Active Multi-Information Source Bayesian Quadrature

Maren Mahsereci

Structurally Constrained Gaussian Processes
GPSS workshop 2019/09/12

Outline

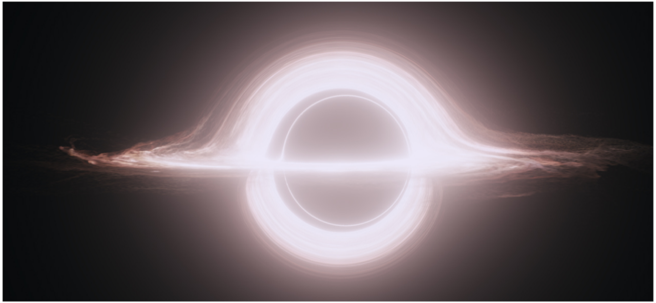
Part I

- Intro to Bayesian quadrature

Part II

- Active multi-information source Bayesian quadrature
  A.Gessner, J.Gonzalez, M.Mahsereci, UAI 2019

Expensive computer simulations

$\alpha \rightarrow$  $= f(\alpha)$

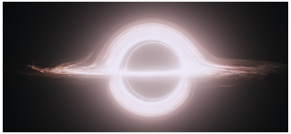
$$\mathbf{E}[f] = \int f(\alpha) p(\alpha) \mathrm{d}\alpha = ?$$

James et al. 2015, "Gravitational lensing by spinning black holes in astrophysics, and in the movie Interstellar" in Classical and Quantum Gravity. DOI: 10.1088/0264-9381/32/6/065001

Expensive Integrals

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)\mathrm{d}\alpha = ?$$

α_1 α_2 α_3 ... α_N

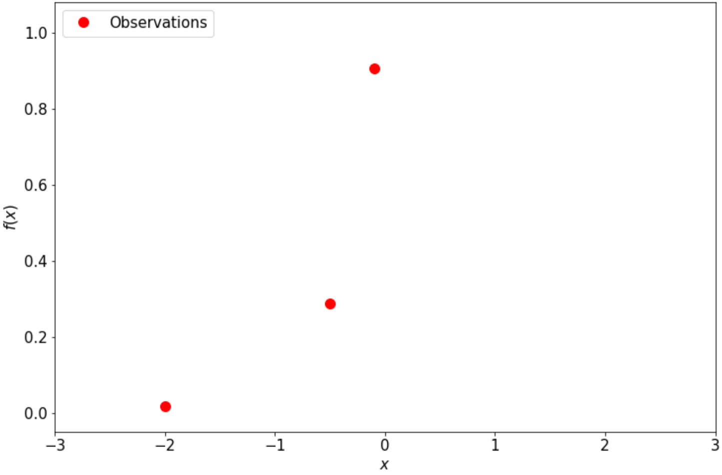


$f(\alpha_1)$ $f(\alpha_2)$ $f(\alpha_3)$ ... $f(\alpha_N)$

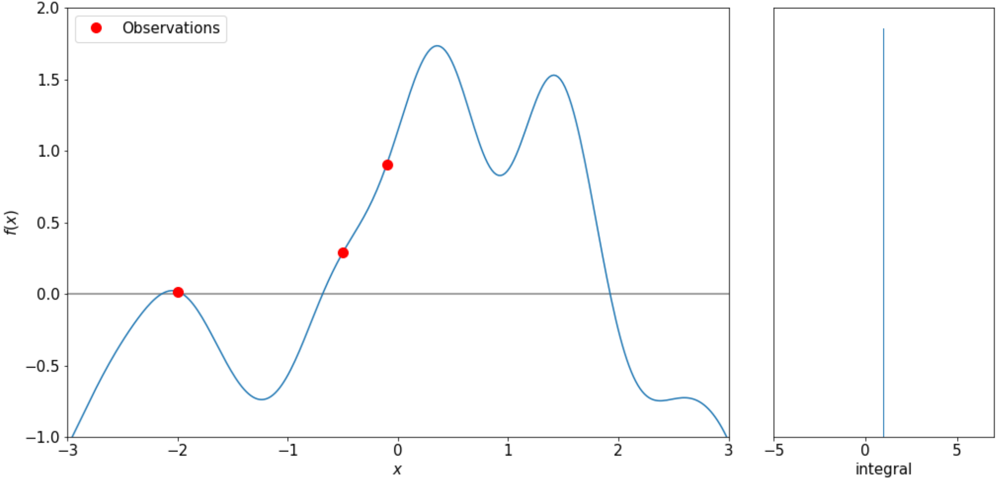
- Evaluations of f ("data") is limited and expensive
- Sampling is prohibitive and sub-optimal.

$\rightarrow$ Want to collect $f(\alpha)$ at the **most informative points α**.

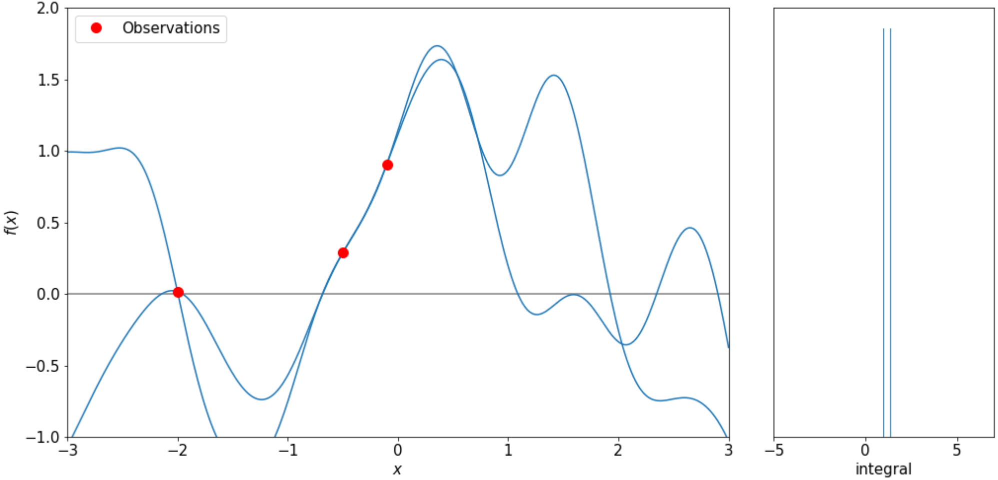
Solve this integral $\int f(x)\mathrm{d}x = ?$



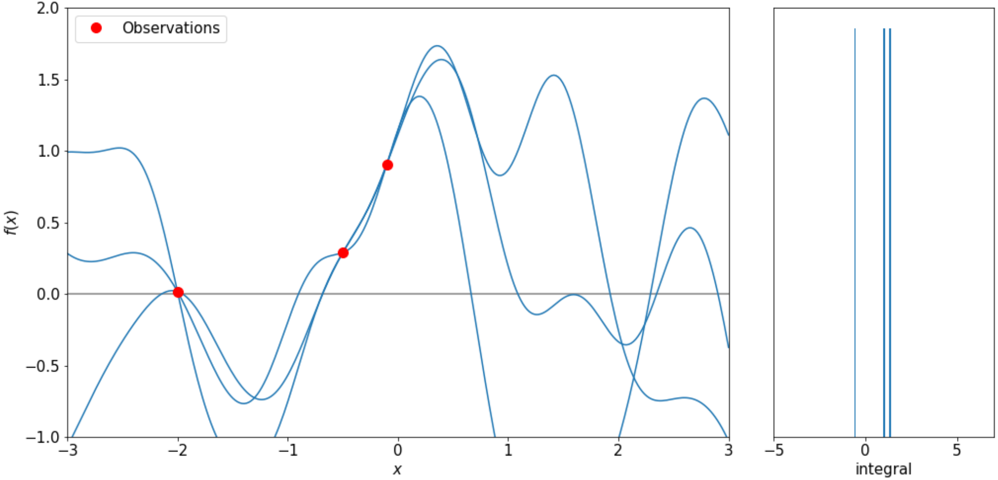
Solve this integral $\int f(x)\mathrm{d}x = ?$



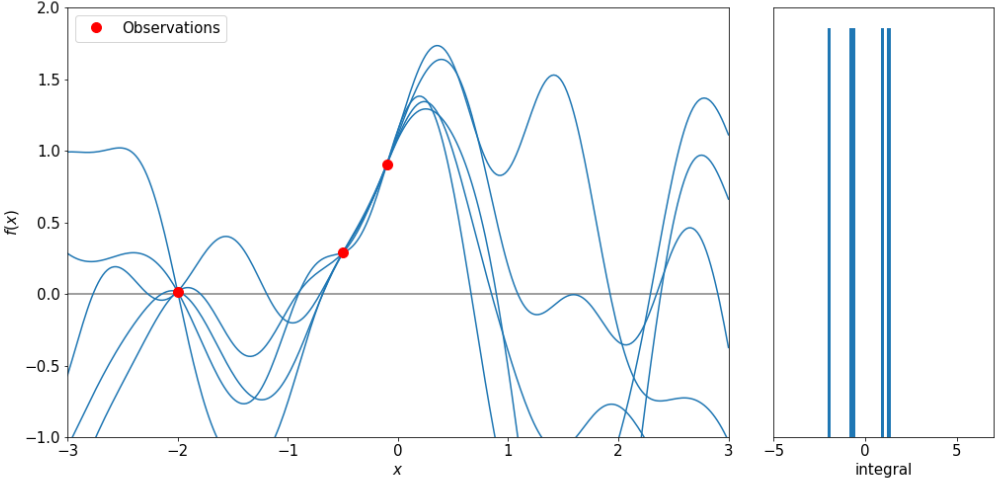
Solve this integral $\int f(x)\mathrm{d}x = ?$



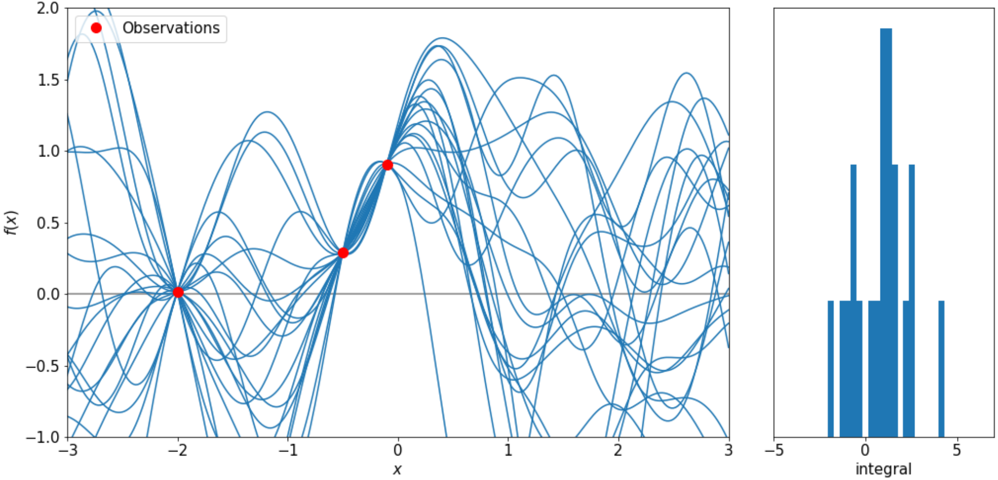
Solve this integral $\int f(x)\mathrm{d}x = ?$



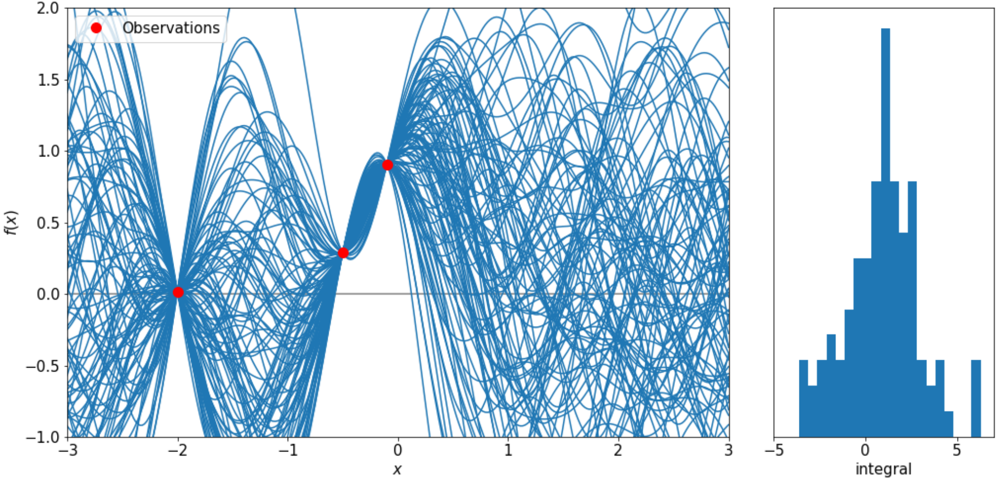
Solve this integral $\int f(x)\mathrm{d}x = ?$



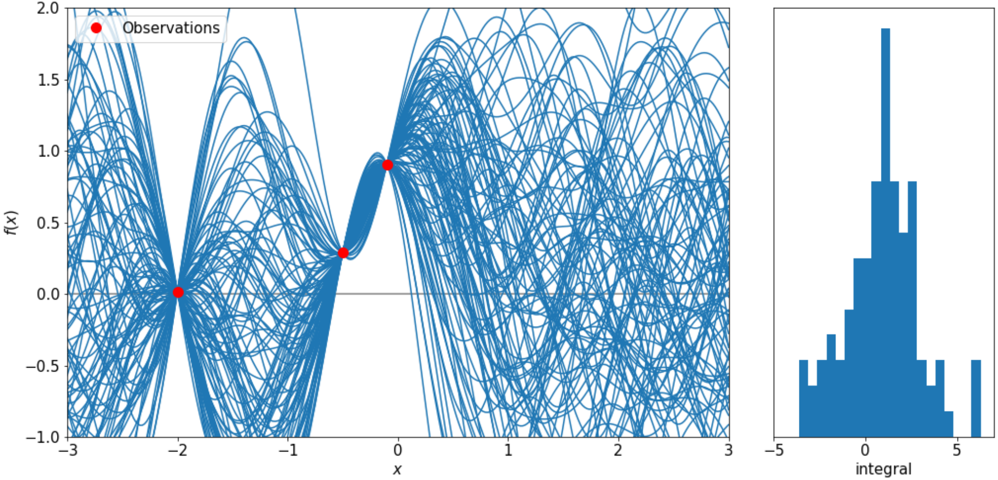
Solve this integral $\int f(x)\mathrm{d}x = ?$



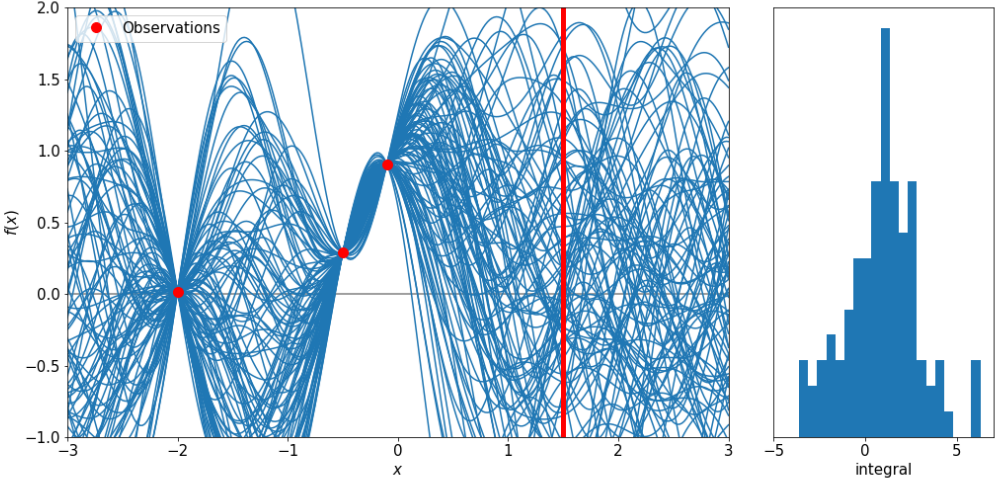
Solve this integral $\int f(x)\mathrm{d}x = ?$



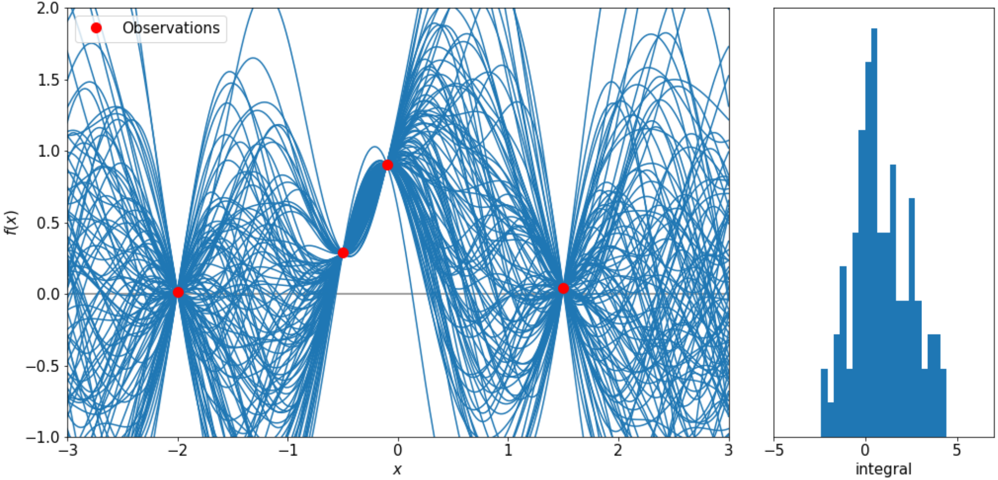
Next point? Use uncertainty of model



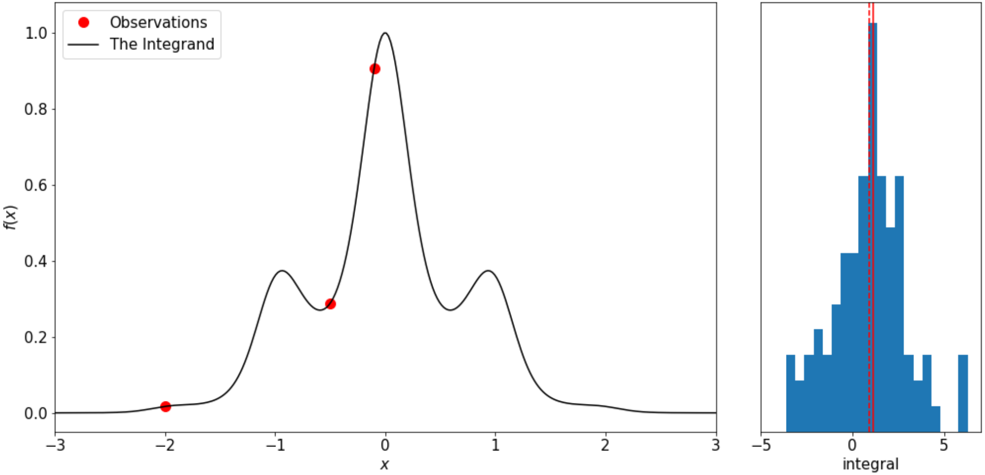
Next point? Use uncertainty of model



Next point? Use uncertainty of model



Ground truth $\int f(x)\mathrm{d}x = ?$ with $f(x) = e^{-x^2-\sin^2(x)}$

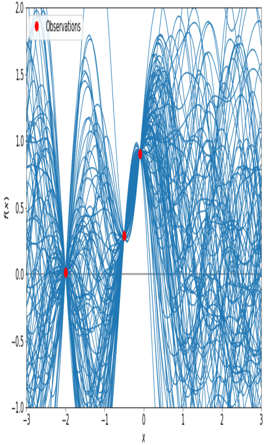


What just happened?

We have build an inutitive integration method:

- we transformed the integration problem into a regression problem.
- we incorporated known structure (smoothness) into the regression model.
- we expressed our uncertainty in unobserved areas.

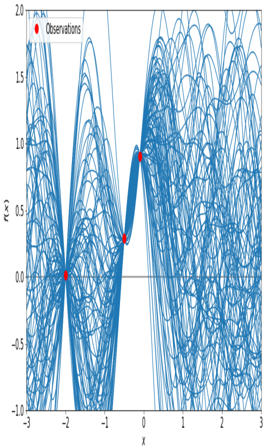
- we integrate the regressor model instead of the true integrand.
- we used the uncertainty of the regressor model to get an uncertainty over the integral value.
- we use the uncertainty to choose informative points.



What just happened?

We have build an inutitive integration method:

procedure BayesQuad(model, some initial observations)
  while stopping criterion not met **do**
    $x \leftarrow \max_{x'}$ acquisition function(x', model)   // new point
    $y \leftarrow f(x)$ // evaluate integrand
    model $\leftarrow$ update_model(x, y)   // fit model
  end while
  return integral over model   // return integral distribution
end procedure



What just happened?

We have build an inutitive integration method:

procedure BayesQuad(model, some initial observations)

while stopping criterion not met **do**

$x \leftarrow \max_{x'}$ acquisition function(x', model) // new point

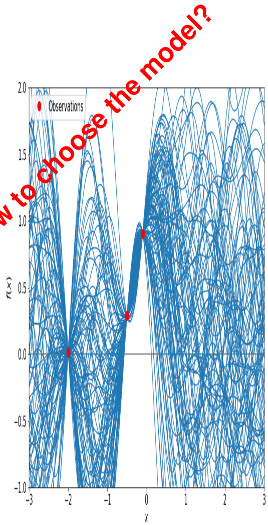
$y \leftarrow f(x)$ // evaluate integrand

model $\leftarrow$ update_model(x, y) // fit model

end while

return integral over model // return integral distribution

end procedure



Bayesian quadrature with Gaussian processes

in a nutshell [Diaconis 1988, O'Hagan 1991, . . .]

Consider the integration problem:

$$Z = \int f(x)p(x)\mathrm{d}x = ?$$

Choose a Gaussian process as regressor model for f. Observe evaluations y of f.

$$f \sim \mathcal{GP}(m, k), \qquad y(x) = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Derive the integral over the GP model. This yields a univariate Gaussian distribution on Z!

$$Z \sim \mathcal{N}(\mathfrak{m}_Z, \mathfrak{v}_Z) \qquad \mathfrak{m}_Z = \int m(x)p(x)\mathrm{d}x, \qquad \mathfrak{v}_Z = \iint k(x, x')p(x)p(x')\mathrm{d}x\mathrm{d}x'$$

Bayesian quadrature with Gaussian processes

in a nutshell [Diaconis 1988, O'Hagan 1991, ...]

Consider the integration problem:

$$Z = \int f(x)p(x)\mathrm{d}x = ?$$

Choose a Gaussian process as regressor model for f. Observe evaluations y of f.

$$f \sim \mathcal{GP}(m, k), \qquad y(x) = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Derive the integral over the GP model. This yields a univariate Gaussian distribution on Z!

$$Z \sim \mathcal{N}(\mathfrak{m}_Z, \mathfrak{v}_Z) \qquad \mathfrak{m}_Z = \int m(x)p(x)\mathrm{d}x, \qquad \mathfrak{v}_Z = \iint k(x, x')p(x)p(x')\mathrm{d}x\mathrm{d}x'$$

$\mathfrak{m}_Z$ and $\mathfrak{v}_Z$ are analytic for certain kernels, e.g., the RBF-kernel!

Bayesian quadrature: joint distribution

and closeness of Gaussians under linear transforms

$$Z = \int f(x)p(x)\mathrm{d}x, \qquad f \sim \mathcal{GP}(m,k), \qquad y(x) = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\begin{bmatrix} Z \\ f(x') \\ y(x) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \int m(s)p(s)\mathrm{d}s \\ m(x') \\ m(x) \end{bmatrix}, \begin{bmatrix} \int k(s,s')p(s)p(s')\mathrm{d}s\mathrm{d}s' & \int k(s,x')p(s)\mathrm{d}s & \int k(s,x)p(s)\mathrm{d}s \\ \int k(x',s)p(s)\mathrm{d}s & k(x',x') & k(x',x) \\ \int k(x,s)p(s)\mathrm{d}s & k(x,x') & k(x,x)+\sigma^2 \end{bmatrix} \right)$$

Nice properties of Gaussians:

- closeness under linear projection: If $x \sim \mathcal{N}(\mu,\Sigma)$ then $Ax \sim \mathcal{N}(A\mu, A\Sigma A^\intercal)$.
- closeness under conditioning and marginalization (see GPSS intro talks!).

Acquisition strategy

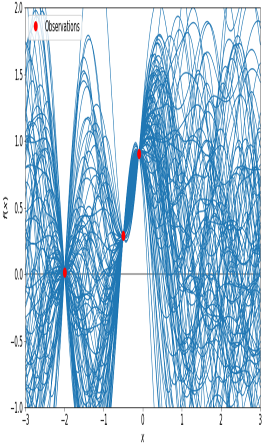
$$x_{next} := \arg\max_x a(x)$$

Select 'most informative points', e.g., point at max...

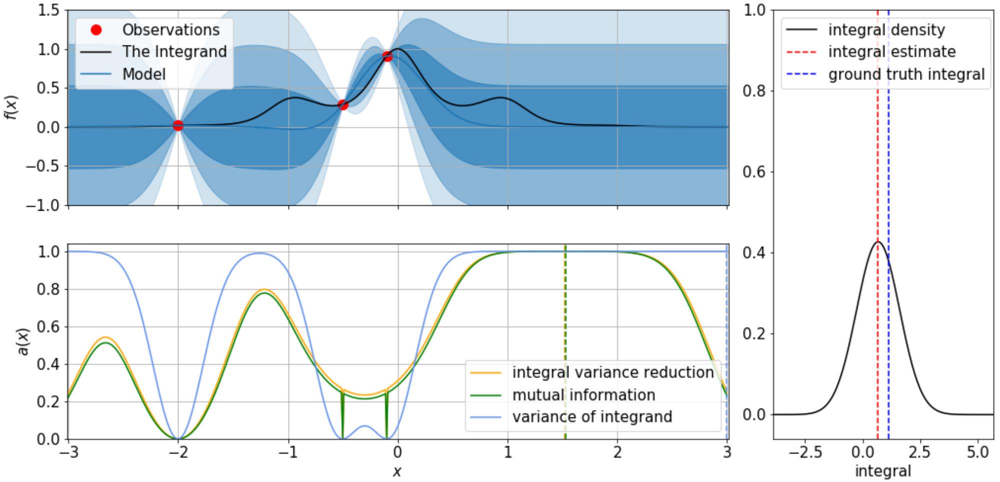
- integrand variance (US): $a(x) := k_{Data}(x, x)$
- variance reduction of integral (IVR): $a(x) := \mathfrak{v}_Z - \mathfrak{v}_Z(x)$
- mutual information (MI): $a(x) := H[Z] + H[y(x)] - H[Z, y(x)]$.

Properties

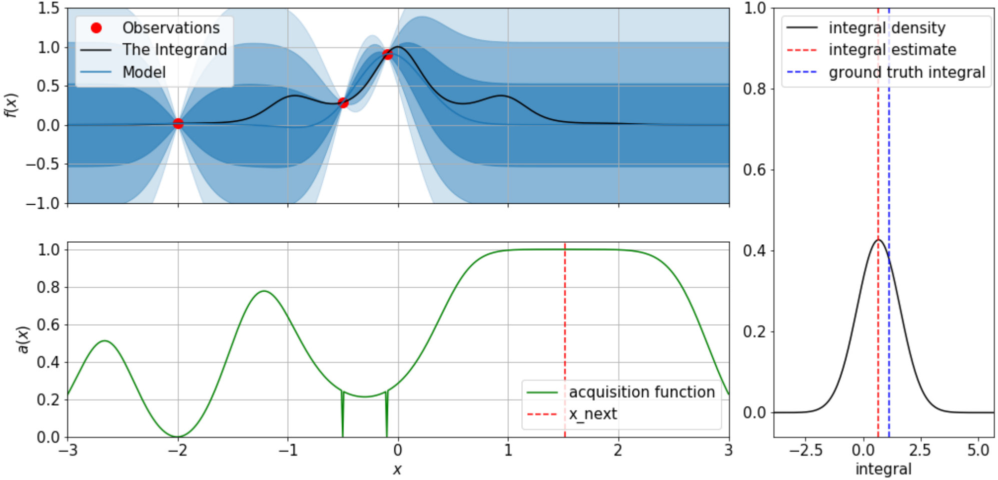
- US is a local, IVR & MI are global properties of the space.
- for Gaussians, both IVF and MI are analytic.
- for Gaussians, IVR and MI monotonic tranforms of each other, and thus share the global optimizer.



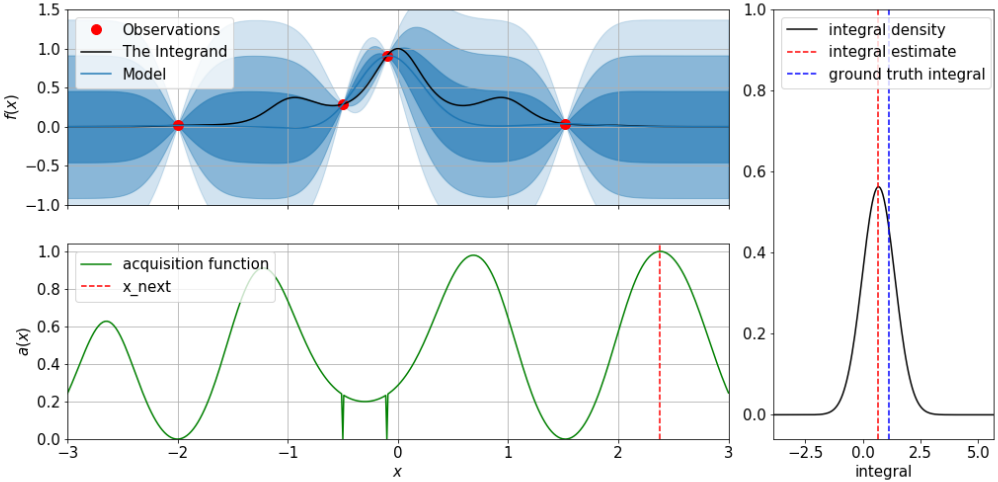
Acquisition strategy: choose next point



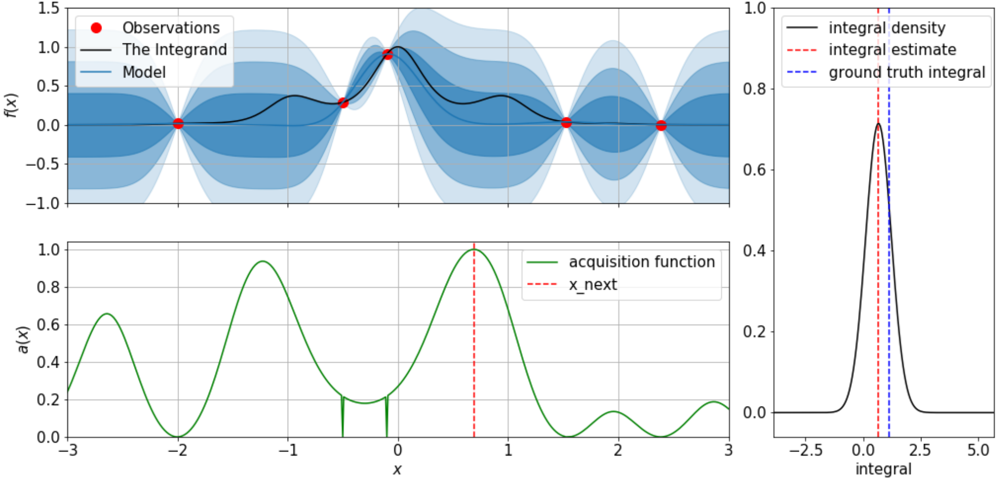
Bayesian quadrature loop *iter* = 0



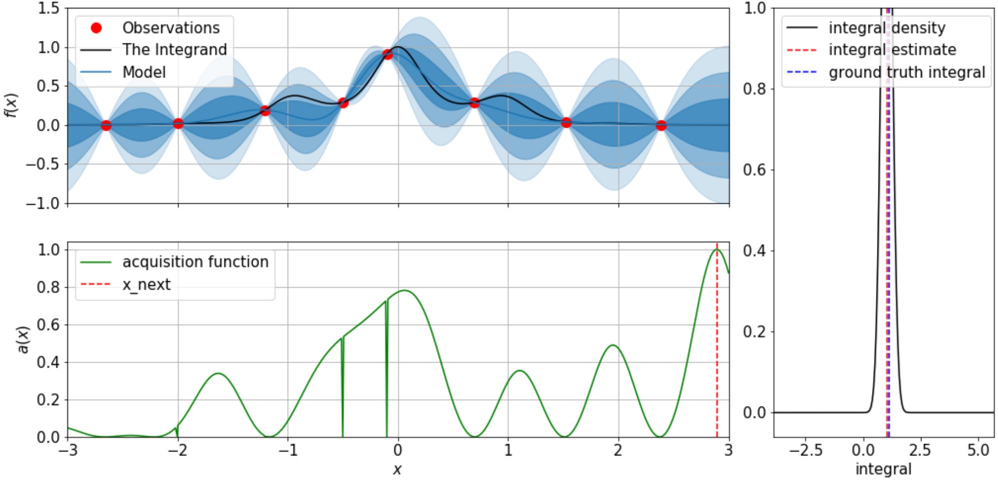
Bayesian quadrature loop   *iter* = 1



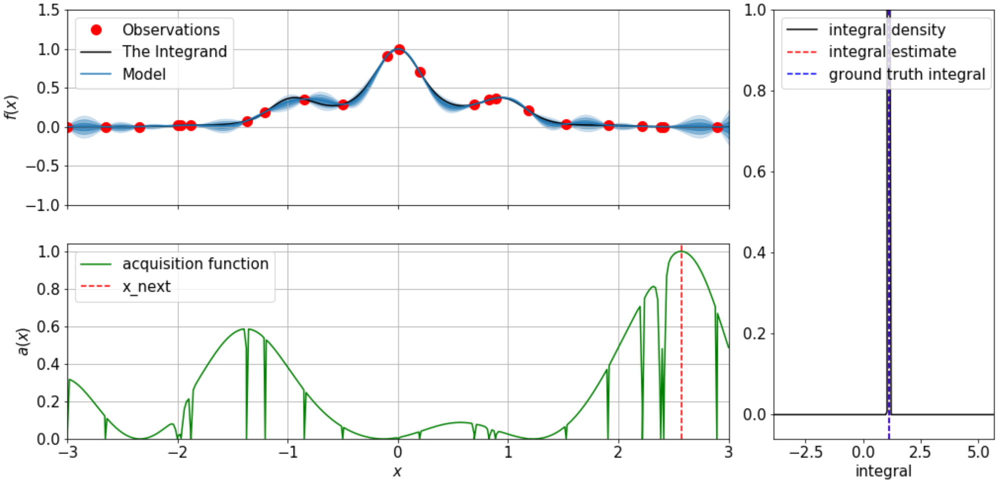
Bayesian quadrature loop *iter* = 2



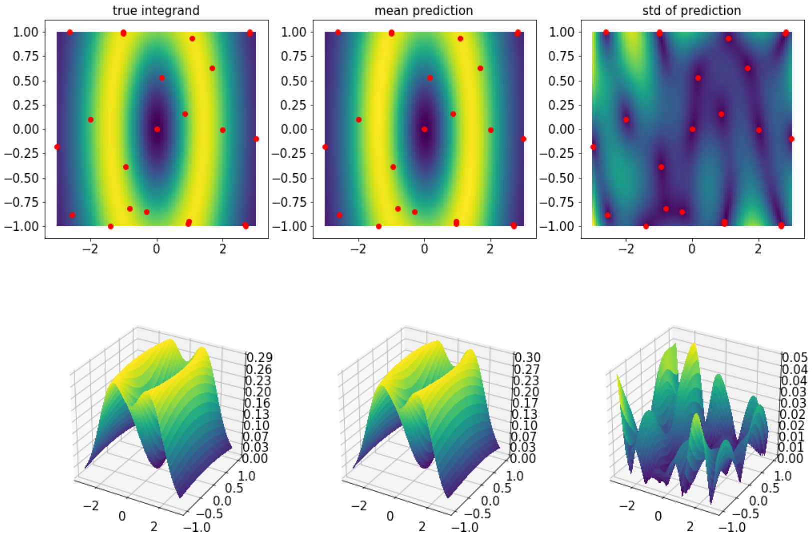
Bayesian quadrature loop *iter* = 5



Bayesian quadrature loop *iter* = 20

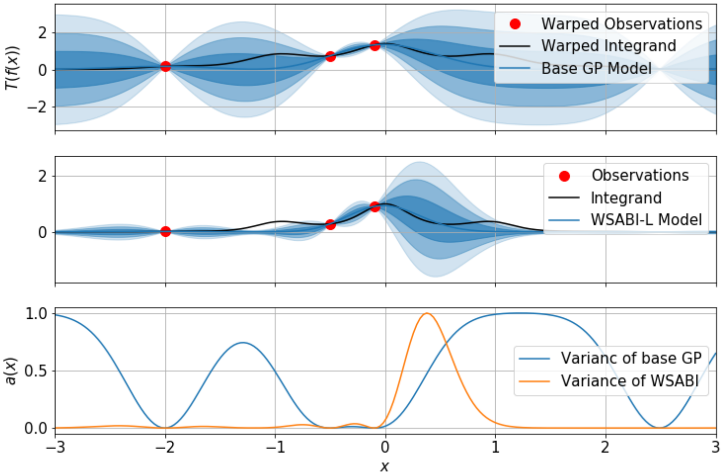


There is more: multi-D extension



There is more: constraints and transformations $f = \frac{1}{2}g^2 + \alpha$

[Gunter et al. NeurIPS 2014]



Outline

Part I

- Intro to Bayesian quadrature

Part II

- Active multi-information source Bayesian quadrature
  A.Gessner, J.Gonzalez, M.Mahsereci, UAI 2019

Outline

Part I

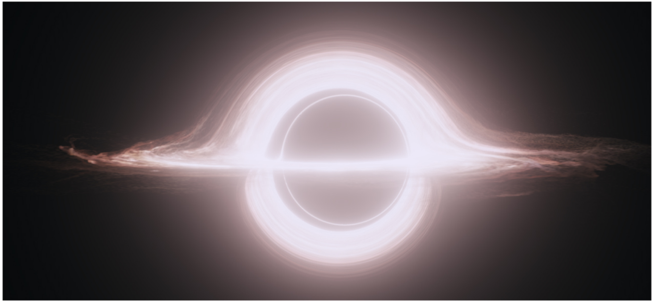
- Intro to Bayesian quadrature

Part II

- Active multi-information source Bayesian quadrature
  A.Gessner, J.Gonzalez, M.Mahsereci, UAI 2019

Expensive computer simulations

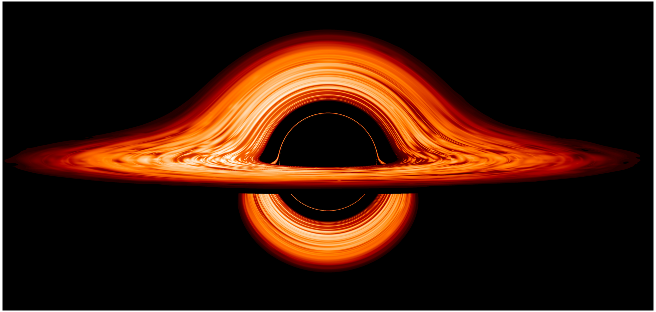
$\alpha \rightarrow$  $= f(\alpha)$

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)\mathrm{d}\alpha = ?$$

James et al. 2015, "Gravitational lensing by spinning black holes in astrophysics, and in the movie Interstellar" in Classical and Quantum Gravity. DOI: 10.1088/0264-9381/32/6/065001, https://www.nasa.gov/

Expensive computer simulations with secondary sources

$\alpha \rightarrow$

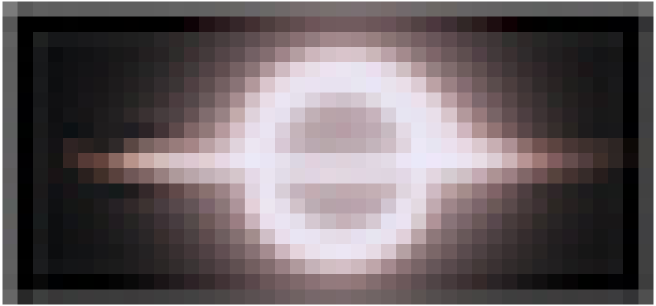


$= f(\alpha)$

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)\mathrm{d}\alpha = ?$$

James et al. 2015, "Gravitational lensing by spinning black holes in astrophysics, and in the movie Interstellar" in Classical and Quantum Gravity. DOI: 10.1088/0264-9381/32/6/065001, https://www.nasa.gov/

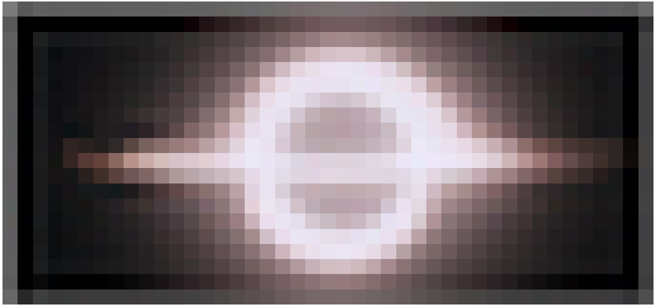
Expensive computer simulations with secondary sources

$\alpha \rightarrow$  $= f(\alpha)$

$$\mathbf{E}[f] = \int f(\alpha) p(\alpha) \mathrm{d}\alpha = ?$$

James et al. 2015, “Gravitational lensing by spinning black holes in astrophysics, and in the movie Interstellar” in Classical and Quantum Gravity. DOI: `10.1088/0264-9381/32/6/065001`, `https://www.nasa.gov/`

Expensive computer simulations with secondary sources

$\alpha \rightarrow$  = $f(\alpha)$

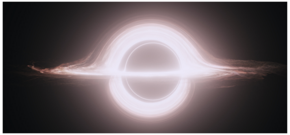
cheaper!

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)\mathrm{d}\alpha = ?$$

James et al. 2015, “Gravitational lensing by spinning black holes in astrophysics, and in the movie Interstellar” in Classical and Quantum Gravity. DOI: 10.1088/0264-9381/32/6/065001, https://www.nasa.gov/

Expensive integrals

α_1 α_2 α_3 ... α_N

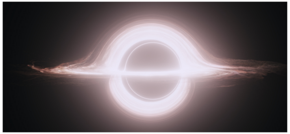


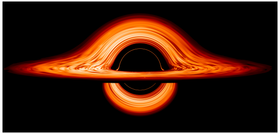
$f(\alpha_1)$ $f(\alpha_2)$ $f(\alpha_3)$ ... $f(\alpha_N)$

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)\mathrm{d}\alpha = ?$$

Expensive integrals with secondary sources

$\alpha_1 \ \dots \ \alpha_N$





$f(\alpha_1) \ \dots \ f(\alpha_N)$

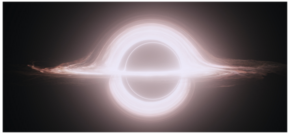
$f_2(\alpha_1) \ \dots \ f_2(\alpha_N)$

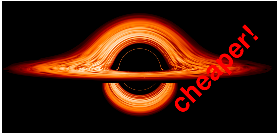
$f_3(\alpha_1) \ \dots \ f_3(\alpha_N)$

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)\mathrm{d}\alpha = ?$$

Expensive integrals with secondary sources

$\alpha_1 \ \dots \ \cancel{\alpha_N}\ \tilde{N} > N$





$f(\alpha_1) \ \dots \ f(\alpha_N)$

$f_2(\alpha_1) \ \dots \ f_2(\alpha_N)$

$f_3(\alpha_1) \ \dots \ f_3(\alpha_N)$

$$\mathbf{E}[f] = \int f(\alpha)p(\alpha)d\alpha = ?$$

Multi-source Bayesian quadrature

with multi-output Gaussian processes [Alvarez et al., 2012, Briol et al. 2019, Xi et al. 2018]

Consider L sources $f_1, f_2, \dots f_L$, w.l.o.g. $L = 1$ is primary source.

$$Z_1 = \int f_1(x)p(x)\mathrm{d}x = ?$$

Choose a multi-output GP as model for all $\mathbf{f} = [f_1, \dots f_L]^\intercal$. Observe evaluations y_l of f_l.

$$\mathbf{f} \sim \mathcal{GP}(\mathbf{m}, \mathbf{k}), \qquad y_l(x) = f_l(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Use intrinsic coregionalization kernel $\mathbf{k}_{ll'}(x, x') = \mathbf{cov}(f_l(x), f_{l'}(x')) := \mathbf{B}_{ll'}\kappa(x, x'), \quad \mathbf{B} \in \mathbb{R}^{L \times L}$ p.d..

Derive the integral over the GP model.

$$Z_1 \sim \mathcal{N}(\mathfrak{m}_{Z_1}, \mathfrak{v}_{Z_1}) \qquad \mathfrak{m}_{Z_1} = \int \mathbf{m}_1(x)p(x)\mathrm{d}x, \qquad \mathfrak{v}_{Z_1} = \int \mathbf{k}_{11}(x, x')p(x)p(x')\mathrm{d}x\mathrm{d}x'$$

Multi-source Bayesian quadrature

with multi-output Gaussian processes [Alvarez et al., 2012, Briol et al. 2019, Xi et al. 2018]

Consider L sources $f_1, f_2, \dots f_L$, w.l.o.g. $L = 1$ is primary source.

$$Z_1 = \int f_1(x)p(x)\mathrm{d}x = ?$$

Choose a multi-output GP as model for all $\mathbf{f} = [f_1, \dots f_L]^\intercal$. Observe evaluations y_l of f_l.

$$\mathbf{f} \sim \mathcal{GP}(\mathbf{m}, \mathbf{k}), \qquad y_l(x) = f_l(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Use intrinsic coregionalization kernel $\mathbf{k}_{ll'}(x, x') = \mathbf{cov}(f_l(x), f_{l'}(x')) := \mathbf{B}_{ll'}\kappa(x, x')$, $\mathbf{B} \in \mathbb{R}^{L\times L}$ p.d..

Derive the integral over the GP model.

$$Z_1 \sim \mathcal{N}(\mathfrak{m}_{Z_1}, \mathfrak{v}_{Z_1}) \qquad \mathfrak{m}_{Z_1} = \int \mathbf{m}_1(x)p(x)\mathrm{d}x, \qquad \mathfrak{v}_{Z_1} = \int \mathbf{k}_{11}(x, x')p(x)p(x')\mathrm{d}x\mathrm{d}x'$$

$\mathfrak{m}_{Z_1}$ and $\mathfrak{v}_{Z_1}$ are analytic if κ is, e.g., the RBF-kernel!

Multi-source Bayesian quadrature: joint distribution

and closeness of Gaussians

$$Z_1 = \int f_1(x)p(x)\mathrm{d}x, \qquad \mathbf{f} \sim \mathcal{GP}(\mathbf{m}, \mathbf{k}), \qquad y_l(x) = f_l(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{bmatrix} Z_1 \\ f_{l'}(x') \\ y_l(x) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \int m_1(s)p(s)\mathrm{d}s \\ m_{l'}(x') \\ m_l(x) \end{bmatrix}, \begin{bmatrix} \int k_{11}(s,s')p(s)p(s')\mathrm{d}s\mathrm{d}s' & \int k_{1l'}(s,x')p(s)\mathrm{d}s & \int k_{1l}(s,x)p(s)\mathrm{d}s \\ \int k_{l'1}(x',s)p(s)\mathrm{d}s & k_{l'l'}(x',x') & k_{l'l}(x',x) \\ \int k_{l1}(x,s)p(s)\mathrm{d}s & k_{ll'}(x,x') & k_{ll}(x,x) + \sigma^2 \end{bmatrix} \right)$$

Nice properties of Gaussians:

- closeness under linear projection: If $x \sim \mathcal{N}(\mu, \Sigma)$ then $Ax \sim \mathcal{N}(A\mu, A\Sigma A^\intercal)$.
- closeness under linear observations: $y_l(x) = H[Z_1, f_1(x'), \dots, y_1(x), \dots]^\intercal$, with H zero row vector, except 1 at $y_l(x)$ entry (see GPSS intro talks!).

Acquisition strategy: rates

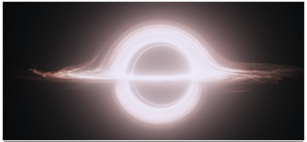
[Gessner et al. 2019]

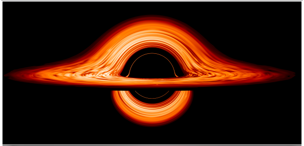
$$l_{next}, x_{next} := \arg\max_{l,x} a_l(x)$$

Select source l and point x with highest rate of information gain:

- assign cost $c_l(x) > 0$ of evaluating source l at x.
- define acquisition rate: $a_l(x) := \frac{a(x)}{c_l(x)}$.
- find point and source $\{l_{next}, x_{next}\}$ that maximizes the rate.

Example: If cost is measured in time, and $a(x)$ is the mutual information, then $a_l(x)$ has unit $\frac{\text{bits}}{\text{second}}$, i.e., a rate of information gain.





...

Active multi-source BayesQuad: loop

[Gessner et al. 2019]

procedure MultiSourceBQ(multi-source-model, some initial observations)

  while stopping criterion not met **do**

    $l, x \leftarrow \arg\max_{x'}$ rate(multi-source-model)   // new point

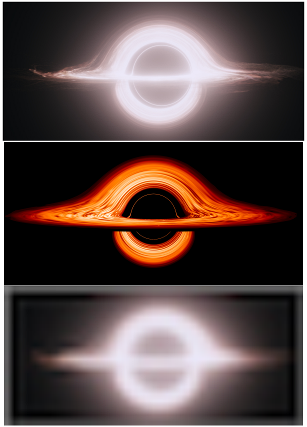
    $y \leftarrow f_l(x)$ // evaluate source l at x

    multi-source-model $\leftarrow$ update_model(l, x, y)   // fit model

  end while

  return integral over model   // return integral distribution

end procedure



...

Acquisition strategy: some pitfalls

[Gessner et al. 2019]

Recall some single-source acquisition strategies:

- mutual information (MI): $a(x) := H[Z] + H[y(x)] - H[Z, y(x)]$.
- variance reduction of integral (IVR): $a(x) := \mathfrak{v}_Z - \mathfrak{v}_Z(x)$
- ...

In single-source BQ any increasing, strictly monotonic transform of $a(x)$ is admissible, since the global optimizer does not change, e.g.,

- negative integral variance (NIV): $a(x) := -\mathfrak{v}_Z(x)$
- integral precision (IP): $a(x) := \mathfrak{v}_Z^{-1}(x)$
- ...

This is not the case in multi-source BQ! Only some translate to meaningful rates (beware of what you encode!)

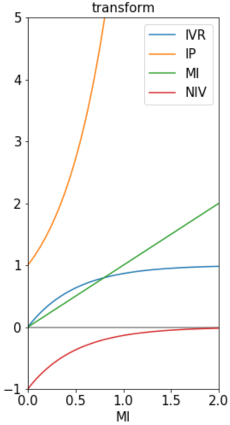
Example: NIV, IP, or any utility that does not encode "improvement".

Acquisition strategy: some pitfalls

[Gessner et al. 2019]

Name	formula	transform
mutual information (MI)	$-0.5 \log\left(1-\rho^2(x)\right)$	a
integral variance reduction (IVR)	$\rho^2(x)$	$1-e^{-2a}$
negative integral variance (NIV)	$\rho^2(x)-1$	$-e^{-2a}$
integral precision (IP)	$(1-\rho^2(x))^{-1}$	e^{2a}
$\dots$	$\dots$	$\dots$

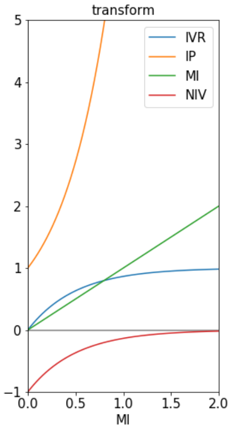
$\rho^2(x) \in [0,1]$ is the correlation between the integral Z_1 and the new observation $y(x)$.



Acquisition strategy: some pitfalls

[Gessner et al. 2019]

Name	formula	transform
mutual information (MI)	$-0.5 \log\left(1-\rho^2(x)\right)$	a
integral variance reduction (IVR)	$\rho^2(x)$	$1-e^{-2a}$
negative integral variance (NIV)	$\rho^2(x)-1$	$-e^{-2a}$
integral precision (IP)	$(1-\rho^2(x))^{-1}$	e^{2a}
...	...	...



$\rho^2(x) \in [0,1]$ is the correlation between the integral Z_1 and the new observation $y(x)$.

Some observations:

- NIV and IP shift the function, IVR does not.
- considering rate $\frac{a(x)}{c(x)}$, points that contribute no information (ρ = 0) will be accessed if the cost is low enough.

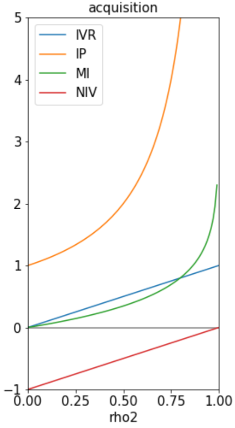
$\rightarrow$ make sure the acquisition function encodes a meaningful value (not just global max.).

Acquisition strategy: some pitfalls

[Gessner et al. 2019]

Name	formula	transform
mutual information (MI)	$-0.5 \log\left(1 - \rho^2(x)\right)$	a
integral variance reduction (IVR)	$\rho^2(x)$	$1 - e^{-2a}$
negative integral variance (NIV)	$\rho^2(x) - 1$	$-e^{-2a}$
integral precision (IP)	$(1 - \rho^2(x))^{-1}$	e^{2a}
$\dots$	$\dots$	$\dots$

$\rho^2(x) \in [0, 1]$ is the correlation between the integral Z_1 and the new observation $y(x)$.



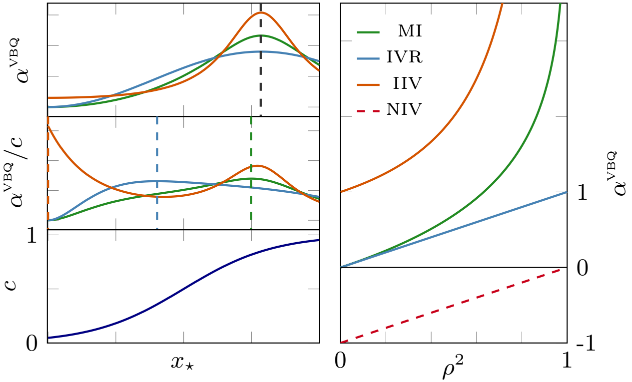
Some observations:

- NIV and IP shift the function, IVR does not.
- considering rate $\frac{a(x)}{c(x)}$, points that contribute no information ($\rho = 0$) will be accessed if the cost is low enough.

$\rightarrow$ make sure the acquisition function encodes a meaningful value (not just global max.).

Acquisition strategy: some pitfalls

[Gessner et al. 2019]

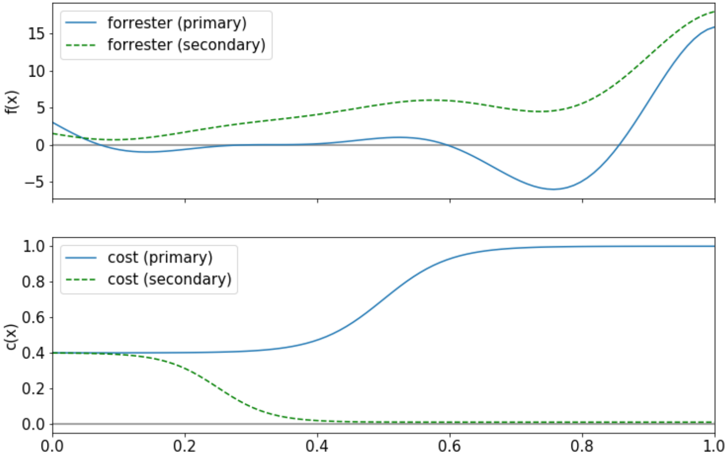


Plot for synthetic $\rho^2(x^*) = 0.95\sin^2(10x^*)$, $x^* \in [0, 0.2]$.

The cost shifts the global maximers of the acquisition function.

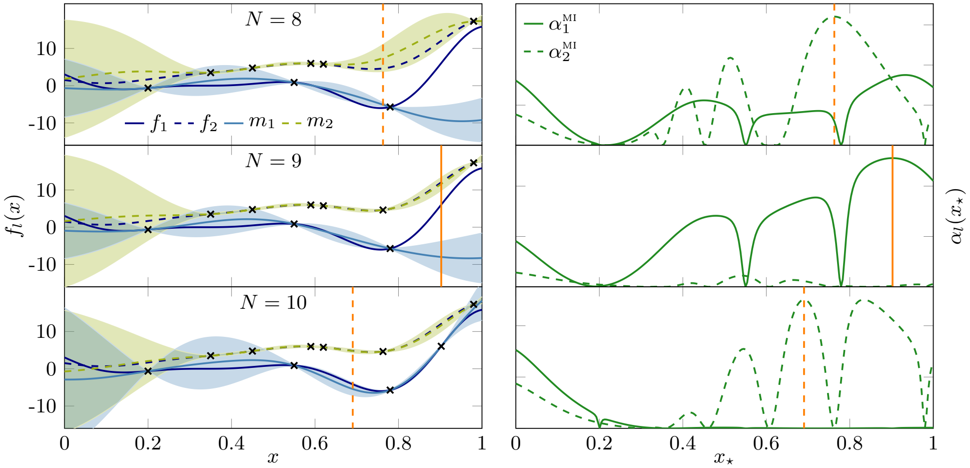
Acquisition demo: 1D Forrester function

[Gessner et al. 2019]



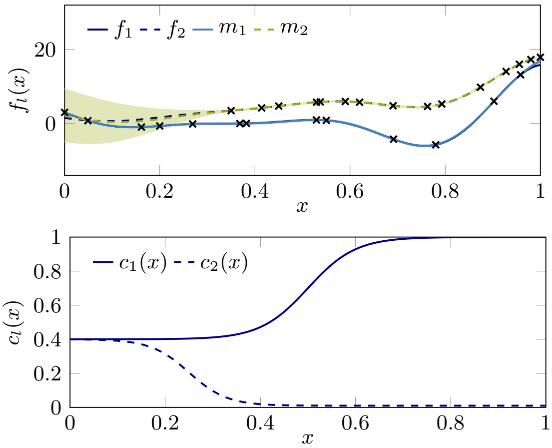
Acquisition demo: 1D Forrester function

[Gessner et al. 2019]



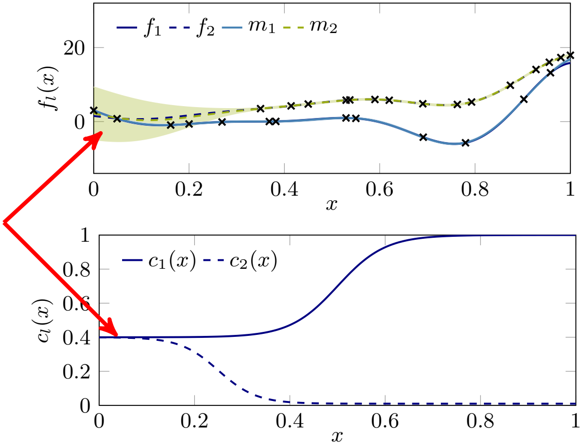
Acquisition demo: 1D Forrester function

[Gessner et al. 2019]



Acquisition demo: 1D Forrester function

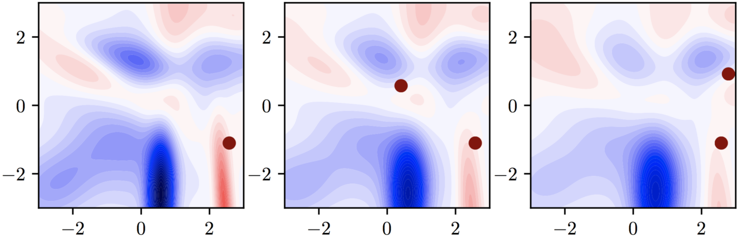
[Gessner et al. 2019]

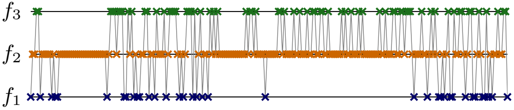


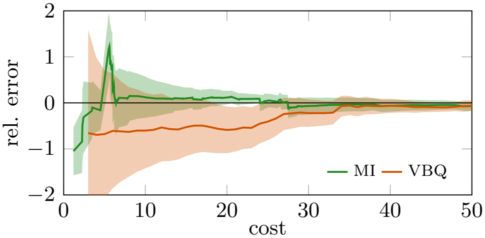
Example: 2D linear combination of Gaussians

[Gessner et al. 2019]

$c_1(x) = 1, \quad c_2(x) = c_2(x) = 0.05$

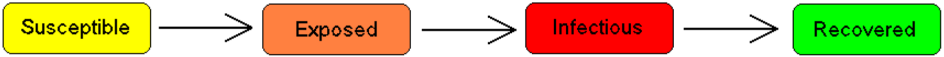






Example: the SEIR epidemiological model [Gessner et al. 2019]

Stochastic discrete-time events of individuals changing infection state (Poisson).



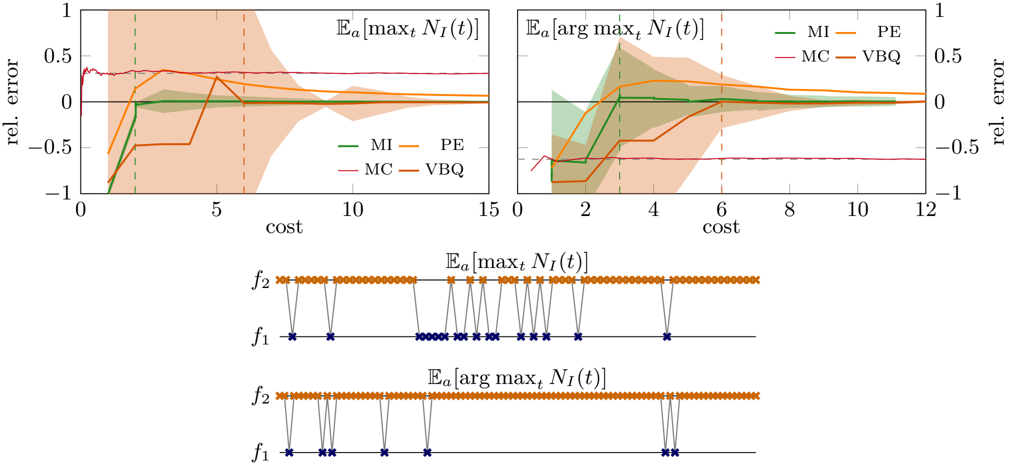
- population: $N = N_S + N_E + N_I + N_R$
- primary source: Gillespie simulation (1 query approx. 16s on laptop)
- secondary source: In the limit of a large population, the average dynamics are governed by a system of ODEs without analytic solution (1 query approx 8e-3s).

- the infection rate a is uncertain
- task1: estimate the expected height of the infection peak $\mathbf{E}_a[\max_t N_I(t)]$.
- task2: estimate the expected time occurrence of the infection peak $\mathbf{E}_a[\arg\max_t N_I(t)]$.

graphic `https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology`

Example: the SEIR model

[Gessner et al. 2019]

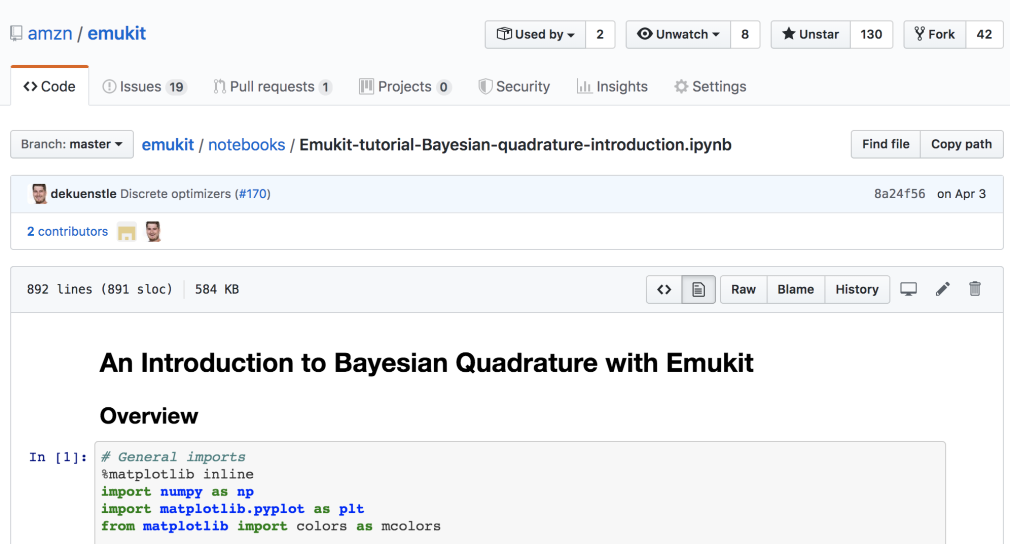


Recap

- BayesQuad has exiting application areas.
- even low dimensional integrals can be hard to solve (there is great work in BayesQuad for higher dimensions)
- the numerical error is non-negligible (important for decision pipelines)
- want to use structure in the problem
- want to use secondary data sources
- code?

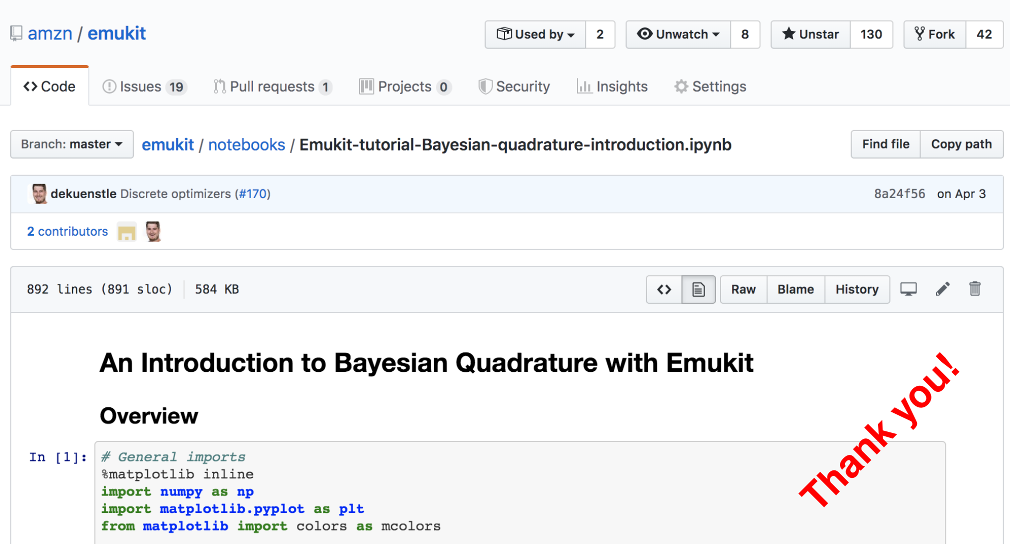
Bayesian quadrature with Emukit

https://github.com/amzn/emukit



Bayesian quadrature with Emukit

https://github.com/amzn/emukit



Some References

- A. Gessner, J. Gonzalez and M. Mahsereci, “Active Multi-Information Source Bayesian Quadrature”, in Uncertainty in Artificial Intelligence (UAI) 2019
- ’M. A. Alvarez, L. Rosasco and N. D. Lawrence, “Kernels for Vector-Valued Functions: A Review”, in Foundations and Trends® in Machine Learning
- X. Xi, F. X. Briol and M. Girolami, “Bayesian Quadrature for Multiple Related Integrals”, in Proceedings of the 35th International Conference on Machine Learning (ICML) 2018
- F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne and D. Sejdinovic, “Probabilistic Integration: A Role inStatistical Computation?”, in Statistical Science 2019
- P. Diaconis, “Bayesian numerical analysis”, in Statistical Decision Theory and Related Topics IV, 1988
- A. O’Hagan, “Bayes-Hermite Quadrature”, in Journal of Statistical Planning and Inference 1991