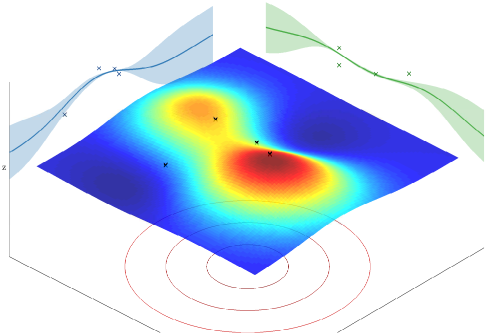
Bayesian Quadrature:

Model-based Approximate Integration



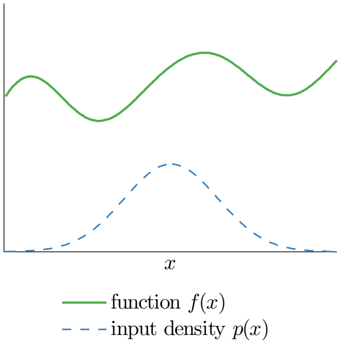
David Duvenaud
University of Cambridge

The Quadrature Problem

- We want to estimate an integral

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

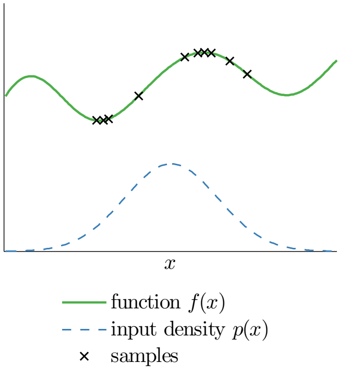
- Most computational problems in inference correspond to integrals:
  - Expectations
  - Marginal distributions
  - Integrating out nuisance parameters
  - Normalization constants
  - Model comparison



Sampling Methods

- Monte Carlo methods:
  Sample from $p(x)$, take empirical mean:

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

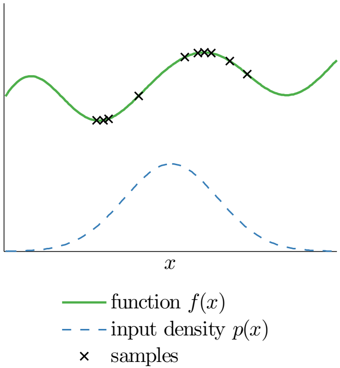


Sampling Methods

- Monte Carlo methods: Sample from $p(x)$, take empirical mean:

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- Possibly sub-optimal for two reasons:

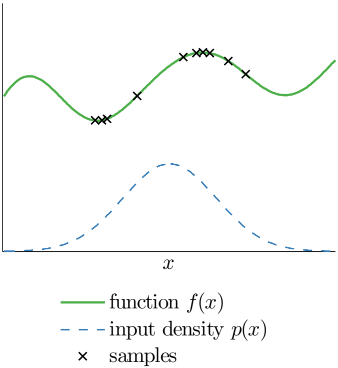


Sampling Methods

- Monte Carlo methods:
  Sample from $p(x)$, take empirical mean:

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

- Possibly sub-optimal for two reasons:
  - Random bunching up

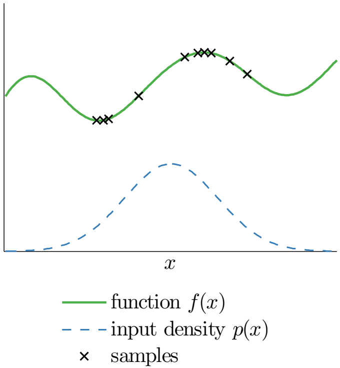


Sampling Methods

- Monte Carlo methods:
  Sample from $p(x)$, take empirical mean:

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- Possibly sub-optimal for two reasons:
  - Random bunching up
  - Often, nearby function values will be similar

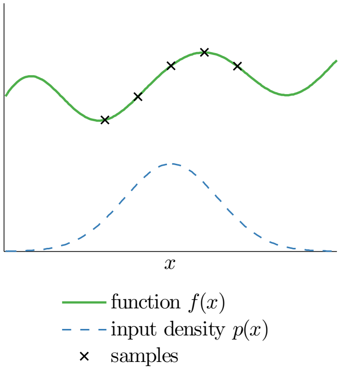


Sampling Methods

- Monte Carlo methods:
  Sample from $p(x)$, take empirical mean:

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

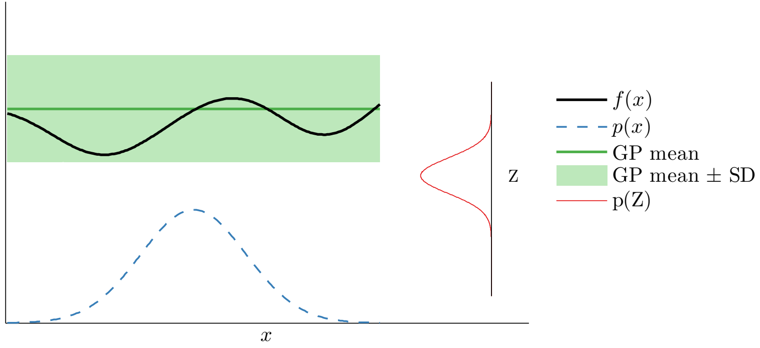
- Possibly sub-optimal for two reasons:
  - Random bunching up
  - Often, nearby function values will be similar
- Model-based and quasi-Monte Carlo methods spread out samples to achieve faster convergence.



Model-based Integration

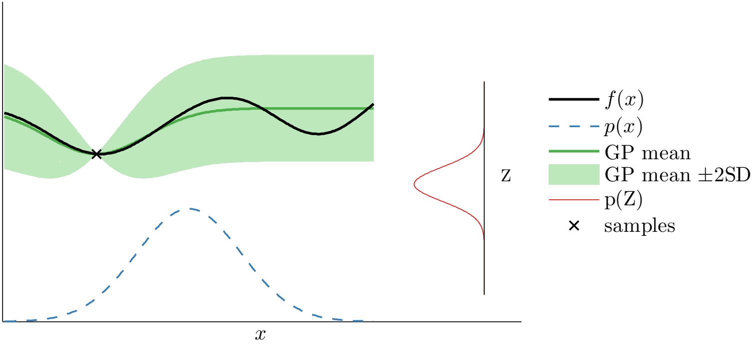
Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



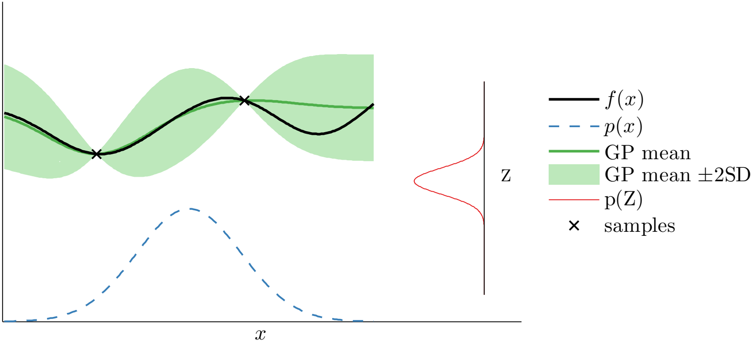
Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



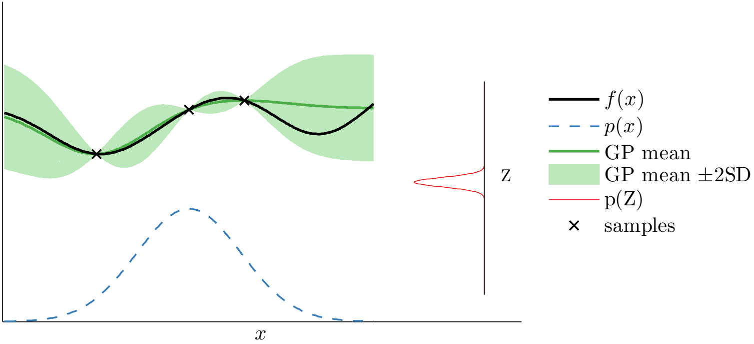
Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



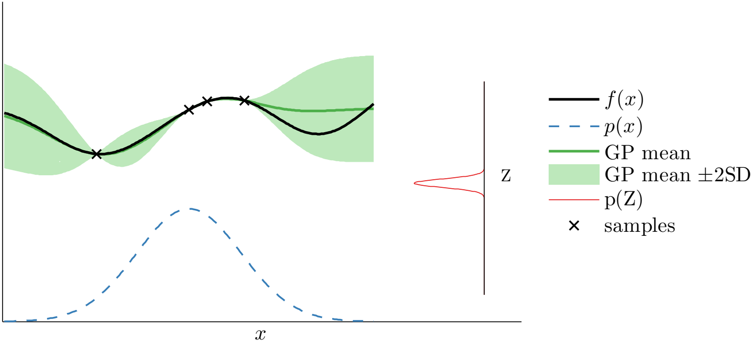
Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



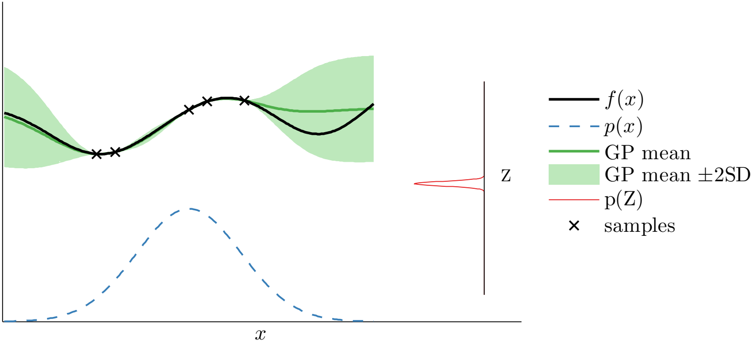
Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



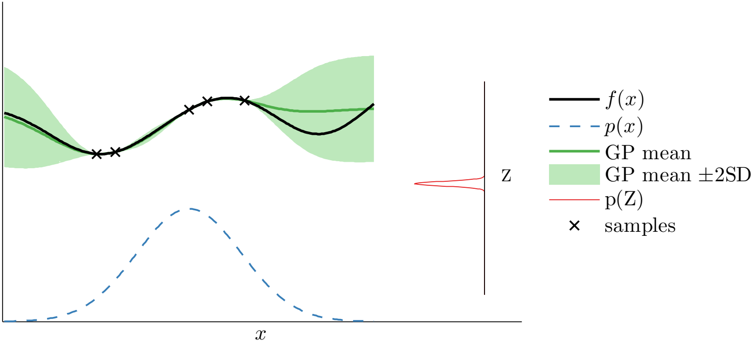
Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



Model-based Integration

- Place a prior on f, for example, a GP
- Posterior over f implies a posterior over Z.



- We'll call using a GP prior Bayesian Quadrature

Bayesian Quadrature Estimator

- Posterior over Z has mean linear in $f(x_s)$:

$$\mathbb{E}_{\mathrm{GP}}\left[Z|f(x_s)\right] = \sum_{i=1}^{N} z^T K^{-1} f(x_i)$$

where $z_n = \int k(x, x_n) p(x) dx$

Bayesian Quadrature Estimator

- Posterior over Z has mean linear in $f(x_s)$:

$$\mathbb{E}_{\mathrm{GP}}\left[Z|f(x_s)\right] = \sum_{i=1}^{N} z^T K^{-1} f(x_i)$$

where $z_n = \int k(x, x_n) p(x) dx$

- Natural to minimize posterior variance of Z:

$$\mathbb{V}\left[Z|f(x_s)\right] = \int\int k(x, x') p(x) p(x') dx dx' - z^T K^{-1} z$$

Bayesian Quadrature Estimator

- Posterior over Z has mean linear in $f(x_s)$:

$$\mathbb{E}_{\text{GP}}\left[Z|f(x_s)\right] = \sum_{i=1}^{N} z^T K^{-1} f(x_i)$$

where $z_n = \int k(x, x_n) p(x) dx$

- Natural to minimize posterior variance of Z:

$$\mathbb{V}\left[Z|f(x_s)\right] = \int\int k(x, x') p(x) p(x') dx dx' - z^T K^{-1} z$$

- Doesn't depend on function values at all!

Bayesian Quadrature Estimator

- Posterior over Z has mean linear in $f(x_s)$:

$$\mathbb{E}_{\mathrm{GP}}\left[Z|f(x_s)\right] = \sum_{i=1}^{N} z^T K^{-1} f(x_i)$$

where $z_n = \int k(x, x_n) p(x) dx$

- Natural to minimize posterior variance of Z:

$$\mathbb{V}\left[Z|f(x_s)\right] = \int\int k(x, x') p(x) p(x') dx dx' - z^T K^{-1} z$$

- Doesn't depend on function values at all!
- Choosing samples sequentially to minimize variance: Sequential Bayesian Quadrature.

Things you can do with Bayesian Quadrature

- Can incorporate knowledge of function (symmetries)

$$f(x,y) = f(y,x) \Leftrightarrow k_s(x,y,x',y') = k(x,y,x',y') + k(x,y',x',y) + k(x',y,x,y') + k(x',y',x,y)$$

Things you can do with Bayesian Quadrature

- Can incorporate knowledge of function (symmetries)

$$f(x, y) = f(y, x) \Leftrightarrow k_s(x, y, x', y') = k(x, y, x', y') + k(x, y', x', y) + k(x', y, x, y') + k(x', y', x, y)$$

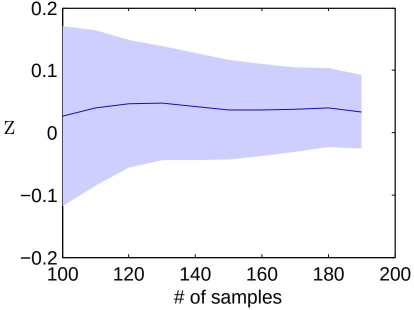
- Can condition on gradients

Things you can do with Bayesian Quadrature

- Can incorporate knowledge of function (symmetries)

$$f(x,y) = f(y,x) \Leftrightarrow k_s(x,y,x',y') = k(x,y,x',y') + k(x,y',x',y) \\ + k(x',y,x,y') + k(x',y',x,y)$$

- Can condition on gradients
- Posterior variance is a natural convergence diagnostic

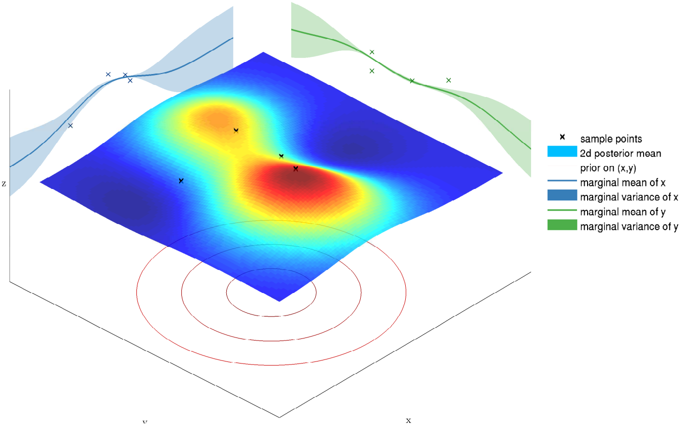


More things you can do with Bayesian Quadrature

- Can compute likelihood of GP, learn kernel
- Can compute marginals with error bars, in two ways:

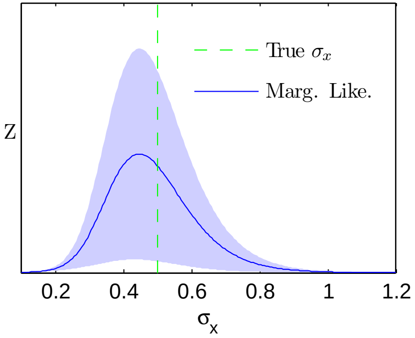
More things you can do with Bayesian Quadrature

- Can compute likelihood of GP, learn kernel
- Can compute marginals with error bars, in two ways:
- Simply from the GP posterior:



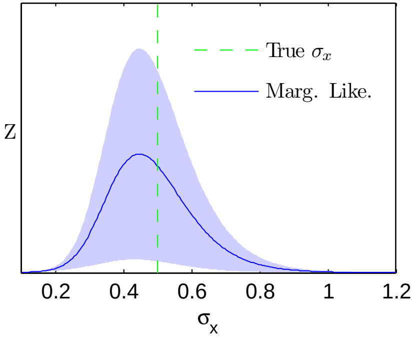
More things you can do with Bayesian Quadrature

- Can compute likelihood of GP, learn kernel
- Can compute marginals with error bars, in two ways:
- Or by recomputing $f_\theta(x)$ for different θ with same x



More things you can do with Bayesian Quadrature

- Can compute likelihood of GP, learn kernel
- Can compute marginals with error bars, in two ways:
- Or by recomputing $f_\theta(x)$ for different θ with same x

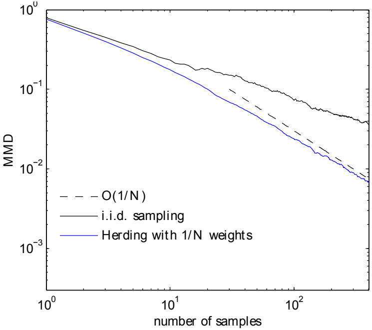


- Much nicer than histograms!

Rates of Convergence

What is rate of convergence of SBQ when its assumptions are true?

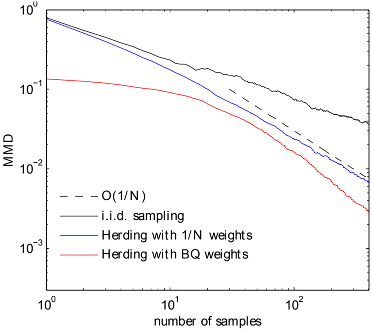
Expected Variance / MMD



Rates of Convergence

What is rate of convergence of SBQ when its assumptions are true?

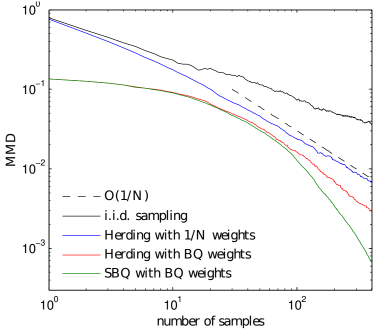
Expected Variance / MMD



Rates of Convergence

What is rate of convergence of SBQ when its assumptions are true?

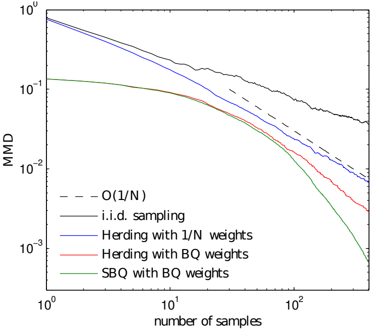
Expected Variance / MMD



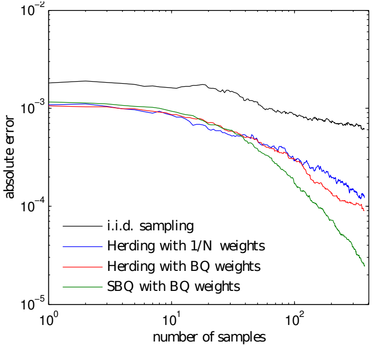
Rates of Convergence

What is rate of convergence of SBQ when its assumptions are true?

Expected Variance / MMD



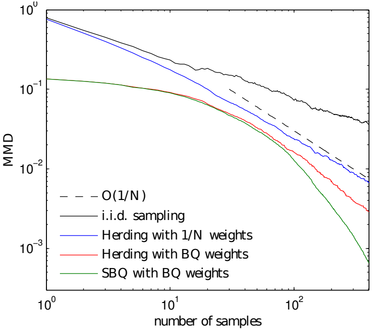
Empirical Rates in RKHS



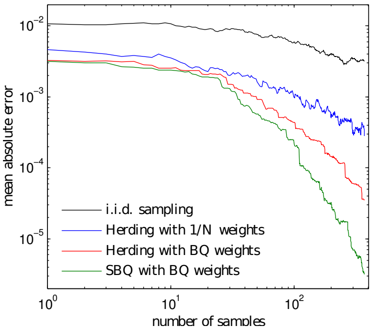
Rates of Convergence

What is rate of convergence of SBQ when its assumptions are true?

Expected Variance / MMD



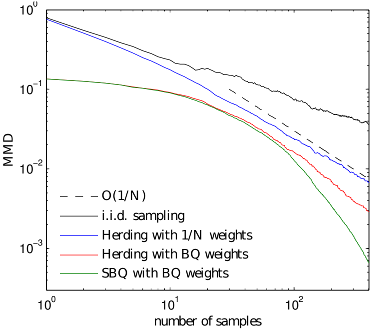
Empirical Rates out of RKHS



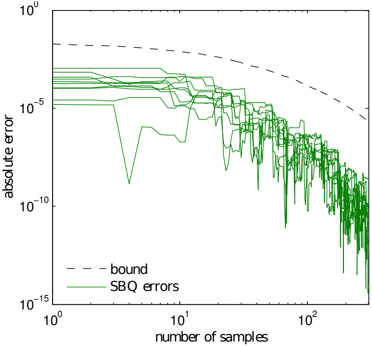
Rates of Convergence

What is rate of convergence of SBQ when its assumptions are true?

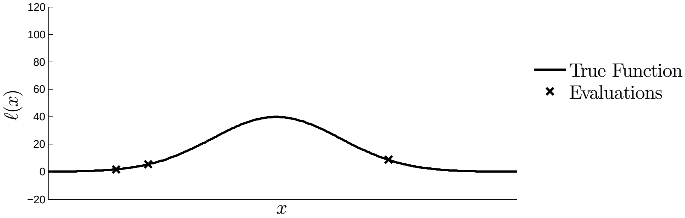
Expected Variance / MMD



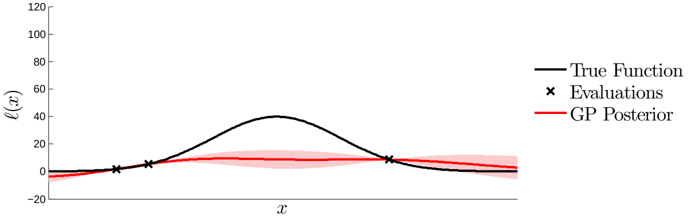
Bound on Bayesian Error



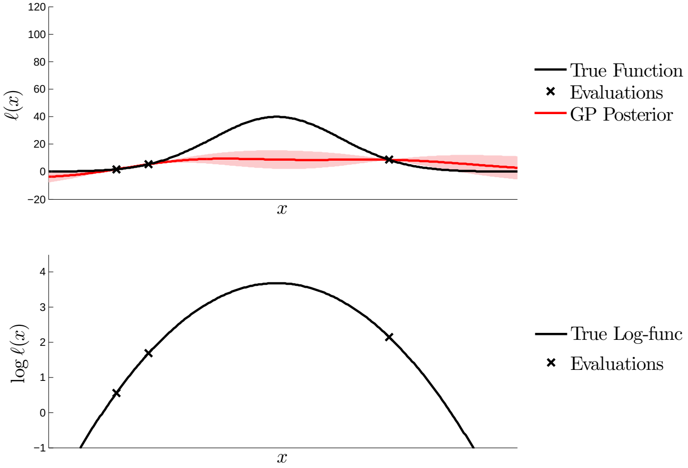
GPs vs Log-GPs for Inference



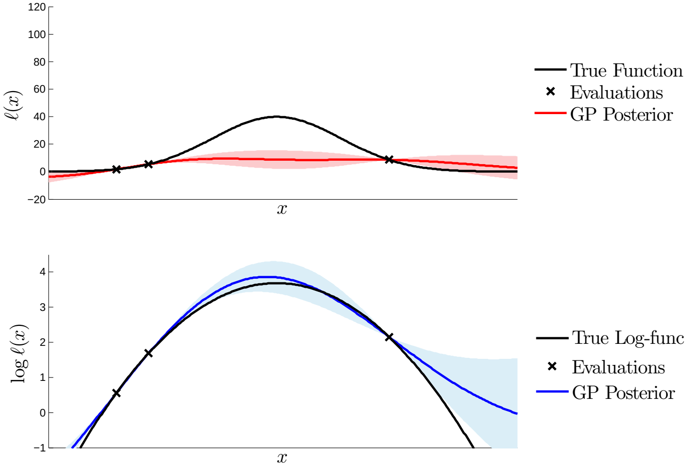
GPs vs Log-GPs for Inference



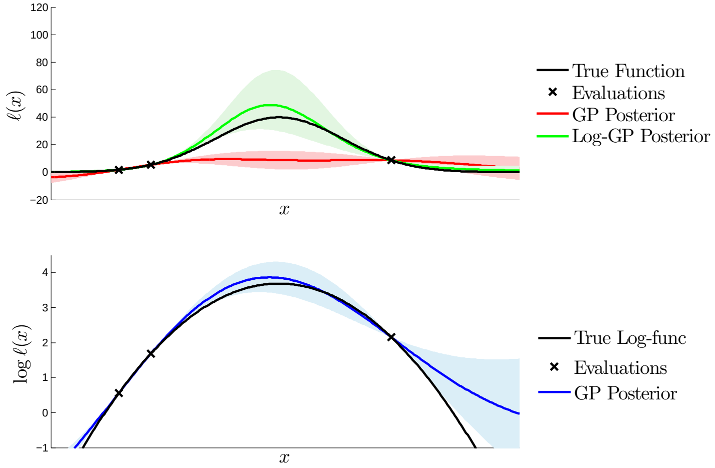
GPs vs Log-GPs for Inference



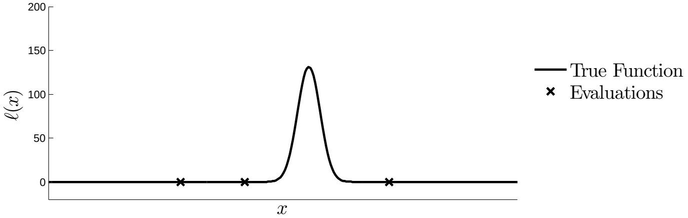
GPs vs Log-GPs for Inference



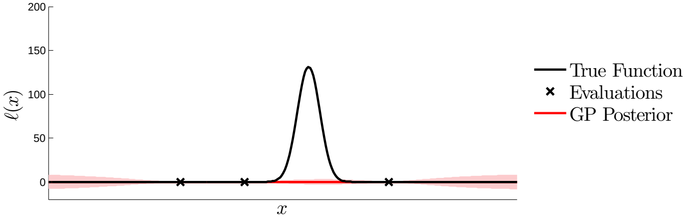
GPs vs Log-GPs for Inference



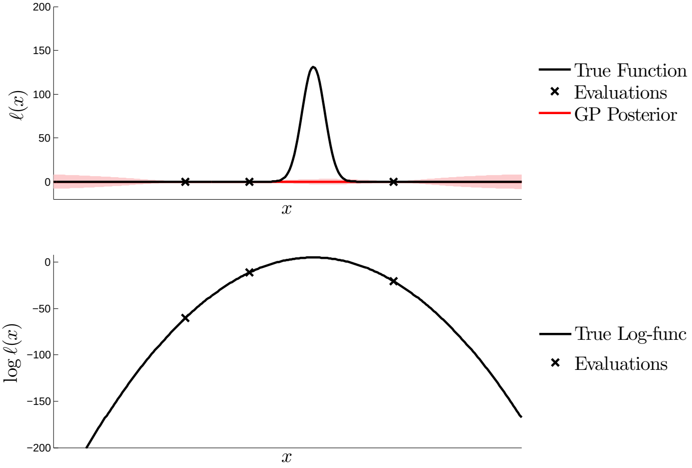
GPs vs Log-GPs



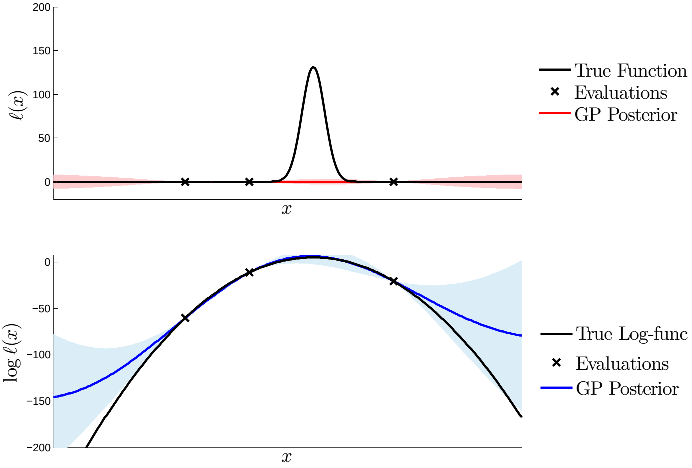
GPs vs Log-GPs



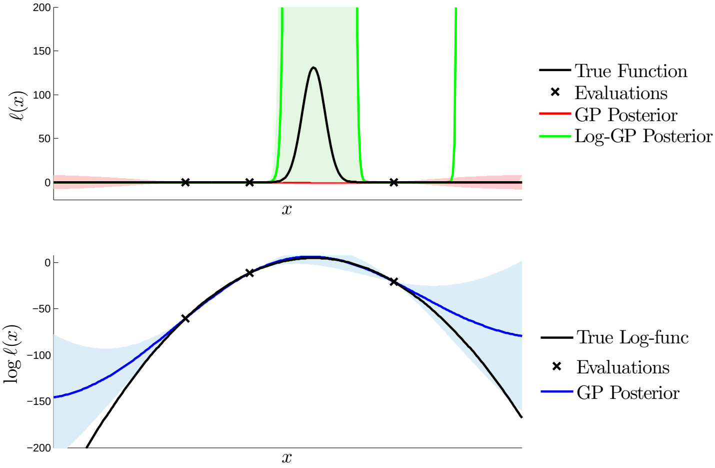
GPs vs Log-GPs



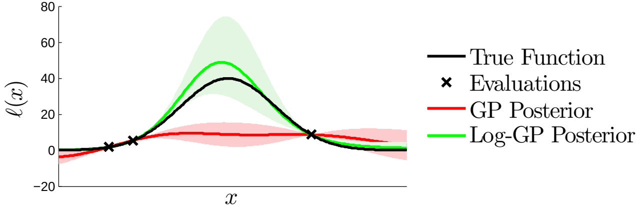
GPs vs Log-GPs

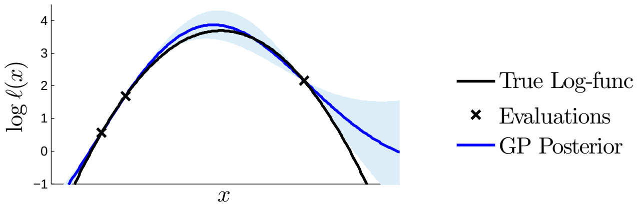


GPs vs Log-GPs

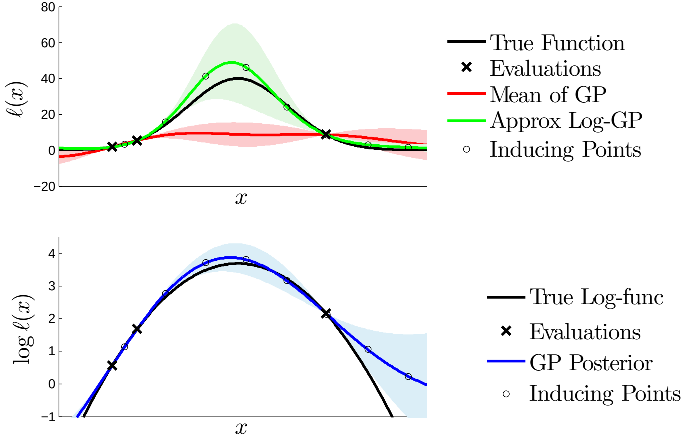


Integrating under Log-GPs





Integrating under Log-GPs



Conclusions

- Model-based integration allows active learning about integrals, can require fewer samples than MCMC, and allows us to check our assumptions.

Conclusions

- Model-based integration allows active learning about integrals, can require fewer samples than MCMC, and allows us to check our assumptions.
- BQ has nice convergence properties if its assumptions are correct.

Conclusions

- Model-based integration allows active learning about integrals, can require fewer samples than MCMC, and allows us to check our assumptions.
- BQ has nice convergence properties if its assumptions are correct.
- For inference, GP is not especially appropriate, but other models are intractable.

Limitations and Future Directions

- Right now, BQ really only works in low dimensions (< 10), when the function is fairly smooth, and is only worth using when computing $f(x)$ is expensive.

Limitations and Future Directions

- Right now, BQ really only works in low dimensions (< 10), when the function is fairly smooth, and is only worth using when computing $f(x)$ is expensive.
- How to extend to high dimensions? Gradient observations are helpful, but a D-dimensional gradient is D separate observations.

Limitations and Future Directions

- Right now, BQ really only works in low dimensions (< 10), when the function is fairly smooth, and is only worth using when computing $f(x)$ is expensive.
- How to extend to high dimensions? Gradient observations are helpful, but a D-dimensional gradient is D separate observations.
- It seems unlikely that we'll find another tractable nonparametric distribution like GPs - should we accept that we'll need a second round of approximate integration on a surrogate model?

Limitations and Future Directions

- Right now, BQ really only works in low dimensions (< 10), when the function is fairly smooth, and is only worth using when computing $f(x)$ is expensive.
- How to extend to high dimensions? Gradient observations are helpful, but a D-dimensional gradient is D separate observations.
- It seems unlikely that we'll find another tractable nonparametric distribution like GPs - should we accept that we'll need a second round of approximate integration on a surrogate model?
- How much overhead is worthwhile? Bounded rationality work seems relevant.

Limitations and Future Directions

- Right now, BQ really only works in low dimensions (< 10), when the function is fairly smooth, and is only worth using when computing $f(x)$ is expensive.
- How to extend to high dimensions? Gradient observations are helpful, but a D-dimensional gradient is D separate observations.
- It seems unlikely that we'll find another tractable nonparametric distribution like GPs - should we accept that we'll need a second round of approximate integration on a surrogate model?
- How much overhead is worthwhile? Bounded rationality work seems relevant.

Thanks!