Recitation 5

March 28, 2025

1 GP Derivative Estimation

Setting. We consider a GP $\mathcal{GP}(\mu, k)$ with the following parameters:

- Mean: $\mu(x) = 0$.
- Kernel: $k(x, x') = \lambda^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$.

In this section, we want to make use of sample information to improve our estimate of the derivative of the function. That is, we calculate the distribution of the derivative of $f(x)$ given data $\{(x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$. Let $\mathbf{X}$ be the set of training points, and $\mathbf{f}$ be the corresponding vector of function evaluations. Since differentiation is a linear functional, the distribution of $\partial f/\partial x$ will also be a Gaussian process. Let $\mathbf{g}$ be the gradients at test points $\mathbf{X}'$. We then have

$$p\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix}\middle|\, \mathbf{X}, \mathbf{X}'\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix}; \begin{bmatrix}\mu\\ L[\mu]\end{bmatrix}, \begin{bmatrix}K & L[K(\mathbf{X}, \cdot)]\\ L[K(\cdot, \mathbf{X})] & L^2[K]\end{bmatrix}\right),$$

where L is the gradient functional,

$$L[f(x_0)] = \left.\frac{\partial f(x)}{\partial x}\right|_{x=x_0}$$

Posterior derivative distribution. For our choice of the mean and kernel functions, we have

1. Derivative prior mean: $L[\mu] = 0$,
2. Derivative prior variance:

$$L^2[K] = \frac{\lambda^2}{\ell^2} \exp\left(-\frac{(x-x')^2}{2\ell^2}\right) \cdot \left(1 - \frac{(x-x')^2}{\ell^2}\right)$$

3. Covariance between derivative and training samples:

$$L[K(\mathbf{X}, x)] = \frac{\lambda^2}{\ell^2} \cdot (\mathbf{X} - x) \exp\left(-\frac{(\mathbf{X}-x)^2}{2\ell^2}\right),$$

   where all operations are element-wise.

Using these results, the posterior distribution for the derivative of $f(x)$ is given by

$$p(\mathbf{g} \mid \mathbf{X}, \mathbf{f}, \mathbf{X}') = \mathcal{N}(\mathbf{g}; \tilde{\mu}, \tilde{\Sigma}),$$

where

$$\tilde{\mu} = L[\mu] + L[K(\mathbf{X}, \mathbf{X}')]^T K^{-1}(\mathbf{f} - \boldsymbol{\mu})$$
$$\tilde{\Sigma} = L^2[K] - L[K(\mathbf{X}, \mathbf{X}')]^T K^{-1} L[K(\mathbf{X}, \mathbf{X}')]$$

Implementation. With $(\lambda, \ell) = (1, 1.5)$, we choose 11 points with uniform gaps in the interval $[-10, 10]$, and sample $f(x)$ for these points. Using these values of $f(x)$, we then calculate the gradient for 1000 points chosen uniformly between $[-12, 12]$. Plot (a) shows the points with the posterior mean function along with a 95% credible interval. Plot (b) shows this posterior mean function alongside the derivative posterior mean function. We also show a 95% credible interval for the derivative posterior.

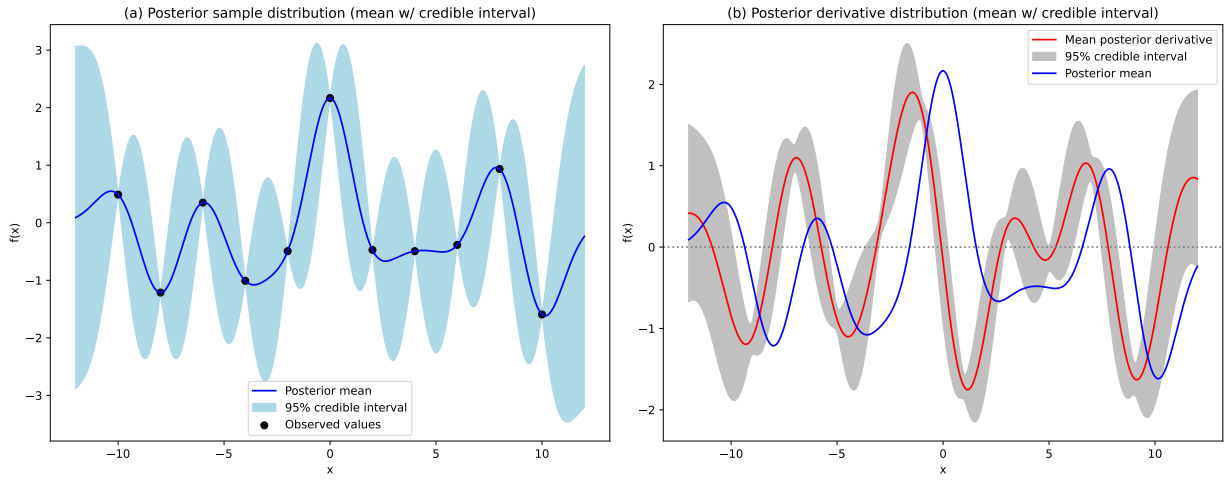


Figure 1: Posterior mean function (part (a)) and corresponding posterior derivative function (part (b)) for observed data.

We observe that the mean posterior derivative corresponds to the derivative of the posterior mean. The mean posterior derivative (red curve) is positive when the posterior mean (blue curve) is increasing, and decreasing when the posterior mean is decreasing. We see that the posterior derivative is zero wherever there is a local optimum for the posterior mean

2 GP Integration Estimation (Bayesian Quadrature / Bayesian Monte Carlo)

2.1 Introduction

Suppose we have an intractable integral that we wish to estimate

$$\int f(x)dx$$

This could be (for example) a hard-to-compute evidence term used to normalize a Bayesian prior.

Often, it is the case that we have an integral we wish to evaluate under a probability, and this will take the form

$$\int f(x)p(x)dx$$

In Bayesian quadrature, we will perform a simple trick: rather than directly trying to compute this integral or directly computing a numerical approximation, we will first create an approximation of $f(x)$. If we choose our approximation carefully, it will be easier to integrate. A common choice is to use a Gaussian Process for this integration.

Gaussian Processes are particularly well suited to this task as they are closed under integration. This means that we will get a Guassian posterior probability for the value of the integral!

As we saw in lecture, the posterior distribution for the estimation of an integral using a GP prior is given by a Gaussian parameterized with mean and covariance as follows:

$$E_{f|D}[\bar{f}_D] = \iint f(x)p(x)dx\, p(f|D)df = \int \left[\int f(x)p(f|D)df\right] p(x)dx = \int \bar{f}_D(x)p(x)dx \tag{1}$$

$$\begin{aligned} V_{f|D}[\bar{f}_D] &= \int \left[\int f(x)p(x)dx - \int \bar{f}(x')p(x')dx'\right]^2 p(f|D)df \\ &= \iiint [f(x) - \bar{f}(x)][f(x') - \bar{f}(x')]p(f|D)df\, p(x)p(x')dxdx' \\ &= \iint \mathrm{Cov}_D(f(x), f(x'))p(x)p(x')dxdx' \end{aligned} \tag{2}$$

As is typical for a Gaussian Process, the posterior mean and covariance are given by

$$\bar{f}_D(x) = k(x, \mathbf{x})K^{-1}\mathbf{f} \tag{3}$$

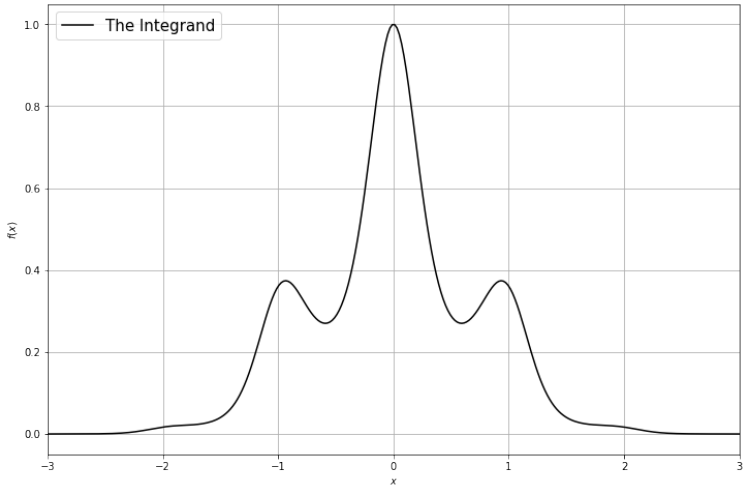
$$\mathrm{Cov}_D(f(x), f(x')) = k(x, x') - k(x, \mathbf{x})K^{-1}k(\mathbf{x}, x') \tag{4}$$

[Rasmussen and Ghahramani]

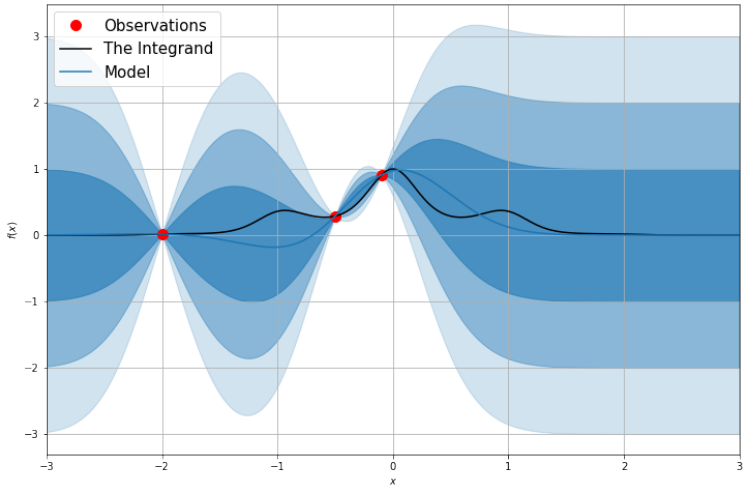
2.2 Visualization

Suppose we wish to compute the following:

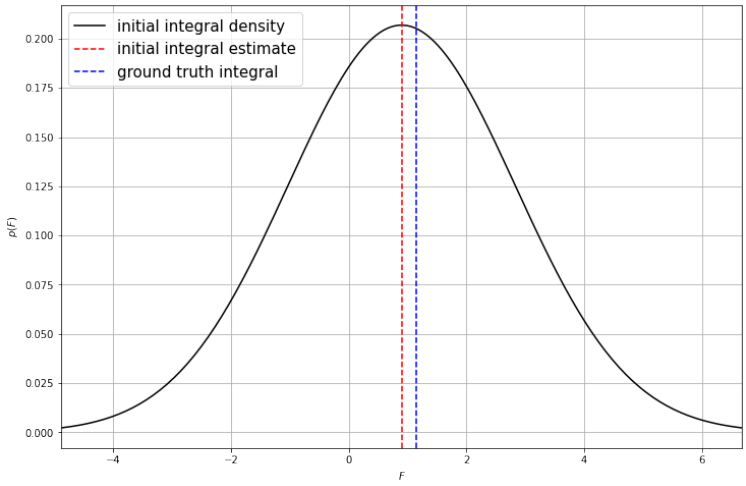
$$\int_{-3}^{3} e^{-x^2-\sin^2(2x)}dx$$



To perform Bayesian Quadratue, we first take a few samples from the function. We can then fit a GP to the function.



Now, we can integrate the Gaussian process to get an estimation for the integral of the true function.

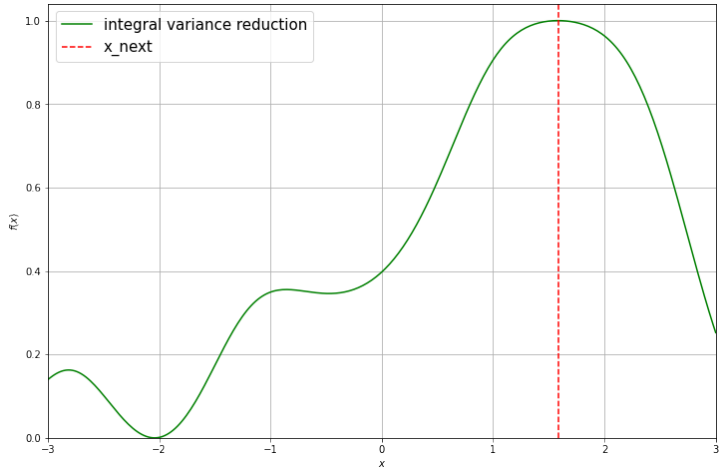


It is typical to perform Bayesian Quadrature in a loop, using active sampling techniques to predict the best next point to observe. One possibility is to use integral variance reduction (IVR), a close cousin of the Knowledge Gradient (KG) acquisition function seen in the lecture. IVR seeks to select the point that will cause the largest difference in the posterior variance

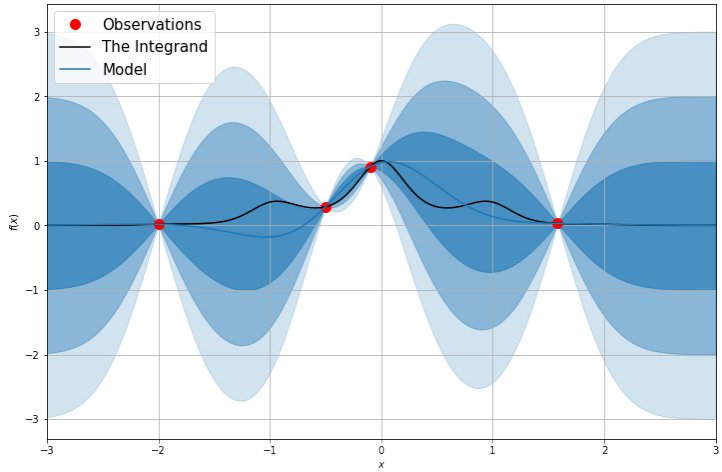
$$a_n^{IVR}(x) = V_{f|D}[\bar{f}_D] - V_{f|D\cup x}[\bar{f}_{D\cup x}]$$

$$= \frac{\left(\int_{\mathcal{X}'} k(x', x)dx'\right)^2}{V_{f|D}[\bar{f}_D]V_{f|D\cup x}[\bar{f}_{D\cup x}]}$$

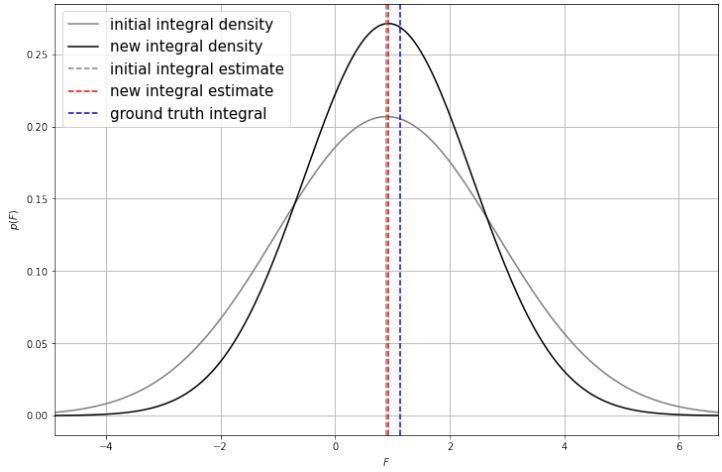
We can visualize this acquisition function:



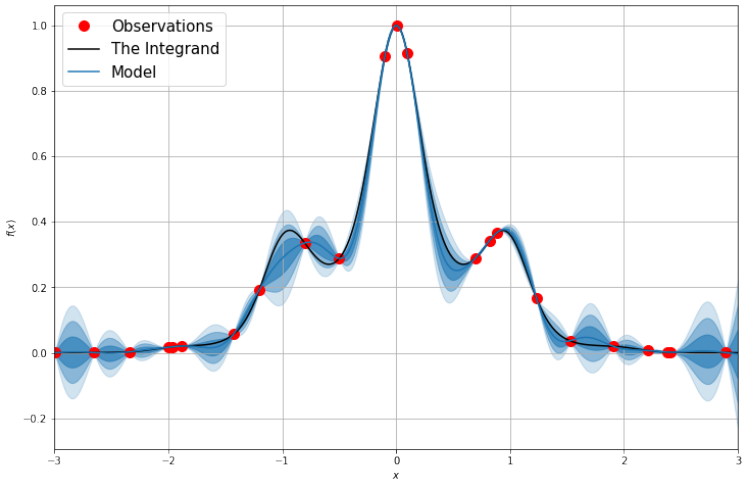
And the resulting GP after updating our belief using this new sample

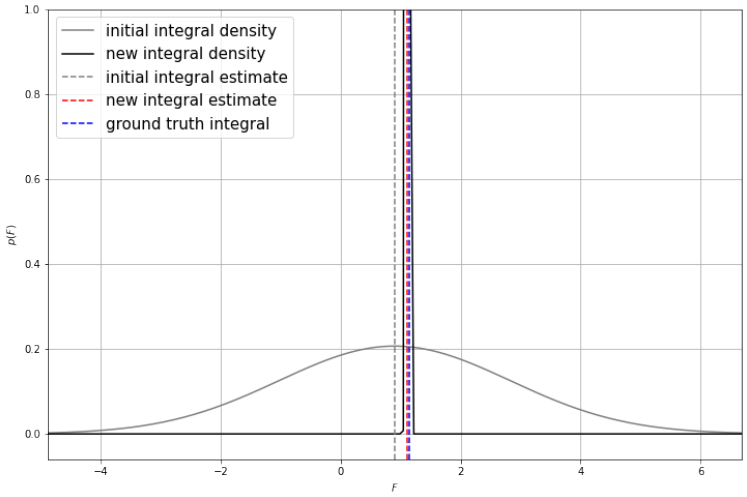


With a corresponding update in our belief on the integral value.



Finally, after several rounds of Bayesian updates we can see that the function approximation and integral approximation have converged to be quite close to the ground truth





[Emukit BQ Tutorial]

Attentive readers will notice that we have used a second difficult-to-calculate integral for our acquisition function. This is often approximated using Monte Carlo sampling.

However, that means that we require a hard integral to approximate a hard integral – a seeming net 0 gain. In practice, our GP is often much easier to sample from than the function it approximates. In such situations, Bayesian Monte Carlo can be a big win. However, if this isn't true (such as in this toy example) we don't gain anything.