Empirical analysis of the re-weighting trick in Bayesian quadrature

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Bayesian Quadrature

Problem Setting

- We want to obtain an numerical approximation of the integral:

$$F := \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathbb{E}_{x\sim p}(f(\mathbf{x}))$$

- Sampling from the integrand is expensive
- The integral is relatively low dimensional (certainly below 10)

Bayesian Quadrature

What is Bayesian quadrature?

Bayesian quadrature views numerical integration as a Bayesian inference task

We want to estimate a distribution on F using observations of the integrand

- We place a prior distribution over the integrand
- Conditions this prior on samples of the integrand to obtain a posterior distribution
- then computes the implied posterior distribution over F

➔ A common choice for the prior distribution over the integrand is a Gaussian process

Definition (Gaussian process)

A stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution

A GP is defined by its mean and covariance function

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

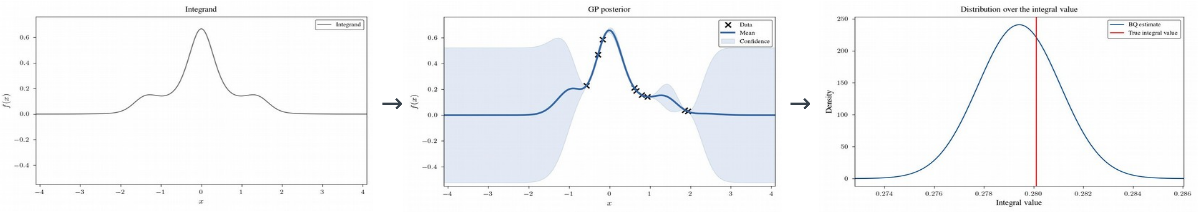
We can condition the GP to data and obtain the posterior mean and covariance function

$$\mathrm{D} = \{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_N, f(\mathbf{x}_N))\}$$

$$m_{\mathcal{D}}(\mathbf{x}) = C(\mathbf{x}, X_N)C(X_N, X_N)^{-1}\mathbf{f}$$

$$C_{\mathcal{D}}(\mathbf{x}, \mathbf{x}') = C(\mathbf{x}, \mathbf{x}') - C(\mathbf{x}, X_N)C(X_N, X_N)^{-1}C(X_N, \mathbf{x}')$$

- We choose a suitable GP prior
- Condition this GP on observations of the integrand
- Integrate the GP instead of the true integrand
- Obtain a distribution over the integral value



Bayesian Quadrature

How to integrate over a Gaussian process?

- Under a GP prior the posterior is (an infinite dimensional joint) Gaussian
- The Integral F is a linear projection (on the direction defined by p(x))
- The posterior distribution over the integral value is therefore also a Gaussian

$$F \sim \mathcal{N}(\mathtt{m}_F, \mathtt{v}_F) \quad \text{where}$$

$$\mathtt{m}_F = \mathbb{E}_x(m_{\mathcal{D}}(\mathbf{x})) = \int_\Omega m_{\mathcal{D}}(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

$$\mathtt{v}_F = \mathbb{E}_{\mathbf{x},\mathbf{x}'}(C_{\mathcal{D}}(\mathbf{x},\mathbf{x}')) = \int_\Omega\int_\Omega C_{\mathcal{D}}(\mathbf{x},\mathbf{x}')p(\mathbf{x})p(\mathbf{x}')d\mathbf{x}d\mathbf{x}'$$

- Under a GP prior the posterior is (an infinite dimensional joint) Gaussian
- The Integral F is a linear projection (on the direction defined by p(x))
- The posterior distribution over the integral value is therefore also a Gaussian

$F \sim \mathcal{N}(\mathtt{m}_F, \mathtt{v}_F)$ where

$$\mathtt{m}_F = \mathbb{E}_x(m_{\mathcal{D}}(\mathbf{x})) = \int_\Omega m_{\mathcal{D}}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$\mathtt{v}_F = \mathbb{E}_{\mathbf{x},\mathbf{x}'}(C_{\mathcal{D}}(\mathbf{x},\mathbf{x}')) = \int_\Omega \int_\Omega C_{\mathcal{D}}(\mathbf{x},\mathbf{x}') p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

In some cases the above integrals can be calculated analytically!

The Re-weighting Trick in Bayesian Quadrature

Analytical Bayesian quadrature for arbitrary probability densities

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

What if we do not have an analytical solution?

- We can rewrite the original integral by introducing a new probability density q

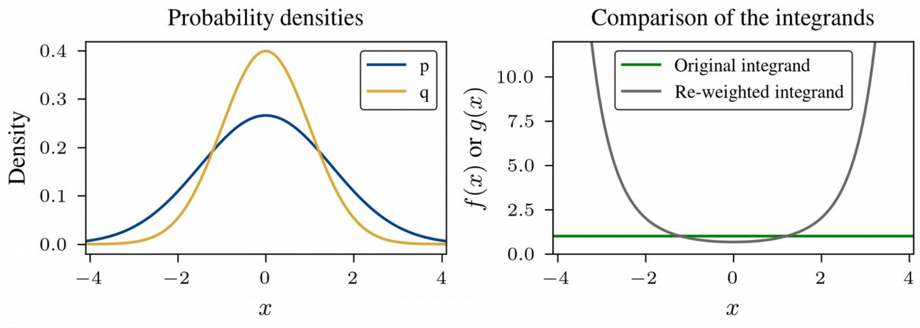
$$F = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \int \frac{f(\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})d\mathbf{x} := \int g(\mathbf{x})q(\mathbf{x})d\mathbf{x},$$

- q can be any probability density as long as: $\forall x \in \Omega : p(x) \neq 0 \implies q(x) \neq 0$

- If we choose q to be the Gaussian or uniform density, we can again obtain analytical results!

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

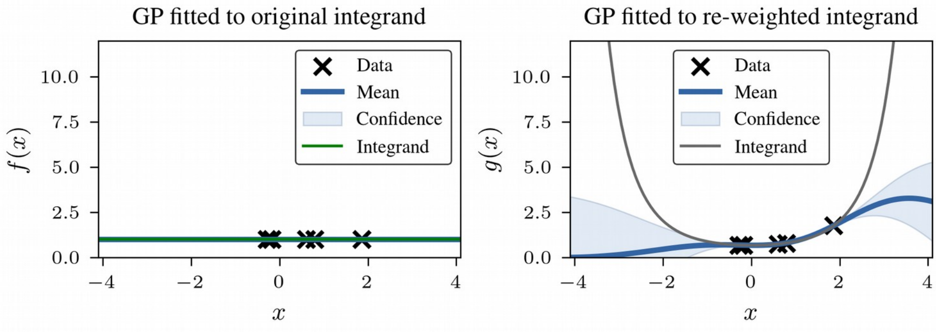
- Does re-weighting affect the performance of the BQ algorithm?



The Re-weighting Trick in Bayesian Quadrature

Drawbacks of re-weighting

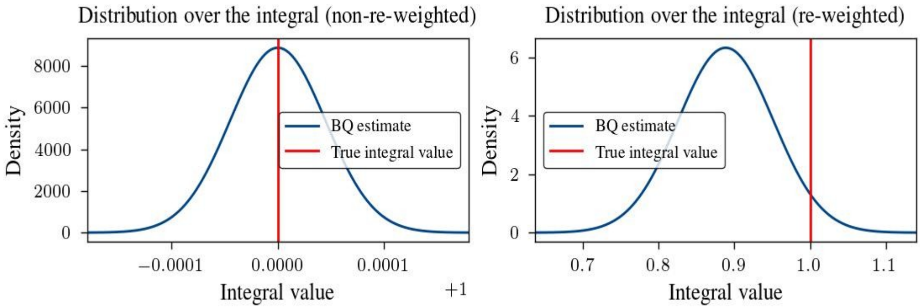
- The GP has a harder time to capture the re-weighted integrand



The Re-weighting Trick in Bayesian Quadrature

Drawbacks of re-weighting

- This results in a less accurate estimation of the integral



The Re-weighting Trick in Bayesian Quadrature

Summary

- BQ transforms the integration problem into a regression problem on the integrand and an often analytical integration problem on the regression model
- The re-weighting trick enables us to use BQ to integrate w.r.t. arbitrary probability densities p
- Depending on the choice of probability density q and its parameters, re-weighting might significantly affect the performance of BQ

Choosing a Suitable q for the Re-weighting Trick

Minimizing performance drop when re-weighting

Can we somehow choose a density q that minimizes the negative effects of re-weighting?

We are limited in the kind of function q can be

- q has to have mass at every position p has mass
- We must have an analytical solution for the combination of kernel and q

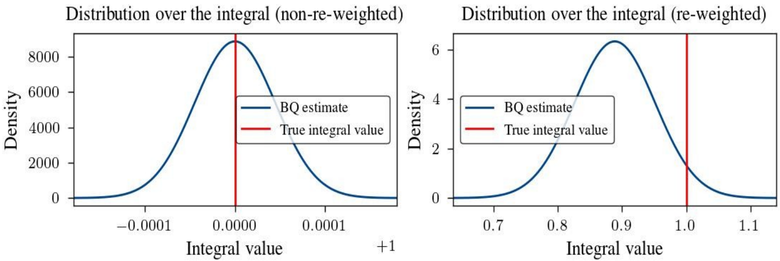
→ q is therefore usually the Gaussian or uniform density

→ The only thing we can freely choose are its parameters

Scores Assessing the Re-weighting Trick

Similarity of BQ estimates

- Ideally we would like to make the BQ estimate of the re-weighted integral as similar as possible to the BQ estimate of the non-re-weighted integral



- This is intractable since we usually do not know the distribution over the integral in the none re-weighted case!

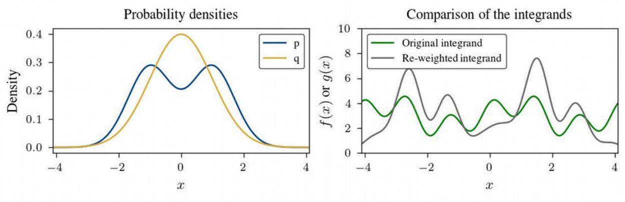
Scores Assessing the Re-weighting Trick

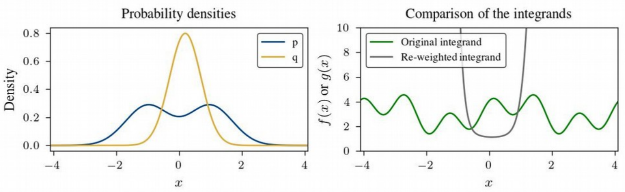
Basic approaches

Low distortion of the integrand

→ Similarity of f and g

→ Transfer of relevant properties from f to g



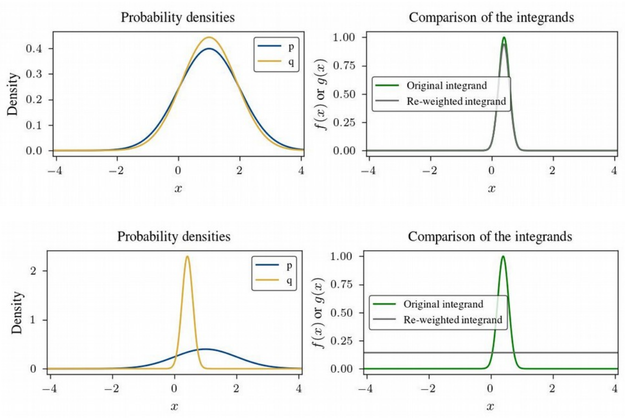


Scores Assessing the Re-weighting Trick

Basic approaches

Suitability of the re-weighted integrand

→ We want the re-weighted integrand to be easy to capture for the GP independent of its similarity to the original integrand



KL Divergence Score

Low distortion of the original integrand

- We try to determine parameters for q that maximize the similarity of f and g

$$\arg\max_{\boldsymbol{\theta}} \text{similarity}(f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x};\boldsymbol{\theta})}, f(\mathbf{x}))$$

- Since evaluating the integrand f is expensive we maximize the similarity of p and q instead

$$\arg\max_{\boldsymbol{\theta}} \text{similarity}(q(\mathbf{x};\boldsymbol{\theta}), p(\mathbf{x}))$$

- But how do we measure similarity?

$$\arg\min_{\boldsymbol{\theta}} \text{KL}(q(\mathbf{x};\boldsymbol{\theta})||p(\mathbf{x})) \quad \text{where} \quad KL(p(\mathbf{x})||q((x))) = \int p(\mathbf{x}) \log(\frac{p(\mathbf{x})}{q(\mathbf{x})}) d\mathbf{x}$$

Shift Score

Low distortion of the original integrand

- We assume $f \sim \mathcal{GP}(m_f, k_f)$

- Re-weighting now implies a non stationary process $g \sim \mathcal{GP}(m_g, k_g)$

- We now construct a score that tries to measure the change in (non)-stationarity of f when re-weighting as:

$$\bar{s}_g(\delta) = \iint_{\Omega \times \Omega} s_g(\mathbf{x}, \mathbf{x}'|\delta) p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

where $$s_g(\mathbf{x}, \mathbf{x}'|\delta) = \left| 1 - \frac{p(\mathbf{x}+\delta)p(\mathbf{x}'+\delta)}{q(\mathbf{x}+\delta)q(\mathbf{x}'+\delta)} \frac{q(\mathbf{x})q(\mathbf{x}')}{p(\mathbf{x})p(\mathbf{x}')} \right|$$

Evidence Score

Suitability of the re-weighted integrand

- We want to measure how suitable the re-weighted integrand is for our GP
- We can construct a score that relies on the evidence of the model
  → We asses how probable is to observe the re-weighted integrand under the GP
- When we consider n samples $(\mathbf{x}, y)$ from g the logarithm of the evidence is given by

$$\log(p(\mathbf{y}|X)) = -\frac{1}{2}\mathbf{y}^T C(X,X)^{-1}\mathbf{y} - \frac{1}{2}\log(|C(X,X)|) - \frac{n}{2}\log(2\pi)$$

- Since we sample from g the evidence requires us to evaluate f

Scores Assessing the Re-weighting Trick

We have constructed three scores that aim to choose suitable parameters for the density q

<u>Low distortion of the integrand</u>

- KL divergence score
  → similarity of p and q
- Shift score
  → small change in stationarity

<u>Suitability of the re-weighted integrand</u>

- Evidence score
  → How probable is the re-weighted integrand under the GP

Empirical Evaluation of the Scores

Are the scores useful?

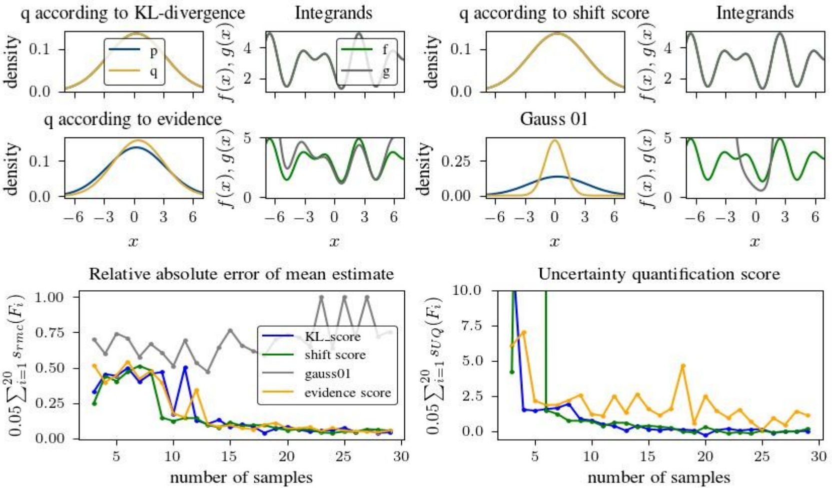
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

How good are the scores that we have constructed?

- We implemented the three scores in Python and evaluate their performance on different test integrals
  → we use EmuKit with GPy for the implementation of BQ
- We use the introduced scores to optimize the parameters of q
- asses the performs of BQ on the integrals re-weighted with the recommended densities

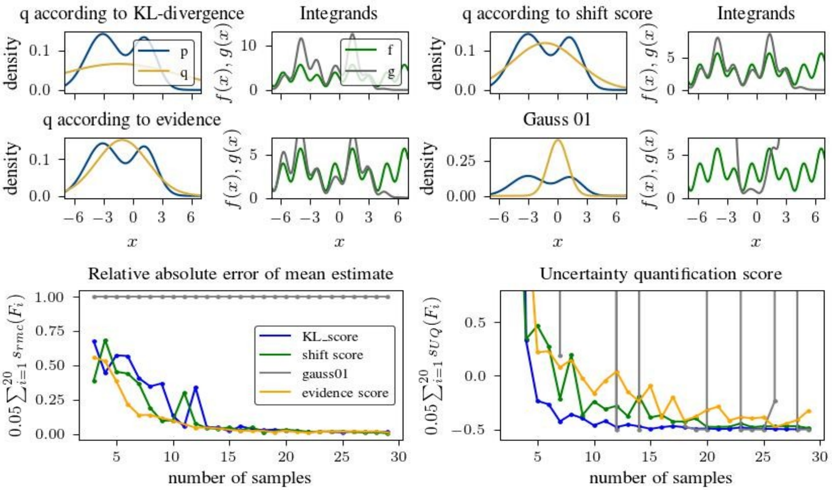
Empirical Evaluation of the Scores

On 20 integrals with different Fourier integrands and Gaussian density p



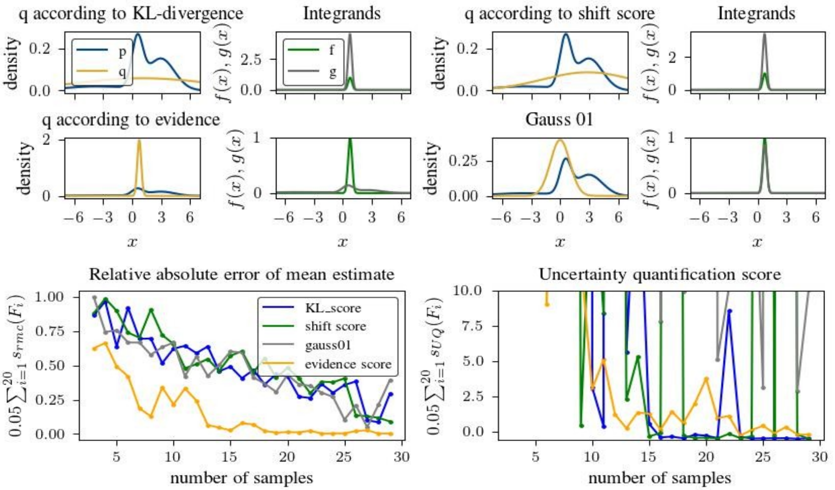
Empirical Evaluation of the Scores

On 20 integrals with different Fourier integrands and Gaussian mixture density p



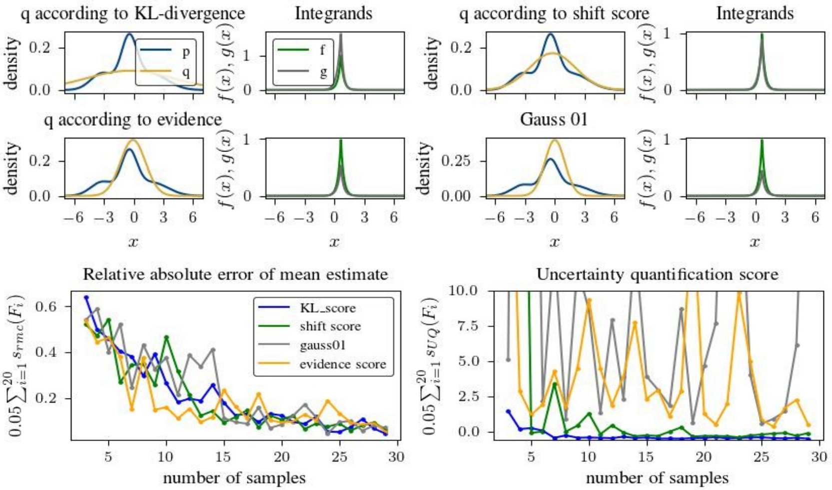
Empirical Evaluation of the Scores

On 20 integrals with different Gaussian peak integrands and Gaussian mixture density p



Empirical Evaluation of the Scores

On 20 integrals with different continuous integrands and Gaussian mixture density p



Limitations

Of the experiments

<u>The evaluation of the scores was relatively limited</u>

→ All test integrands used Gaussian mixture densities
(Scores would work for arbitrary p)

→ Only looked at 1D integrals

→ Only GPs with RBF kernel

→ Results are only as accurate as the metrics we used to evaluate them
(especially the uncertainty quantification should be treated with caution)

Key Findings

Summary

- Depending on the choice of probability density q and its parameters, re-weighting might significantly affect the performance of BQ
- All three proposed scores seem to perform reasonably well for the test integrands
- The evidence score gives better results in special cases, but often performs similar to the other two scores
- KL divergence score and shift score perform relatively similar, but KL divergence score seems more robust and easier to optimize

Shift Score

Low distortion of the original integrand

- Instead of assessing the similarity between p and q, we try to assess how much we 'change' the integrand under the re-weighing trick

- We construct a measure of (non)-stationarity as:

$$s(\mathbf{x}, \mathbf{x}'|\delta) := \left| \frac{C(\mathbf{x}, \mathbf{x}') - C(\mathbf{x}+\delta, \mathbf{x}'+\delta)}{C(\mathbf{x}, \mathbf{x}')} \right|$$

- If we assume that f is a draw from a stationary GP the re-weighting trick implies a non-stationary process g and we get:

Definition (Stationarity)

A stochastic process $\{\eta(x)\}_{x\in\Omega}$ is said to be stationary when its distribution is unchanged by an index shift, that is $\{\eta(x)\}_{x\in\Omega} = \{\eta(x+\delta)\}_{x+\delta\in\Omega}$ for arbitrary δ. This implies that all its statistics obey this property as well.

$$s_g(\mathbf{x}, \mathbf{x}'|\delta) = \left| 1 - \frac{p(\mathbf{x}+\delta)p(\mathbf{x}'+\delta)}{q(\mathbf{x}+\delta)q(\mathbf{x}'+\delta)} \frac{q(\mathbf{x})q(\mathbf{x}')}{p(\mathbf{x})p(\mathbf{x}')} \right|$$ resp. $$\bar{s}_g(\delta) = \iint_{\Omega\times\Omega} s_g(\mathbf{x}, \mathbf{x}'|\delta) p(\mathbf{x}) p(\mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

Shift Score

Low distortion of the original integrand

- Instead of assessing the difference between p and q directly, we try to assess how much we ‘change’ the integrand under the re-weighing trick

- We assume that f is a draw from a stationary GP

$$f(\mathbf{x}) \sim \mathcal{GP}(m_f, C_f)$$

- The re-weighting trick now implies a non-stationary process:

$$g(\mathbf{x}) := f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})} \sim \mathcal{GP}(m_g, C_g)$$

$$C_g(\mathbf{x}, \mathbf{x'}) = \frac{p(\mathbf{x})p(\mathbf{x'})}{q(\mathbf{x})q(\mathbf{x'})} C_f(\mathbf{x}, \mathbf{x'})$$

- We construct a measure of (non)-stationarity as:

$$s(\mathbf{x}, \mathbf{x'}|\delta) := \left| \frac{C(\mathbf{x}, \mathbf{x'}) - C(\mathbf{x}+\delta, \mathbf{x'}+\delta)}{C(\mathbf{x}, \mathbf{x'})} \right|$$

- For C_g we get:

$$s_g(\mathbf{x}, \mathbf{x'}|\delta) = \left| 1 - \frac{p(\mathbf{x}+\delta)p(\mathbf{x'}+\delta)}{q(\mathbf{x}+\delta)q(\mathbf{x'}+\delta)} \frac{q(\mathbf{x})q(\mathbf{x'})}{p(\mathbf{x})p(\mathbf{x'})} \right|$$

resp.

$$\bar{s}_g(\delta) = \iint_{\Omega\times\Omega} s_g(\mathbf{x}, \mathbf{x'}|\delta) p(\mathbf{x}) p(\mathbf{x'}) d\mathbf{x} d\mathbf{x'}$$

Empirical Evaluation of the Scores

How to measure performance of the BQ estimate

How do we measure the performance of BQ on the re-weighted integrals?

Accuracy of the mean estimate

→ absolute value of the relative difference between the posterior mean estimate and the true integral value

$$s_{rmc}(F) = \left| \frac{F - \mathtt{m}_F}{F} \right|$$

Calibration of the distribution

→ expected logarithmic density ratio

$$s_{UQ}(F) = \mathbb{E}\left(\log \frac{p(F_{GP}|\mathcal{D})}{p(F_{GP}|\mathcal{D} = F)}\right) = \frac{1}{2}\left(\frac{(F - \mathtt{m}_F)^2}{\mathtt{v}_F} - 1\right)$$

Base Slide?

Keep this subtitle?