Bayesian optimisation for likelihood-free cosmological inference

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

with the Aquila Consortium
www.aquila-consortium.org

October 22nd, 2018

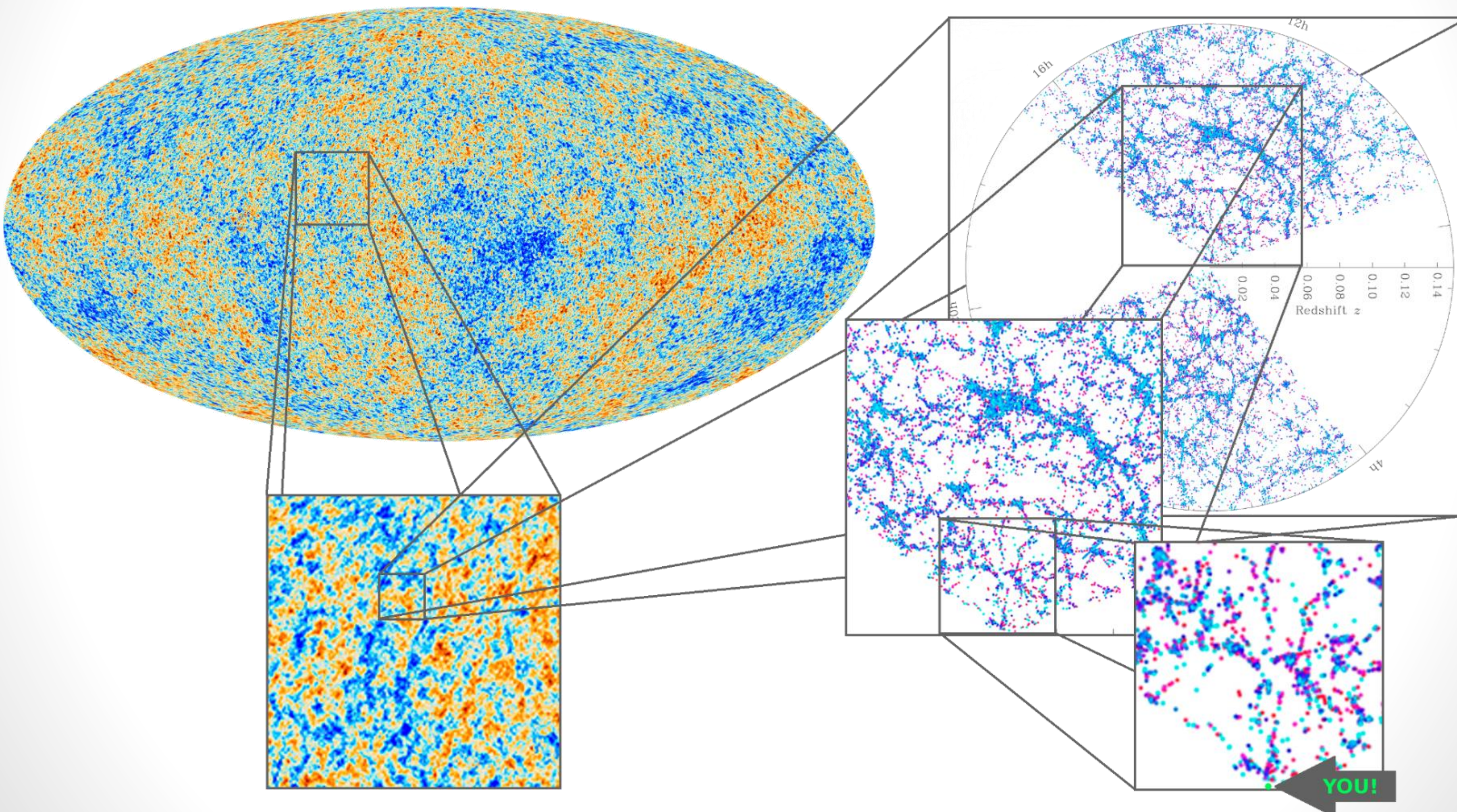
Phys. Rev. D 98, 063511 (2018), arXiv:1805.07152

ICIC Imperial Centre for Inference & Cosmology

Imperial College London

The big picture: the Universe is highly structured

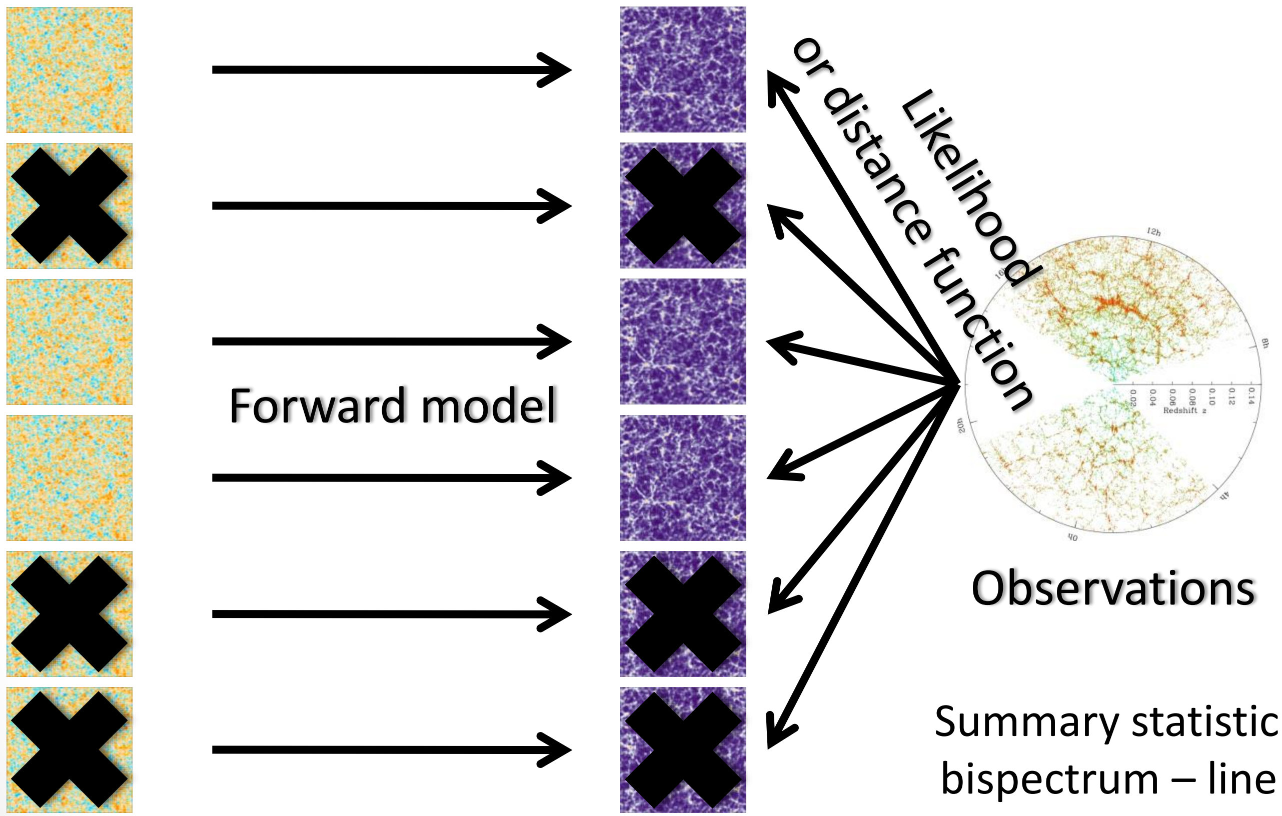
You are here. Make the best of it…



Planck collaboration (2013-2015)
M. Blanton and the Sloan Digital Sky Survey (2010-2013)

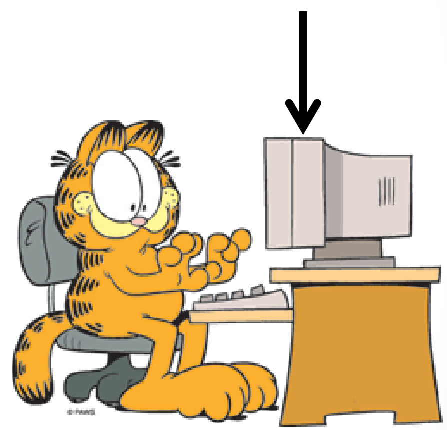
Bayesian forward modeling: the ideal scenario

Forward model = N-body simulation + Halo occupation + Galaxy formation + Feedback + …



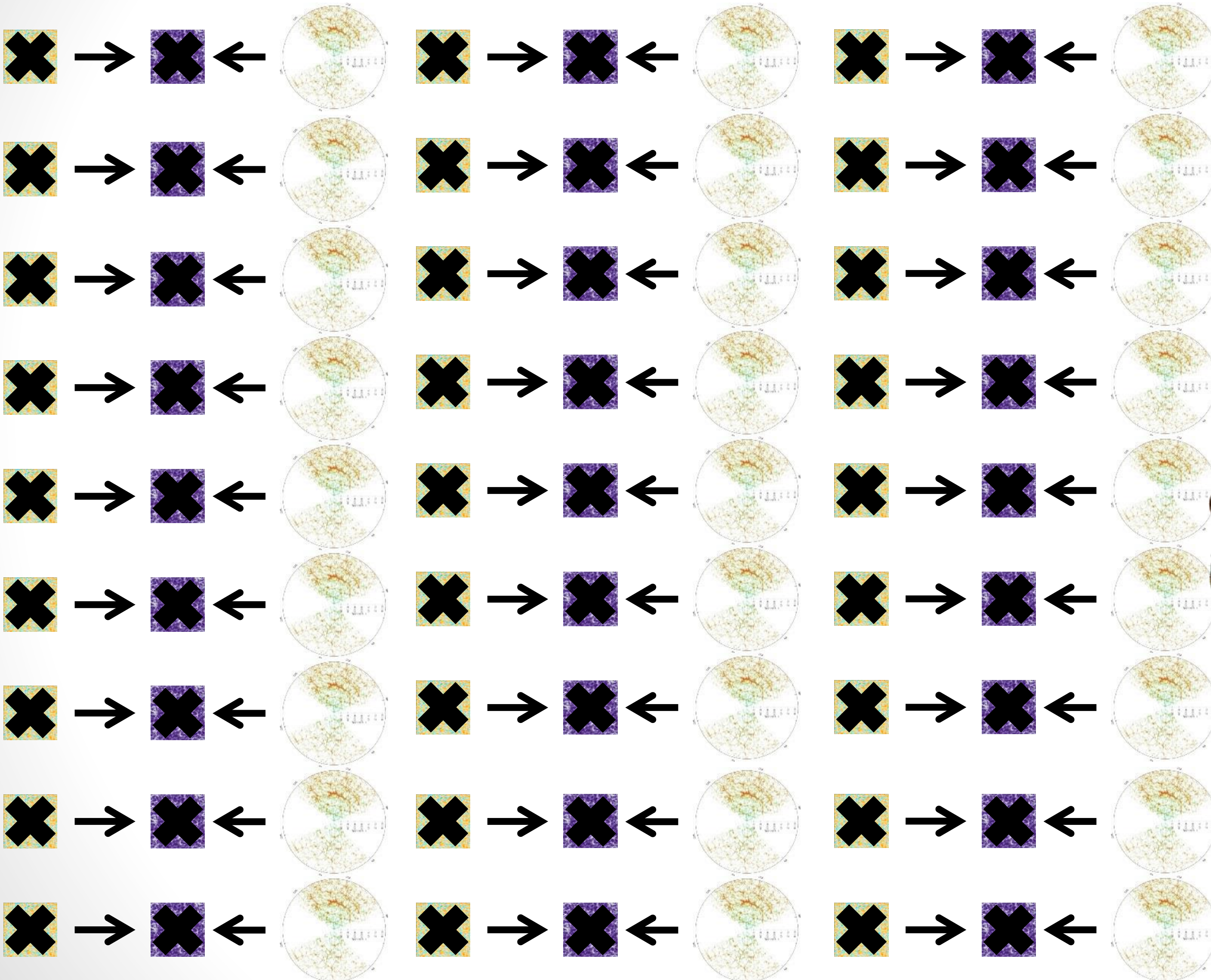
All possible ICs

All possible FCs



Summary statistic = power spectrum – bispectrum – line correlation function – clusters – voids…

Bayesian forward modeling: the challenge

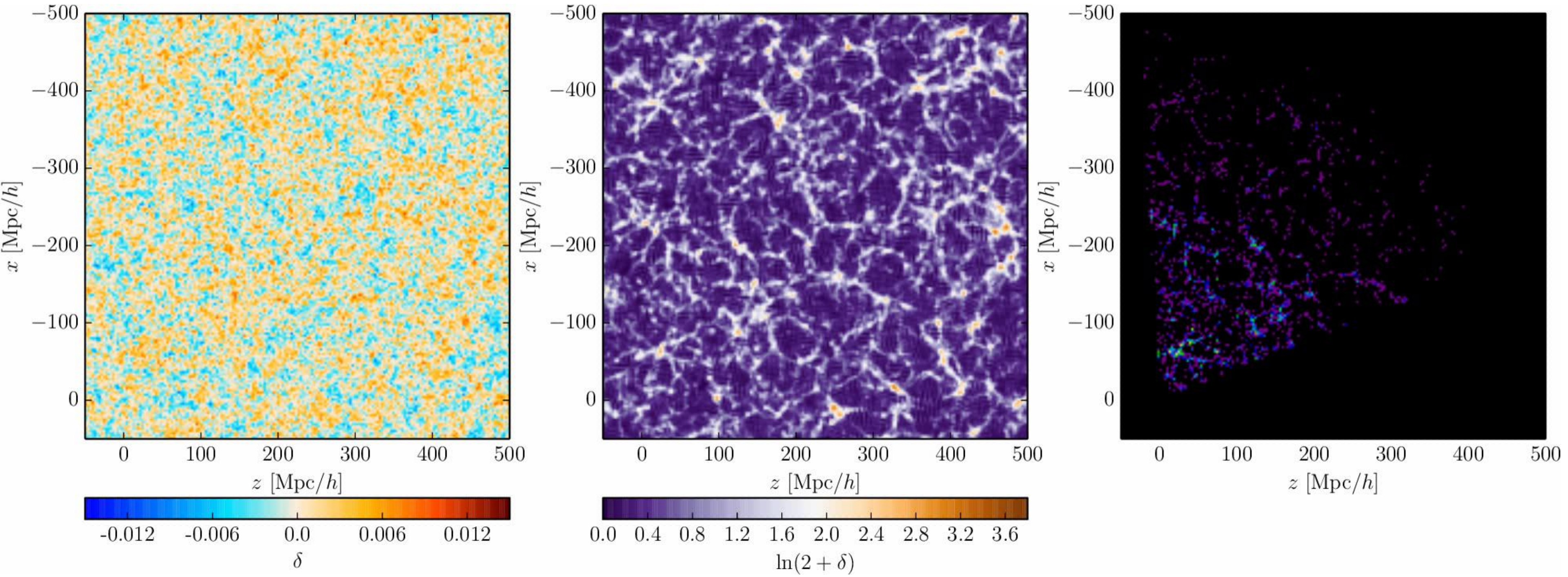


The (true) likelihood lives in

$d \approx 10^7$!

Likelihood-based solution: BORG at work

uses Hamiltonian Monte Carlo (HMC) to explore the exact posterior



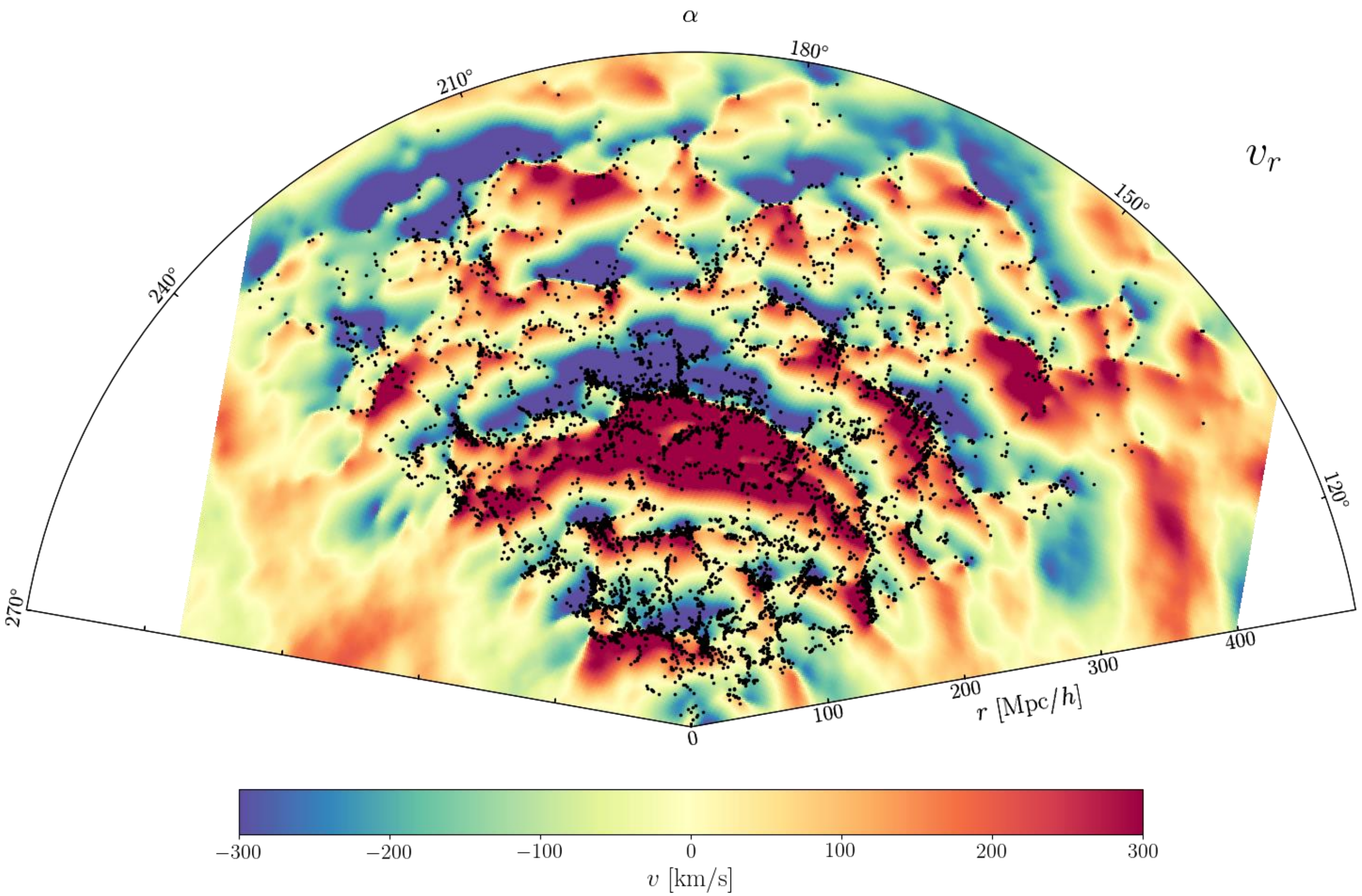
Initial conditions | **Final conditions** | **Observations**

334,074 galaxies, ≈ 17 million parameters, 3 TB of primary data products,
12,000 samples, ≈ 250,000 data model evaluations, 10 months on 32 cores

Jasche, FL & Wandelt 2015, arXiv:1409.6308

All data products are publicly available:
https://github.com/florent-leclercq/borg_sdss_data_release, doi: 10.5281/zenodo.1455729

Radial velocity field in the equatorial plane

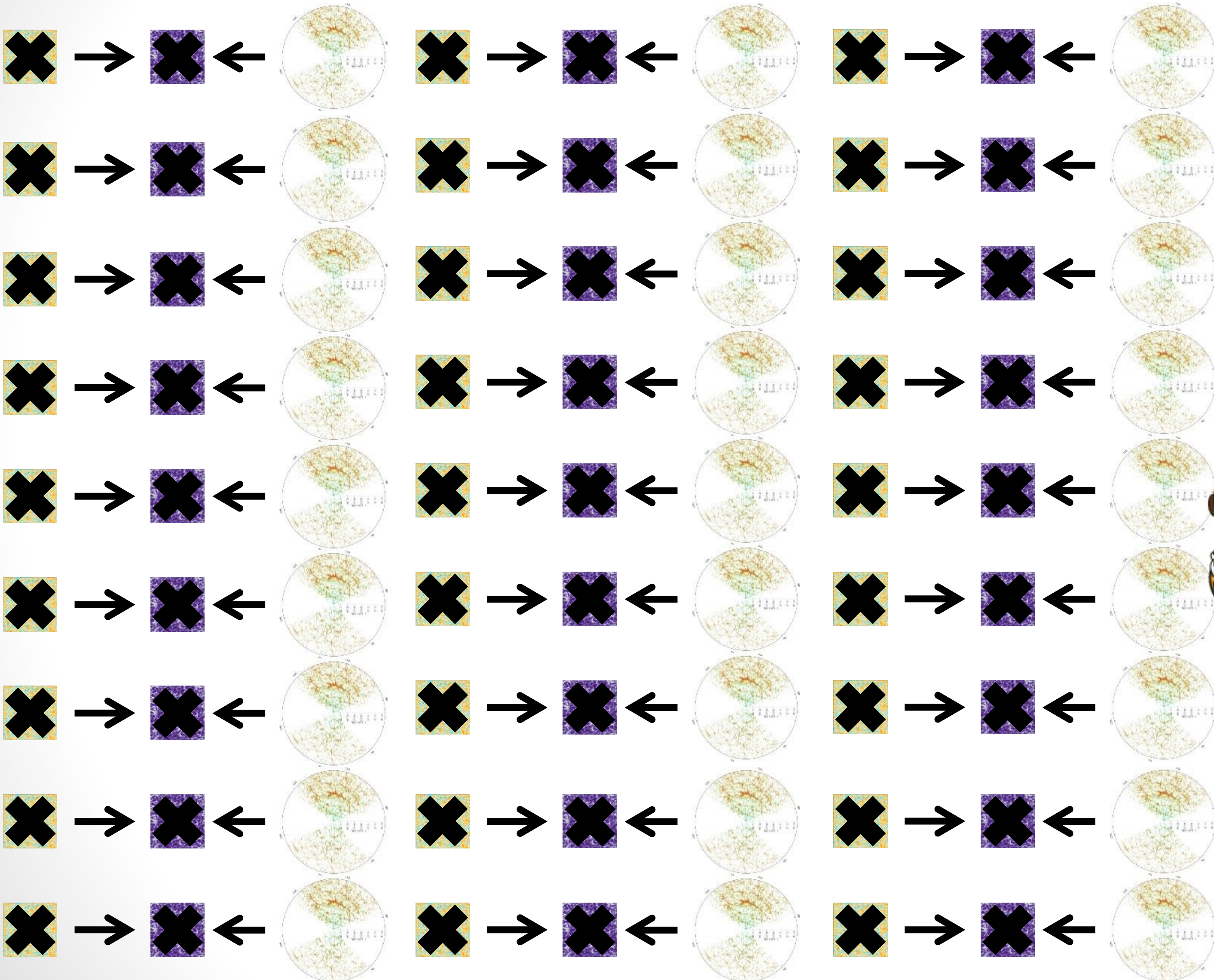


Much more about **cosmic web analysis** next Monday!

(29/10/2018, 11:00-12:00, Amphi Darboux, IHP)

FL, Jasche, Lavaux, Wandelt & Percival 2017, arXiv:1601.00093

Let's go back to the challenge…



Approximate Bayesian Computation (ABC)

- Statistical inference for models where:
  1. The likelihood function is intractable
  2. Simulating data is possible

- **General idea**: find parameter values for which the distance between simulated data and observed data is small

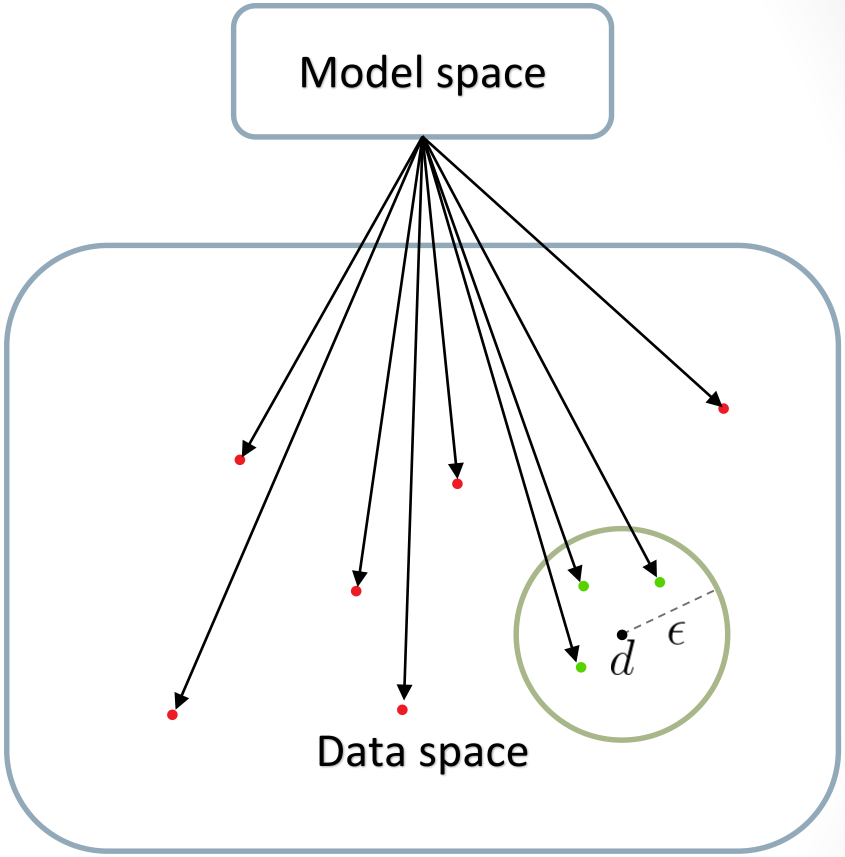
$$p(\theta|d) \Longrightarrow p(\theta|\tilde{d}) \quad \text{where } \mathrm{d}(\tilde{d}(\theta), d) \text{ is small}$$

- **Assumptions**:
  - Only a small number of parameters are of interest
  - But the process generating the data is a very general “black box”: a noisy non-linear dynamical system with an unrestricted number of hidden variables

Likelihood-free rejection sampling

- Iterate many times:
  - Sample θ from a proposal distribution $q(\theta)$
  - Simulate $\tilde{d}(\theta)$ according to the data model
  - Compute distance $\mathrm{d}(\tilde{d}(\theta), d)$ between simulated and observed data
  - Retain θ if $\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon$, otherwise reject
- Effective likelihood approximation:

$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon\right)$$



ϵ can be adaptively reduced (Population Monte Carlo)

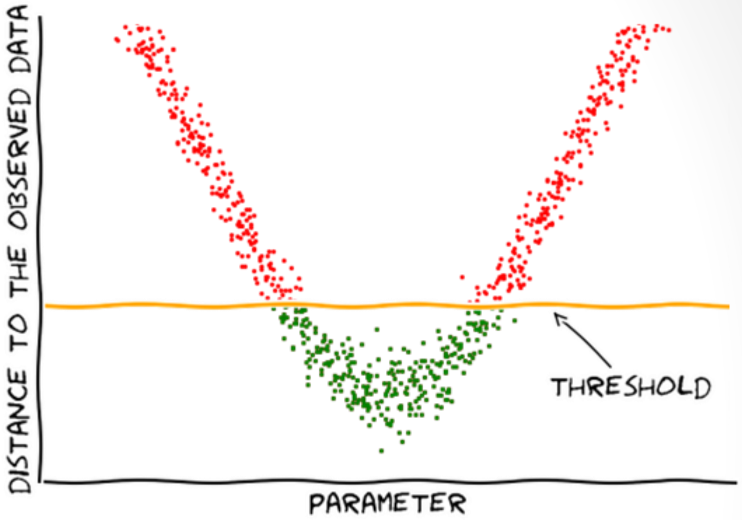
Why is likelihood-free rejection so expensive?

1. It rejects most samples when ϵ is small

2. It does not make assumptions about the shape of $L(\theta)$

3. It uses only a fixed proposal distribution, not all information available

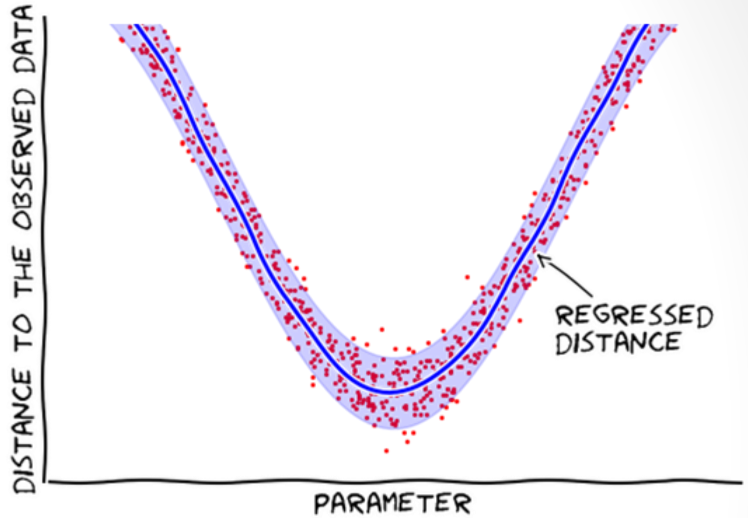
4. It aims at equal accuracy for all regions in parameter space



$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon\right)$$

Proposed solution:

BOLFI: *Bayesian Optimisation for Likelihood-Free Inference*



1. It rejects most samples when ϵ is small

   ⇨ **Don't reject samples: learn from them!**

2. It does not make assumptions about the shape of $L(\theta)$

   ⇨ **Model the distances, assuming the average distance is smooth**

3. It uses only a fixed proposal distribution, not all information available

   ⇨ **Use Bayes' theorem to update the proposal of new points**

4. It aims at equal accuracy for all regions in parameter space

   ⇨ **Prioritize parameter regions with small distances to the observed data**

Gutmann & Corander JMLR 2016, arXiv:1501.03291

Related recent work in cosmology:

Alsing & Wandelt 2018, arXiv:1712.00012
(linear data compression for ABC)

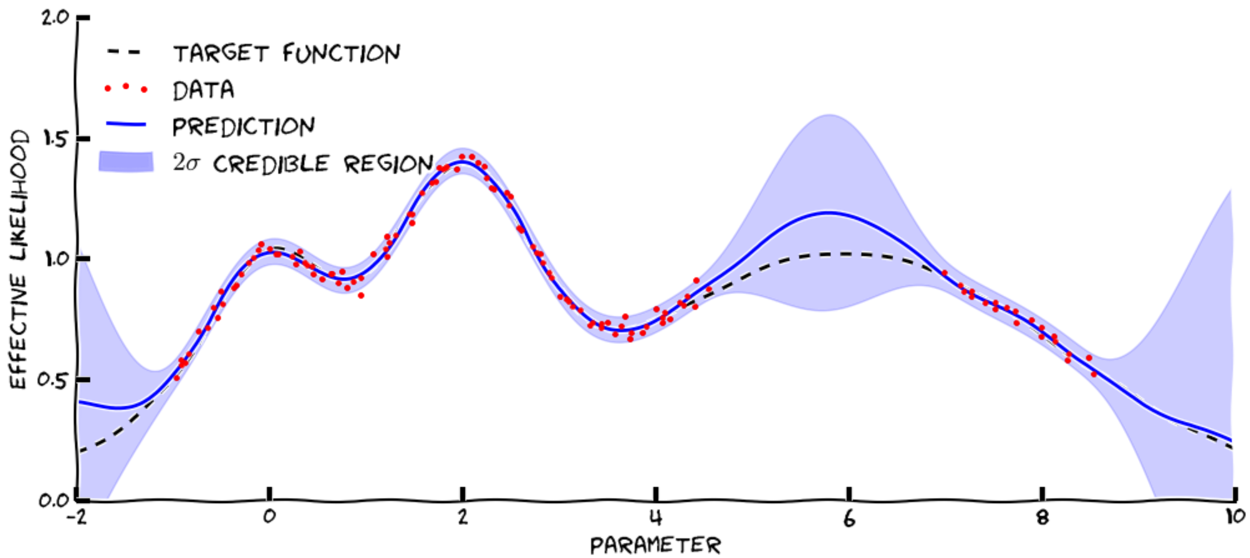
Alsing, Wandelt & Feeney 2018, arXiv:1801.01497
(density estimation for ABC – DELFI)

Charnock, Lavaux & Wandelt 2018, arXiv:1802.03537
(information-maximizing neural networks)

Hahn, Beutler *et al.* 2018, arXiv:1803.06348
(likelihood fitting before parameter inference)

Torrado & Liddle, in prep.
(Bayesian quadratures for slow $L(\theta)$)

Regressing the effective likelihood (points 1 & 2)



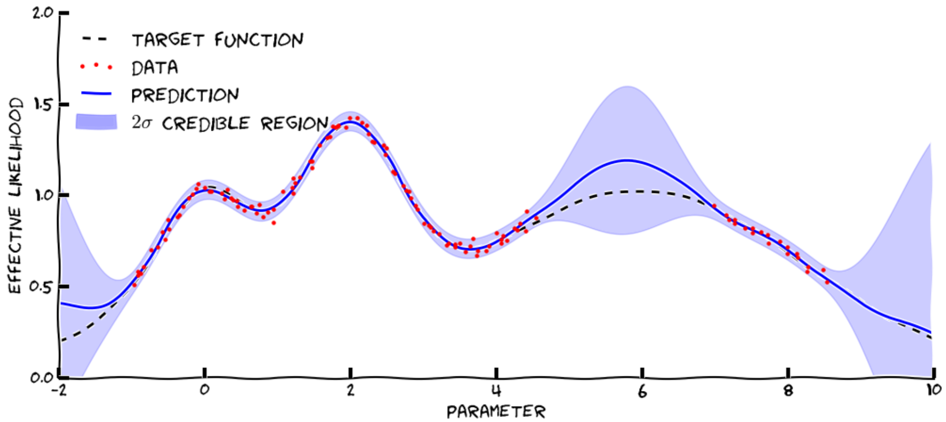
1. "It rejects most samples when ϵ is small"

- Keep all values (θ_i, d_i) $\quad \mathrm{d}_i = \mathrm{d}(\tilde{d}(\theta_i), d)$

2. "It does not make assumptions about the shape of $L(\theta)$"

- Model the conditional distribution of distances given this training set

Gaussian process regression (a.k.a. kriging)



- Why?
  - It is a **general purpose regressor**: it will be able to deal with a large variety of complex/non-linear features of likelihood functions.
  - It provides not only a prediction, but also the **uncertainty of the regression**.
  - It allows to **extrapolate** in regions where we have no data points.

$$p(\mathbf{f}|\mathbf{X}) \propto \exp\left[-\frac{1}{2}\sum_{mn}(f(\mathbf{x}_m)-\mu(\mathbf{x}_m))^\intercal K(\mathbf{x}_m,\mathbf{x}_n)(f(\mathbf{x}_n)-\mu(\mathbf{x}_n))\right]$$

$$K(\mathbf{x}_m,\mathbf{x}_n) = \underbrace{C_1}_{K_\mathrm{C}(C_1)} \times \underbrace{\exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}_m-\mathbf{x}_n}{C_2}\right)^2\right]}_{K_\mathrm{RBF}(C_2)} + \underbrace{C_3\delta_\mathrm{K}^{mn}}_{K_\mathrm{GN}(C_3)}$$

The prediction and uncertainty for a new point is:

$$p(f_\star|\mathbf{x}_\star,\mathbf{X},\mathbf{f}) \propto \exp\left[-\frac{1}{2}\left(\frac{f_\star-\alpha(\mathbf{x}_\star)}{\sigma(\mathbf{x}_\star)}\right)^2\right]$$

$$\alpha(\mathbf{x}_\star) = \mu(\mathbf{x}_\star) + K(\mathbf{x}_\star,\mathbf{x}_m)^\intercal K^{-1}(\mathbf{x}_m,\mathbf{x}_n)(\mathbf{f}-\mu(\mathbf{X}))_n$$

$$\sigma(\mathbf{x}_\star)^2 = K(\mathbf{x}_\star,\mathbf{x}_\star) - K(\mathbf{x}_\star,\mathbf{x}_m)^\intercal K^{-1}(\mathbf{x}_m,\mathbf{x}_n)K(\mathbf{x}_\star,\mathbf{x}_n)$$

Hyperparameters C_1, C_2, C_3 are automatically adjusted during the regression.

Rasmussen & Williams 2006

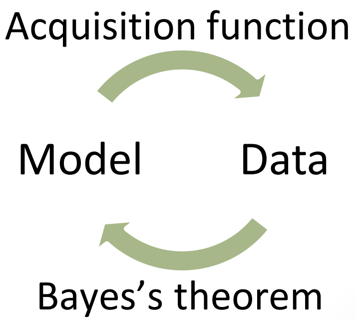
Data acquisition (points 3 & 4)

3. "It uses only a fixed proposal distribution, not all information available"

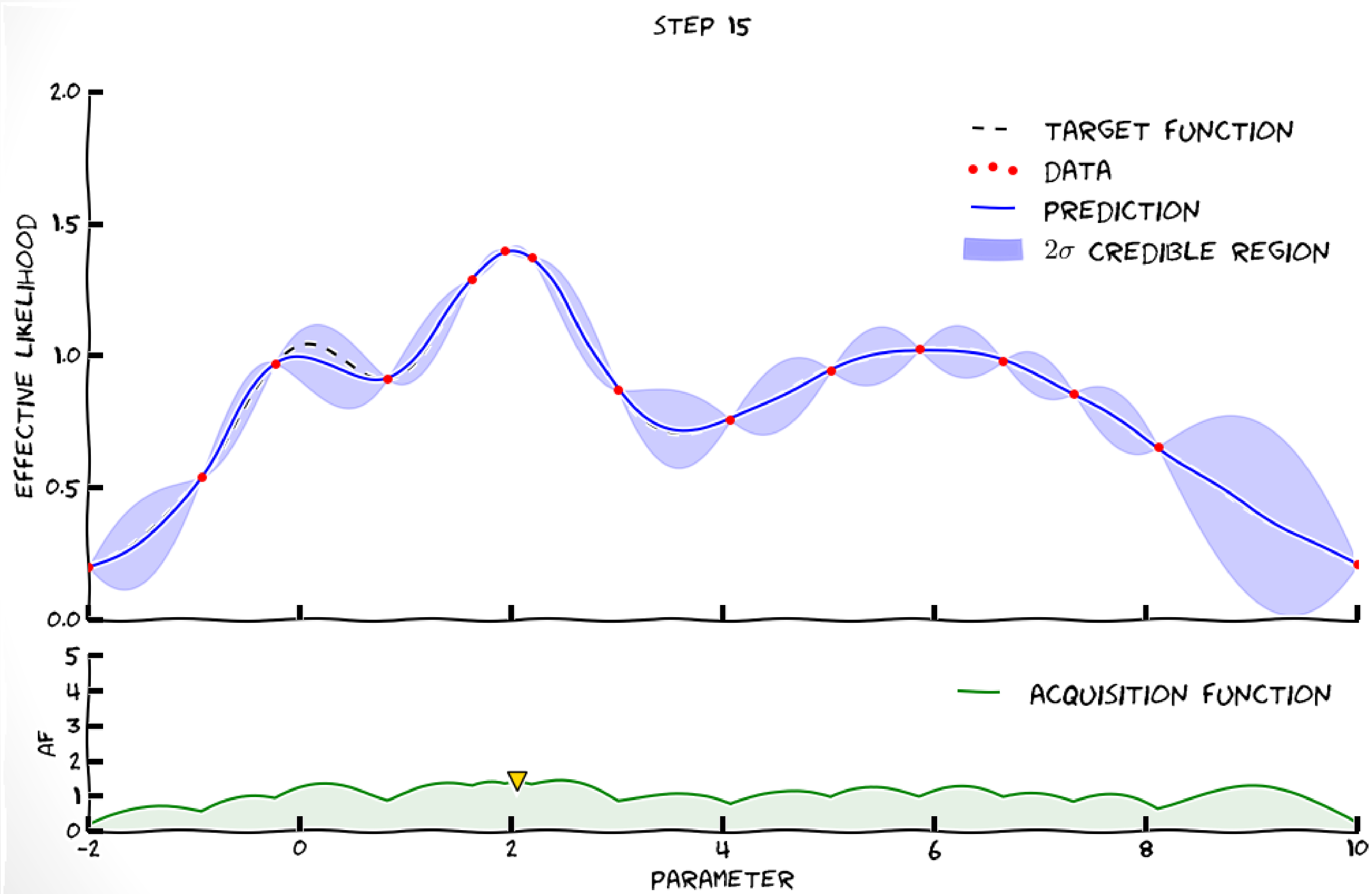
- Samples are obtained from sampling an **adaptively-constructed proposal distribution**, using the regressed effective likelihood

4. "It aims at equal accuracy for all regions in parameter space"

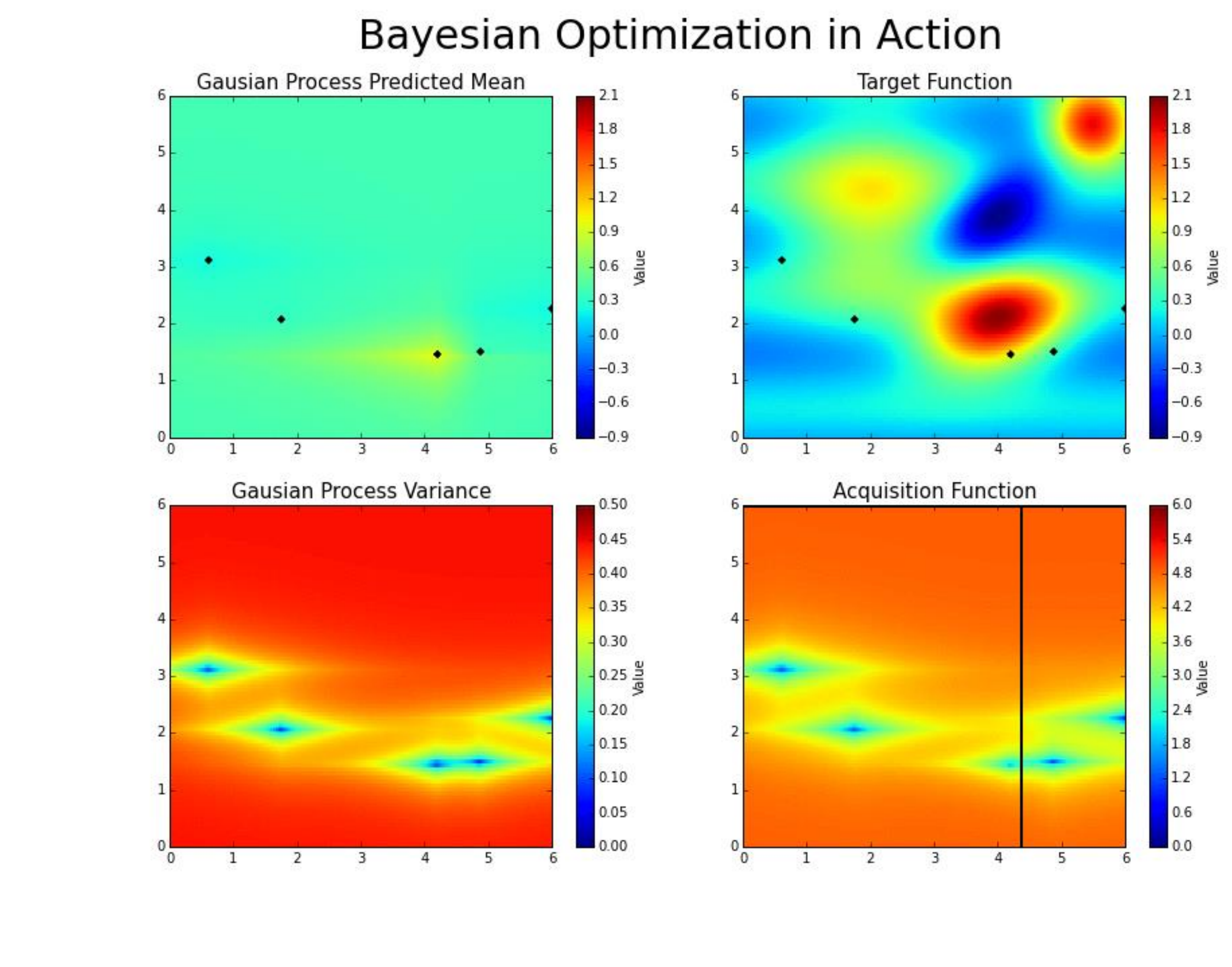
- The **acquisition function** finds a compromise between exploration (trying to find new high-likelihood regions) & exploitation (giving priority to regions where the distance to the observed data is already known to be small)
- **Bayesian optimisation** (decision making under uncertainty) can then be used



Data acquisition



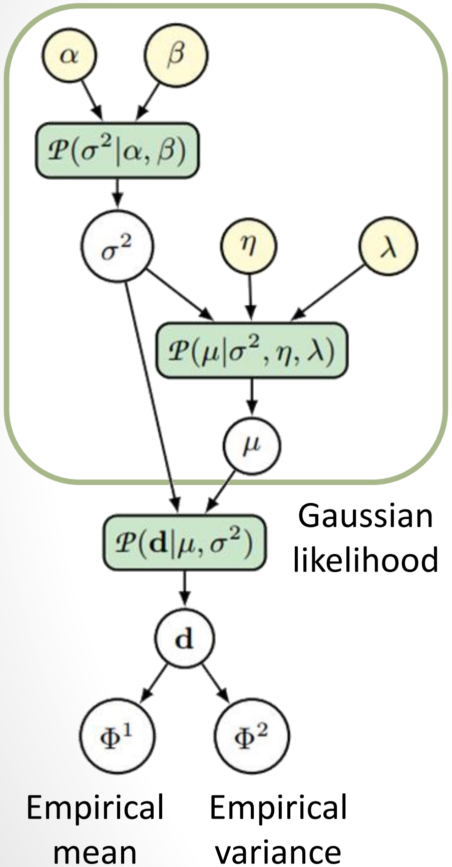
In higher dimension…



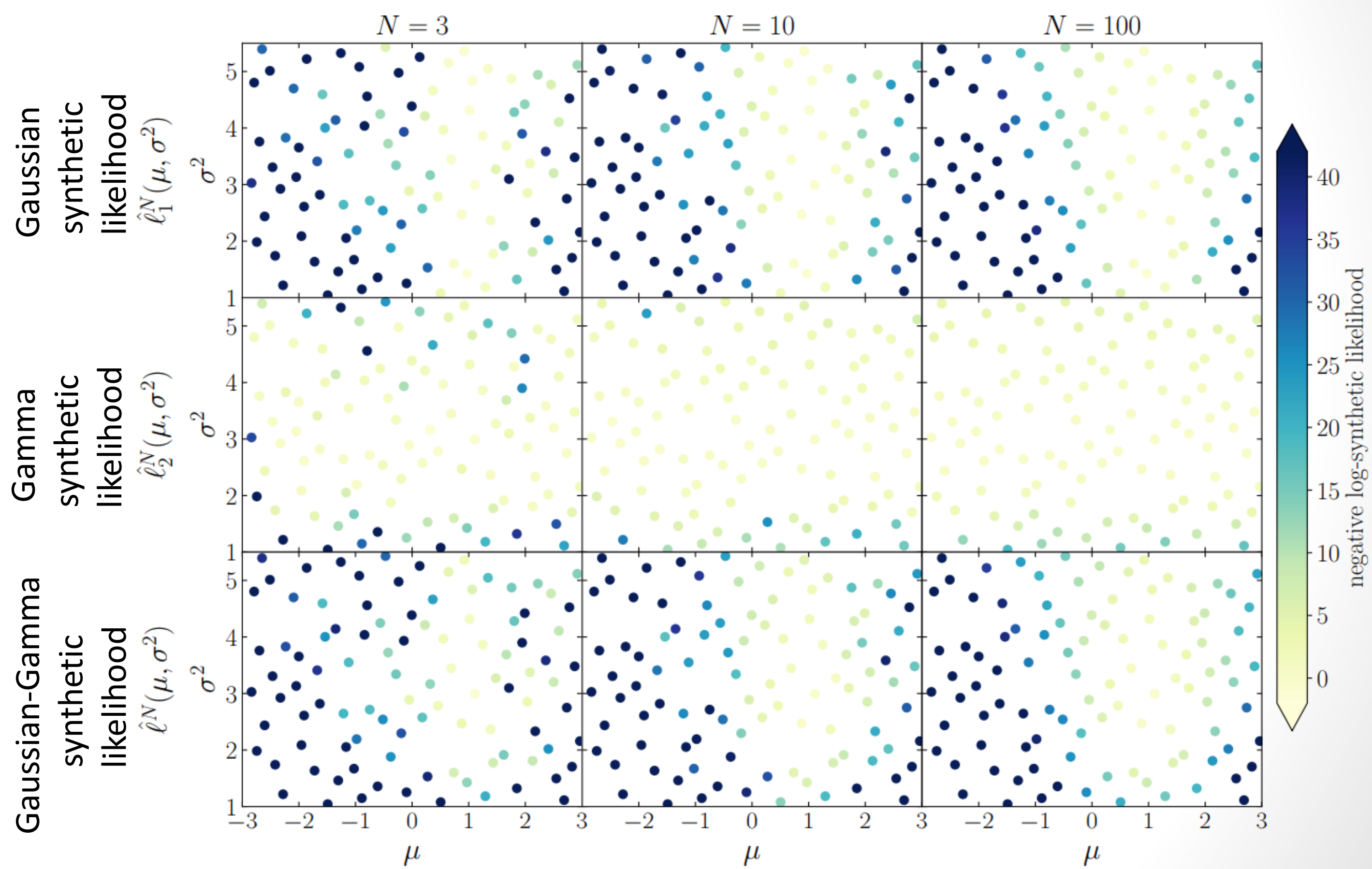
F. Nogueira, https://github.com/fmfn/BayesianOptimization

Toy example: Summarising Gaussian signals

Gaussian-inverse-Gamma conjugate prior

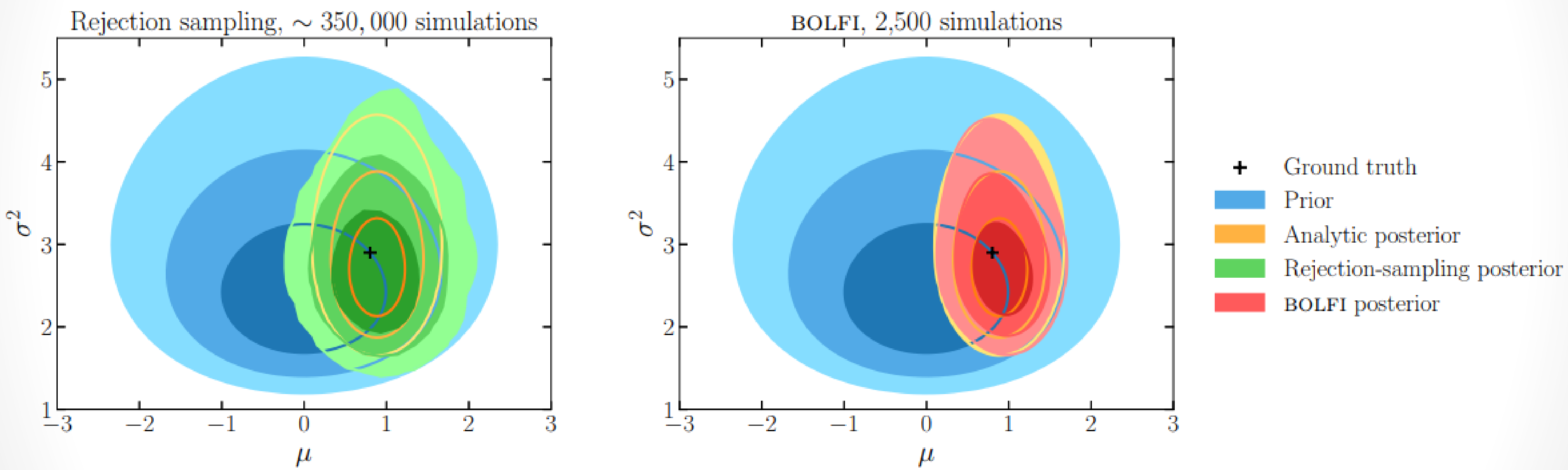


Gaussian-Gamma synthetic likelihood



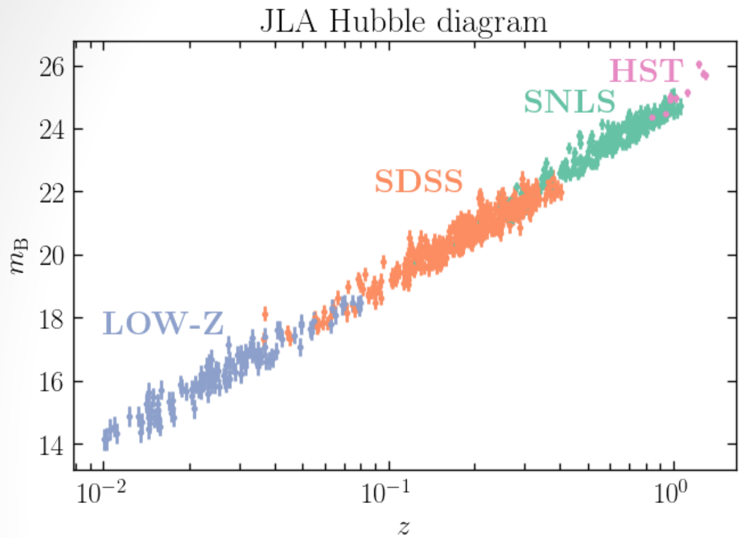
FL 2018, arXiv:1805.07152

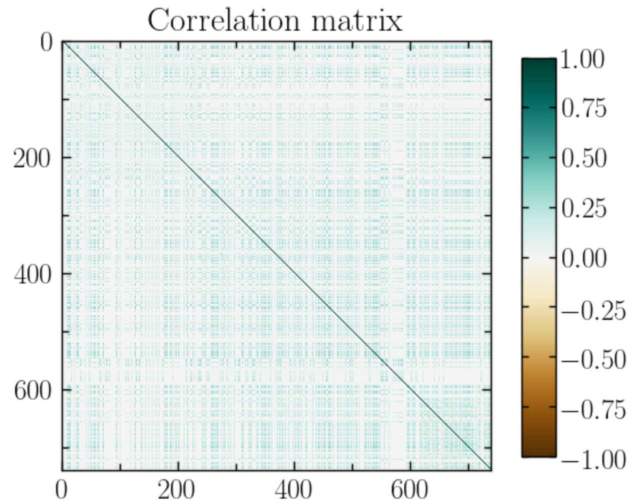
Toy example: Summarising Gaussian signals



FL 2018, arXiv:1805.07152

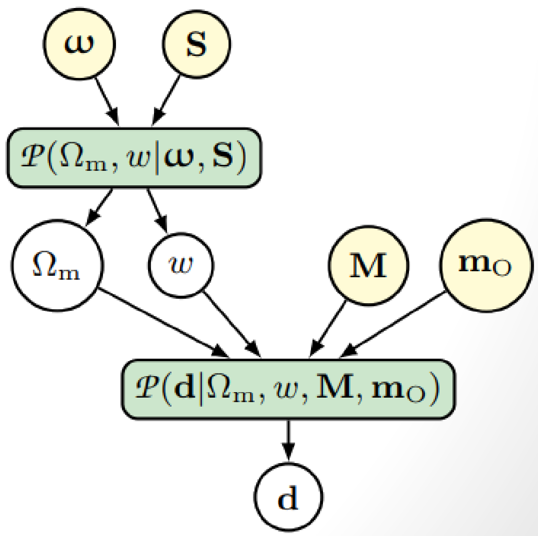
Application: Analysis of the JLA supernova sample





Betoule *et al.* 2014, arXiv:1401.4064

- 6-parameter model:
  2 cosmological parameters + 4 nuisance parameters



$$m_\mathrm{B} = 5\log_{10}\left[\frac{D_\mathrm{L}(z)}{10\ \mathrm{pc}}\right] + \widetilde{M}_\mathrm{B}(M_\mathrm{stellar}, M_\mathrm{B}, \delta M) - \alpha X_1 + \beta C$$

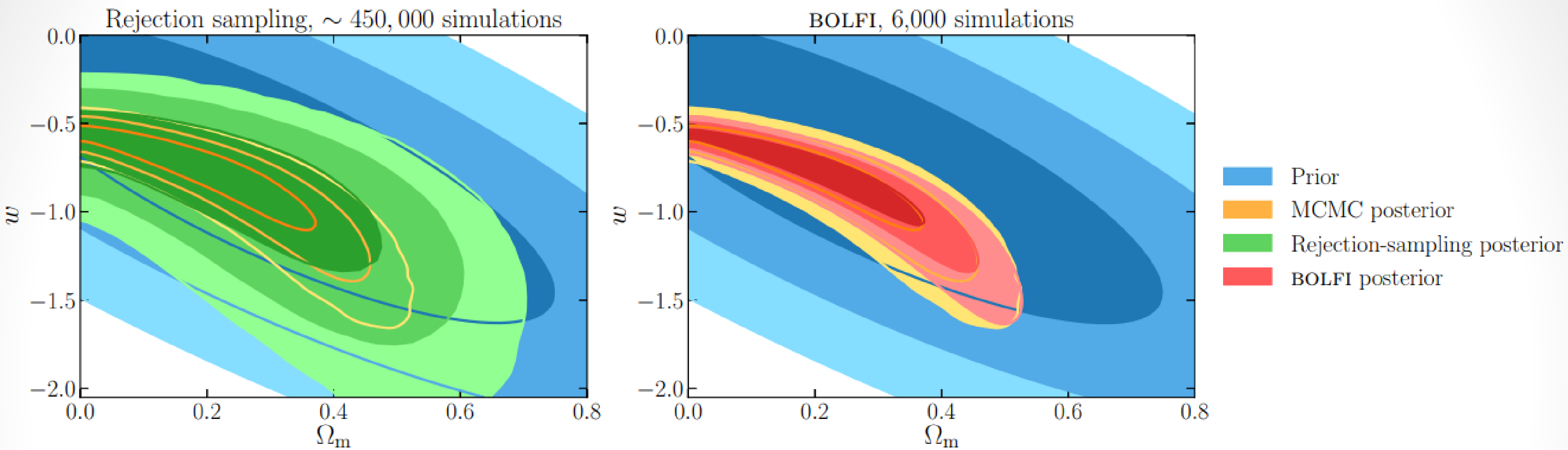
$$\widetilde{M}_\mathrm{B}(M_\mathrm{stellar}, M_\mathrm{B}, \delta M) = M_\mathrm{B} + \delta M\,\Theta\left(M_\mathrm{stellar} - 10^{10}\mathrm{M}_\odot\right)$$

$$D_\mathrm{L}(z) = \frac{(1+z)\,\mathrm{c}}{H_0}\int_0^z \frac{\mathrm{d}z'}{E(z')}$$

$$E(z) \equiv \sqrt{\Omega_\mathrm{m}(1+z)^3 + (1-\Omega_\mathrm{m})(1+z)^{3(w+1)}}$$

FL 2018, arXiv:1805.07152

Application: Analysis of the JLA supernova sample



- The **number of required simulations is reduced** by:
  - 2 orders of magnitude with respect to likelihood-free rejection sampling (for a much better approximation of the posterior)
  - 3 orders of magnitude with respect to exact Markov Chain Monte Carlo sampling

FL 2018, arXiv:1805.07152

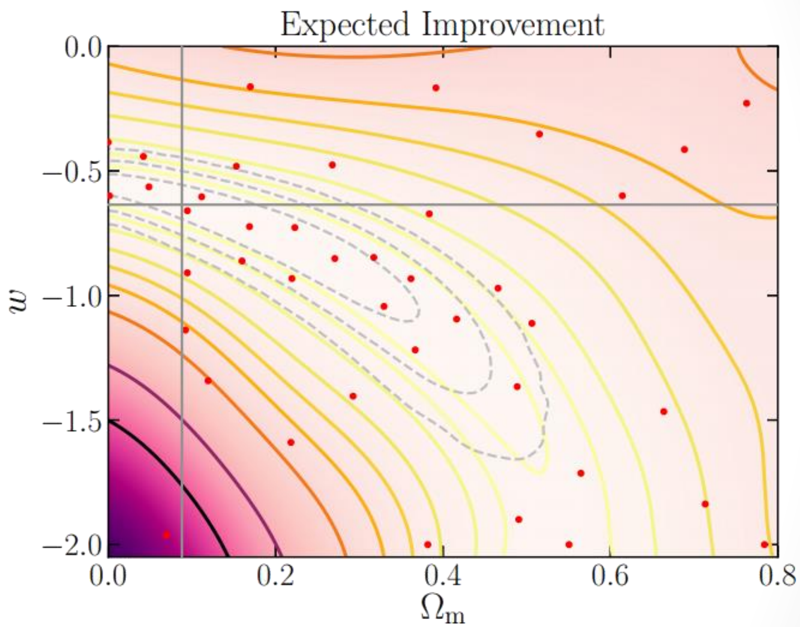
Standard acquisition functions are suboptimal

- Goal for Bayesian optimisation: find the optimum (assumed unique) of a function
- Example of acquisition function : the **Expected Improvement**

Gaussian cdf, Gaussian pdf

$$\mathrm{EI}(\boldsymbol{\theta}_\star) \equiv \underbrace{\sigma(\boldsymbol{\theta}_\star)}_{\text{Exploration}} \underbrace{\left[z\Phi(z) + \phi(z)\right]}_{\text{Exploitation}}$$

$$z \equiv \frac{\min(\mathbf{f}) - \mu(\boldsymbol{\theta}_\star)}{\sigma(\boldsymbol{\theta}_\star)}$$



- Drawbacks:
  - Do not take into account prior information
  - Local evaluation rules
  - Too greedy for ABC

Järvenpää *et al.* 2017, arXiv:1704.00520

FL 2018, arXiv:1805.07152

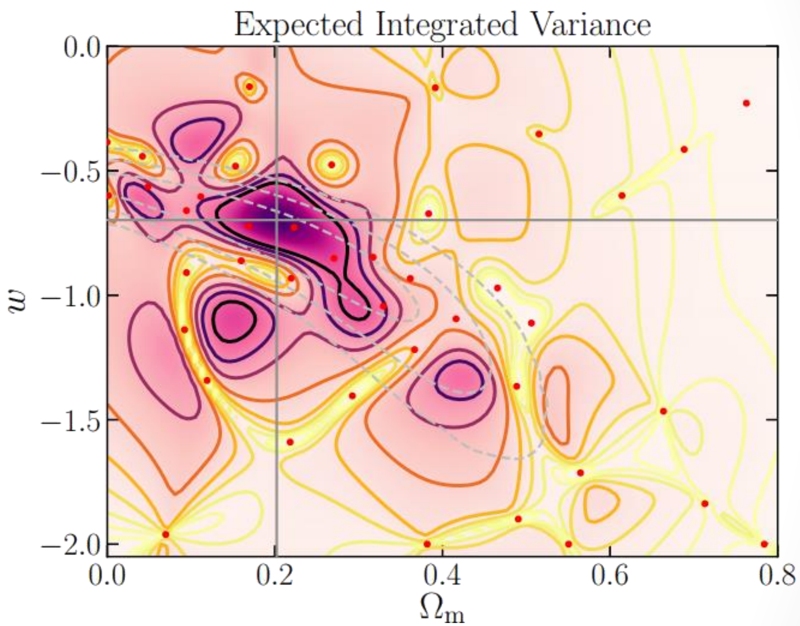
The optimal acquisition function for ABC

- Goal for ABC: minimise the expected uncertainty in the estimate of the approximate posterior over the future evaluation of the simulator
- The optimal acquisition function : the **Expected Integrated Variance**

$$\mathrm{EIV}(\boldsymbol{\theta}_\star) = \int \frac{\mathcal{P}(\boldsymbol{\theta})^2}{4} \underbrace{\exp\left[-\mu(\boldsymbol{\theta})\right]}_{\text{Exploitation}} \underbrace{\left[\sigma^2(\boldsymbol{\theta}) - \tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)\right]}_{\text{Exploration}} \, \mathrm{d}\boldsymbol{\theta}$$

Integral Prior

$$\tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) \equiv \frac{\mathrm{cov}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)}{\sigma^2(\boldsymbol{\theta}_\star)}$$



- Advantages:
  - Takes into account the prior
  - Non-local (integral over parameter space): more expensive… but much more informative
  - Exploration of the posterior tails is favoured when necessary
  - Analytic gradient

Järvenpää *et al.* 2017, arXiv:1704.00520 (expression of the EIV in the non-parametric approach)

FL 2018, arXiv:1805.07152 (expression of the EIV in the parametric approach)

Summary

Inference with generative cosmological models

Exact statistical inference Approximate physical model	?	Approximate statistical inference Exact physical model

- A likelihood-based method for principled analysis of galaxy surveys:
  Hamiltonian Monte Carlo (BORG)… (more this week)
- A likelihood-free method for models where the likelihood is intractable but simulating is possible:
  Regression of the distance + Bayesian optimisation (BOLFI)
  - The **number of required simulations** is reduced by several orders of magnitude.
  - The optimal acquisition rule for ABC can be derived: the **Expected Integrated Variance**.
  - The approach will allow to **ask targeted questions to cosmological data**, including all relevant physical and observational effects.