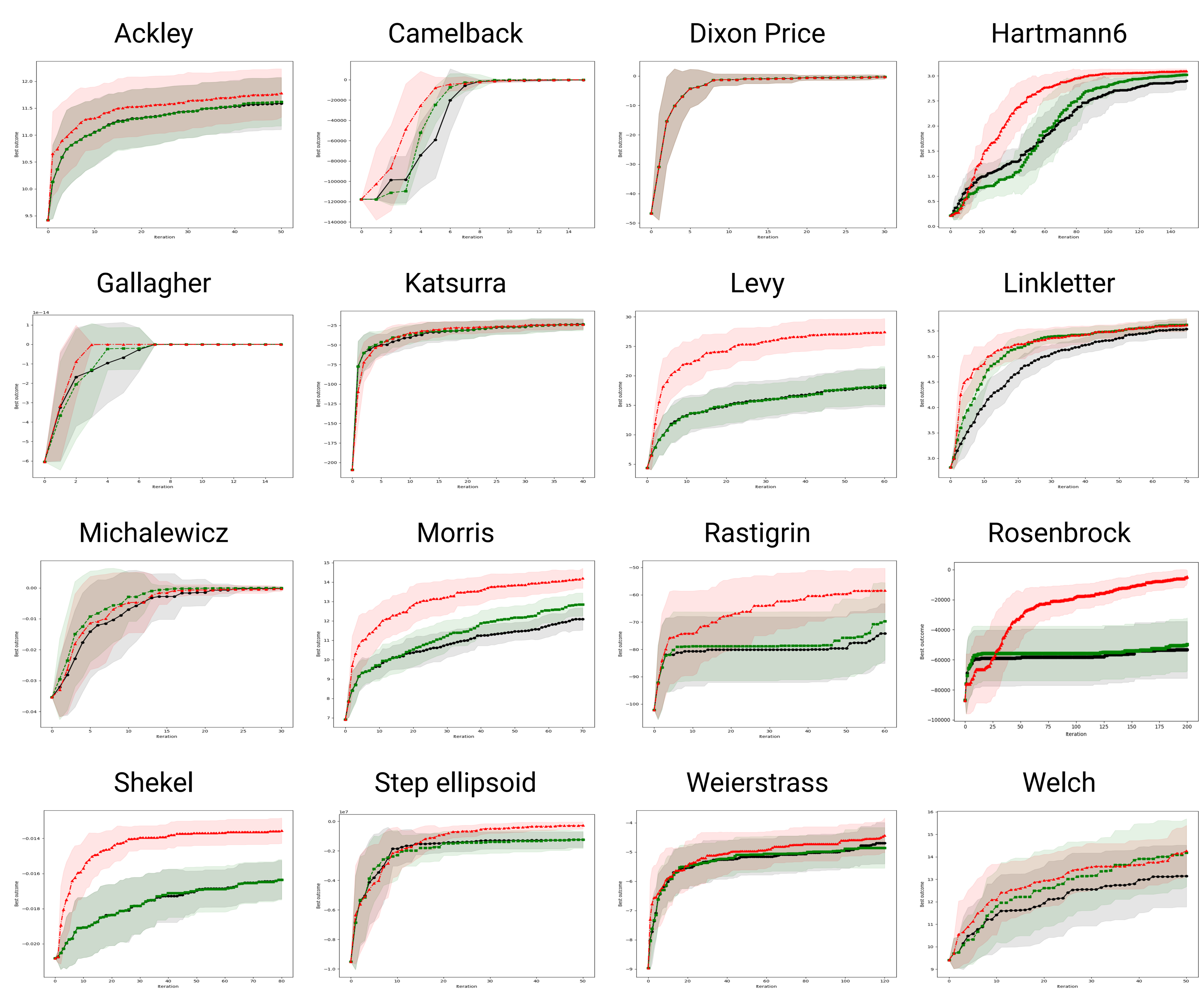
$$C_k = 1 - \frac{\sigma(S_{k-n})}{\mu(S_{k-n})}$$

   ⇒ If the estimated HSIC index of X_k is high and **reliable** enough, then X_k contributes significantly to the kernel.

4. Results

⇒ Global evaluation on benchmark functions :
— **Standard Kernel** ●
— **Sensitivity-based Kernel** ■
— **Gradual Sensitivity-based Kernel** ▲

⇒ Each point represents the mean of maximum outcome values on 100 subsets at a given iteration, surrounded by the standard deviation.



5. Conclusion

— The BO using *GSBK* outperforms the BO using the baseline or naive sensitivity-based kernel, depending on parameters influence.
— *GSBK* proves to be particularly efficient on complex problem, while the two other methods struggle to reach the optimum.

This work is funded by the **AMPERE project - CNRS 80 PRIME**