Introduction to Bayesian Optimization

and more generally to the Design & Analysis of Computer Experiments (DACE)

Julien Bect

Université Paris-Saclay, CNRS, CentraleSupélec
Laboratoire des Signaux et Systèmes (L2S), Gif-sur-Yvette, France
https://l2s.centralesupelec.fr/

GdR MASCOT-NUM, INSMI, CNRS
https://www.gdr-mascotnum.fr/

15th Peyresq Summer School
in Signal and Image Processing
Online, 2021, June 20–26

Setting: computer experiments

$x \in \mathbb{X}$

$Z \in \mathbb{R}$

- Consider a computer model for
  - a system to be designed (engineering),
  - a physical or biological phenomenon...

Setting: computer experiments

$x \in \mathbb{X}$

$Z \in \mathbb{R}$

- Consider a computer model for
  - a system to be designed (engineering),
  - a physical or biological phenomenon...
- "Computer experiment"
  - 1 experiment = run the program for some $x \in \mathbb{X}$ and obtain one output value $Z \in \mathbb{R}$ (or $\mathbb{R}^p$, or...)
  - Assumed to be time-consuming.
  - Can be deterministic or stochastic.

$Z = f(x)$ $\qquad$ $Z \sim P_x$

$f : \mathbb{X} \to \mathbb{R}$ $\qquad$ $(P_x)_{x \in \mathbb{X}}$

Setting: computer experiments

$x \in \mathbb{X}$

$Z \in \mathbb{R}$

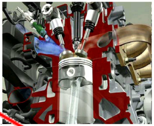
- Consider a computer model for
  - a system to be designed (engineering),
  - a physical or biological phenomenon...
- "Computer experiment"
  - 1 experiment = run the program for some $x \in \mathbb{X}$ and obtain one output value $Z \in \mathbb{R}$ (or $\mathbb{R}^p$, or...)
  - Assumed to be time-consuming.
  - Can be deterministic or stochastic.
- Statistical tasks (DACE)
  - Design: choose x_1, x_2, ... ✓
  - Analysis: process the results Z_1, Z_2, ... ✓

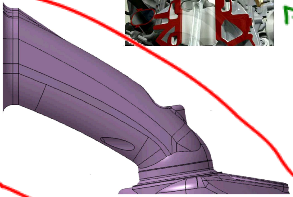
    $\Rightarrow$ various possible goals

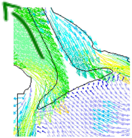
Example 1: intake port design (Renault)

Context: automotive industry

- intake port design
- complex simulation chain
  (3D CAD, meshing, PDE solving)
- source: PhD thesis of
  Villemonteix (2008)







Goal: bi-objective optimization

- maximize engine performance
- minimize emission of pollutants

Features

- several hours / computation
  on dedicated high-end servers
- $\sim$ 5–10 geometric parameters to optimize

$x \in \mathbb{X} \subset \mathbb{R}^d$, $5 \le d \le 10$

Example 2: the BEMUSE case

Context : nuclear safety

- loss-of-coolant accident (LOCA)
  $\rightsquigarrow$ thermal-hydraulic computations
- BEMUSE: international benchmark
  (de Crécy et al., NED, 2008)
- software: CATHARE
  (CEA, IRSN, EDF, FRAMATOME)

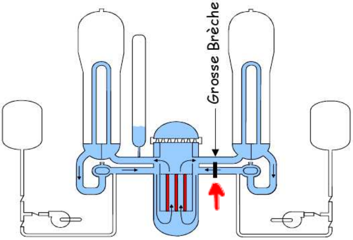
Features

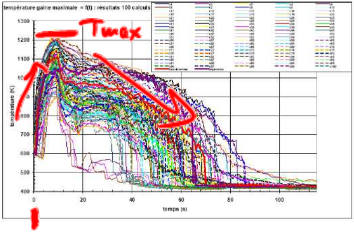
- QoI: maximal temperature T_{max}
- $\approx$ 10 minutes / computation
- 53 uncertain parameters ($\to$ random)

$x \in \mathbb{R}^{53}$

Some possible goals

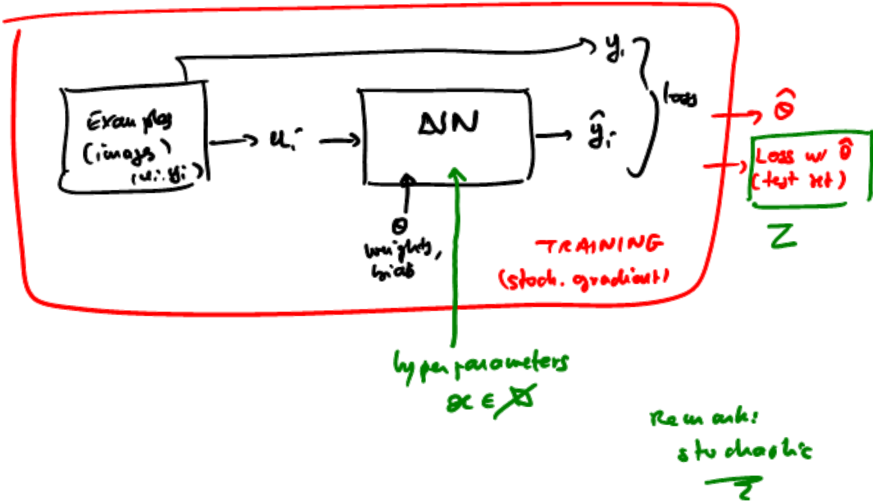
- estimate a quantile of T_{max}
- sensitivity analysis





(B. Iooss, J. Nat. Fiabilité, 2010)

Example 3: hyper-parameter tuning in ML



Exploratory designs

- Space-filling designs: "filling" the input space $\mathbb{X} \subset \mathbb{R}^d$
  - various criteria (distance-based, discrepancies, etc.)
  - full space vs low-dimensional projections

Exploratory designs

- Space-filling designs: "filling" the input space $\mathbb{X} \subset \mathbb{R}^d$
  - various criteria (distance-based, discrepancies, etc.)
  - full space vs low-dimensional projections
- Example: maximin Latin Hypercube Designs (maximin LHDs)

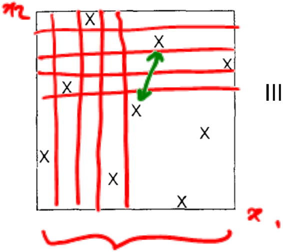
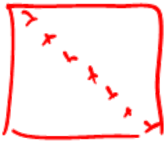


Illustration from Morris & Mitchell (1995):
a maximin LHD in $[0,1]^2$, size $n = 9$



Exploratory designs

- Space-filling designs: "filling" the input space $\mathbb{X} \subset \mathbb{R}^d$
  - various criteria (distance-based, discrepancies, etc.)
  - full space vs low-dimensional projections

- Example: maximin Latin Hypercube Designs (maximin LHDs)

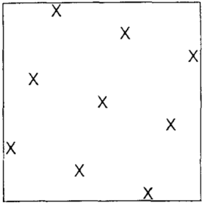


Illustration from Morris & Mitchell (1995):
a maximin LHD in $[0,1]^2$, size $n = 9$

$f : \mathbb{X} \to \mathbb{R}$

$\hat{f} : \mathbb{X} \to \mathbb{R}$ $\quad \hat{f} \approx f$

- Suitable for global approximation
  - a.k.a. "meta-modeling", a.k.a. "surrogate modeling"...

Sequential designs

- "Localized" quantities of interest, e.g.:
  - Optimization: minima and/or minimizers, Pareto set...
  - Reliability: level sets, probabilities of failure, quantiles...

Sequential designs

- "Localized" quantities of interest, e.g.:
  - Optimization: minima and/or minimizers, Pareto set...
  - Reliability: level sets, probabilities of failure, quantiles...

- Local knowledge through sequential design (a.k.a. active learning)

$\mathcal{X} \subset \mathbb{R}^2$

✏ Start from an initial space-filling DoE of size n_0 (here $n_0 = 9$)

✏ Choose X_{n_0+1} using $Z_1, \ldots, Z_{n_0}$

✏ Choose X_{n_0+2} using $Z_1, \ldots, Z_{n_0+1}$

✏ ...

- Fully-sequential versus batch-sequential design

The Bayesian approach to sequential design

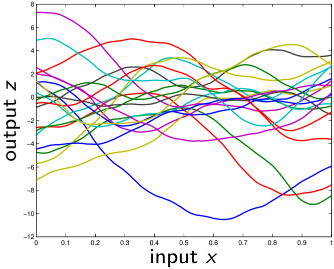
- Probabilistic modeling of knowledge / uncertainty
  - Prior knowledge about the computer model $\rightsquigarrow$ prior distrib. P_0
  - Posterior distrib. P_n, $\mathrm{P}_{n+1}\ldots$ $\rightsquigarrow$ used to select X_{n+1}, $X_{n+2}\ldots$

The Bayesian approach to sequential design

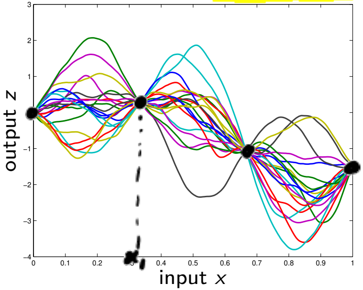
- Probabilistic modeling of knowledge / uncertainty
  - Prior knowledge about the computer model $\rightsquigarrow$ prior distrib. P_0
  - Posterior distrib. P_n, $\mathrm{P}_{n+1} \ldots \rightsquigarrow$ used to select X_{n+1}, $X_{n+2} \ldots$

$\mathbb{X} \subset \mathbb{R}^1$

Prior distribution



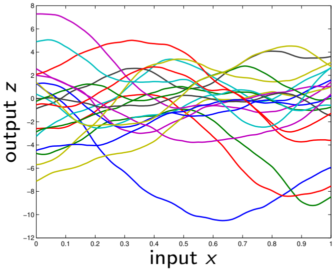
Posterior after $n = 4$ runs



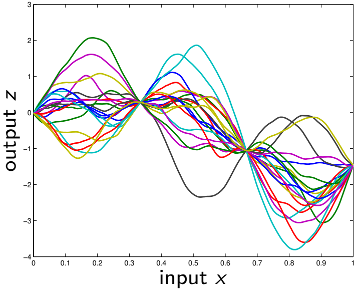
The Bayesian approach to sequential design

- Probabilistic modeling of knowledge / uncertainty
  - Prior knowledge about the computer model $\rightsquigarrow$ prior distrib. P_0
  - Posterior distrib. $\mathrm{P}_n, \mathrm{P}_{n+1} \ldots \rightsquigarrow$ used to select $X_{n+1}, X_{n+2} \ldots$

Prior distribution



Posterior after $n = 4$ runs



- Prior on unknown function $\Rightarrow$ non-parametric Bayes
  - Notation: ξ = random function that represents the unknown f

Gaussian processes

REAL - VALUED

- Definition: $\xi \sim \mathcal{GP}(m, k)$ if, $\forall n \geq 1$, $\forall x_1, \ldots, x_n \in \mathbb{X}$,

$$\begin{pmatrix} \xi(x_1) \\ \vdots \\ \xi(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & k(x_i, x_j) & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{pmatrix} \right)$$

- Terminology
  - $m : \mathbb{X} \to \mathbb{R}$: mean function
  - $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$: covariance function
- Existence (Daniell-Kolmogorov theorem)
  - $\xi \sim \mathcal{GP}(m, k)$ iff k is symmetric and positive-definite

Gaussian processes (cont'd)

- Simplified notations: $\xi(\underline{x}_n) \sim \mathcal{N}\left(m(\underline{x}_n),\ k(\underline{x}_n, \underline{x}_n)\right)$

$$\underline{x}_n = (x_1, \dots x_n) \in \mathbb{X}^n$$

$$\xi(\underline{x}_n) = \begin{pmatrix} \xi(x_1) \\ \vdots \\ \xi(x_n) \end{pmatrix} \qquad k(\underline{x}_n, \underline{x}'_n) = \left(k(x_i, x'_j)\right)_{i,j}$$

- Teminology from geostatistics
  - $m \equiv 0$ (or known mean): simple kriging ✓
  - $m \equiv \mu \in \mathbb{R}$, $\mu \sim \mathcal{U}_{\mathbb{R}}$: ordinary kriging ✓
  - $m = \sum_j \beta_j \varphi_j$, $\beta_j \overset{\text{iid}}{\sim} \mathcal{U}_{\mathbb{R}}$: universal kriging ✓
- Remark: complex-valued GPs can be defined too.

Posterior distribution

- Assume $m \equiv 0$ (simple kriging) for simplicity, and

$$Z_i \;=\; \xi(x_i) + \tau_i U_i, \quad U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad 1 \le i \le n.$$

- then...

$$\boxed{\xi \mid Z_1, \dots Z_n \;\sim\; GP\left(\hat{\xi}_n, k_n\right)}$$

with

$$\begin{cases} \hat{\xi}_n(x) = \lambda_n^T(x)\, Z_n & Z_n = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \\ k_n(x,x') = k(x,x') - \lambda_n(x)^T (K_n + \Delta_n)^{-1} \lambda_n(x') \end{cases}$$

$$\lambda_n(x) = (\underbrace{K_n}_{k(x_n,\, x_n)} + \underbrace{\Delta_n}_{\operatorname{diag}(\tau_i^2)})^{-1}\, k(x_n, x)$$

- Remark: similar equations hold for ordinary & universal kriging

Posterior distribution (cont'd)

- The noiseless case ("exact measurements")
  - The equations remain valid when $\tau_i = 0$ for (some or) all i.
  - Then $\widehat{\xi}_n$ interpolates the observations.
  - Commonly used for deterministic computer experiments.

Posterior distribution (cont'd)

- The noiseless case ("exact measurements")
  - The equations remain valid when $\tau_i = 0$ for (some or) all i.
  - Then $\widehat{\xi}_n$ interpolates the observations.
  - Commonly used for deterministic computer experiments.

- More teminology from geostatistics
  - Posterior mean $\widehat{\xi}_n(x)$: kriging predictor
  - Posterior variance $\sigma_n^2(x) \triangleq k_n(x, x)$: kriging variance

Posterior distribution (cont'd)

- The noiseless case ("exact measurements")
  - The equations remain valid when $\tau_i = 0$ for (some or) all i.
  - Then $\widehat{\xi}_n$ interpolates the observations.
  - Commonly used for deterministic computer experiments.

- More teminology from geostatistics
  - Posterior mean $\widehat{\xi}_n(x)$: kriging predictor
  - Posterior variance $\sigma_n^2(x) \triangleq k_n(x, x)$: kriging variance

- $\widehat{\xi}_n$: one interpolation/regression method, many names
  - noiseless case w/ $k(x, x') = g(\|x - x'\|)$

    $\Leftrightarrow$ RBF interpolation — book: Wendland (2004)
  - noisy (regression) case

    $\Leftrightarrow$ smoothing splines — books: Wahba (1990); Gu (2013)

(Remark: the correspondance with RBF interpolation and smoothing splines holds in full if we consider the universal kriging case, with a generalized notion of—conditionally positive definite—covariance function.)

Illustration

STK demos (https://github.com/stk-kriging/stk)

`stk_example_kb01`

- Ordinary kriging in 1D, with noiseless data

`stk_example_kb01n`

- Ordinary kriging in 1D, with noisy data

`stk_example_kb03`

- Ordinary kriging in 2D

`stk_example_kb05`

- Generation of conditioned sample paths

Practical GP modeling: bird's-eye view

Practical GP modeling involves various additional steps:

- Choosing the (a family of. . .) GP model
  - mean function
  - covariance function
- Selecting ("estimating") suitable hyper-parameters
  - for the covariance function
  - for the noise model (regression case only)
  - for the mean function (if applicable)
- Assessing the goodness of fit
  - LOO cross-validation plot

Choosing the mean function

- Standard "default" choices
  - $m \equiv 0$ (simple kriging) + empirical output centering
  - $m \equiv \mu \sim \mathcal{U}_{\mathbb{R}}$: ordinary kriging $\rightarrow$ used in this lecture

Choosing the mean function

- Standard "default" choices
  - $m \equiv 0$ (simple kriging) + empirical output centering
  - $m \equiv \mu \sim \mathcal{U}_{\mathbb{R}}$: ordinary kriging → used in this lecture
- Some other possible choices (universal kriging framework)
  - polynomial trend
    - e.g., Le Riche & Picheny (2021) recommend the general use of a quadratic trend for Bayesian optimisation applications
  - periodic trend
  - multi-fidelity / calibration: using a cheap approximation

$$\xi(x) \;=\; \delta + \alpha\, f_{\text{cheap}}(x; \theta) + \xi^{\textit{centered}}(x)$$

(with δ, α, θ: hyper-parameters)

Stationary covariance functions

- stationarity: $k(x, x') = \tilde{k}(x - x'), \quad x, x' \in \mathbb{X} \subset \mathbb{R}^d$

Stationary covariance functions

- stationarity: $k(x, x') = \tilde{k}(x - x')$, $\quad x, x' \in \mathbb{X} \subset \mathbb{R}^d$
- Theorem (Bochner): k is a real, continuous and stationary covariance function iff...

$$\tilde{k}(h) = \int_{\mathbb{R}^d} e^{i\omega^T h} \mu(d\omega)$$

$\hookrightarrow$ bounded measure on $\mathbb{R}^d$
+ symmetric

- Special case: isotropic / geometrically anisotropic

$$\tilde{k}(h) = \sigma^2 \, \Omega\left(\sum_{j=1}^{d} \frac{h_j^2}{\ell_j^2} \right)$$

The Matérn family of covariance functions

- k is an isotropic Matérn covariance function if the spectral density S is of the Student-t type: $\exists \nu > 0$,

$$S(\omega) \;\propto\; \left(1 + \frac{1}{2\nu} \|\omega\|^2\right)^{-\left(\nu + \frac{d}{2}\right)}$$

Named after Bertil Matérn. Popularized by M. L. Stein (1999).

The Matérn family of covariance functions

- k is an isotropic Matérn covariance function if the spectral density S is of the Student-t type: $\exists \nu > 0$,

$$S(\omega) \;\propto\; \left(1 + \frac{1}{2\nu}\|\omega\|^2\right)^{-\left(\nu + \frac{d}{2}\right)}$$

Named after Bertil Matérn. Popularized by M. L. Stein (1999).

- Tunable regularity !

Theorem

$$\xi \sim \mathcal{GP}(0,\, \mathrm{Matern}_\nu) \text{ is } k\text{-times differentiable}$$
$$\text{in the mean-square sense iff } \nu > k.$$

(The regularity parameter can also be shown to control the smoothness of the sample paths of ξ in the scale of L^2 [redacted] spaces; cf. Scheuerer, 2010.)

SOBOLEV

The Matérn family of covariance functions (cont'd)

- Special cases
  - $\nu = \frac{1}{2}$: $r(h) = \exp(-h)$ ✓
  - $\nu = \frac{3}{2}$: $r(h) = (1+h)\exp(-h)$
  - ...
  - $\nu \to +\infty$: $r(h) \to \exp(-\frac{1}{2}h^2)$ ✓

$\nu = p + \frac{1}{2} \rightarrow$ analytical form

Range and regularity parameters: illustration

STK demos (`https://github.com/stk-kriging/stk`)

`stk_example_misc01`

- Several correlation functions from the Matern family

`stk_example_kb07`

- Simulation of sample paths with various values of ν
- Simulation of sample paths with various values of ρ

Choosing the hyper-parameters

- Most commonly used: the maximum likelihood approach
  - $\widehat{\theta}^{\mathrm{ML}} = \operatorname{argmax} \ell_n(\theta)$, where ℓ_n is the log-likelihood:

$$\begin{aligned} -2\ell_n(\theta) &= n \ln(2\pi) + \ln \det(K_n) \\ &\quad + \left(\underline{Z}_n - m(\underline{x}_n)\right)^{\mathrm{t}} \left(K_n + \Delta_n\right)^{-1} \left(\underline{Z}_n - m(\underline{x}_n)\right) \end{aligned}$$

  - with θ: all the hyper-parameters of m, k and τ^2.

Choosing the hyper-parameters

- Most commonly used: the maximum likelihood approach
  - $\widehat{\theta}^{\mathrm{ML}} = \operatorname{argmax} \ell_n(\theta)$, where ℓ_n is the log-likelihood:

$$\begin{aligned} -2\ell_n(\theta) \;=\; & n \ln(2\pi) + \ln \det(K_n) \\ & + \left(\underline{Z}_n - m(\underline{x}_n)\right)^{\mathrm{t}} \left(K_n + \Delta_n\right)^{-1} \left(\underline{Z}_n - m(\underline{x}_n)\right) \end{aligned}$$

  - with θ: all the hyper-parameters of m, k and τ^2.

- Other approaches
  - Restricted ML (ReML): supports *generalized* cov. functions
  - Hierachical Bayes, with a prior on θ
    - Maximum a posteriori (MAP)
    - Fully Bayes ($\Rightarrow$ numerical integration, e.g., MCMC or SMC)

Goodness-of-fit diagnostic: LOO-CV plots

- Set $\widehat{\xi}_n^{(-i)}(x) \triangleq \mathsf{E}\left(\xi(x) \mid Z_{i'},\, i' \neq i\right)$, for all $i \leq n$.
- LOO-CV plot: scatter plot of Z_i versus $\widehat{\xi}_n^{(-i)}(X_i)$.

STK demo (https://github.com/stk-kriging/stk)

`stk_example_kb10`

- LOO cross-validation plots (including residuals)
- "Borehole function", $d = 8$, $n = 10d = 80$, ReML

Note: hyper-parameters often kept fixed → "virtual LOO" formulas.