Bayesian optimization to improve cooling rate in LEReC

Bayesian Optimization (BO)

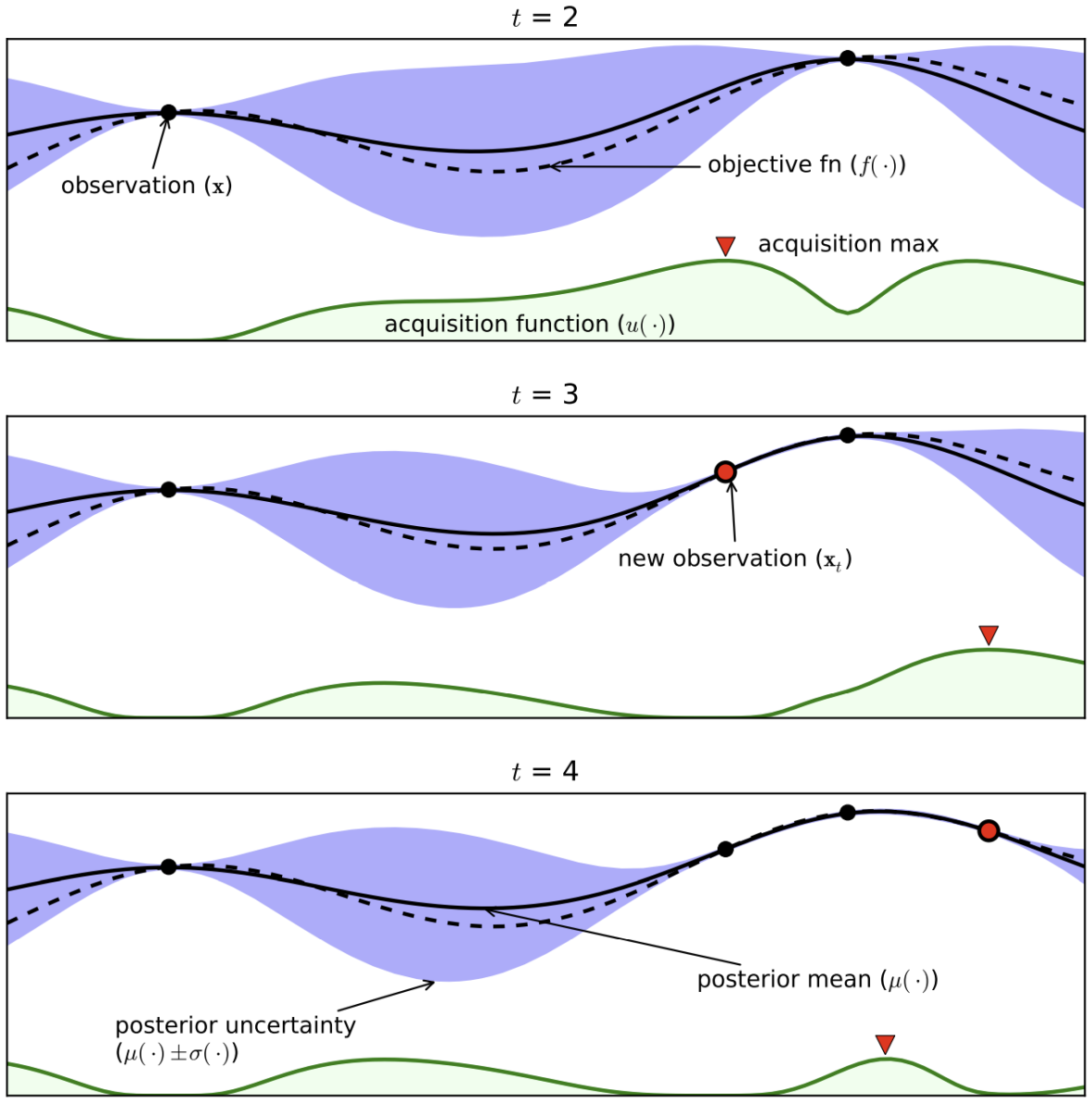
- Optimize a nonlinear function f(x) over a compat set A: $\max_{\mathbf{x} \in \mathcal{A} \subset \mathbb{R}^d} f(\mathbf{x})$

- In many real-world learning problems, evaluating the objective function is expensive or even impossible, and the derivatives and convexity properties are unknown.
- Bayesian optimization is a powerful strategy for finding the extrema of objective functions that are expensive to evaluate.
- It is called Bayesian because it uses the famous "Bayes' theorem"

$$P(f|\mathcal{D}_{1:t}) \propto P(\mathcal{D}_{1:t}|f)P(f)$$

- This step is also called estimating the objective function with a ***surrogate function;*** Gaussian Process (GP) is one of the most used priors over functions.
- To sample efficiently, Bayesian optimization uses an ***acquisition function*** to determine the next location x(t+1) ∈ A to sample.
- It has an automatic trade-off between exploration and exploitation; It has the nice property that it aims to minimize the number of objective function evaluations. Moreover, it is likely to do well even there are multiple local maxima.

Bayesian Optimization



- The acquisition is high where the GP predicts a high objective (exploitation) and where the prediction uncertainty is high (exploration).
- Areas with both attributes are sampled first.
- Note that the area on the far left remains unsampled. Although it has high uncertainty, it is (correctly) predicted to offer little improvement over the highest observation.

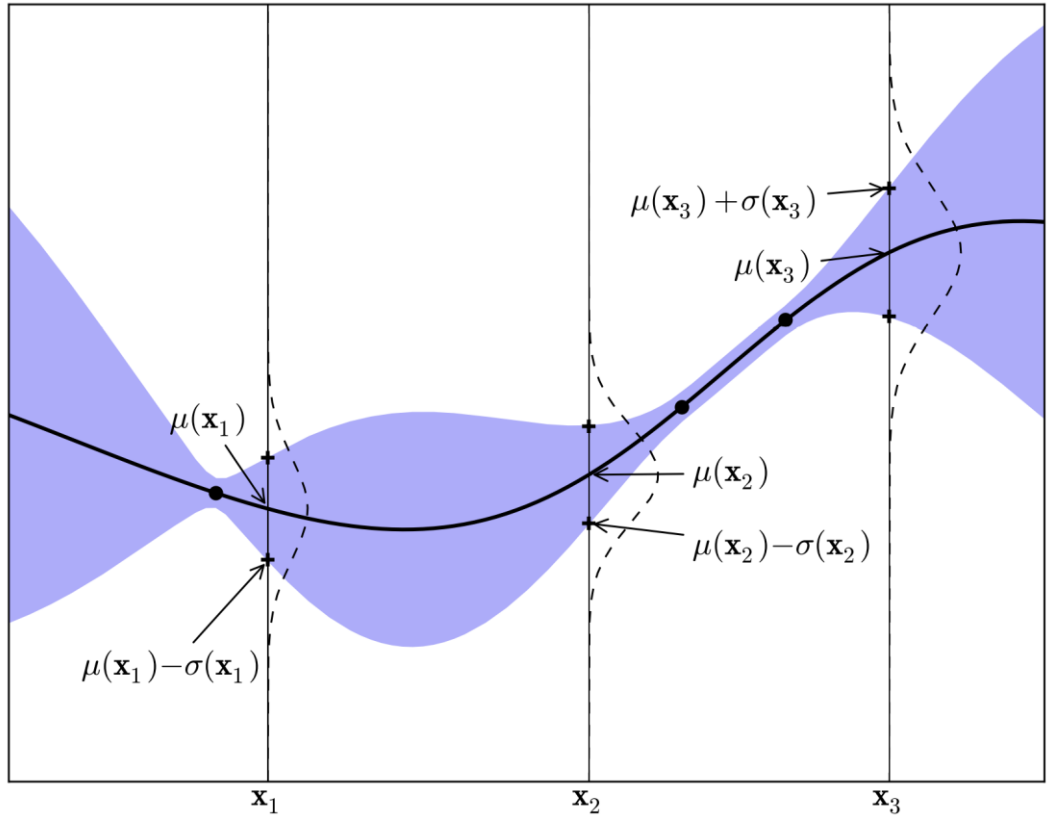
$$\mathrm{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \kappa\sigma(x)$$

Gaussian Process (GP)

- A GP is an extension of the multivariate Gaussian distribution to an infinite dimension stochastic process for which any finite combination of dimensions will be a Gaussian distribution.
- Just as a Gaussian distribution is a distribution over a random variable, completely specified by its mean and covariance, a GP is a distribution over functions, completely specified by its mean function, m and covariance function, k:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

- Sampling from a GP:
  1. The solid black line is the GP surrogate mean
  2. The shaded area shows the mean plus and minus the variance.
  3. The superimposed Gaussians correspond to the GP mean and standard deviation (μ(·) and σ(·)) of prediction at the points x1:3.



Data-informed GP, Physics-informed GP

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

- For convenience, we usually assume the prior mean is the zero-function m(x)=0.

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_t) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_t, \mathbf{x}_1) & \dots & k(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix}$$

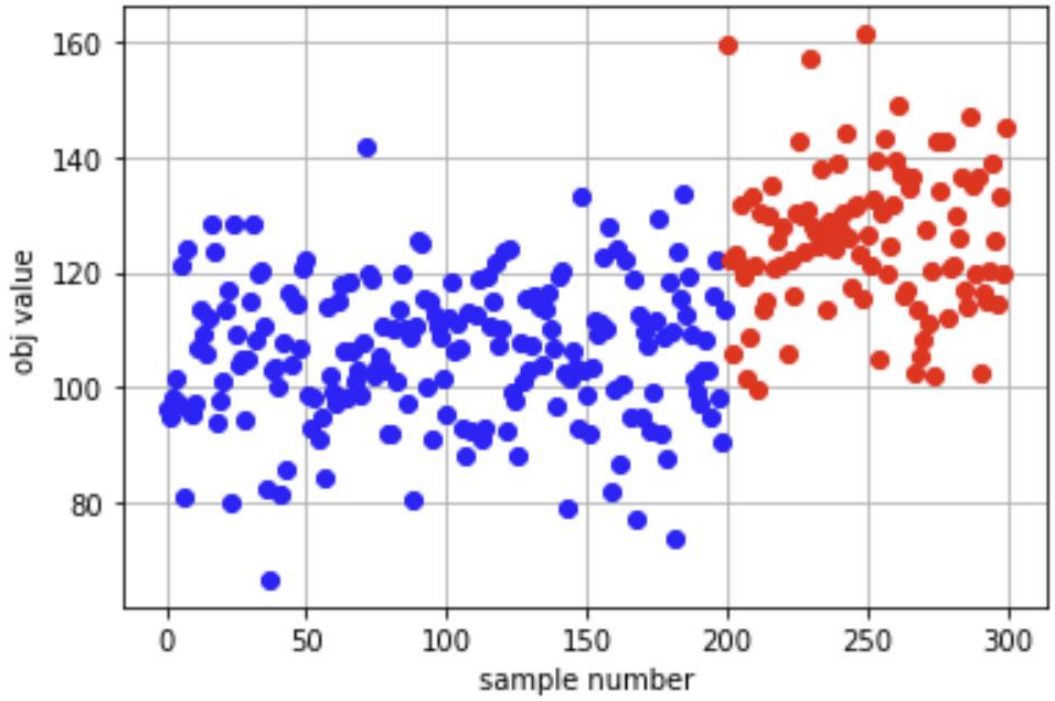
- A very popular choice is the squared exponential function: Accurately estimate of the precision matrix Sigma is very important.

$$k_{\mathrm{RBF}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T \Sigma (\boldsymbol{x} - \boldsymbol{x}')\right]$$

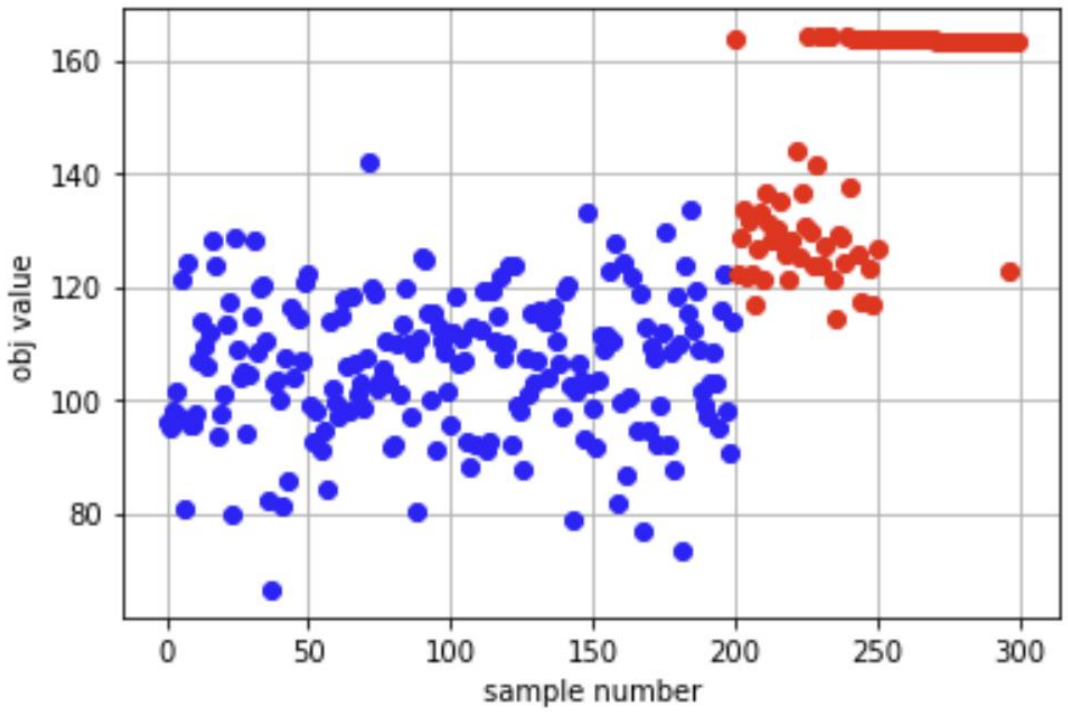
- Data-informed GP, by fitting the data repeatedly to estimate the Sigma matrix.
- Physics-informed GP, by evaluating the Hessian matrix around the optimal point, then calculate the Sigma directly: $\Sigma = -H/2$.

Testing problem – 16 dimensions maximization

Data-informed



Physics-informed



Contextual Gaussian Process

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

- Can be used to model the dynamic environment factors.

- A fact about kernels, new kernels can be created:
  1. Multiplying two kernels means the resulting kernel will have high value only if both of the two base kernels have a high value.
  2. Adding two kernels means the resulting kernel will have high value if either of the two base kernels have a high value.

- For example, in LEReC:
  1. The correctors can be modeled using a data kernel or physics-informed kernel.
  2. The changing ion intensity will be a dynamic environment factor to be modeled using a separate kernel.
  3. The combined kernel will be the addition of the two.

BO applications in the accelerator field

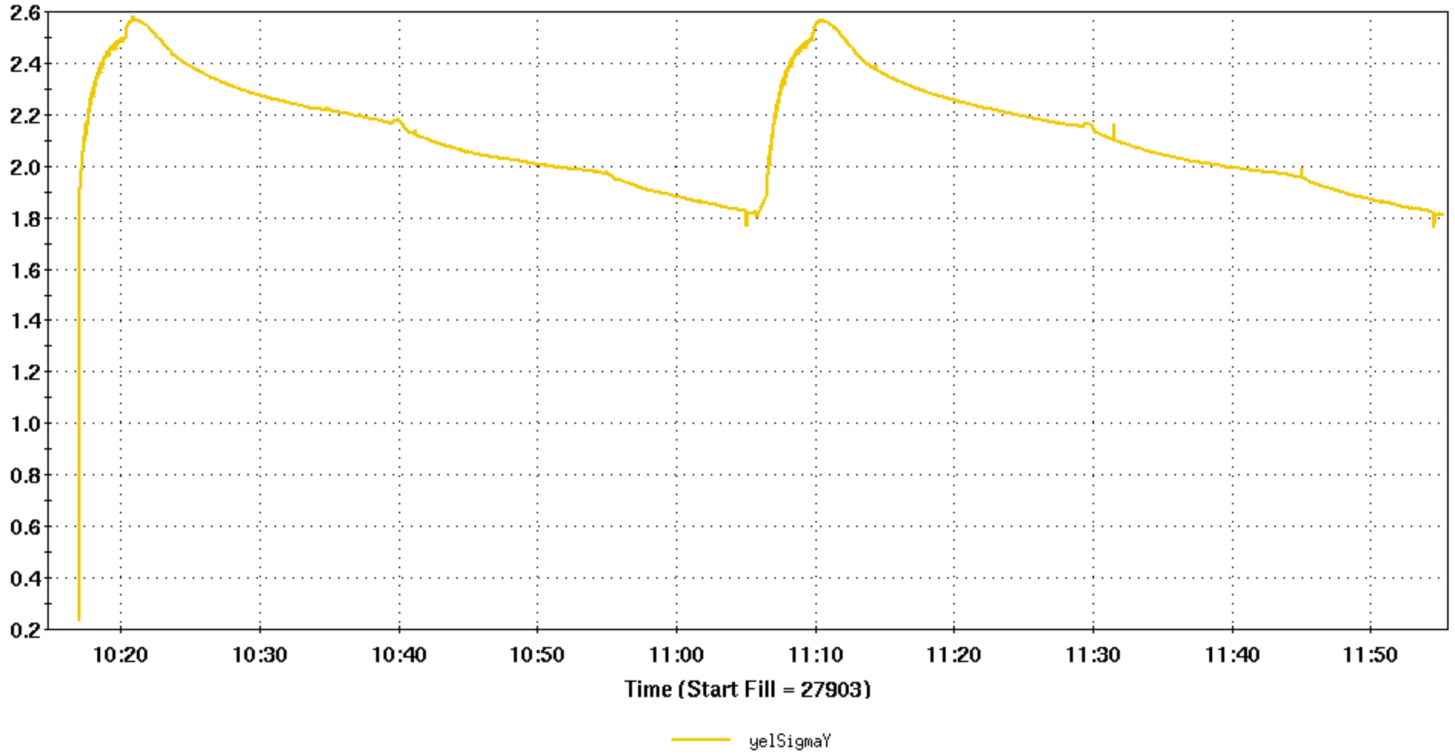
In SLAC, Bayesian optimization [1, 2] has been used to tune quadrupole magnets settings in the LCLS Free Electron Laser (FEL). Comparisons have been made with hand-tuning and existing Nelder-Mead optimization method. Moreover, work [2] showed that by building correlated kernels, the GP converges faster in simulation scenarios where the input dimensions are correlated.

Physics-informed GP [3] was proposed and used to optimize the electron beam loss rate on the SPEAR3 storage ring, and the results showed that the GP with a physics-informed kernel converges faster than the GP with a data-informed kernel and the Nelder-Mead simplex optimizer.

In work [4], a variant of Bayesian optimization called SafeLineBO was proposed, which divides global problem into sequential subproblems that can be solved efficiently without violating safety constraints. Then the algorithm was compared with the simple parameter scanning and Nelder-Mead method to tune the FEL outputs of SwissFEL with up to 40 parameters.

[1] M. McIntire, T. Cope, D. Ratner, et al., “Bayesian Optimization of FEL Performance at LCLS”, in Proceedings of the 7th International Particle Accelerator Conference, Jun 2016.
[2] J. Duris, D. Kennedy, A. Hanuka, et al., “Bayesian Optimization of a Free-Electron Laser”, in Physical Review Letters 124, 124801, Mar. 2020.
[3] A. Hanuka, J. Duris, J. Shtalenkova, et al., “Online tuning and light source control using a physics-informed Gaussian process”, in Proceedings of the 33rd Conference on Neural Information Processing Systems, Nov. 2019.
[4] J. Kirschner, A. Adelmann, N. Hiller,et al., “Bayesian Optimisation for Fast and Safe Parameter Tuning of SwissFEL”, in Proceedings of the 39th International Free-Electron Laser Conference, Nov. 2019.

Motivations for the LEReC



- Cooling rate is defined as the transverse beam size decreasing rate.
- Maximum cooling rate is unknown.
- Current script tunes the correctors to make electrons in the center whenever they drift away above a distance. But it may not be the optimum scheme.
- The plan is to tune the correctors using Bayesian optimization.

Sampling plan

- For either data-informed GP or physics-informed GP, sampling some data is necessary.

- For the data-informed GP…

- For the physics-informed GP, need to estimate the Hessian around the optimal point.
  1. For each corrector, scan it in its range (where there are cooling) with a step size while keeping other correctors fixed at the initial value.
  2. For each setting, keep it running for around 1 or 2 minutes, measure the beam size and calculate the cooling rate, then move to the next setting.
  3. Repeat it for every corrector. The goal is to estimate an approximate optimum point in the joint distribution of the correctors.

- Start from the yellow cooling section, start from the first few correctors.

A blackbox function y (unknown expression and derivatives)

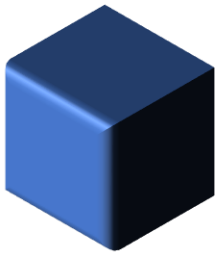
History data set:
{x1, x2, …, xn}
{y1, y2, …, yn}

The goal is to get min(y)

Min(y)
Some goal to optimize

Direct query is very expensive

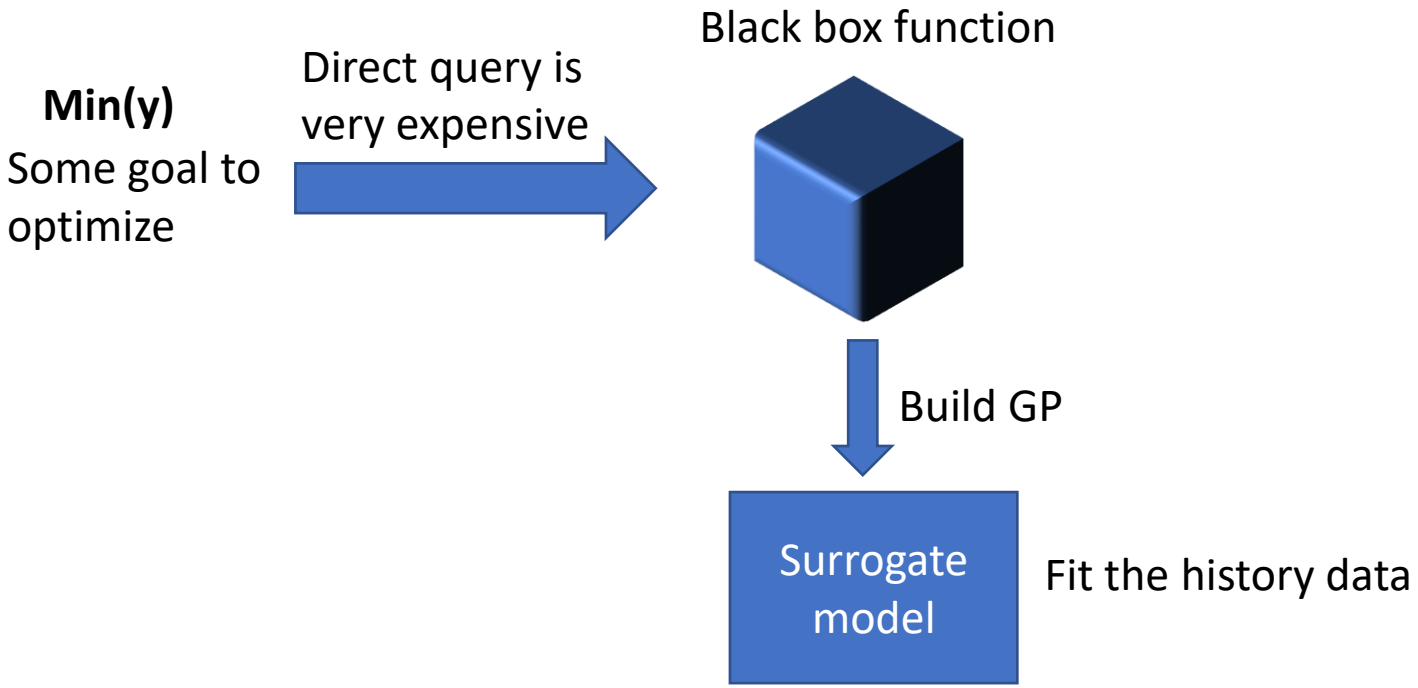
Black box function



A blackbox function y, (unknown expression and derivatives)

History data set:
{x1, x2, …, xn}
{y1, y2, …, yn}

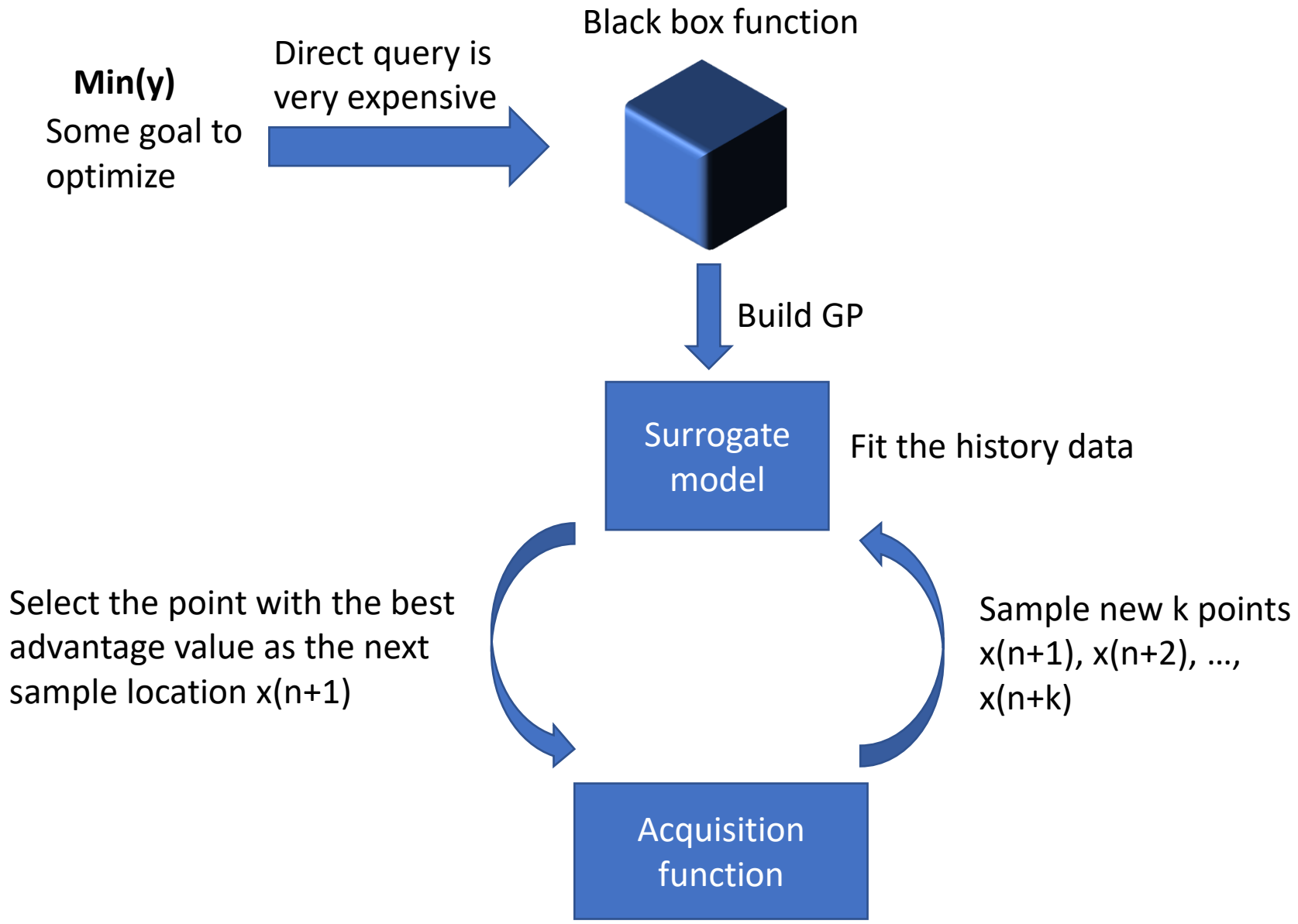
The goal is to get min(y)



A blackbox function y, (unknown expression and derivatives)

History data set:
{x1, x2, …, xn}
{y1, y2, …, yn}

The goal is to get min(y)

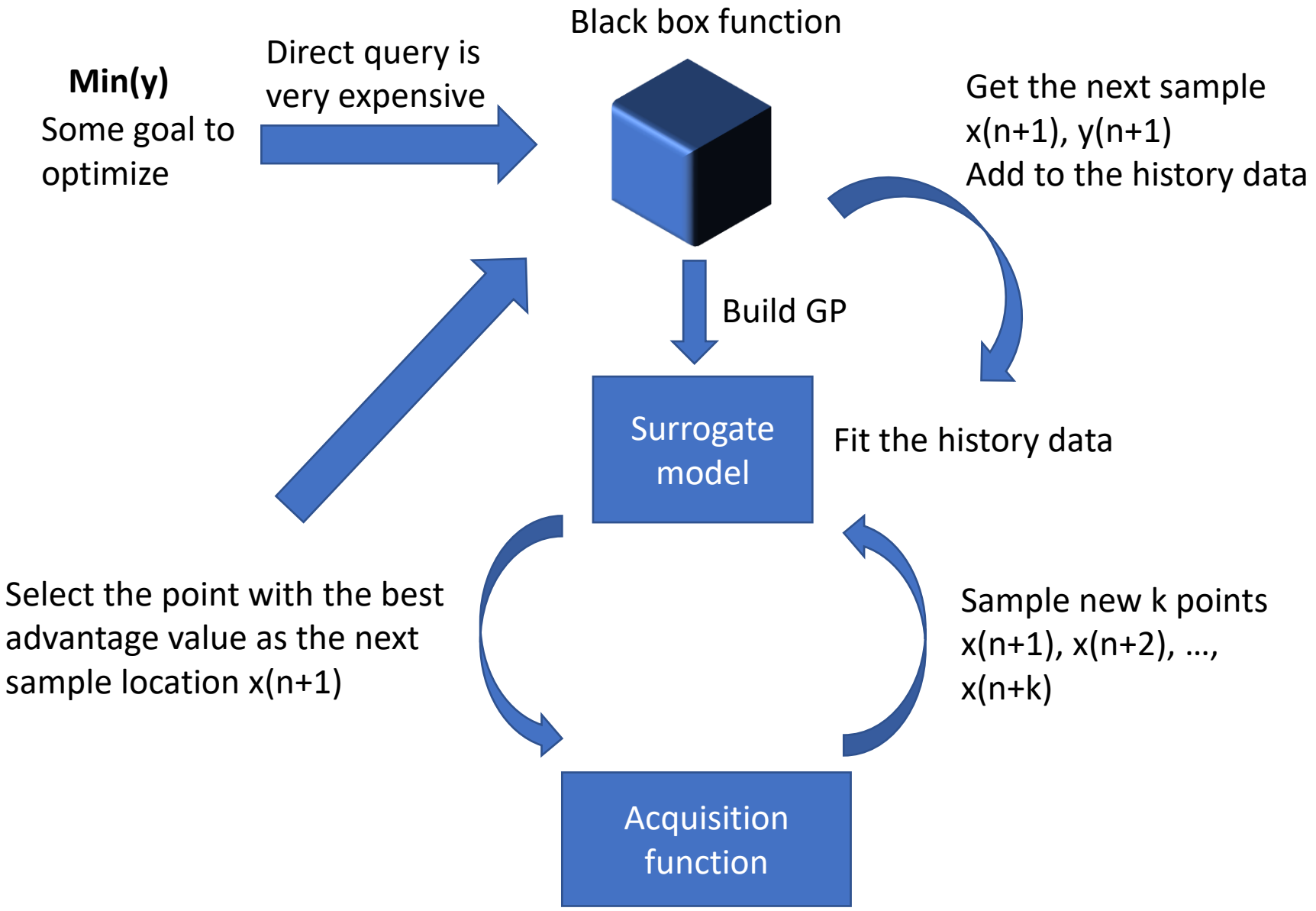


A blackbox function y, (unknown expression and derivatives)

History data set:
{x1, x2, …, xn}
{y1, y2, …, yn}

The goal is to get min(y)

Data-informed GP



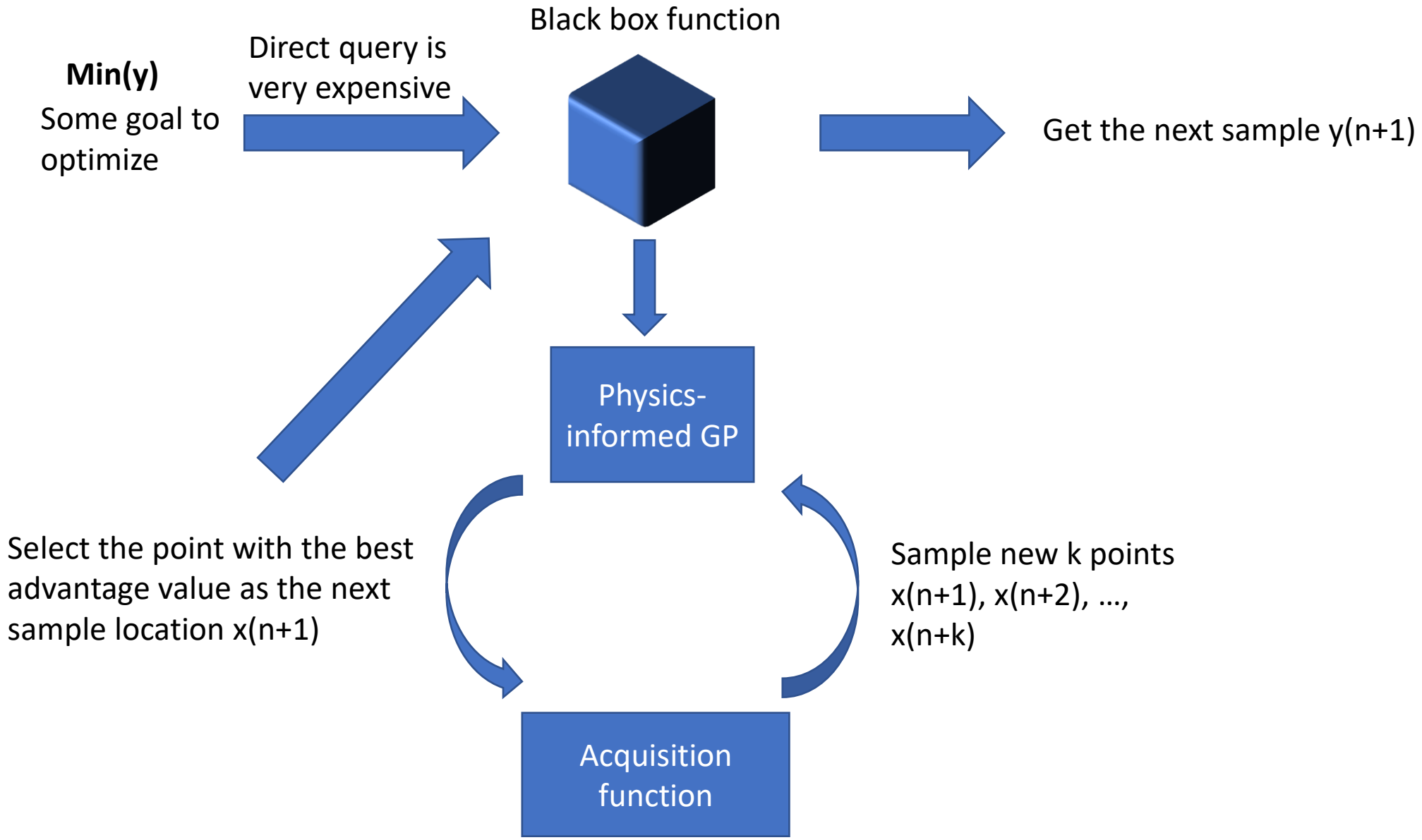
Physics-informed GP

- Gaussian process: gp(m(x), k(x, x')).
- Kernel function k(x, x') describes correlations between points in the objective space.
- RBF kernel:

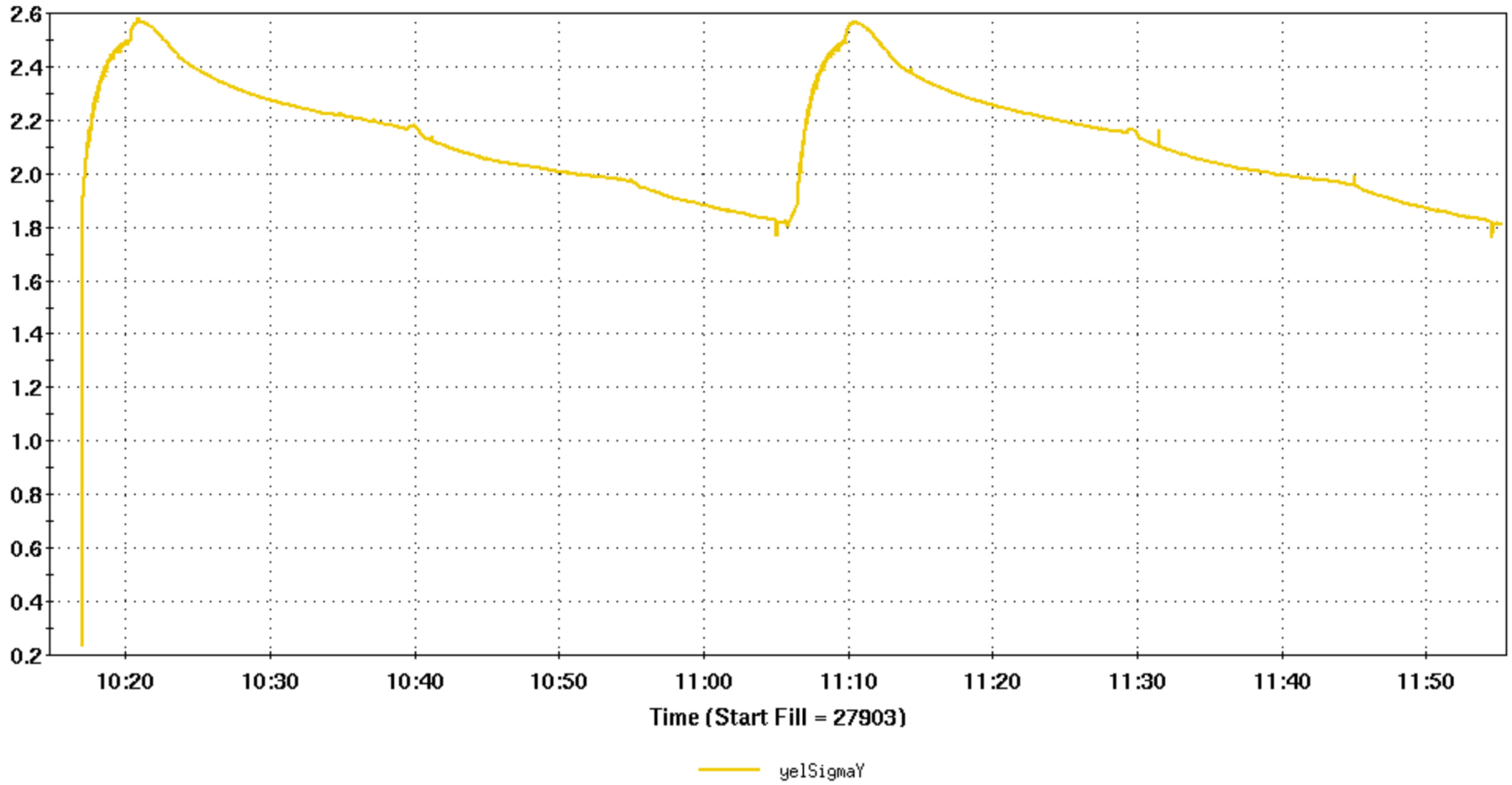
$$k_{\mathrm{RBF}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T \Sigma (\boldsymbol{x} - \boldsymbol{x}')\right]$$

- Accurately estimate the precision matrix sigma is important.
- Physics-informed calculation of the sigma matrix: $\Sigma = -H/2$., H is the Hessian matrix around the optimal point.
- Better capture the relationship between data. No need to fit the data anymore.

Physics-informed GP



Benefits


2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
10:20
10:30
10:40
10:50
11:00
11:10
11:20
11:30
11:40
11:50
Time (Start Fill = 27903)
yelSigmaY