Universitetet i Stavanger

Fast Bayesian inference with Gaussian Processes

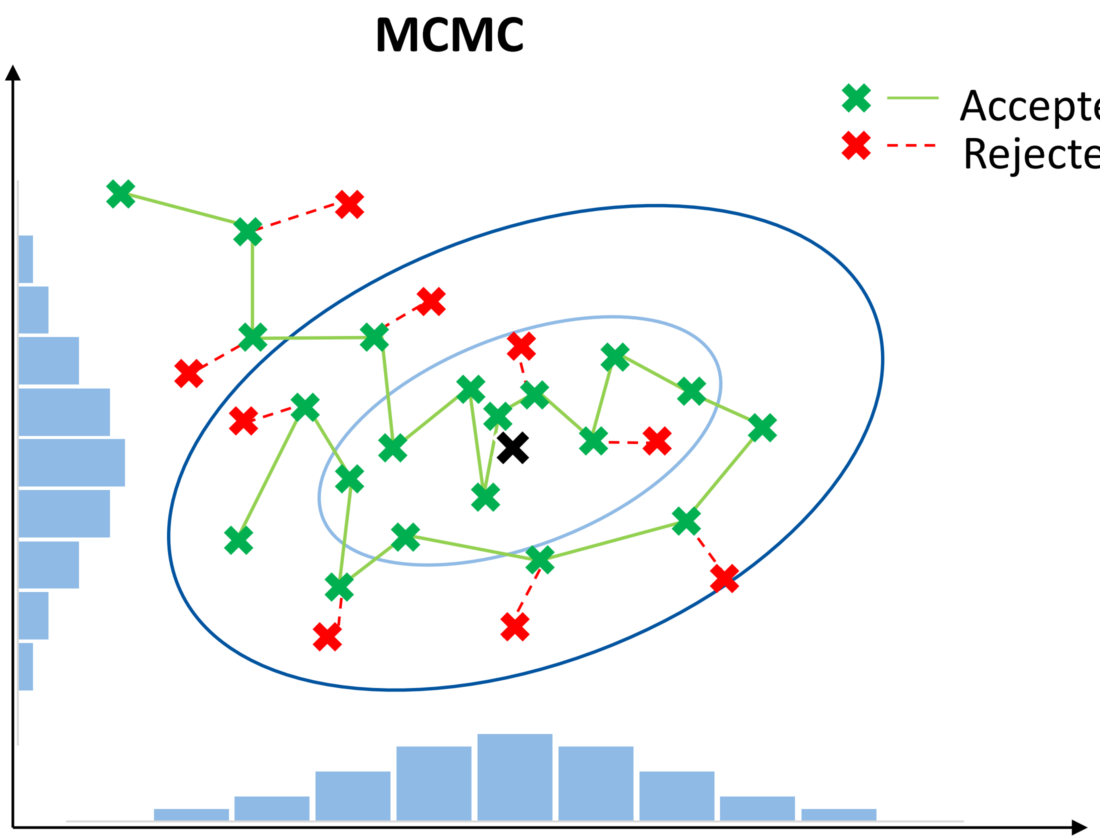
LIDA Toulouse 2022

arXiv:2211.02045
https://github.com/jonaselgammal/GPry

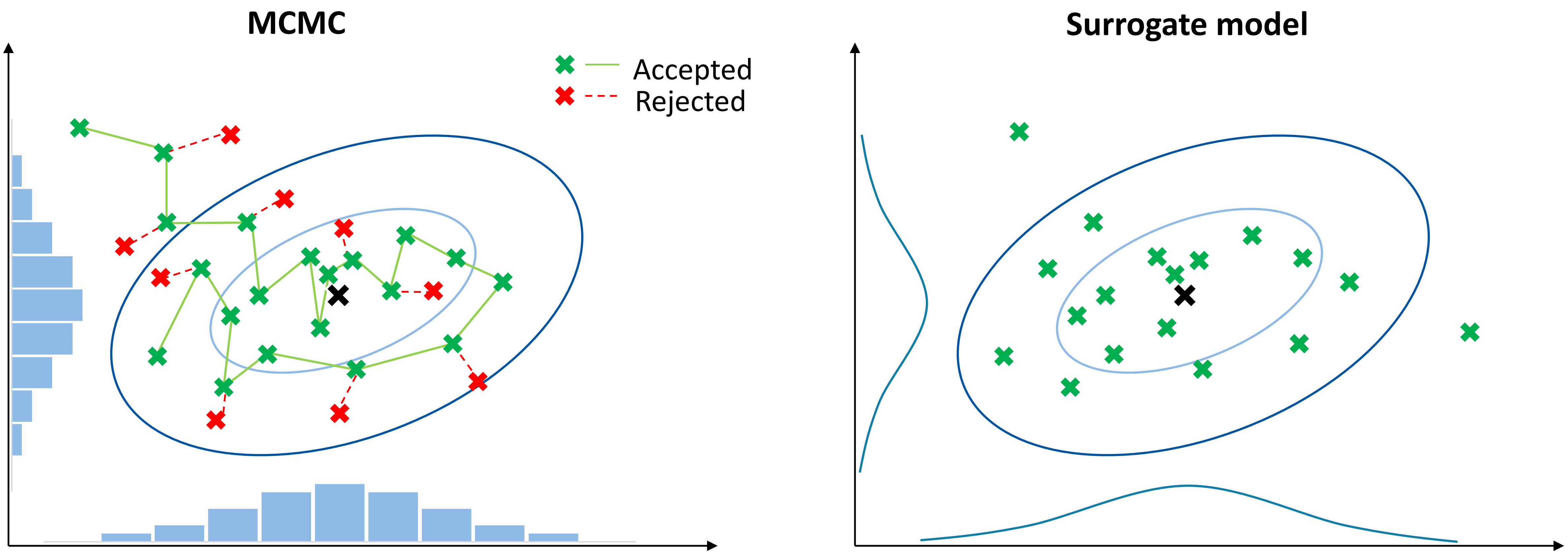
JONAS EL GAMMAL (UNIVERSITY OF STAVANGER)

WITH J. TORRADO, N. SCHÖNEBERG, R. BUSCICCHIO, G. NARDINI, C. FIDLER

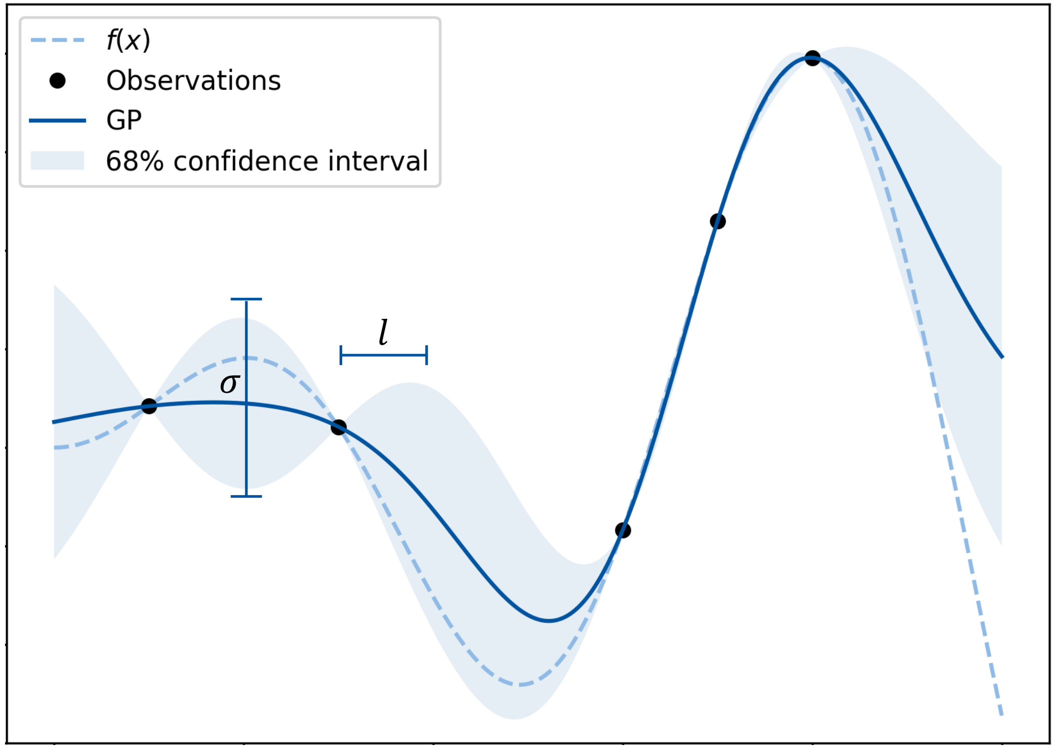
1. Idea



1. Idea

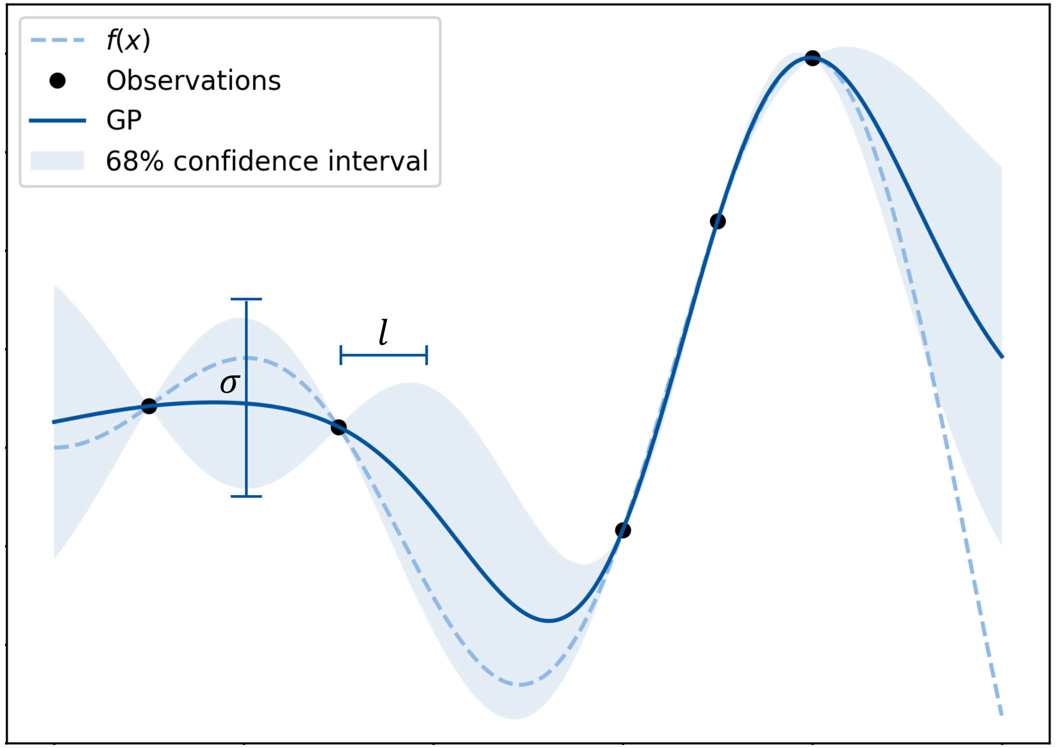


2. Gaussian Process Surrogate

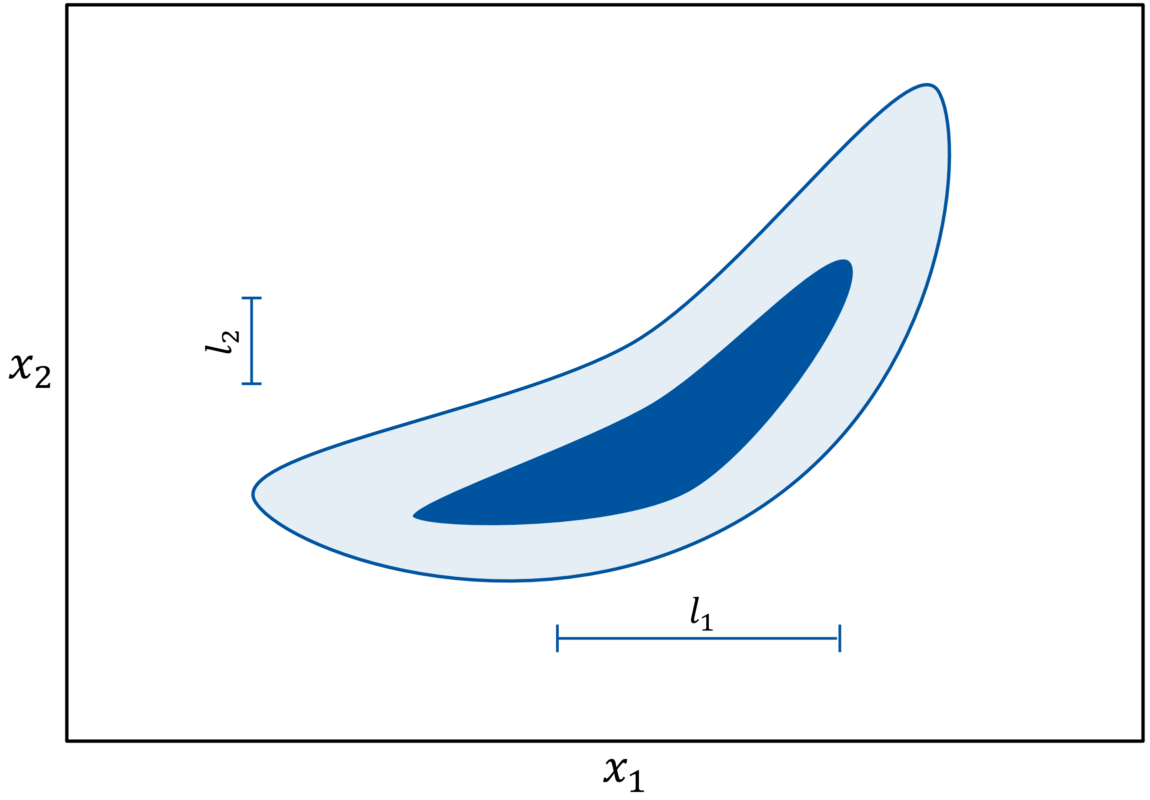


$$k(x, x') = \sigma^2 \cdot \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

2. Gaussian Process Surrogate

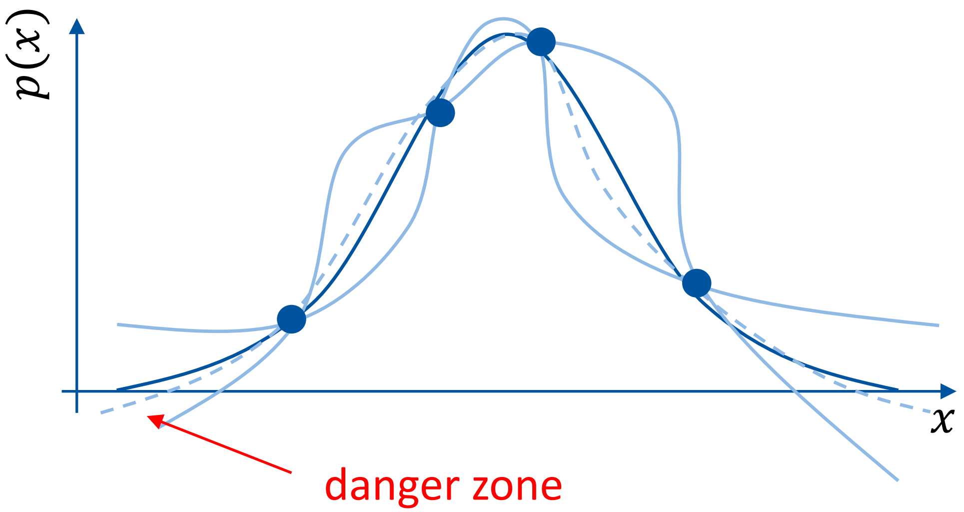


$$k(x, x') = \sigma^2 \cdot \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

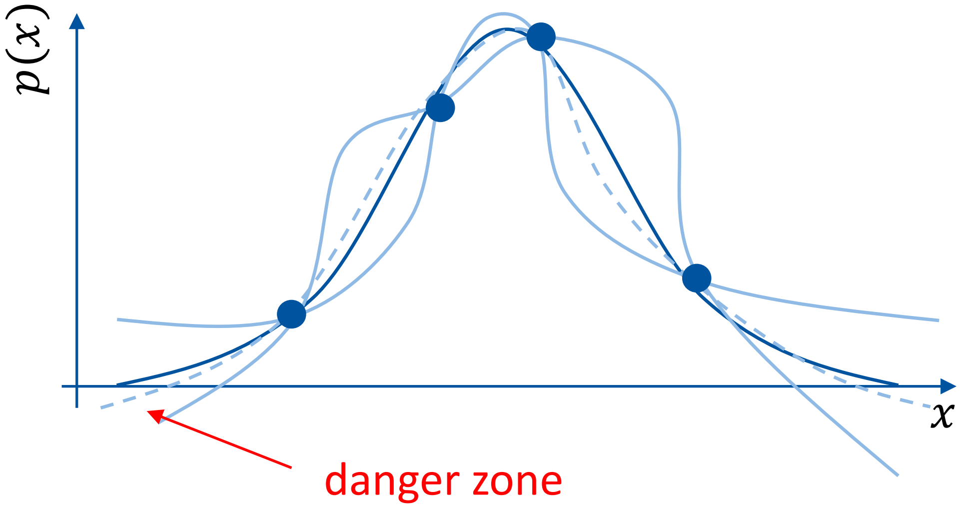


$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \cdot \exp\left(-\sum \frac{(x_i - x_i')^2}{2l_i^2}\right)$$

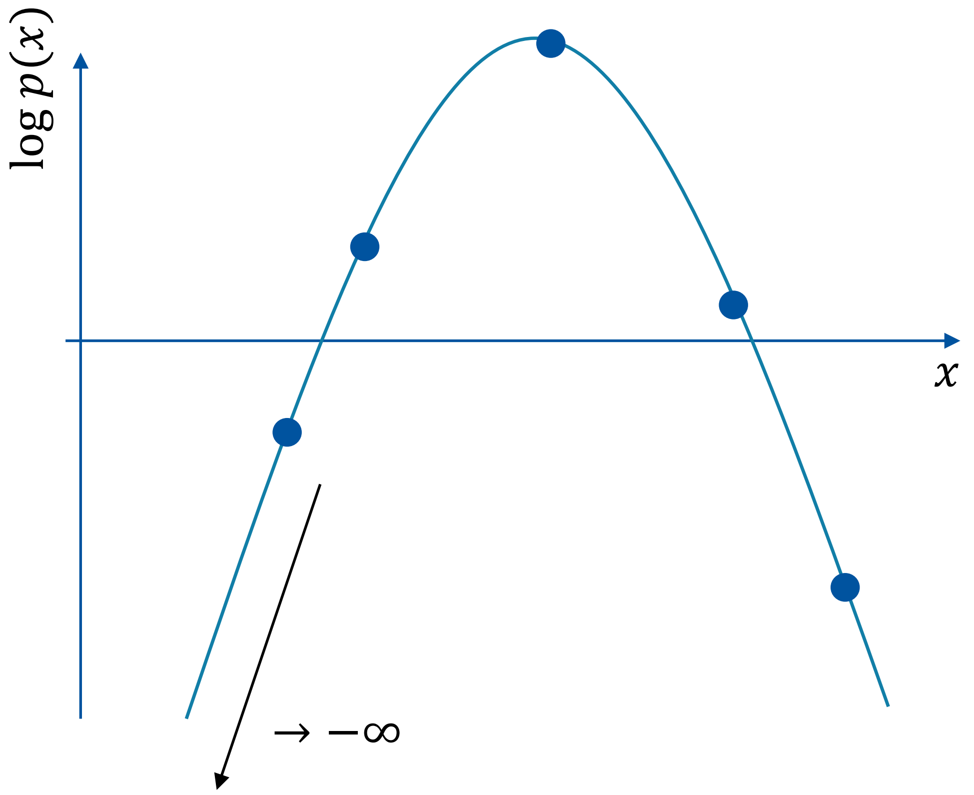
3. Region of interest



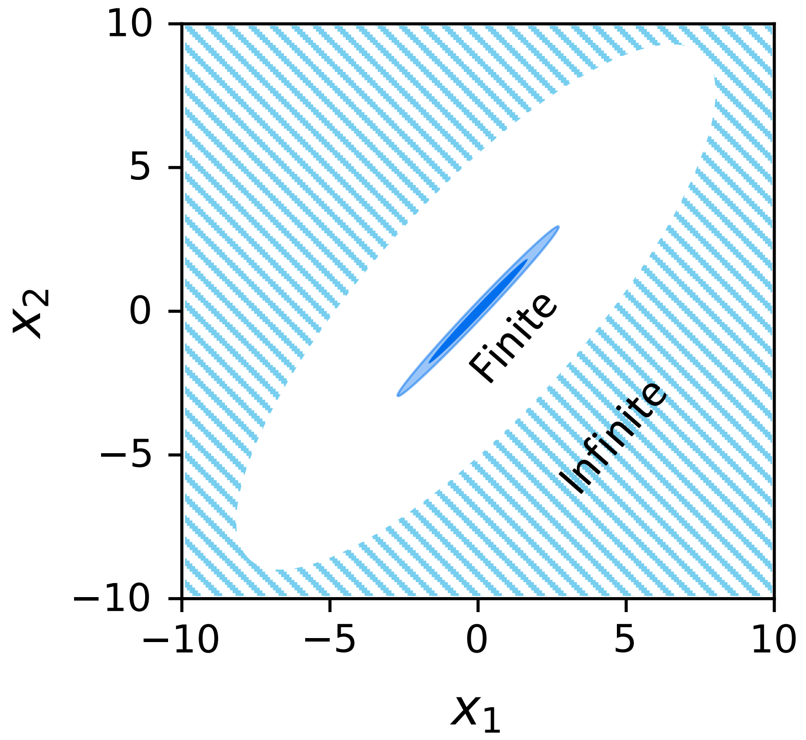
3. Region of interest



⇒Interpolate **log-posterior** to enforce positivity

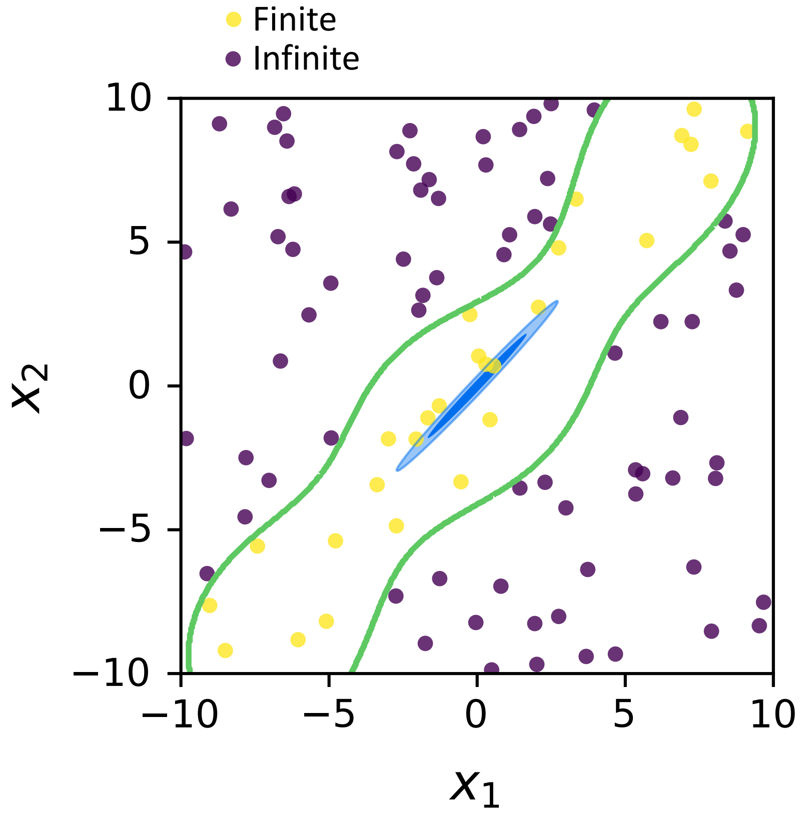


3. Region of interest



Solution: SVM Classifier

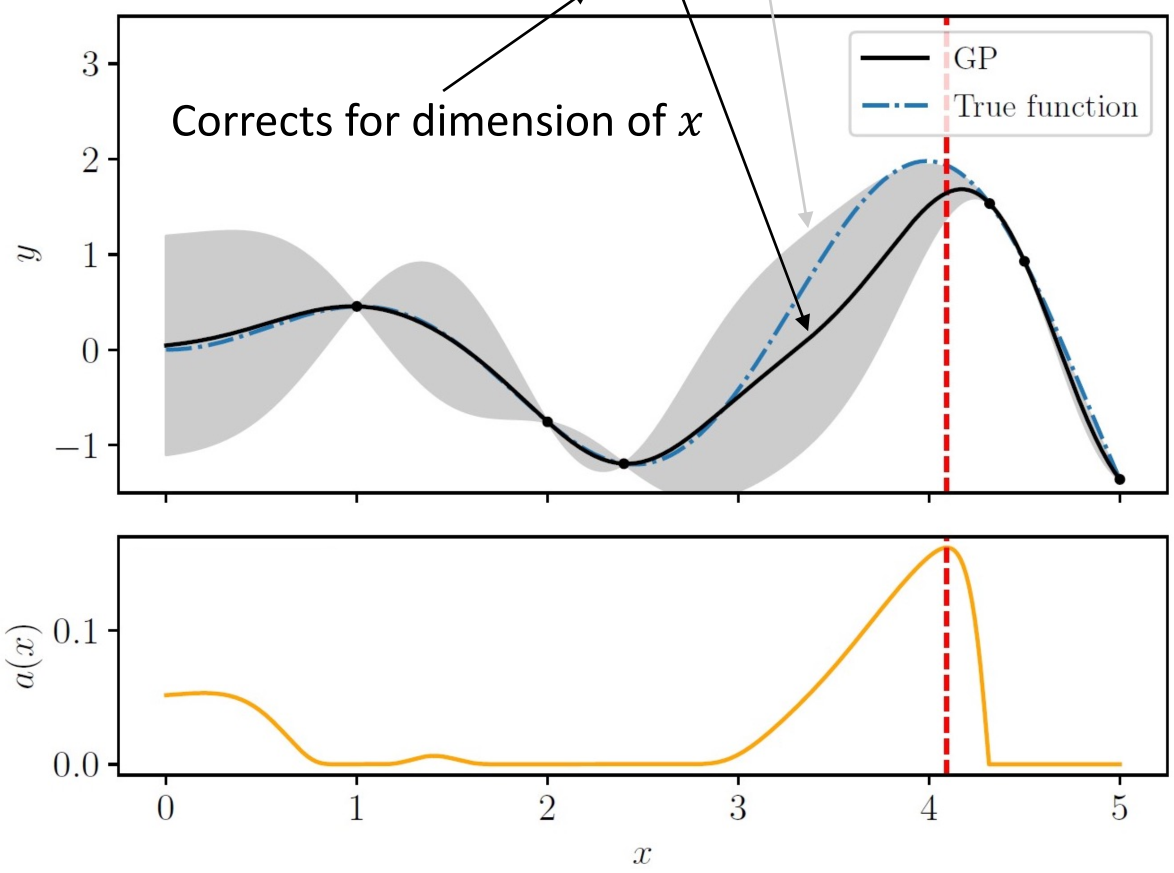
Multiply μ with $-\infty$ where SVM classifies as "infinite"



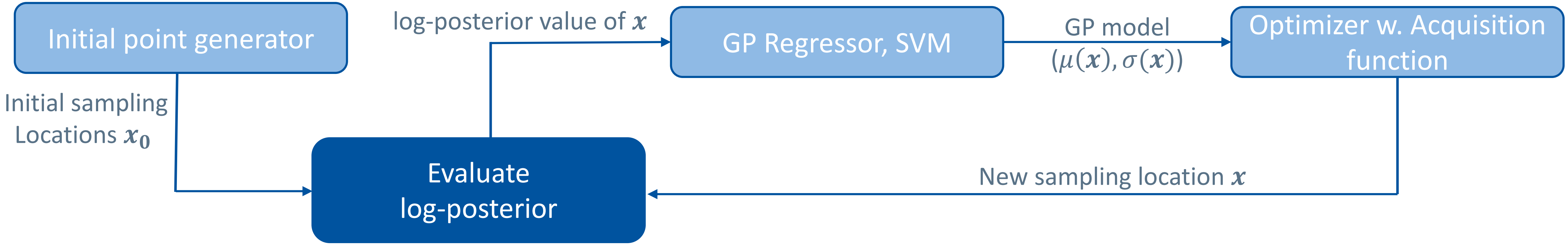
4. Active sampling

Propose samples by maximizing an **acquisition function**

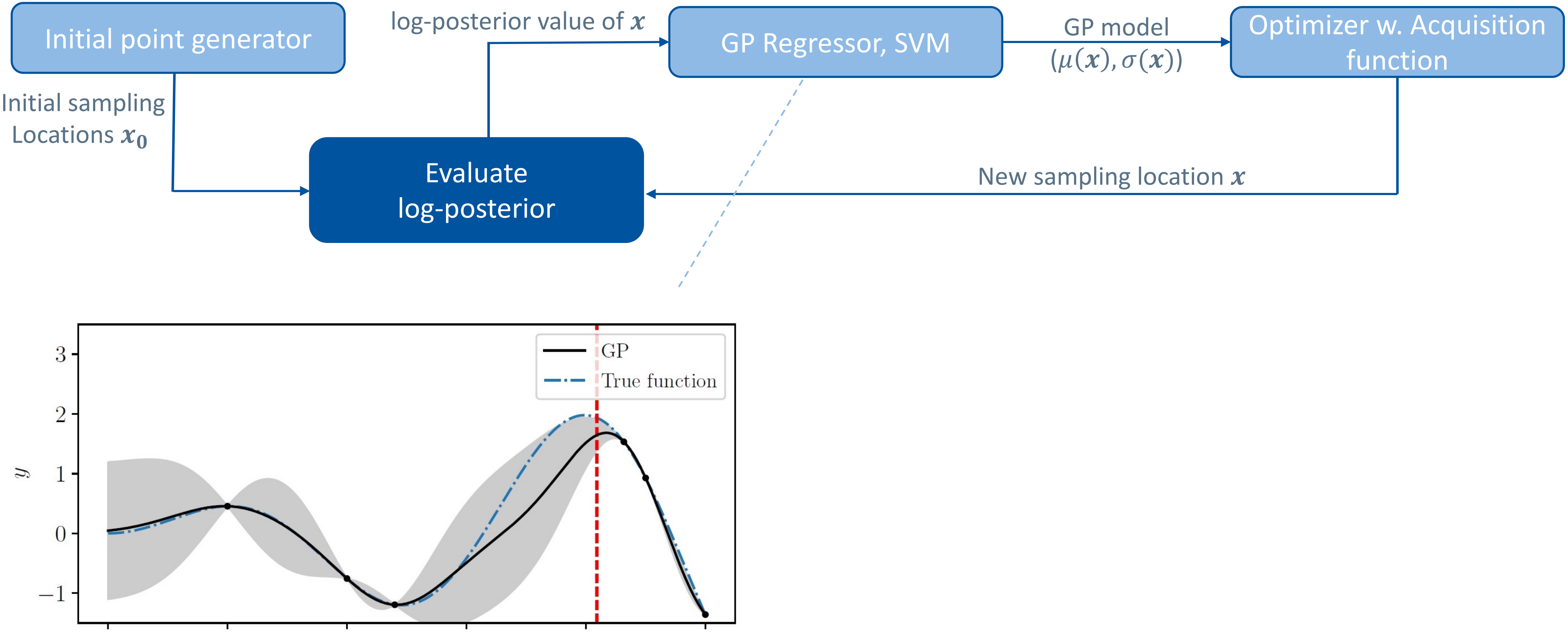
$$a(x) = \exp(2\zeta \cdot \bar{\mu}) \cdot \sigma_{\bar{\mu}}(x)$$



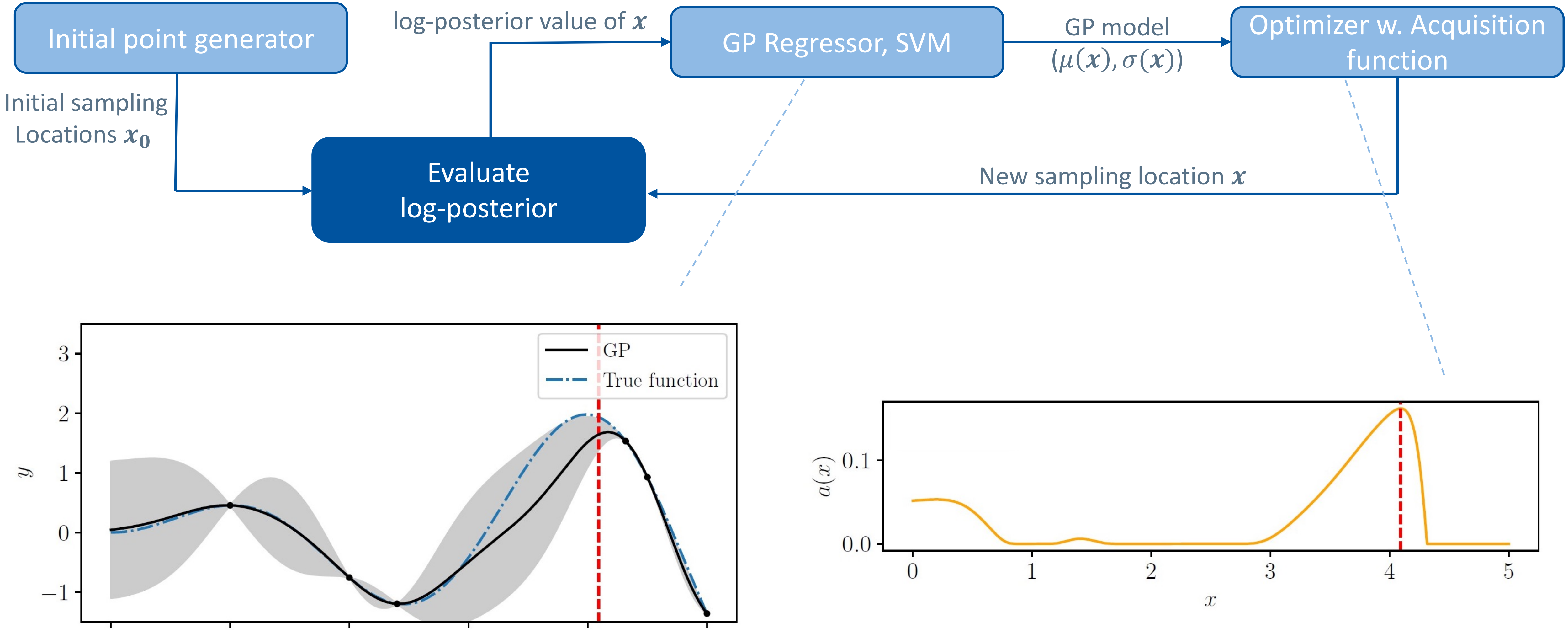
5. The Algorithm



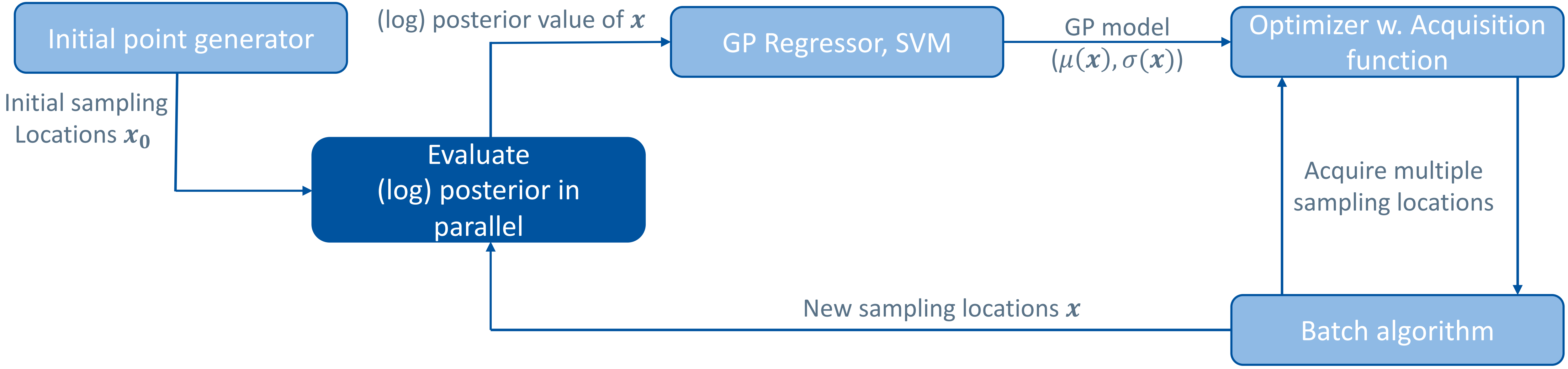
5. The Algorithm



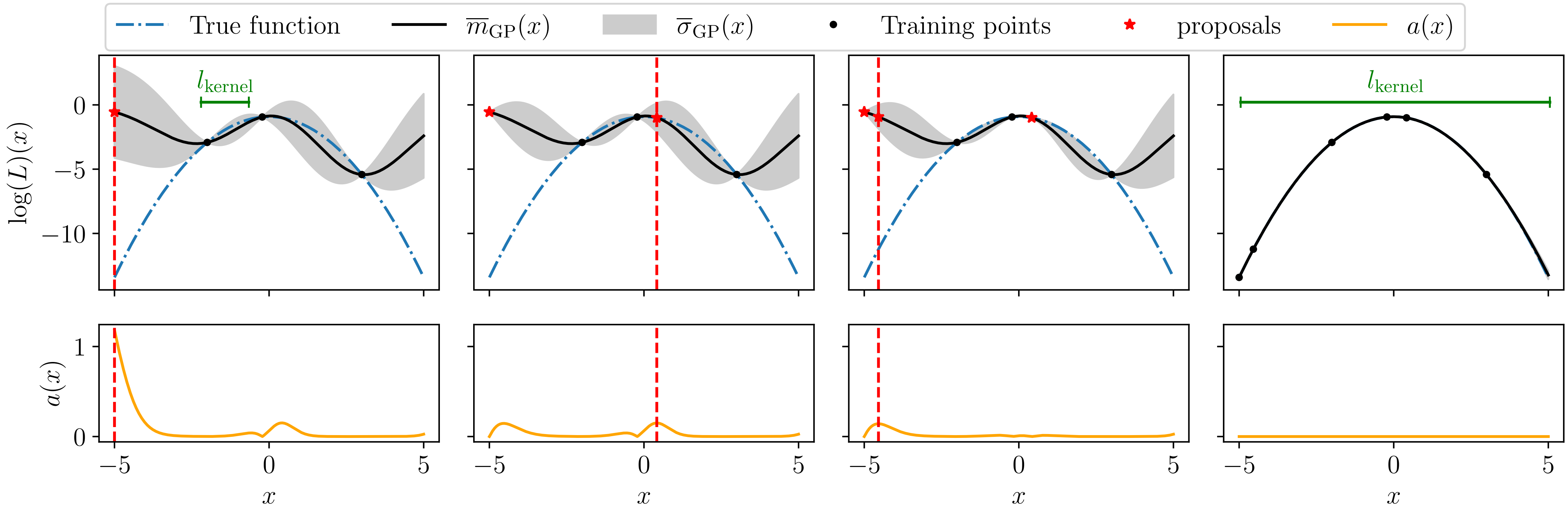
5. The Algorithm



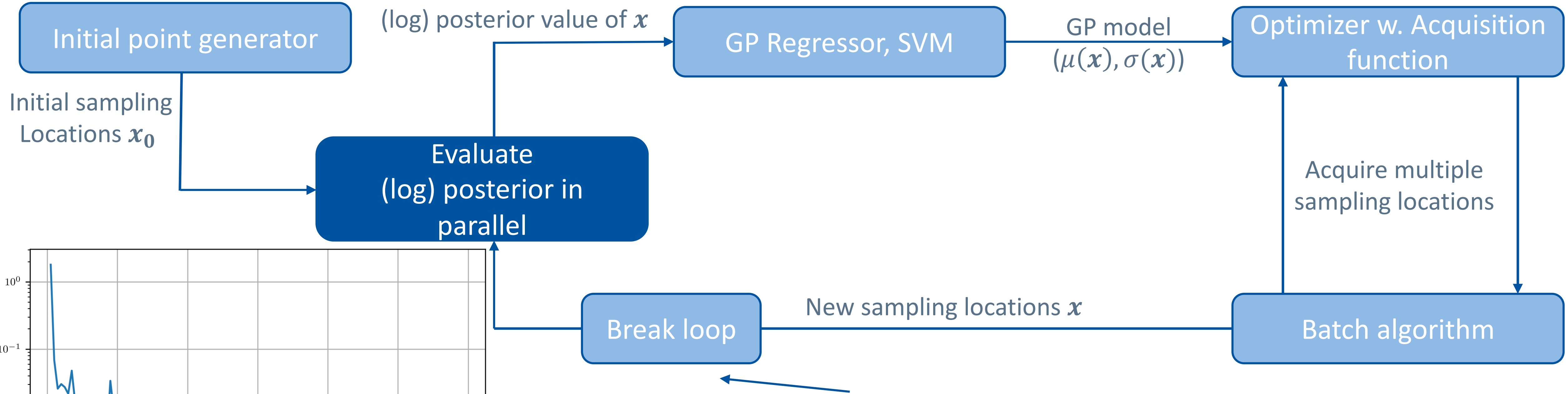
5. The Algorithm



5. The Algorithm

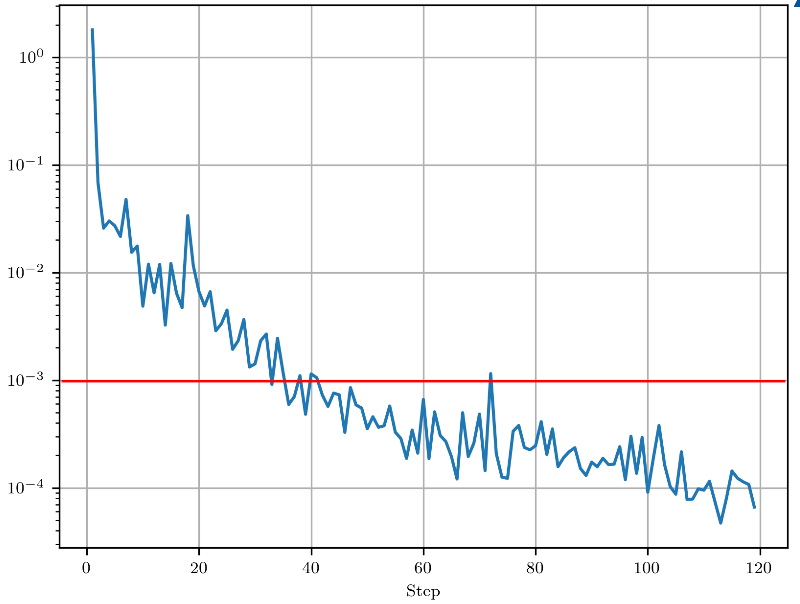


5. The Algorithm



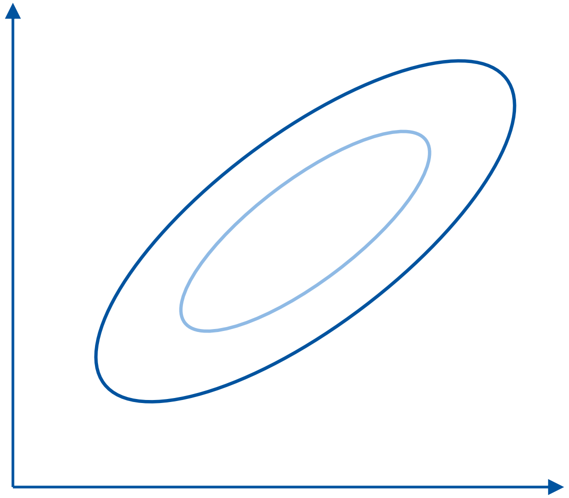
Break if

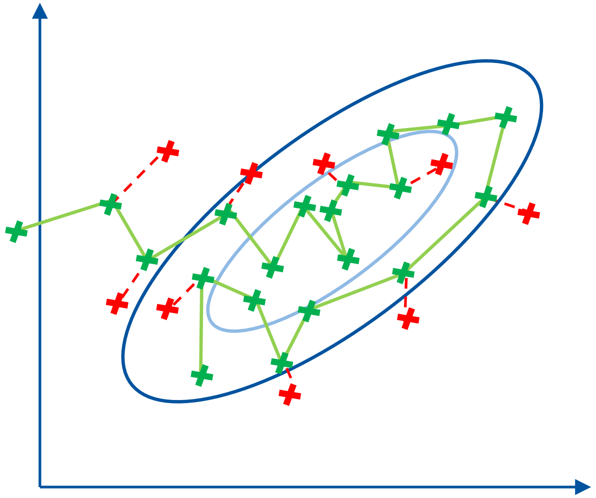
$$\left|\log p_{\text{true}} - \log p_{\text{predict}}\right| < \alpha + \beta * |\log p_{\max} - \log p_{\text{true}}|$$

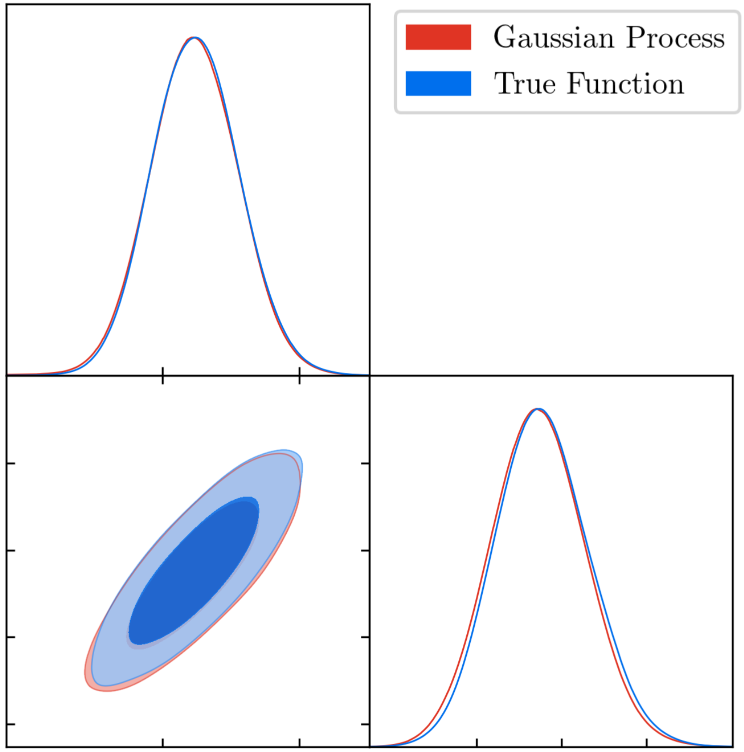


6. Marginalised quantities

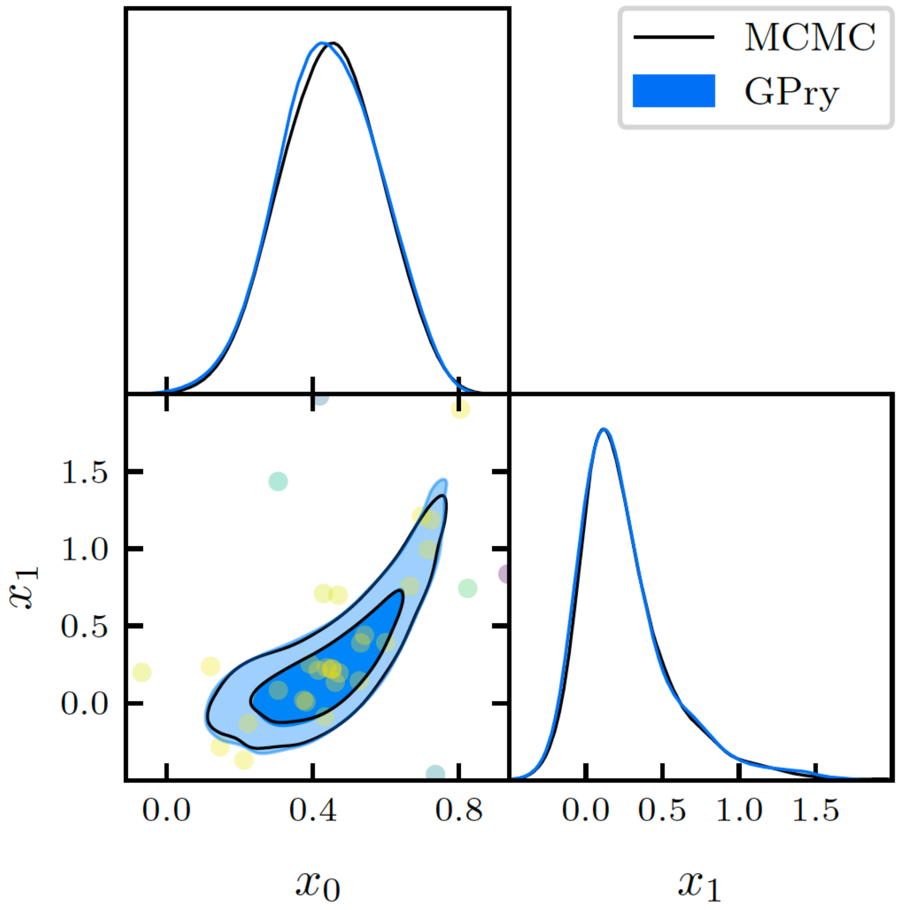
GP model $(\mu(x), \sigma(x))$ → MC sampler → Plotting etc.



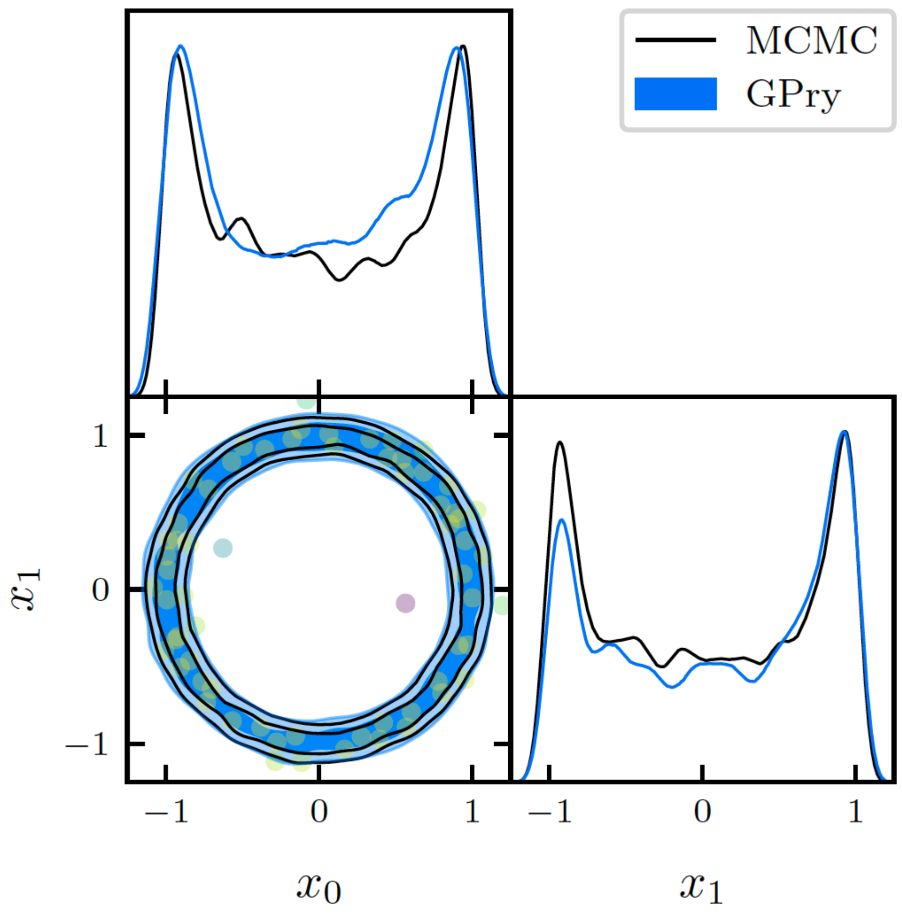




7. Experiments

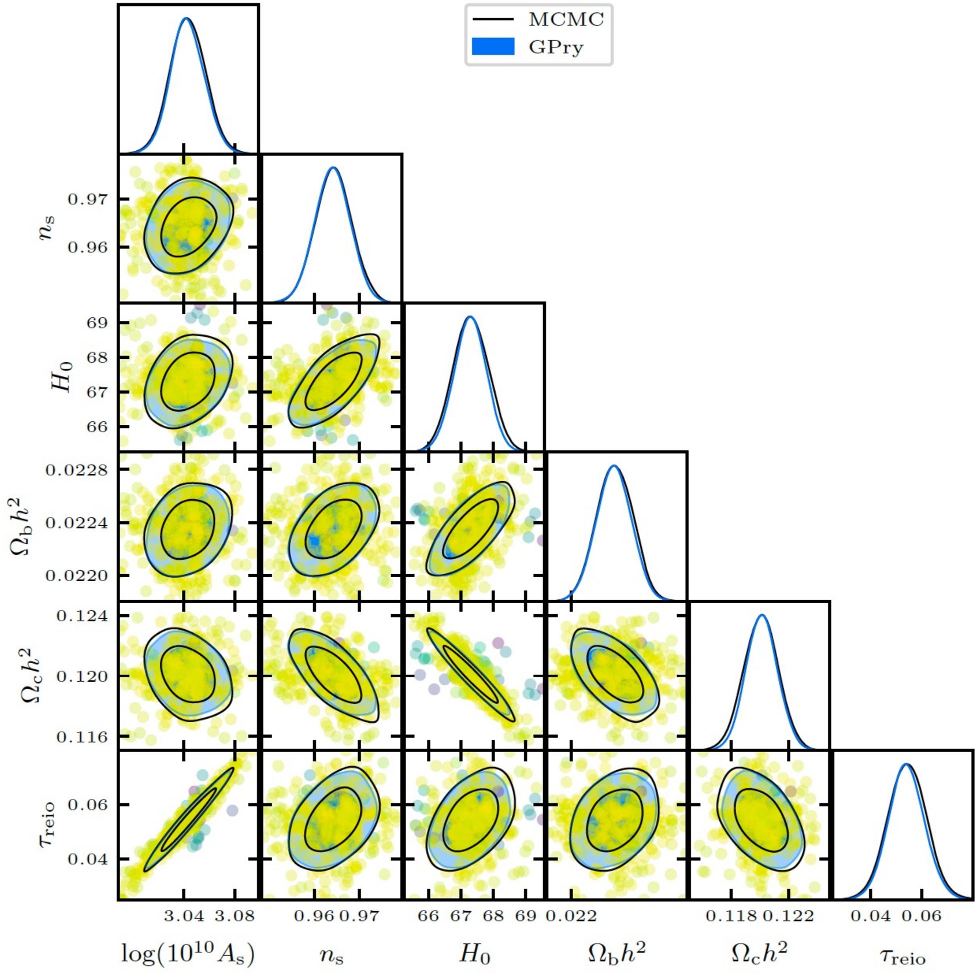


40 posterior evaluations

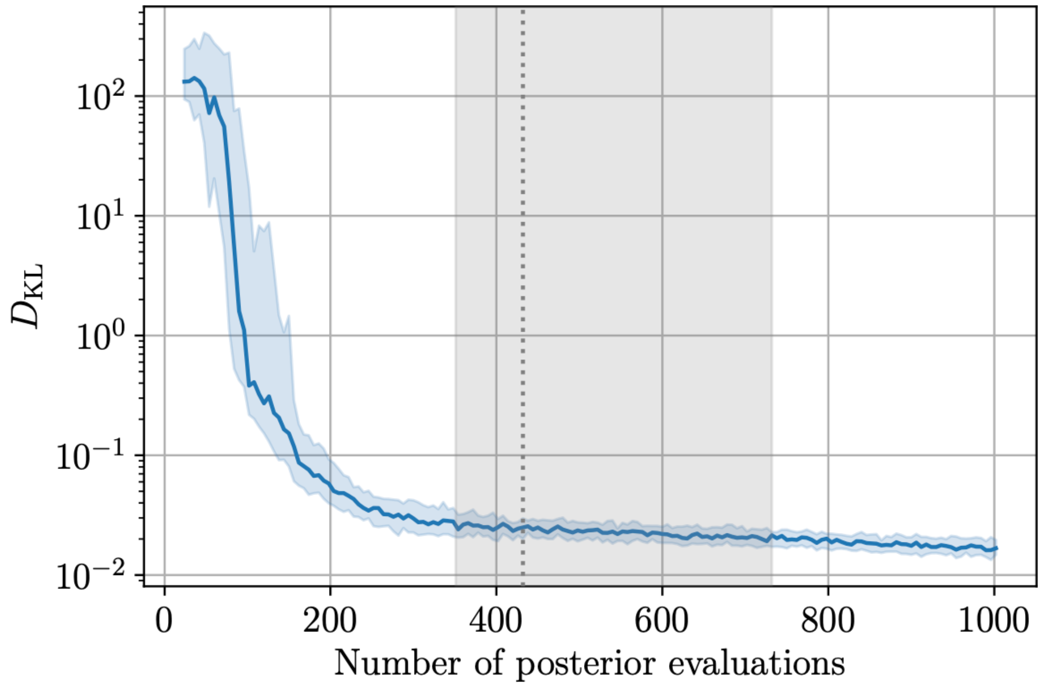


68 posterior evaluations

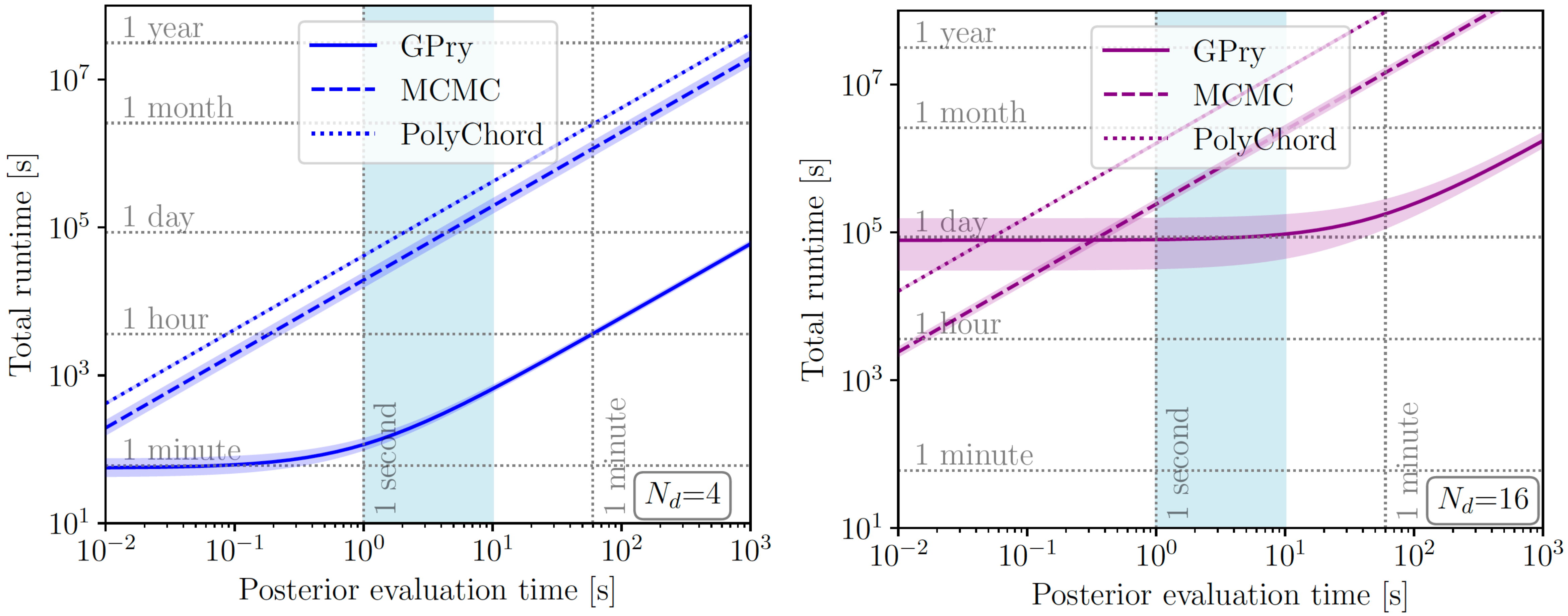
7. Experiments



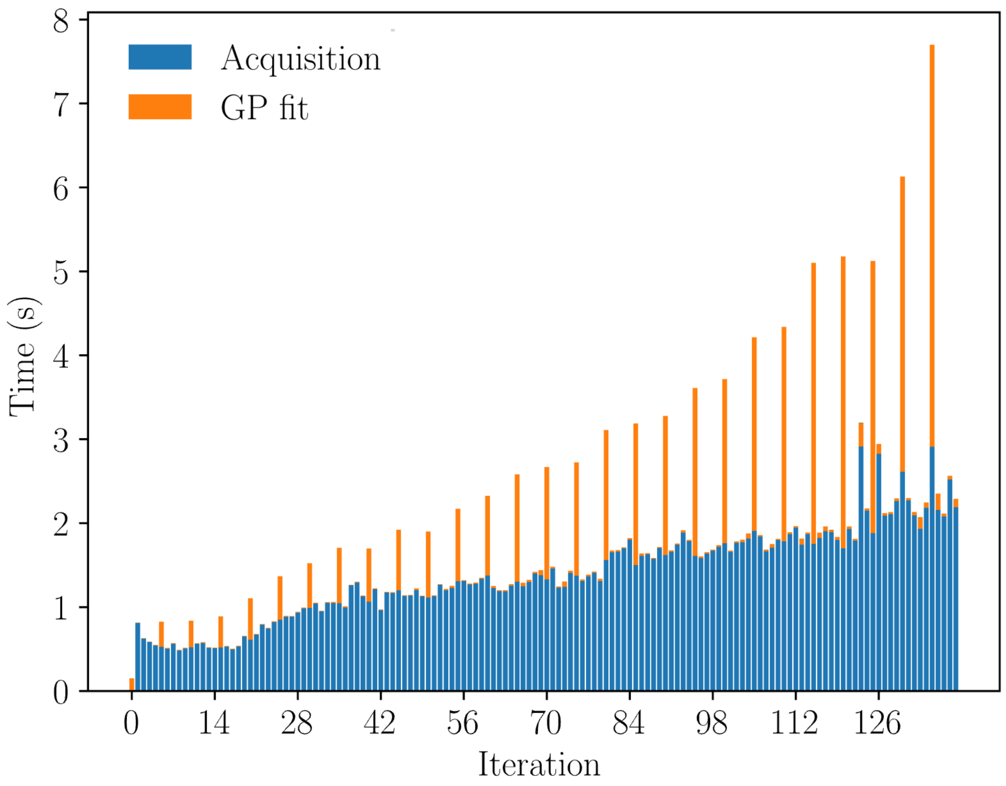
420 posterior evaluations



8. Performance

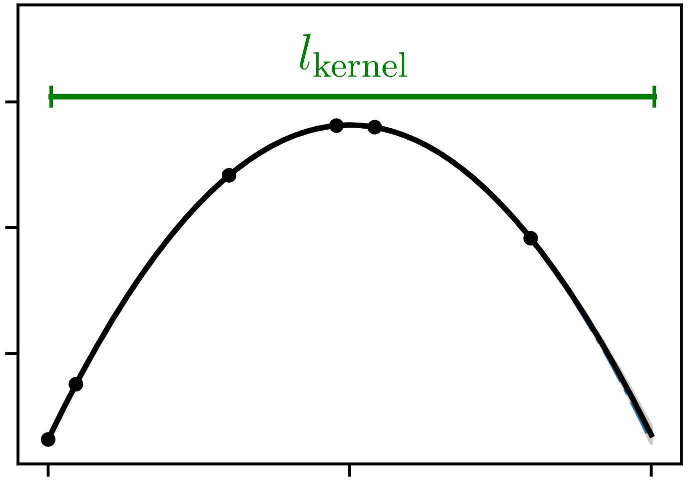


9. Limitations



Overhead

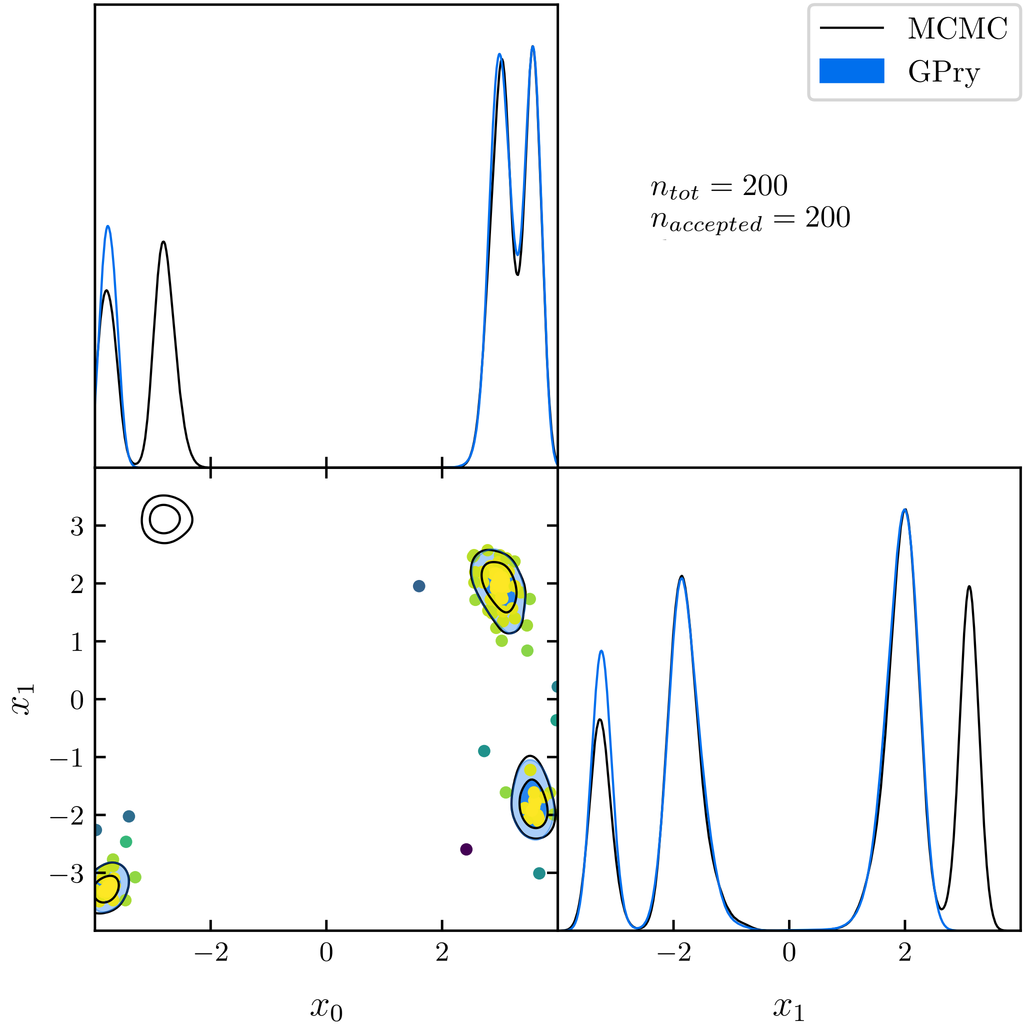
9. Limitations



Overhead

Overfitting

9. Limitations



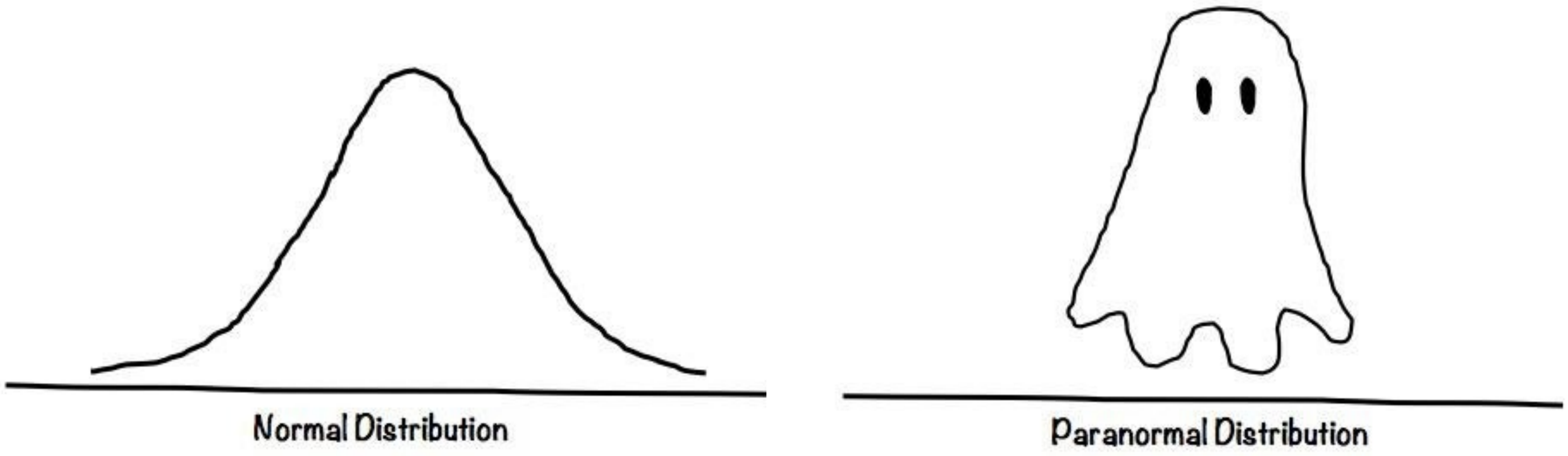
Overhead

Overfitting

Multimodality

We work on solving those problems…

Thank you!

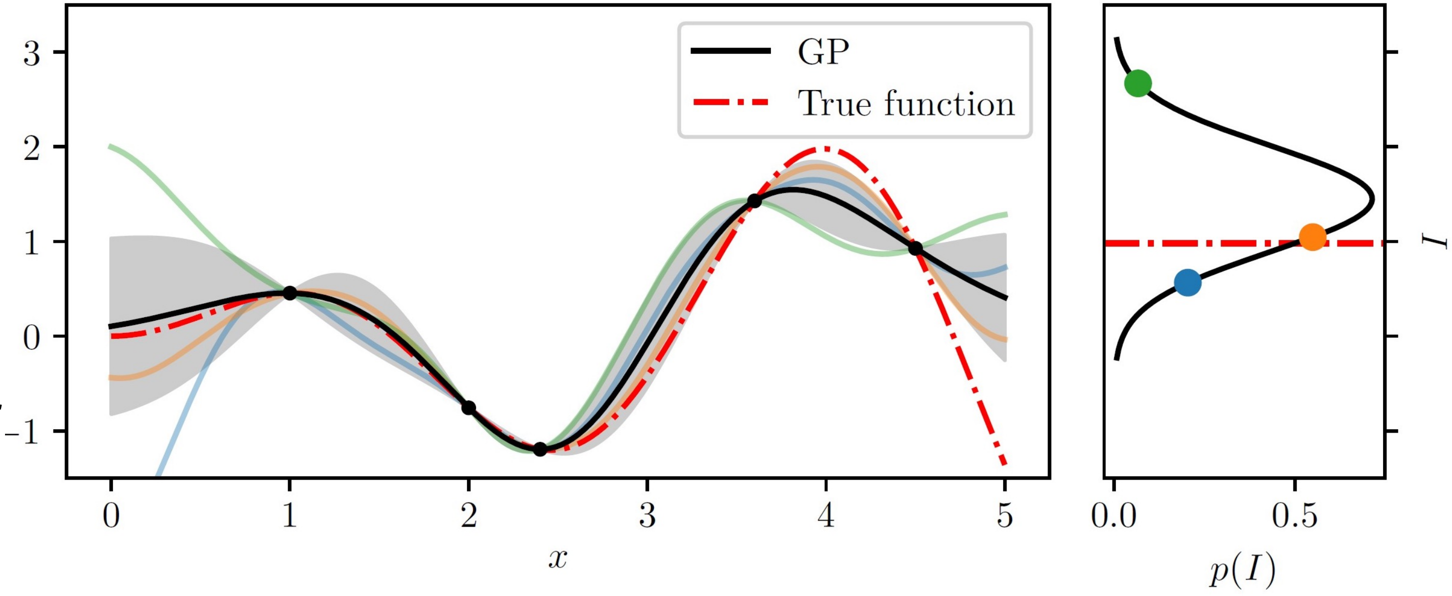


https://www.memedroid.com/memes/detail/3518248/Normal-vs-paranormal-distribution

Backup

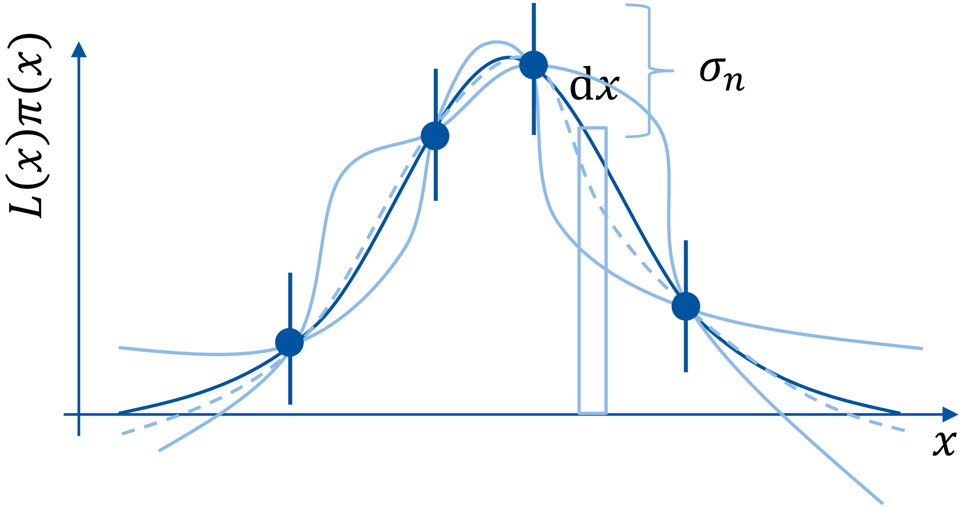
3. Active sampling

- To get marginalised quantities we want to integrate
  $$\int L(x)\pi(x)\,\mathrm{d}x$$
- With a GP we can get a model for
  $$L(x)\pi(x) \sim \mathcal{GP}(0, k(x, x'))$$
- We can integrate that model by integrating $\int \mu(x)\,\mathrm{d}x = \int \bar{f}(x)\,\mathrm{d}x$
- We can use $\mu(x)$ and $\sigma(x) = \sqrt{\mathrm{cov}(f_*(x,x))}$ to find the next most informative point to sample



3. Active sampling

⇒ At each step maximize an **acquisition function**



$L(x)\pi(x)$ is always positive

$\Rightarrow a(x) = \mu(x) \cdot \sigma(x)$

$L(x)\pi(x)$ has high dynamic range

⇒ Sample log-posterior:

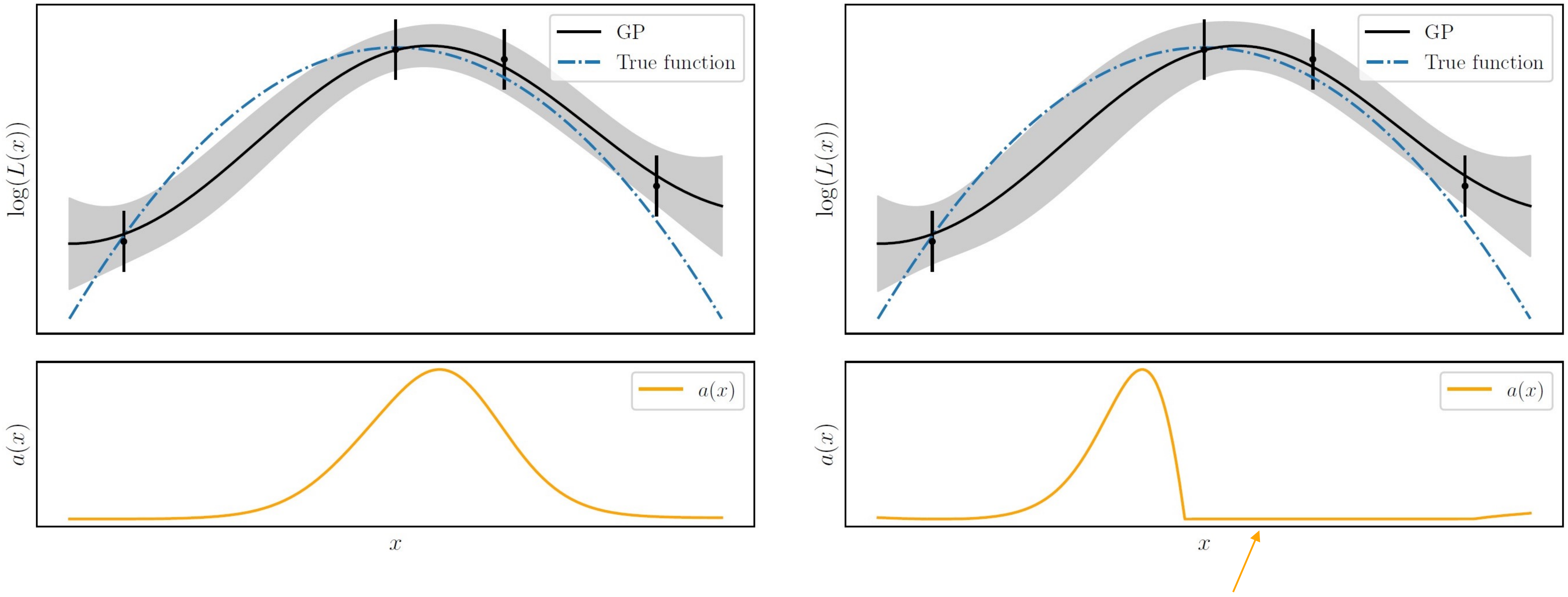
$$a(x) = \exp(2 \cdot \bar{\mu}) \cdot \sigma_{\bar{\mu}}(x)$$

$\bar{\mu}$ = Mean of GP fit to log-posterior

Correction factor ζ and statistical noise σ_n

$$a(x) = \exp(2\zeta \cdot \bar{\mu}) \cdot (\sigma_{\bar{\mu}}(x) - \sigma_n)$$

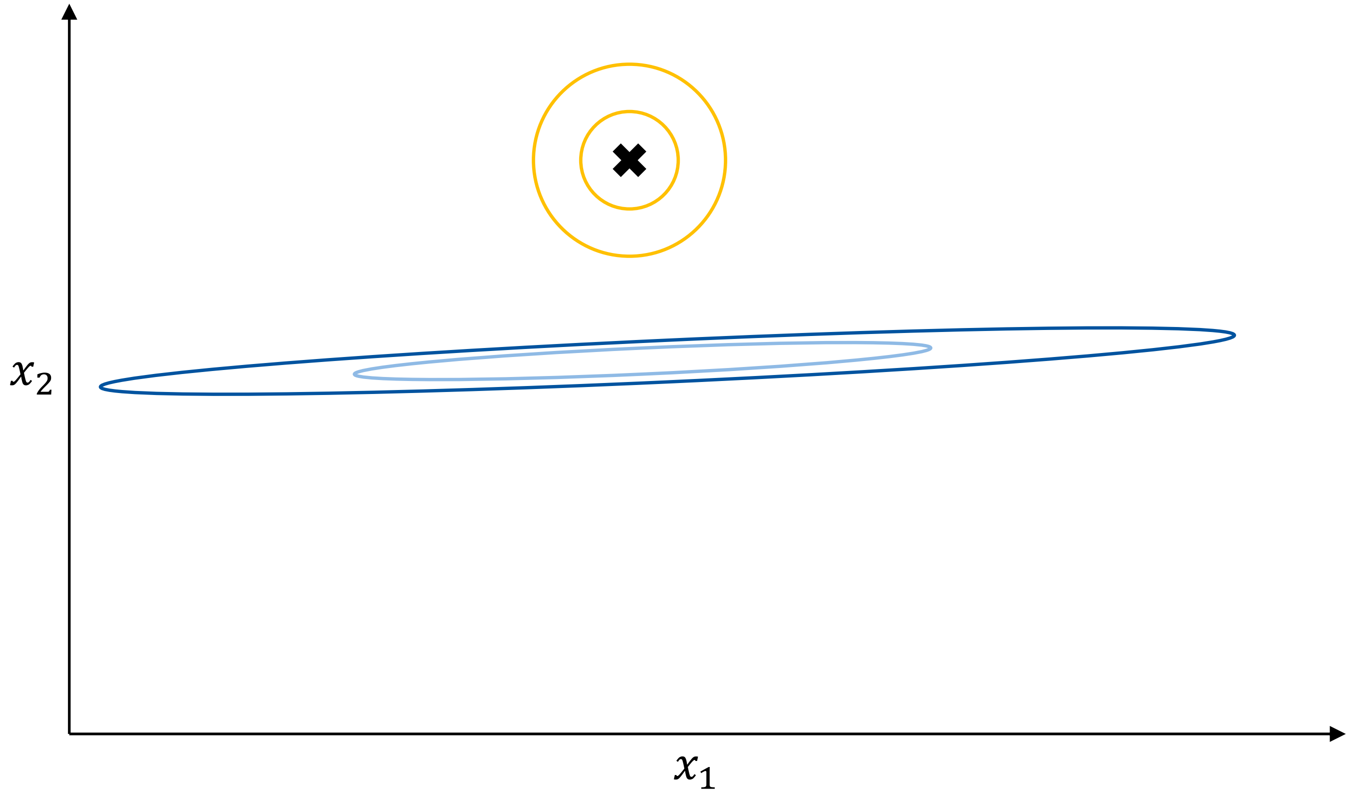
The acquisition function



Flat over large areas ⇒ We take the log of the acquisition function when actually optimizing it

Preprocessing

Problem 1: Different scales



Do two things:

1. Scale the priors such that they occupy the unit hypercube (every parameter is in $[0,1]$)

2. Make kernel asymmetric

$$k(x, x') = \sigma^2 \cdot \prod_{i=1}^{d} \exp\left(-\frac{(x_i - x_i')^2}{2l^2}\right)$$

More hyperparameters to fit ($d + 1$) but robust!

Preprocessing

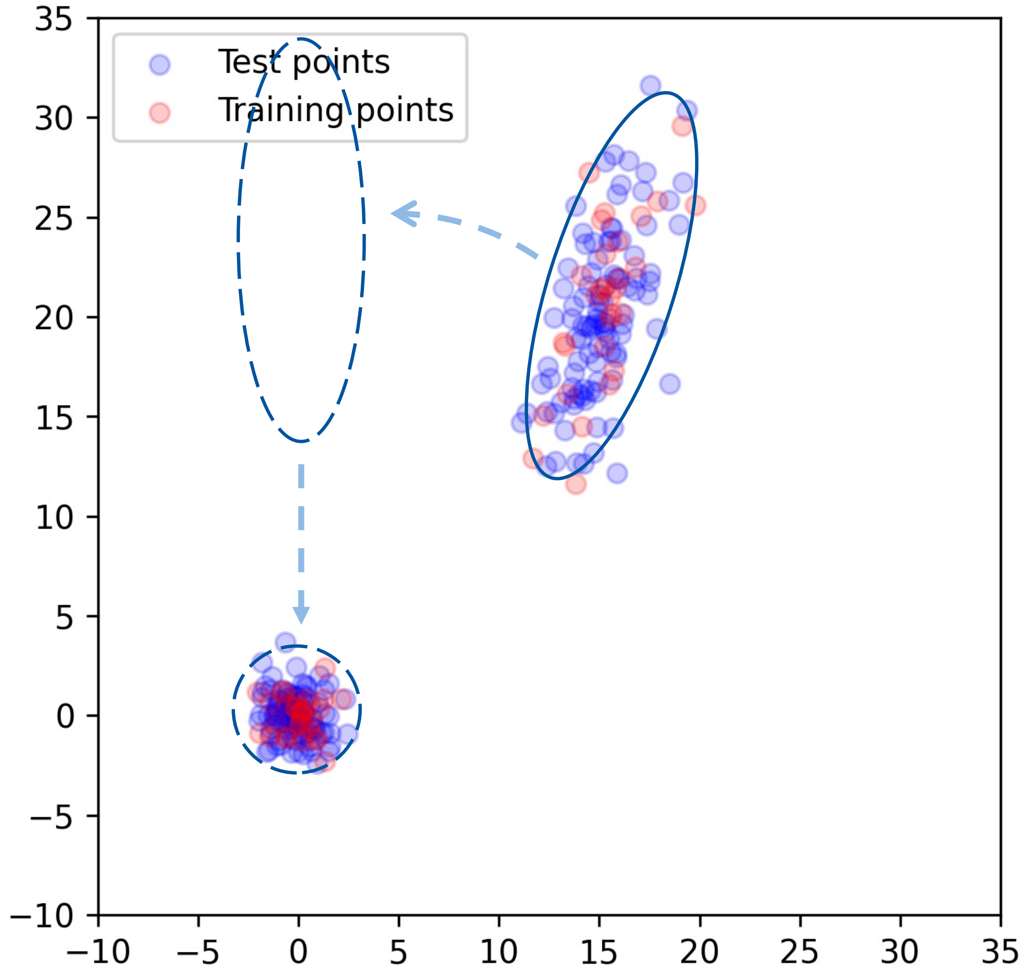
Alternative: Whitening

$$x_i \rightarrow x_i{}' = \frac{R_{ij}(x - \hat{\mu})_j}{\hat{\Sigma}_{ii}}$$

with

- $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
- $\hat{\Sigma}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \hat{\mu}_j\right)\left(x_{ik} - \hat{\mu}_k\right)$
  (empirical mean and covariance along each dimension)
- $\hat{\Sigma} = R\Lambda\mathrm{R}^{\mathrm{T}}$ with Λ diagonal

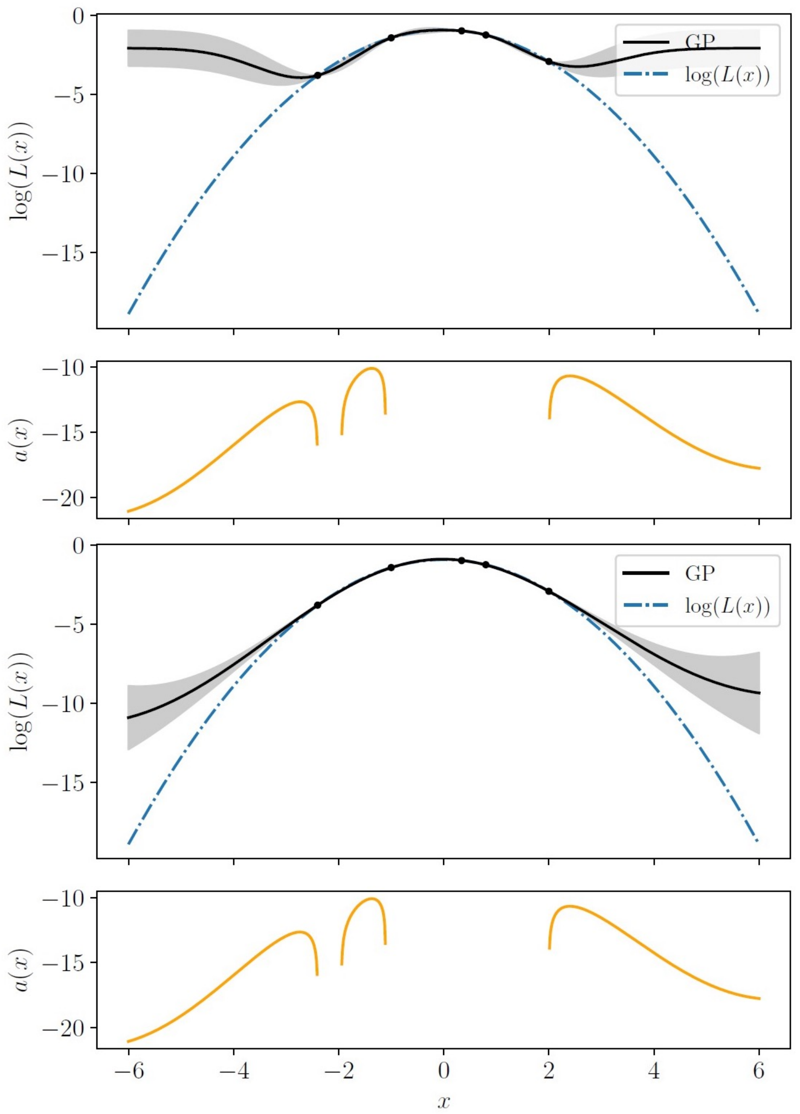
$\Rightarrow$ No need to make kernel asymmetric but less robust



Preprocessing

What about log-posterior values?

- Transform such that they have zero mean and unit variance
- Encourages exploration when lots of high values of the log-posterior
- Encourages exploitation when lots of low values of the log-posterior



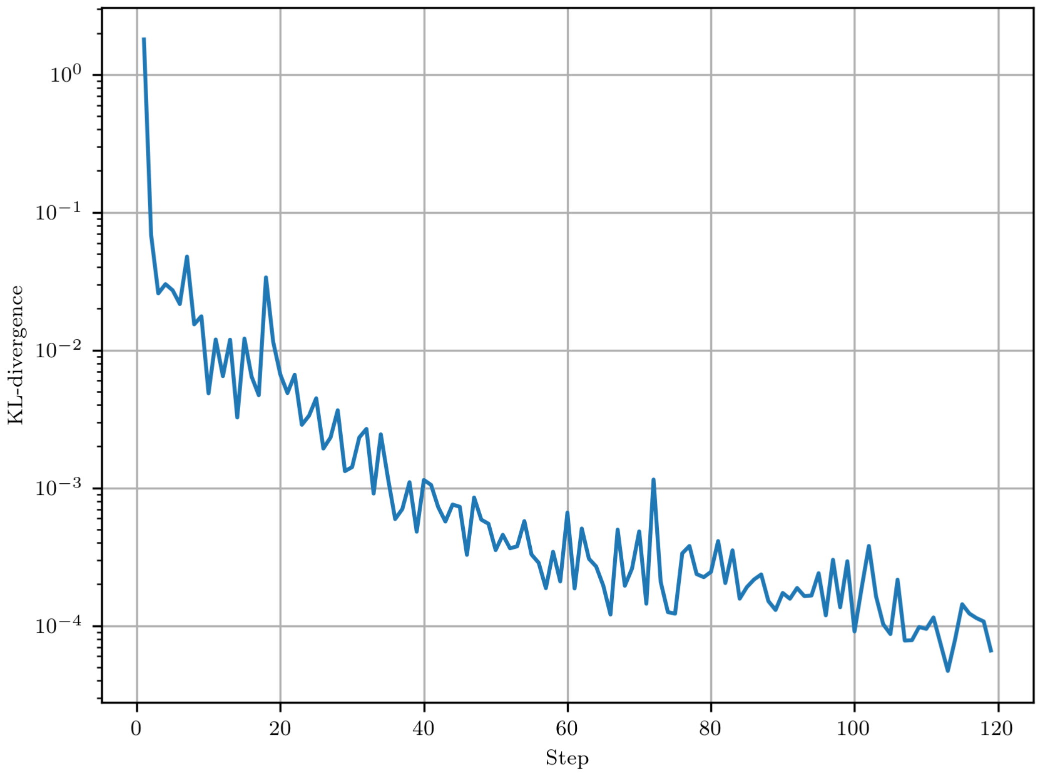
Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(P_{n+1}||P_n) = \sum_{x\in\chi} P_{n+1}(x)\log\left(\frac{P_{n+1}(x)}{P_n(x)}\right)$$

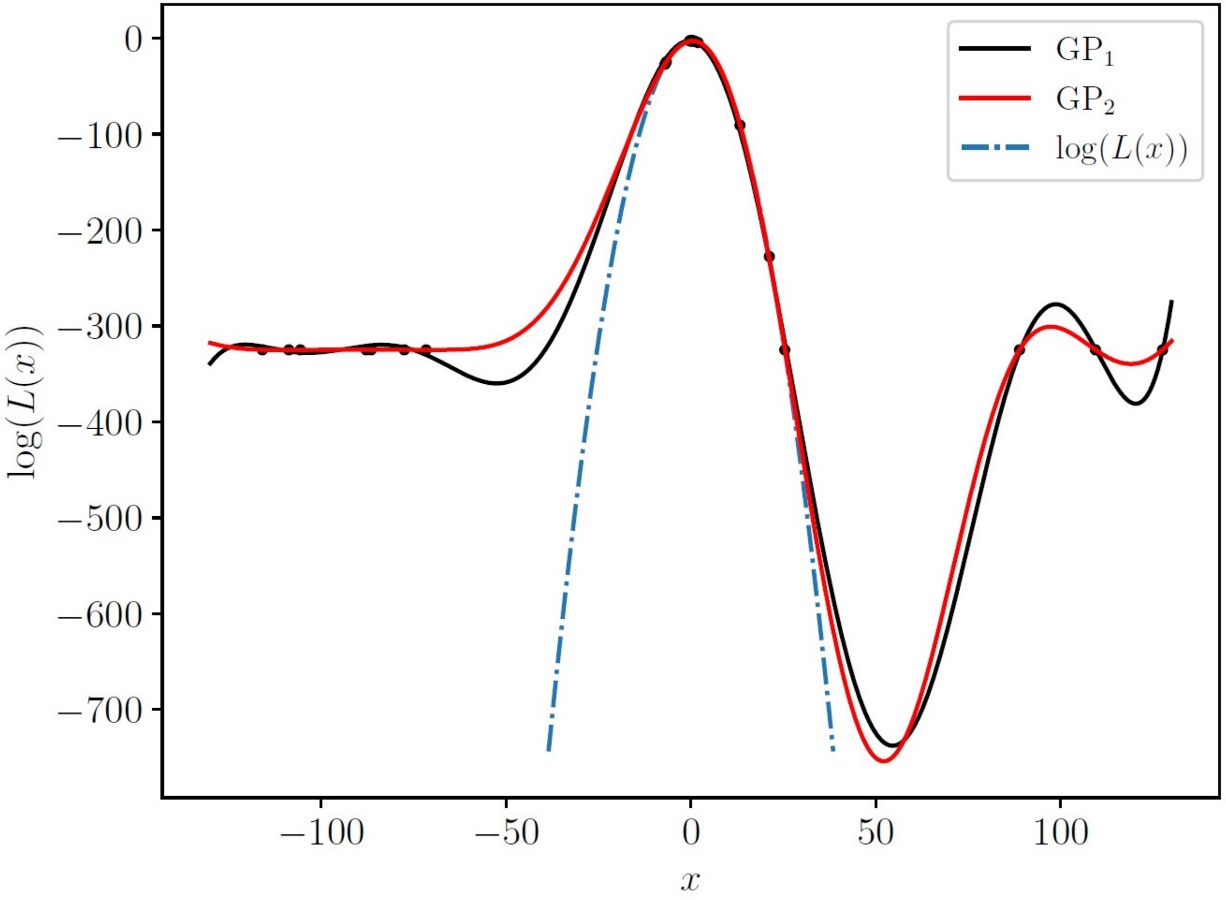
In case of a multivariate Gaussian this is just

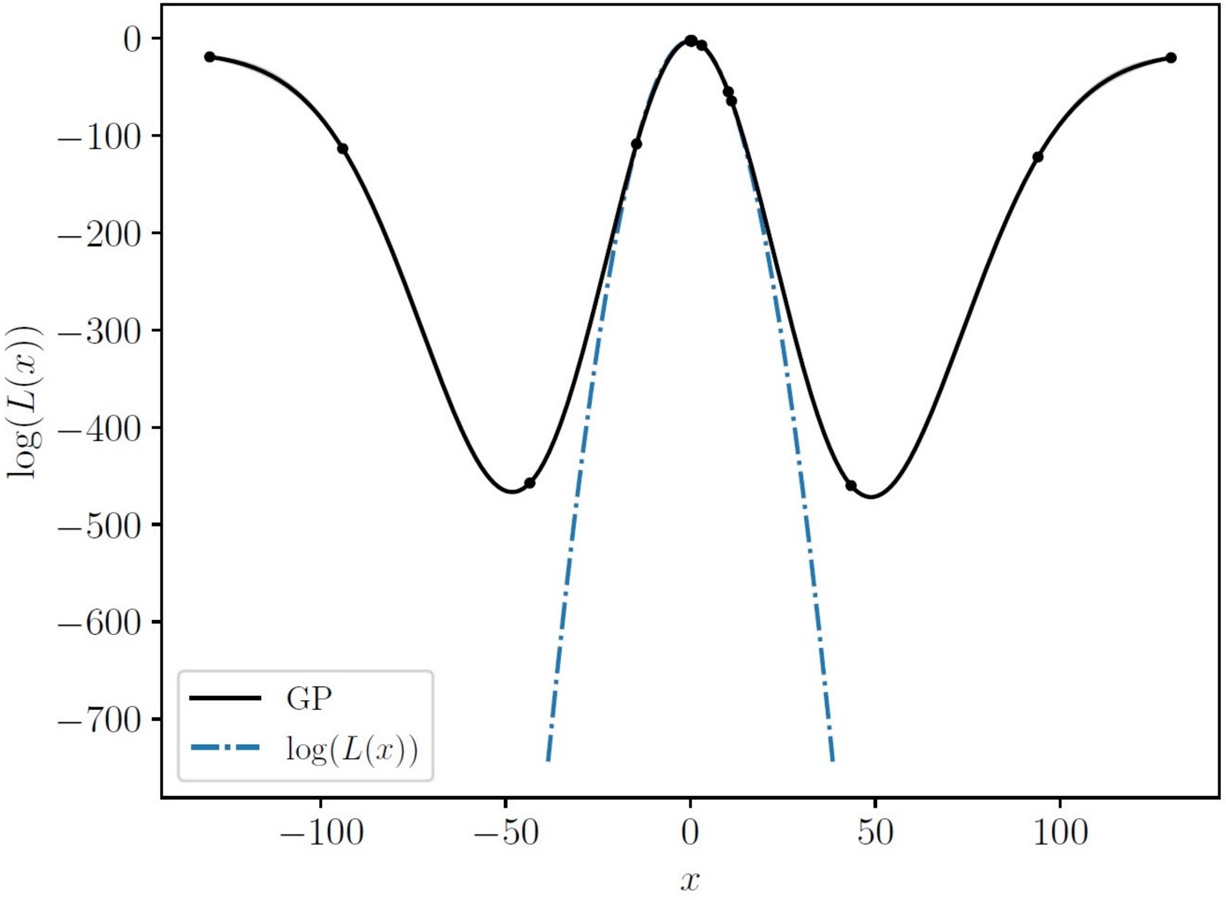
$$D_{\mathrm{KL}}(P||Q) = \frac{1}{2}\left[\log\frac{|\Sigma_q|}{|\Sigma_p|} - d + \mathrm{tr}\left(\Sigma_q^{-1}\Sigma_p\right) + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^{\mathrm{T}}\Sigma_q^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)\right]$$



For now: Take empirical mean and covariance of the training points **training points**

The problem with infinity



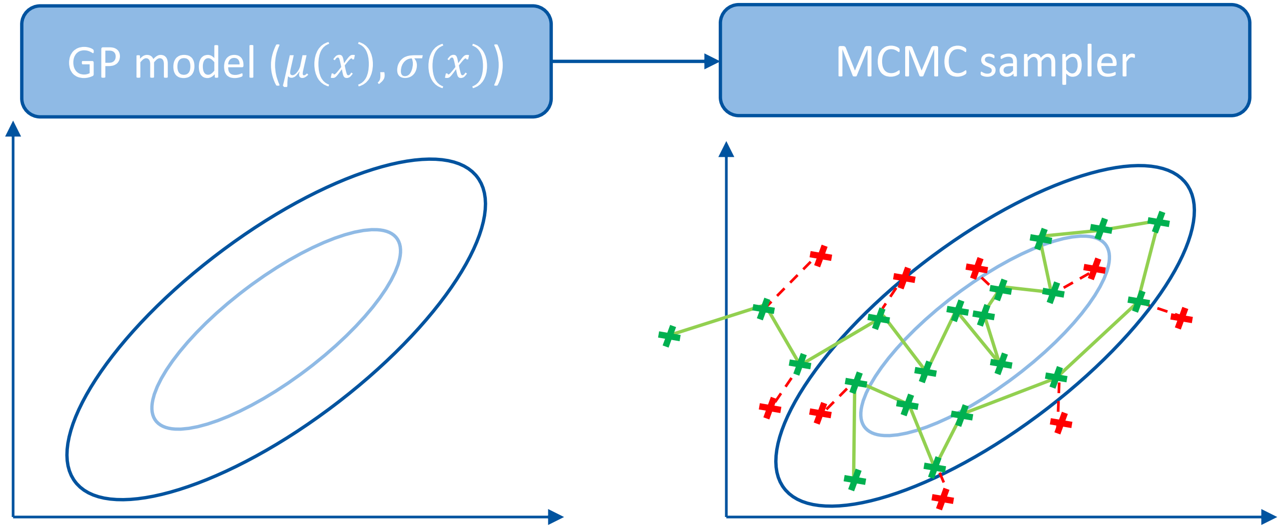


Are we preserving Bayesianity?

We are violating Bayesianity at two points:

$$\Rightarrow \log\big(p(y|X)\big) = -\frac{1}{2}y^T K^{-1} y - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi)$$

We are maximizing this with MLII. Correct Bayesian way:
Sampling the posterior distribution but very expensive!



$\rightarrow$ Ignoring $\sigma_{\mathrm{GP}}(x)$

Correct:
Get $\mathrm{var(posterior)}$ by solving

$$\iint k_{f,\mathrm{D}}(x, x')\, \mathrm{d}x\, \mathrm{d}x'$$

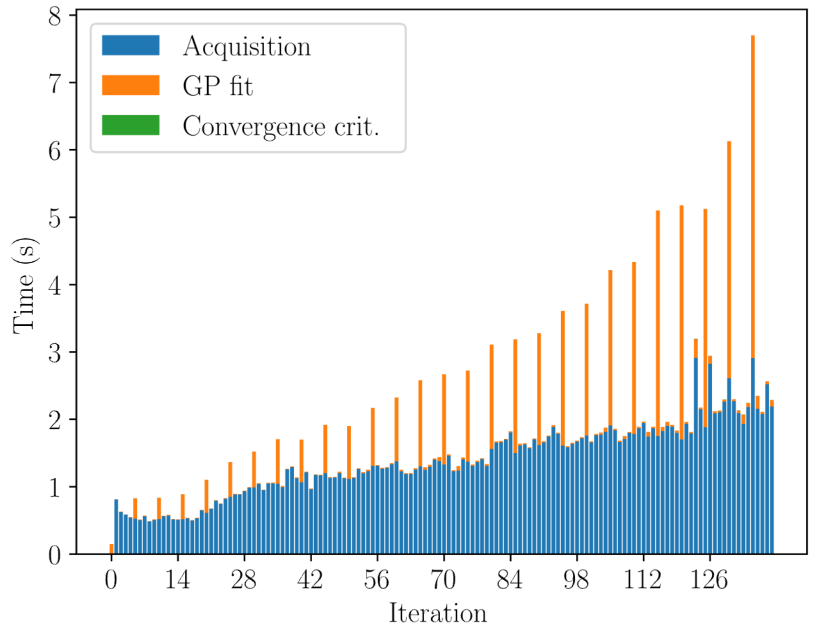
But super expensive!

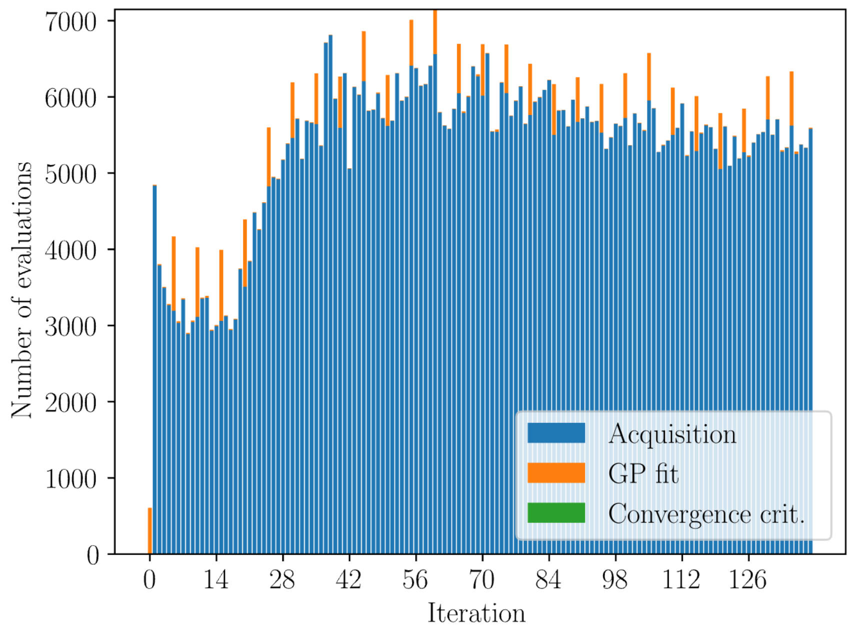
Overhead

8 dimensions

2 Kriging believer steps/iteration

In total 300 accepted samples





Refitting GP hyperparameters requires many inversions of the kernel matrix, scales $\mathcal{O}(N_{\text{samples}}^3)$

Experiments