Model-based Design for GPs

Yunran Chen Bo Liu Phuc Nguyen Xiaojun Zheng

September 7, 2021

Motivation

- Model-free designs choose space-filling X_n. Without knowing much about the response surface you intend to model a priori, a space-filling design (i.e. LHS, maximin, minimax) represents a good choice.
- Recall if the response surface is linear, observations at boundaries are optimal, because they maximize leverage, minimize s.e. of $\hat{\beta}$.
- For a Gaussian Process (GP) response surface, we can choose samples optimal in some statistical sense.

"The best time to plan an experiment is after you've done it." – R. A. Fisher

- Sequential design helps avoid over-leveraging of prior beliefs before data collection.

Outline

- One-batch approach
  - Maximum entropy design (Maxent)
  - Minimum predictive uncertainty (IMSPE)
- Sequential approach
  - Active learning MacKay (ALM)
  - Active learning Cohn (ALC)

Maxent: Theory

- Model assumption:

$$
\begin{aligned}
Y &= f(X) + \epsilon \\
\epsilon_i &\sim N(0, \tau^2 g) \\
f &\sim GP(0, \tau^2 K), K^{ij} = C_\theta(x_i, x_j) \\
\Rightarrow Y \mid X, g, \tau^2, \boldsymbol{\theta} &\sim N(0, \tau^2 (K + gI))
\end{aligned}
$$

- Maximize the entropy of the marginal of Y w.r.t X_n:

$$
-E\{\log p(Y|X_n, g, \tau^2, \boldsymbol{\theta})\}
$$

  $\Rightarrow$ Equivalent to maximizing $|K_n|$

- Most informative for Bayesian learning

Maxent: Algorithm

```
library(plgp)
maxent <- function(n, m, theta=0.1, g=0.01, T=100000)
 {
  if(length(theta) == 1) theta <- rep(theta, m)
  X <- matrix(runif(n*m), ncol=m)
  K <- covar.sep(X, d=theta, g=g)
  ldetK <- determinant(K, logarithm=TRUE)$modulus

  for(t in 1:T) {
    row <- sample(1:n, 1)
    xold <- X[row,]
    X[row,] <- runif(m)
    Kprime <- covar.sep(X, d=theta, g=g)
    ldetKprime <- determinant(Kprime, logarithm=TRUE)$modulus
    if(ldetKprime > ldetK) { ldetK <- ldetKprime
    } else { X[row,] <- xold }
  }
  return(X)
}
```

Figure: Naive algorithm for maxent

Maxent: Strengths and Limitations

Figure: Maxent design under isotropic variance (left) and varying lengthscales (right)

- Strengths: adjust spread of different dimensions, theoretical guarantee.
- Limitations: design points cluttered at boundaries, few unique settings

Maxent: Strengths and Limitations

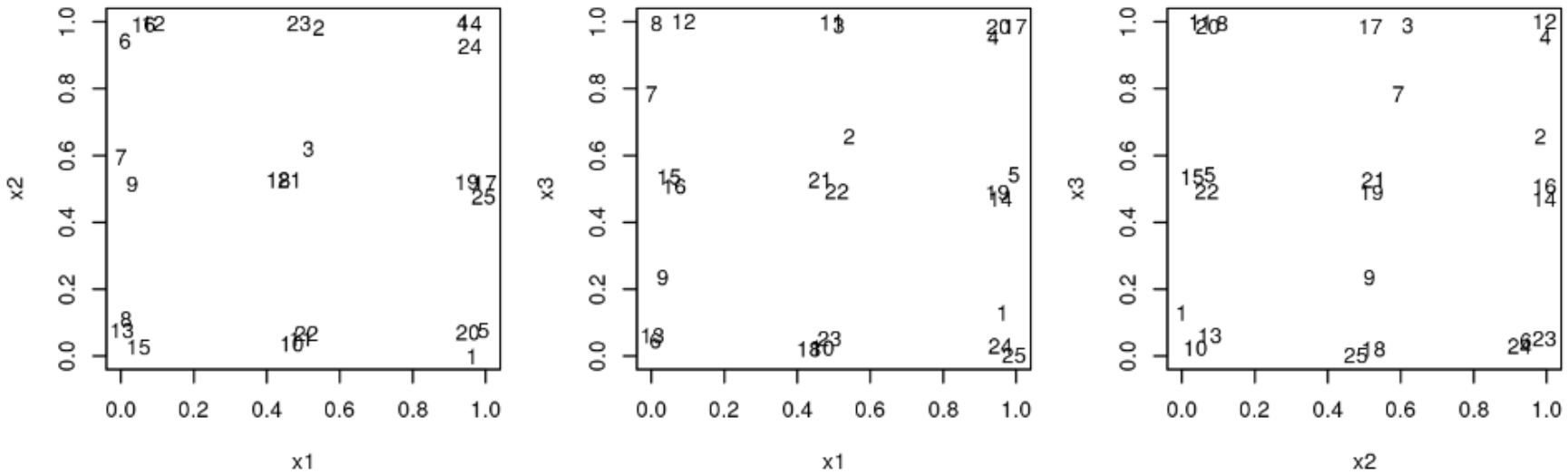

Figure: Projections of pairs of inputs involved in a 3d maximum entropy design

- Strengths: adjust spread of different dimensions, theoretical guarantee.
- Limitations: design points cluttered at boundaries, few unique settings, projections into lower dimensions don't have uniformity.

Fast GP Update

- Update $\log |K_{n+1}|$

$$\begin{aligned} \log |K_{n+1}| &= \log |K_n| + \log v_n(x_{n+1}) \\ \text{where } v_n(x_{n+1}) &= 1 + g - k_n(x_{n+1})^T K_n^{-1} k_n(x_{n+1}) \\ &= \frac{\sigma_n^2(x_{n+1})}{\hat{\tau}_n^2} \\ \Rightarrow O(n^2) & \end{aligned}$$

- Update scale-free predictive variance $v_{n+1}(x)$
- Update precision matrix

IMSPE: Motivation

Space-filling: spread points over the input space of interest (not related to prediction)
IMSPE: enhance prediction accuracy, interested in a sub-region of input space (local IMSPE, weighted IMSPE)

IMSPE: Model Assumption

- Model assumption:

$$Y(x) = \sum_{i=1}^{p} f_i(x)\beta_i + Z(x) = f^T(x)\beta + Z(x), \tag{1}$$

where $Z(x)$ is a GP with Gaussian correlation function $R(.)$.

- Minimize mean-sqared prediction error (MSPE):

$$\begin{aligned} \text{MSPE}\left(\boldsymbol{x}_0, \boldsymbol{X} \mid \sigma_Z^2, \boldsymbol{\rho}\right) &= E_Y\left\{(Y(\boldsymbol{x}_0) - \widehat{y}(\boldsymbol{x}_0))^2\right\} \\ &= \sigma_Z^2\left(1 - \begin{bmatrix}\boldsymbol{f}_0^T & \boldsymbol{r}_0^T\end{bmatrix}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{F}^T \\ \boldsymbol{F} & \boldsymbol{R}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{f}_0 \\ \boldsymbol{r}_0\end{bmatrix}\right), \end{aligned}$$

where F is known regressor, R_ρ is Gaussian correlation matrix, y^n is training outputs.

- generalized A-optimality: min trace of inverse of info matrix

IMSPE: Local IMSPE and W-IMSPE

- (known ρ)A local IMSPE: integrate out x_0

$$IMSPE(X|\sigma_Z^2, \rho) = \int_{[0,1]^d} MSPE(x_0, X|\sigma_Z^2, \rho) dx_0 \tag{2}$$

Specially, if GP has constant mean (i.e. F are 1's), depends on ρ only

$$\min IMSPE(\ldots|\sigma_Z^2, \rho) = \min IMSPE(\ldots|1, \rho)$$

- (unknown ρ) A weighted IMSPE (W-IMSPE):

$$W(X|\pi) = \int_{[0,1]^d} IMSPE(X|1, \rho)\pi(\rho) d\rho \tag{3}$$

IMSPE: Computation

- Closed form: IMSPE with a known ρ; rectangular input space and certain covariance kernels.
- Numeric approximation: W-IMSPE no available closed form
- Approximation: quasi Monte Carlo numerical integration based on a low discrepancy sequence

$$W(X|\pi) = \int_{[0,1]^d} IMSPE(X|1,\rho)\pi(\rho)d\rho \tag{4}$$

- (Leatherman et. al., 2018)

$$W_a(X|\pi) = \frac{1}{2^k}\sum_{j=1}^{2^k} IMSPE(X|1,\rho_j)\pi(\rho_j), \tag{5}$$

where ρ_j is 2^k-point Sobol sequence.
Modification: 1) adaptive k; 2) PSO algorithm to choose starting point.

IMSPE: Computation

(Gramacy, retrieved 2021)

- other reference grids such as poor-man's quadrature or random reference grid.
- cons: not off-boundary; discrete or mixed continuous-discrete optimization
- using random reference grid for non-regular space
- improvements: a larger reference set, more stochastic exchange proposals, sequential design adaption, etc.

IMSPE: Comparison with space-filling design

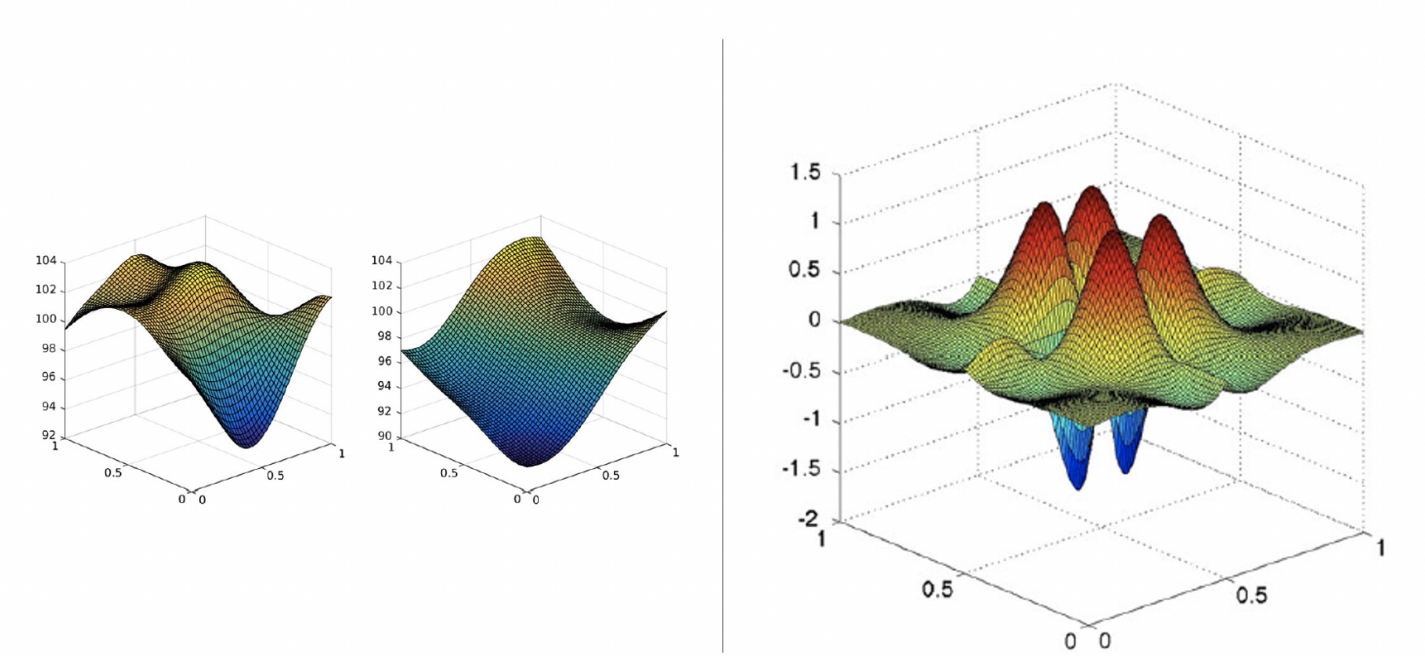

Figure: Example surfaces

- Smooth "stationary" surfaces, IMSPE-based methods are recommended
- Functions with pronounced non-stationary activity near the "middle" of the input domain: space-filling LHDs and MaxPro are recommended

IMSPE: Comparison with space-filling design

- Similar to maxent or maximin
- avoid boundary of input space (sites at boundary don't cover space efficiently)
- In higher input dimension, more "off the boundary"

Sequential design/ Active learning

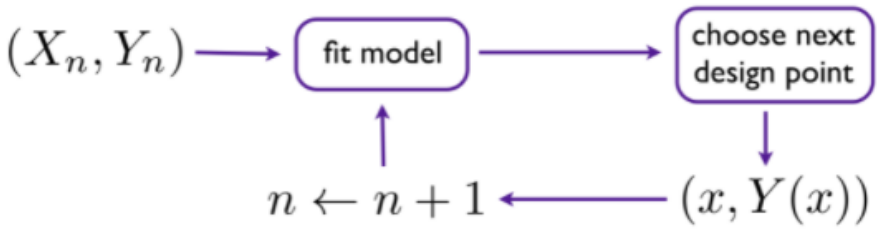

Figure: Diagram of Sequential design/active learning

1. **Assume** a flexible surrogate, e.g., a GP model with unknown. hyperparamters
2. **Require** outputs $y \sim f(x)$, a choice of initial design size n and final size N, and criterion $J(x)$ to choose to next point.

Then

1. Fit the surrogate (hyperparameters) using $D_n = (X_n, Y_n)$, e.g., via MLE.
2. Choose $X_{n+1} = \text{argmax}_{x \in \mathcal{X}} J(x)|D_n$.
3. Observe the response by running a new simulation to get y_{n+1}, and update D_{n+1}.

Why using sequential design ?

1. **More practical**: In many situations, selecting one design point at a time works better than static, single-batch design
    1. Single-batch design is sensitive to hyperparameters, while in sequential design, could **update hyperparameters** for each run.
    2. Data measurements are relatively **expensive or slow**, and we want to know where to look next so as to learn as much as possible.
    3. There is an **immense amount of data** and we wish to select a subset of data points that is most useful for our purposes.
2. The sample size N need **not** to be **fixed**, and **information gain** for each new data point is available.
    1. **Omit** the data points that are expected to be **least informative**
    2. Form a **stopping rule**, so that we could decide whether to gather more data, given a desired exchange rate of information gain per measurement (Lindley 1956).

Active Learning Mackay (ALM)

Setup:

1. Start with a LHS in 2d of size $n_0 = 12$ with $f(x_1, x_2) = x_1 \cdot e^{-x_1^2 - x_2^2} + N(0, 0.01^2)$.

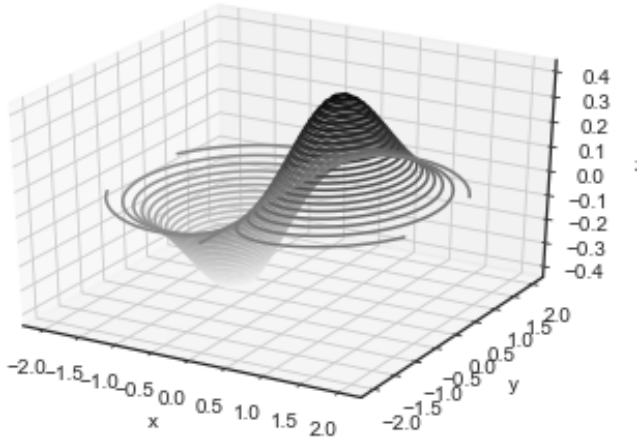

Figure: Function f

2. Create a testing grid and saves true (noiseless) responses at those locations.
3. Calculate RMSE to see out-of-sample progress over iterations of design acquisition

Criterion $J(x)$ is predictive variance $\sigma_n^2(x)$ in ALM.

Multi-start Scheme

1. Predictive variance produces sausage-shaped error-bars, so it must have many local maxima.
2. The number of local maxima could grow linearly in sample size n. Optimizing globally over that surface presents challenges
3. Use the library-based local solver in R, "optim" with method="L-BFGS-B".

Thus, we adopt the multi-start scheme.

Multi-start Scheme

We "design" a collection of starting locations placed in parts of the input space known to have high variance.

```
xnp1.search <- function(X, gpi, obj=obj.alm, ...)
 {
  start <- mymaximin(nrow(X), 2, T=100*nrow(X), Xorig=X)
  xnew <- matrix(NA, nrow=nrow(start), ncol=ncol(X) + 1)
  for(i in 1:nrow(start)) {
    out <- optim(start[i,], obj, method="L-BFGS-B", lower=0,
      upper=1, gpi=gpi, ...)
    xnew[i,] <- c(out$par, -out$value)
  }
  solns <- data.frame(cbind(start, xnew))
  names(solns) <- c("s1", "s2", "x1", "x2", "val")
  return(solns)
 }
```

Figure: Function for searching x_{n+1}

Active Learning Mackay (ALM)

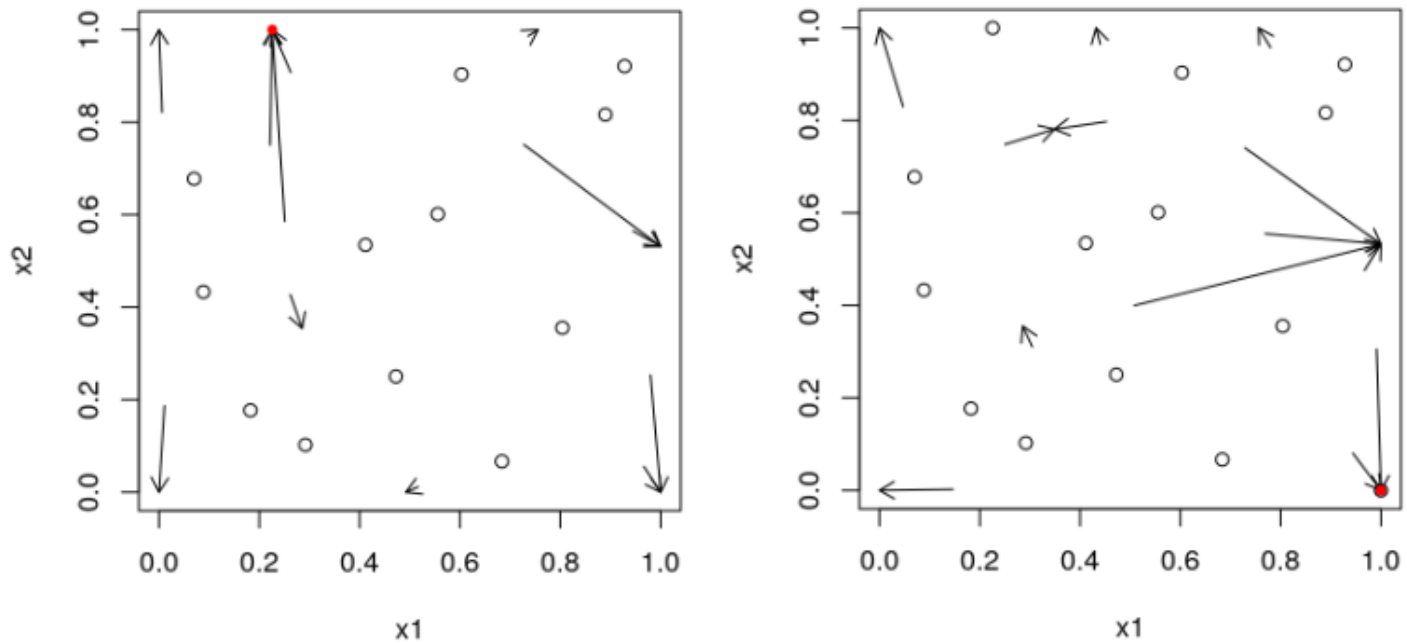

Figure: First/second iteration of ALM search. Each arrow represents an origin and outcome of multi-start exploration of predictive variance. Variance-maximizing location is indicated as a red dot.

Active Learning Mackay (ALM)

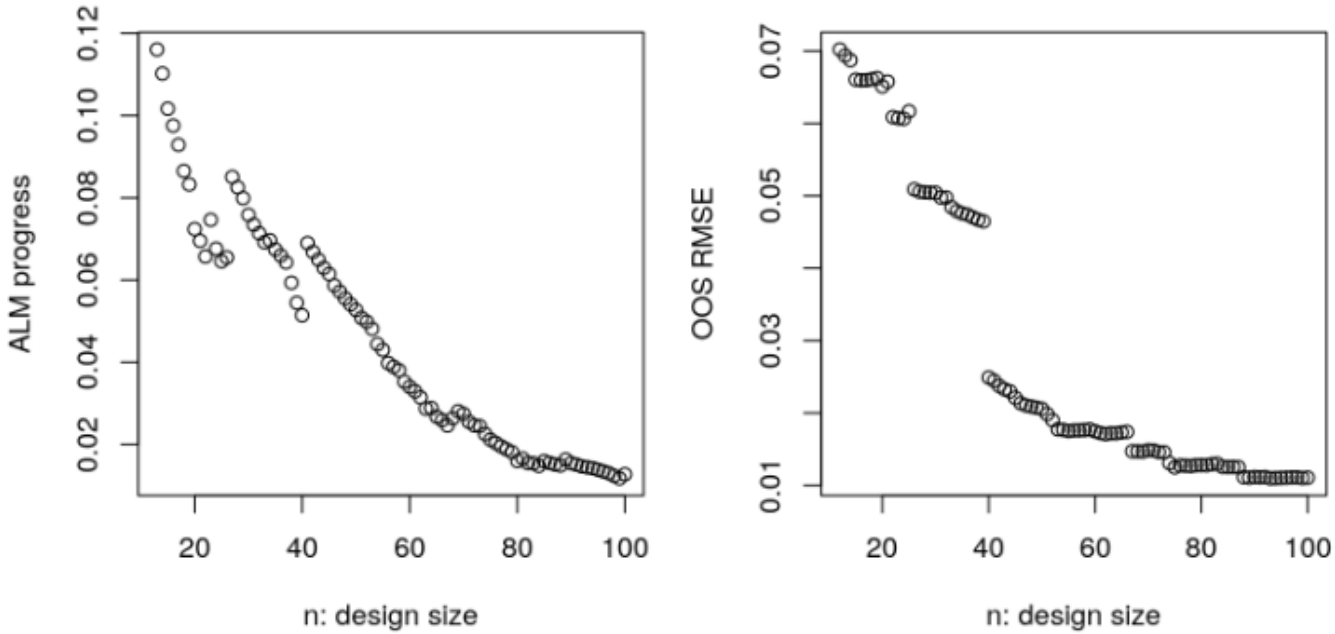

Figure: Maximum variance (left, lower is better) and out-of-sample RMSE (right) over 100 ALM acquisitions.

Progress metrics and RMSE are starting to level off in the later 35 iterations or so.

Active Learning Mackay (ALM)

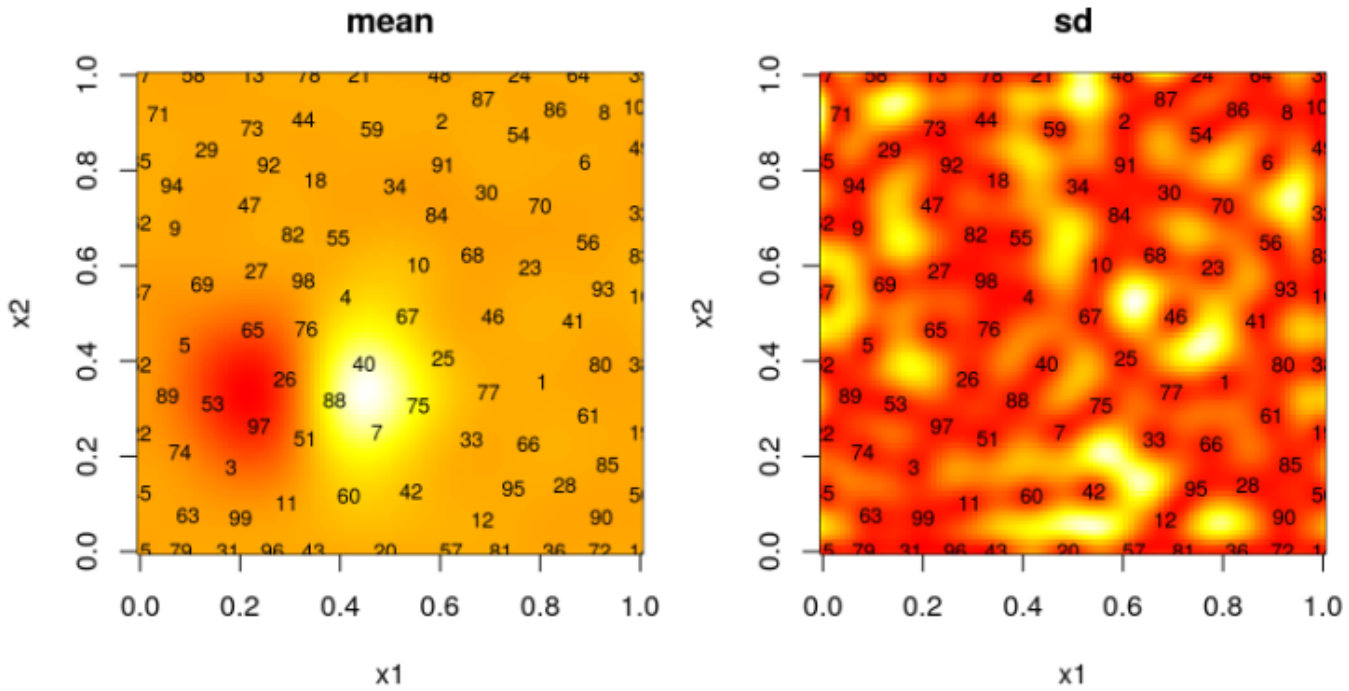

Figure: Predictive mean (left) and standard deviation (right) after ALM-based sequential design.

Observe dense coverage along the boundary since variance is high along the boundary.

Active Learning Mackay (ALM)

1. ALM can misbehave, especially if the starting design is unlucky to miss strong signal in the data
2. ALM doesn't recognize that acquisitions impact predictive equations globally.
3. Potentially ignoring a fatter regions where uncertainty may cumulatively be much larger.
4. Variance is high along the boundary because there are fewer data points nearby, so we end up with lots of points on the boundary.

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...
- ...but how much does it help to add a point at that spot?

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...
- ...but how much does it help to add a point at that spot?
- Might be better to consider how much reduction in posterior variance can be obtained by adding an extra point.

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...
- ...but how much does it help to add a point at that spot?
- Might be better to consider how much reduction in posterior variance can be obtained by adding an extra point.
- Question: where should the reduction be measured?

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...
- ...but how much does it help to add a point at that spot?
- Might be better to consider how much reduction in posterior variance can be obtained by adding an extra point.
- Question: where should the reduction be measured?
- Two extremes:

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...
- ...but how much does it help to add a point at that spot?
- Might be better to consider how much reduction in posterior variance can be obtained by adding an extra point.
- Question: where should the reduction be measured?
- Two extremes:
  - Global: integrate over the whole space;

Active Learning Cohn (ALC)

- High posterior variance at some points is definitely an issue...
- ...but how much does it help to add a point at that spot?
- Might be better to consider how much reduction in posterior variance can be obtained by adding an extra point.
- Question: where should the reduction be measured?
- Two extremes:
  - Global: integrate over the whole space;
  - Local: at specific reference point(s).

Active Learning Cohn (ALC)

- Cohn (1994) suggests such an acquisition heuristic in a nonparametric regression context for neural networks.
- Seo et al. (2000) adapt Cohn's ideas to Gaussian Process and called it ALC.

Active Learning Cohn (ALC)

- How does it work?
- Recall that predictive variance follows

$$\sigma_n^2 = \hat{\tau}_n^2[1 + \hat{g}_n - k_n^{\mathrm{T}}(x)K_n^{-1}k_n(x)], \text{where } k_n(x) \equiv C_{\hat{\theta}_n}(X_n, x).$$

- The deduced variance

$$\tilde{\sigma}_{n+1}^2 = \hat{\tau}_n^2[1 + \hat{g}_n - k_{n+1}^{\mathrm{T}}(x)K_{n+1}^{-1}k_{n+1}(x)], \text{where } k_{n+1}(x) \equiv C_{\hat{\theta}_n}(X_{n+1}, x).$$

- The ALC criterion is the average reduction in variance from $n \to n+1$ measured through a choice of x_{n+1}:

$$\Delta\sigma_n^2(x_{n+1}) = \int_{\mathcal{X}} [\sigma_n^2(x) - \tilde{\sigma}_{n+1}^2(x)]\,\mathrm{d}x.$$

Active Learning Cohn (ALC)

- The criterion must be solved in each iteration of sequential design.

$$x_{n+1} = \arg\min \tilde{\sigma}^2_{n+1}(x), x \in \mathcal{X}.$$

- Closed form when $\mathcal{X}$ is rectangular.
- Often in practice approximated by a reference set.

Active Learning Cohn (ALC)

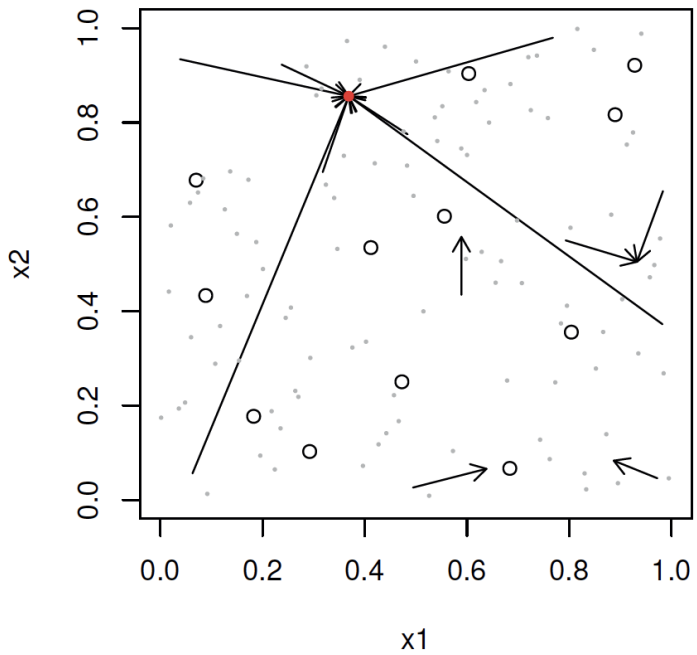

Figure: First iteration of ALC search. Gray dots denote reference locations.

Active Learning Cohn (ALC)

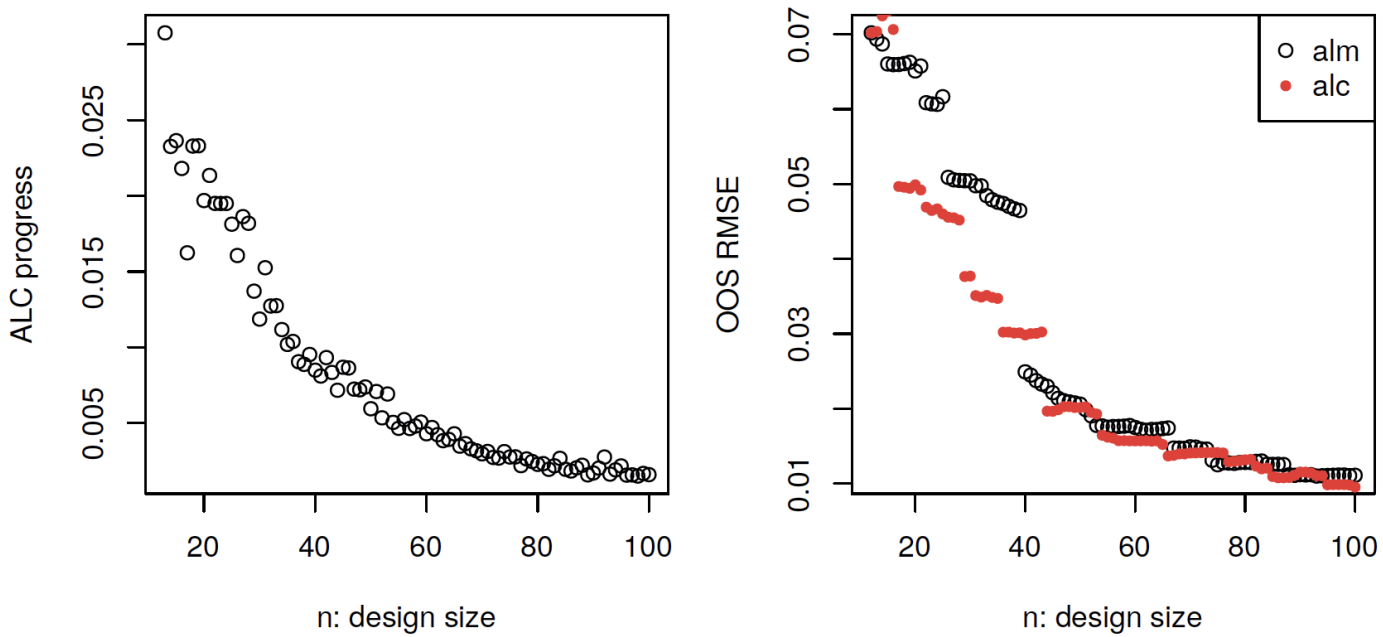

Figure: Progress in ALC sequential design in terms of integrated reduction in variance (left, lower is better) and out-of-sample RMSE (right), with comparison to ALM.

Active Learning Cohn (ALC)

Recall what the true function looks like.

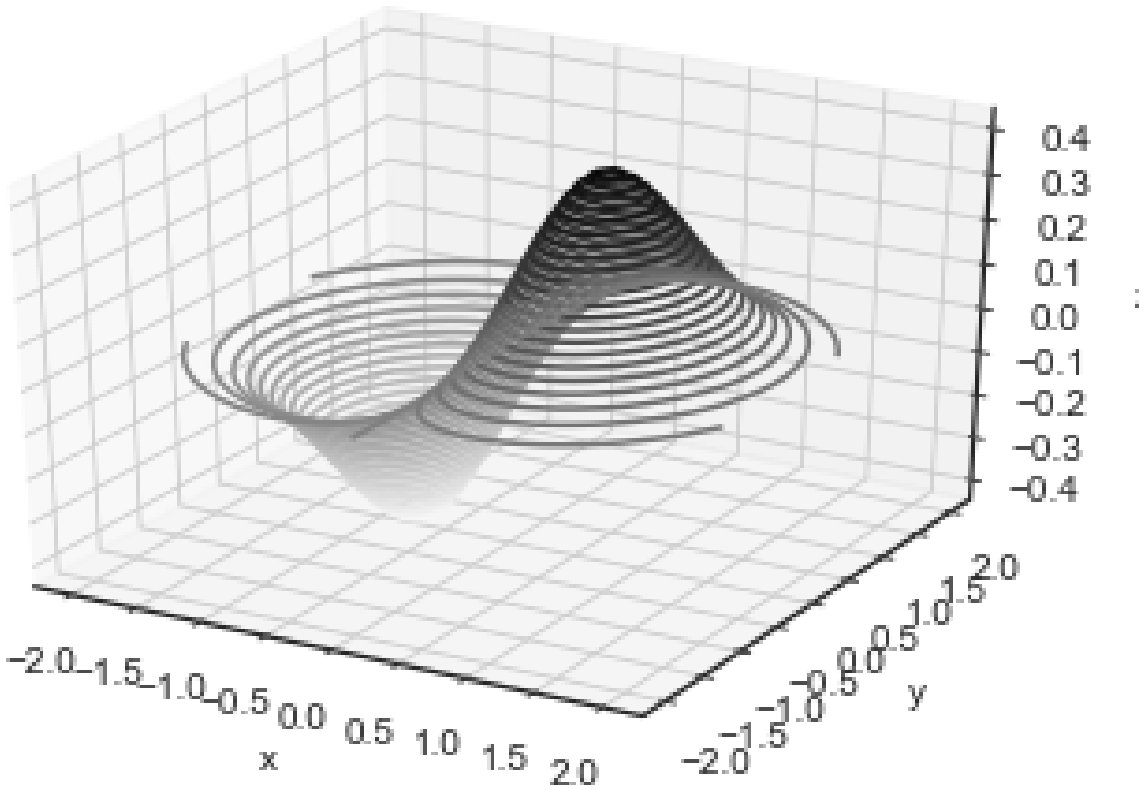

Figure: Function f.

Active Learning Cohn (ALC)

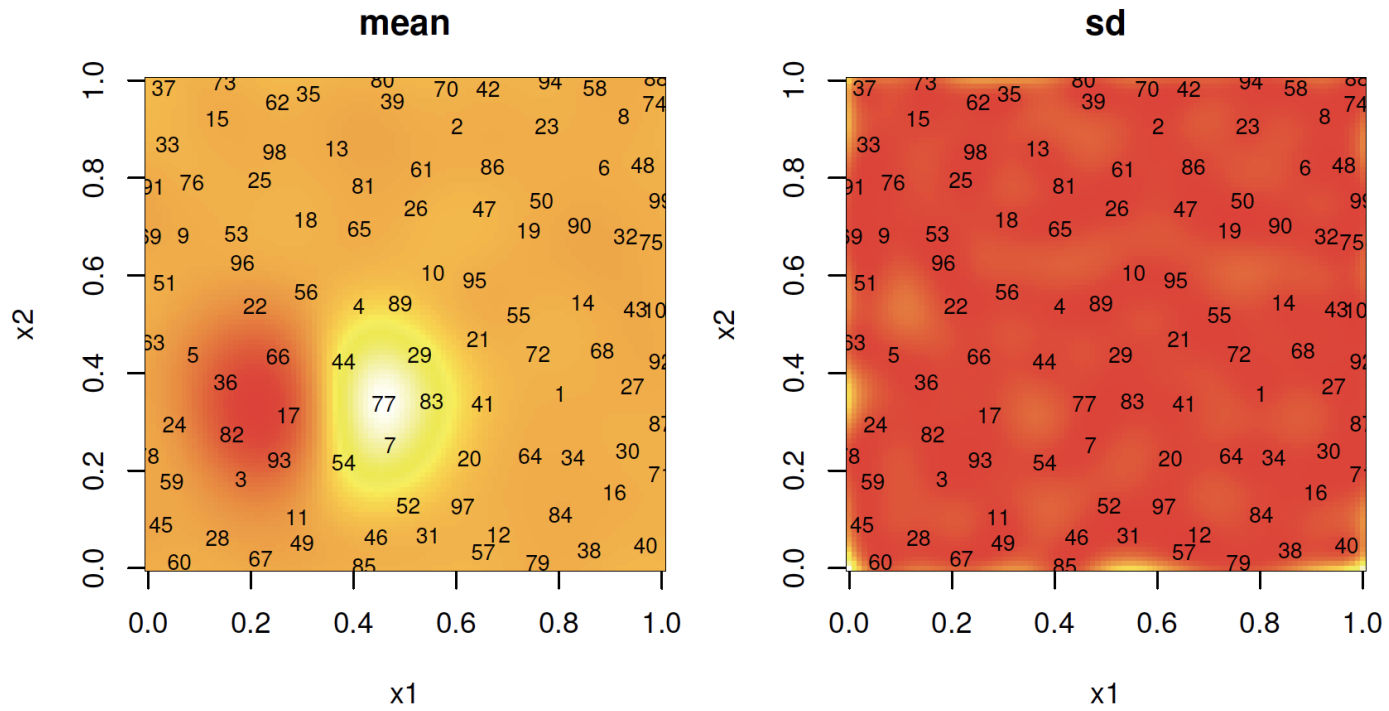

Figure: Predictive mean (left) and standard deviation (right) after ALC-based sequential design.

Other Sequential Criteria - Fisher Information

- Thinking about the hyperparameters - what point can we select in order to estimate the hyperparameters more accurately?
- Criterion: maximize the Fisher Information.
- Does not lead to designs with the most accurate predictors.
- Hybrid approach:
  - FI - learn hyperparameters
  - ALC - prediction

Other Sequential Criteria - Fisher Information

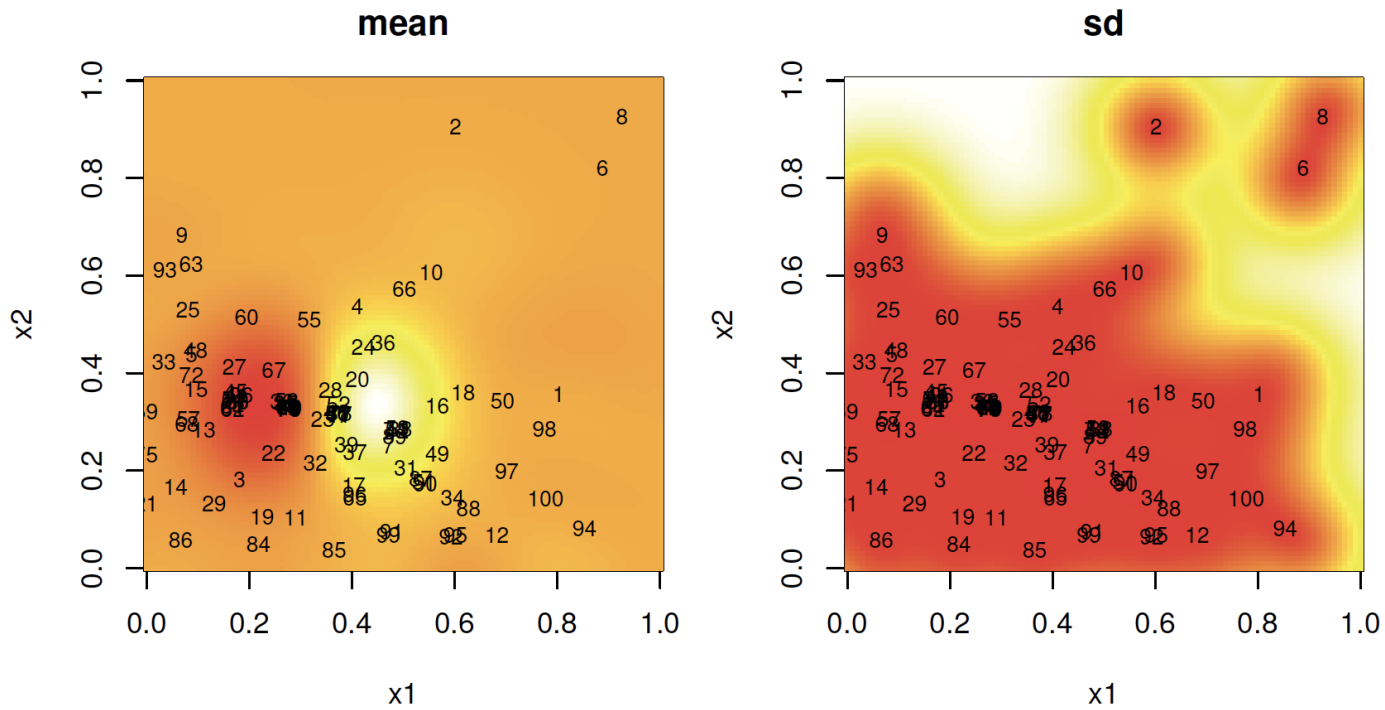

Figure: Predictive mean (left) and standard deviation (right) after FI-based sequential design.

Other Sequential Criteria - Fisher Information

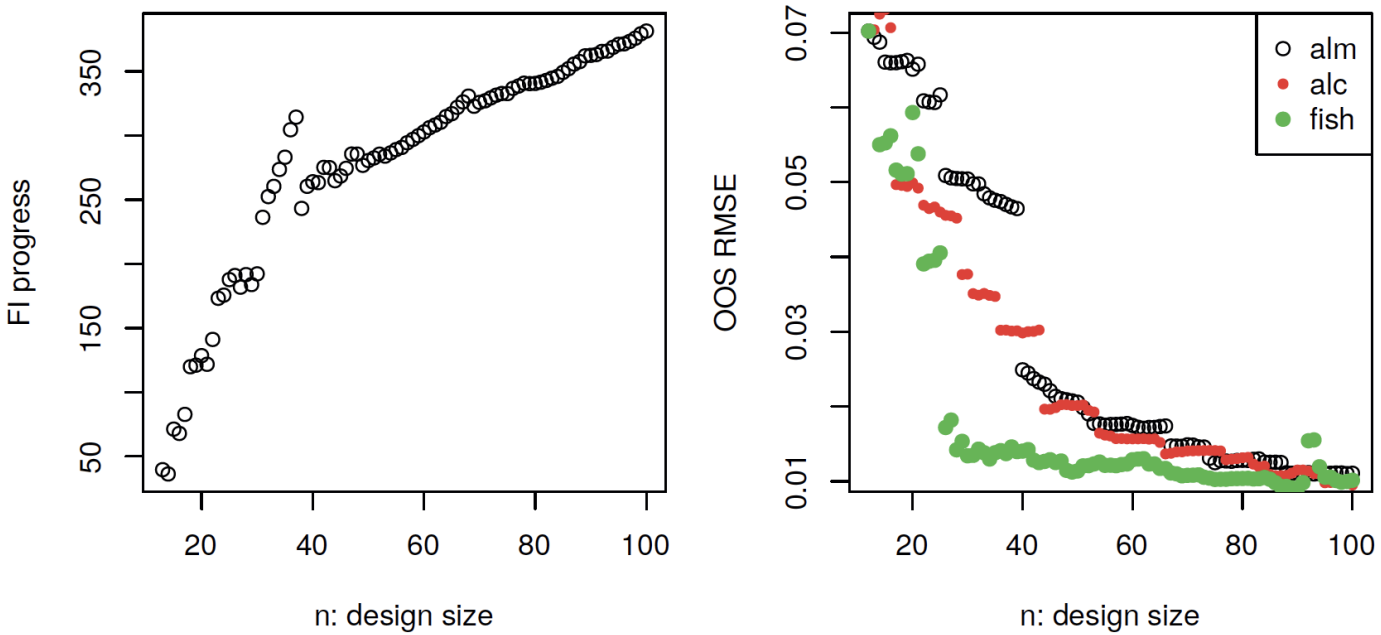

Figure: Progress in terms of FI (left, higher is better) and out-of-sample RMSE as compared to previous heuristics.