Benchmarking field-level inference from galaxy surveys

Hugo Simonfroy,
Ph.D. student supervised by *Arnaud de Mattia, François Lanusse*

CEA - Saclay

EDSU Tools, 2024/06/07

The universe recipe (so far)

Cosmological principle + Einstein equation + Inflation

$$\frac{H}{\boxed{H_0}} = \sqrt{\boxed{\Omega_r + \Omega_b + \Omega_c + \Omega_\kappa + \Omega_\Lambda}}$$

instantaneous expansion rate

energy content

$$\boxed{\delta_L} \sim \mathcal{G}(0, \boxed{\mathcal{P}})$$

initial field

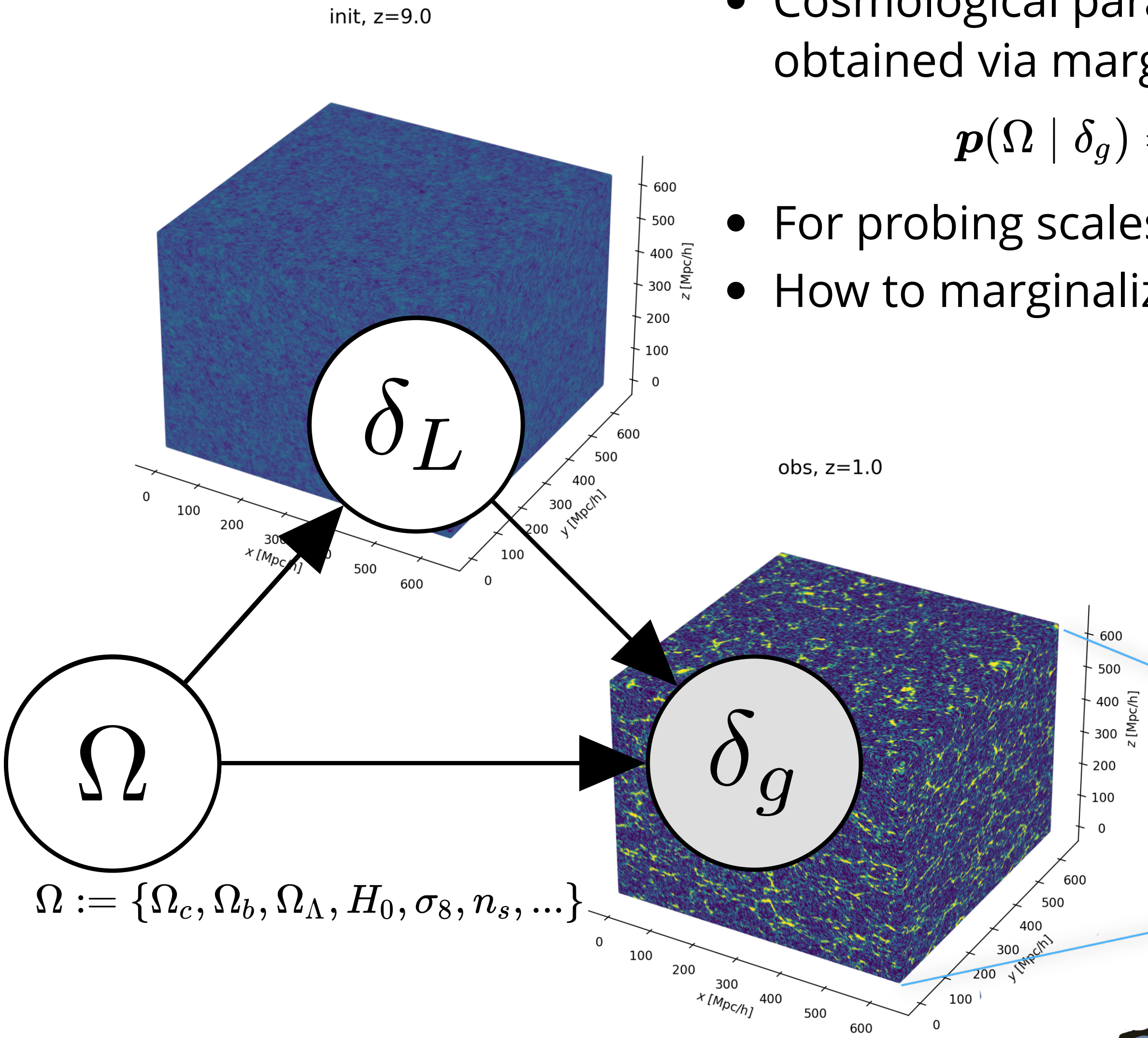
primordial power spectrum

$$\boxed{\sigma_8} := \sigma(\delta_L * \Pi_8)$$

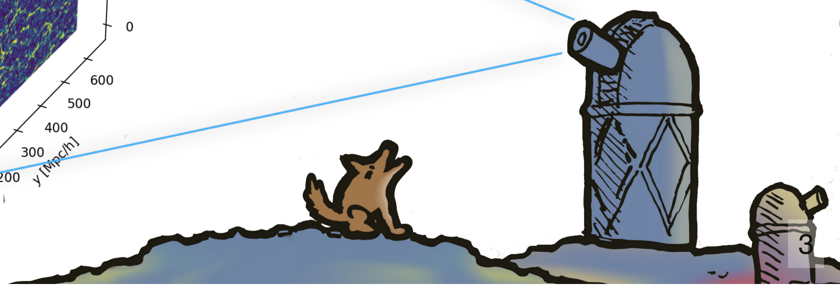
std. of fluctuations smoothed at $8\ \mathrm{Mpc/h}$

A high-dimensional inference problem

- Cosmological parameter inference obtained via marginalizing full posterior
$$\boldsymbol{p}(\Omega \mid \delta_g) = \int \boldsymbol{p}(\Omega, \delta_L \mid \delta_g)\, \mathrm{d}\delta_L$$
- For probing scales of interest $\dim(\delta_L) \simeq 1024^3$
- How to marginalize over such a random field?

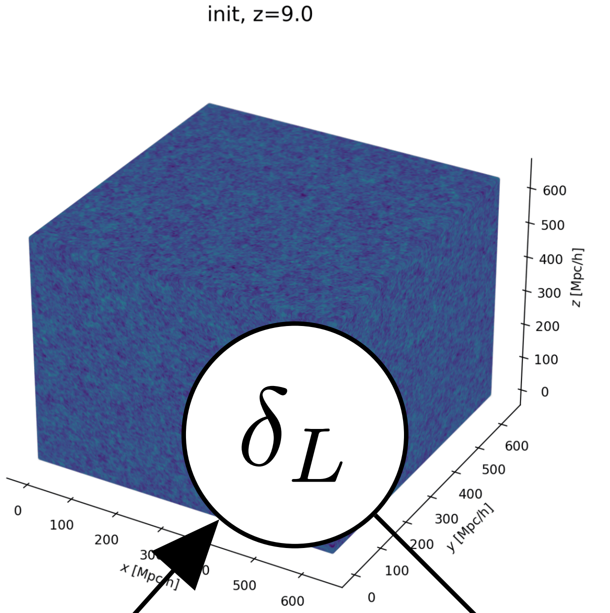


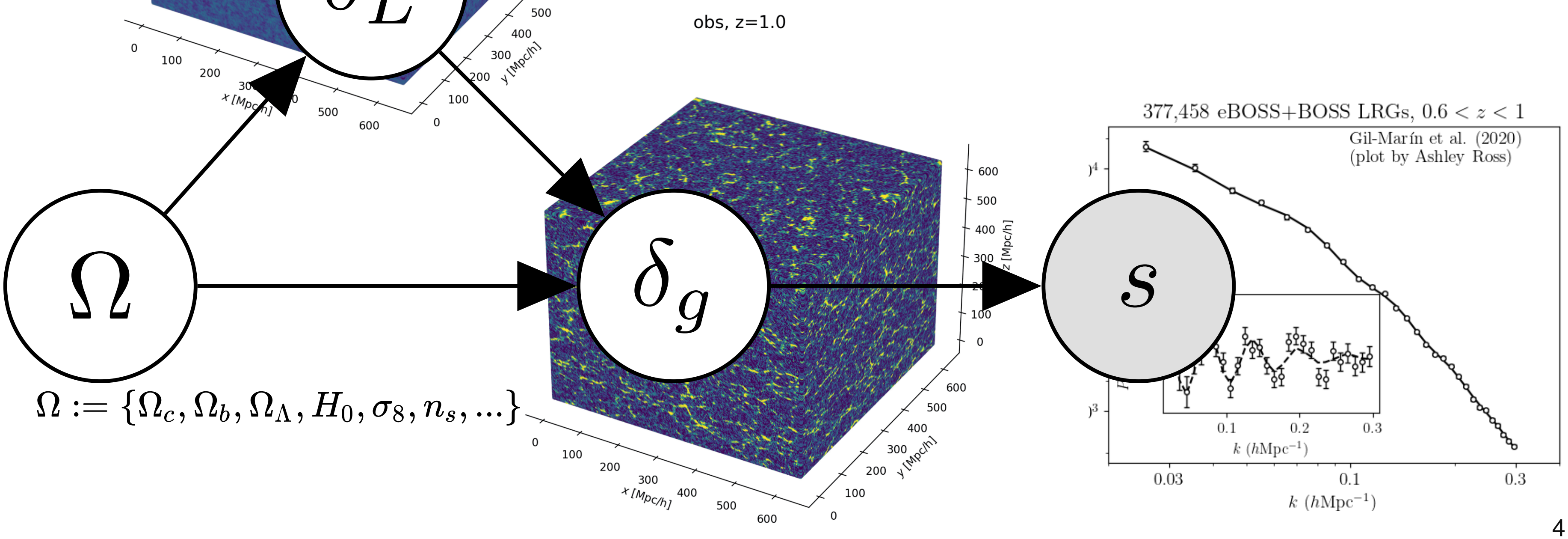
$$\Omega := \{\Omega_c, \Omega_b, \Omega_\Lambda, H_0, \sigma_8, n_s, \ldots\}$$



Use summary statistics

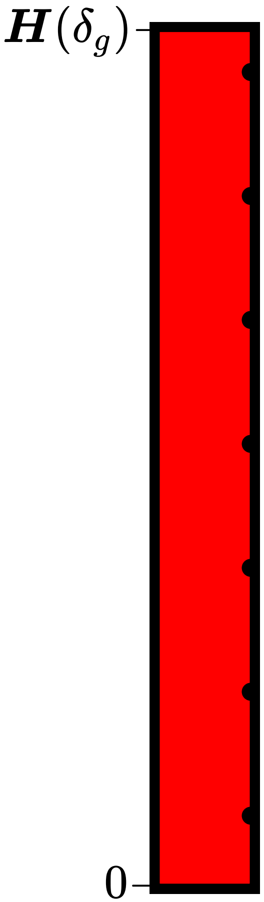
- Find some summary stat $\boldsymbol{s}$ such that $\boldsymbol{p}(\Omega \mid \boldsymbol{s})$
  - is (more) tractable
  - $\simeq \boldsymbol{p}(\Omega \mid \delta_g)$ (least information loss)





We gotta pump this information up

- At large scales, matter density field almost Gaussian so power spectrum is almost lossless compression.
- To prospect smaller non-Gaussian scales, let's use:



Field-level

CNN...

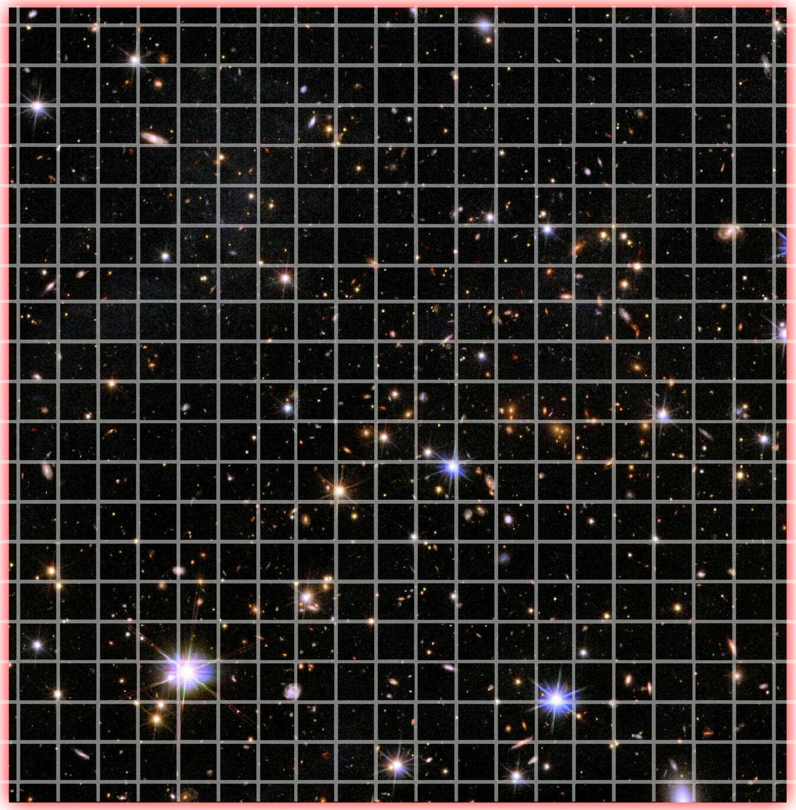
WST...

Halo, Peak, Void, Split, Hole...

3PCF, Bispectrum

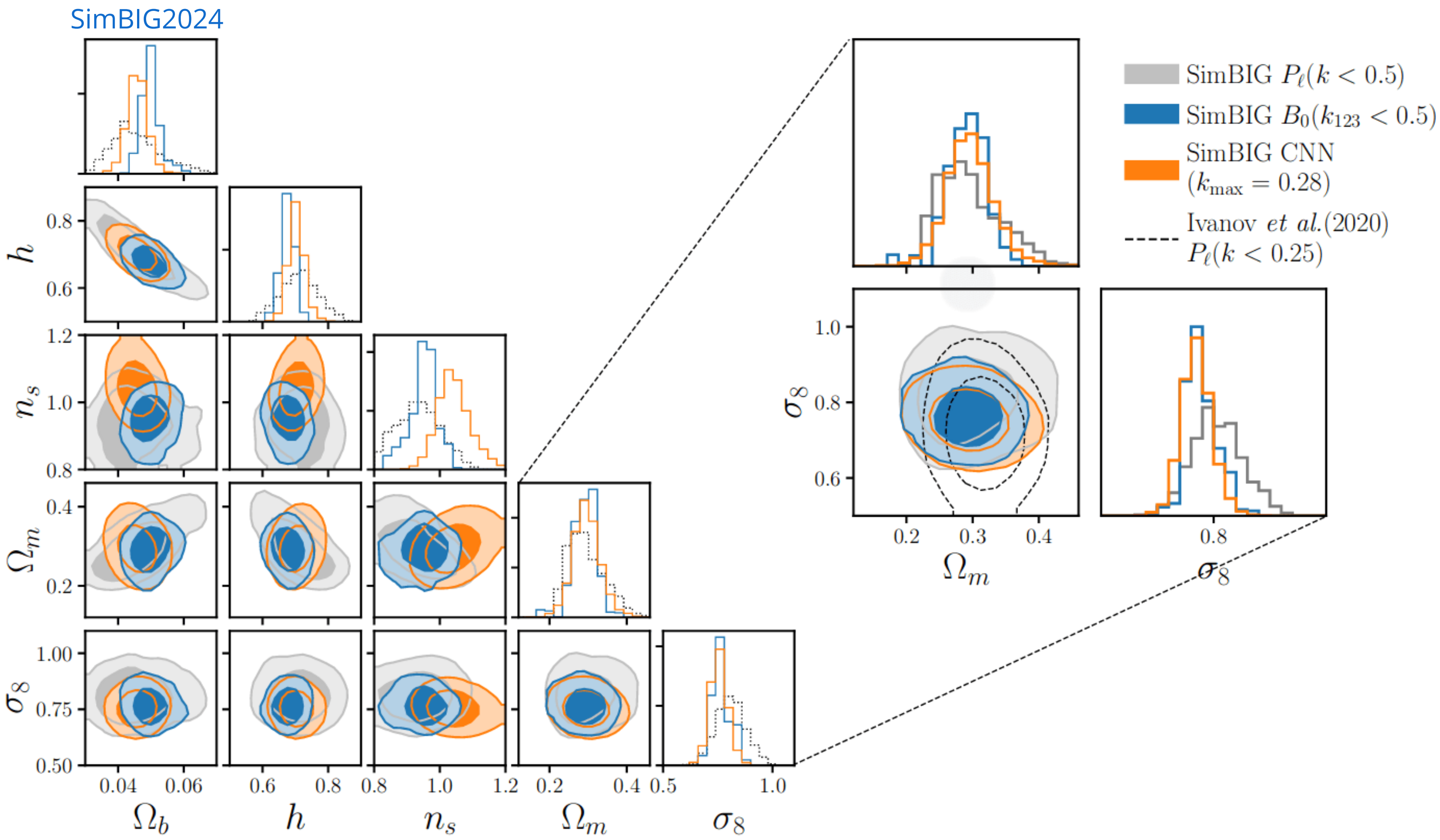
2PCF, Power spectrum

1D-PDF



Euclid's view of Perseus

More information is better, but how much better?



Let's push one step further and use the whole field:

field-level inference

Some useful programming tools

- **JAX**
  - GPU acceleration
  - Just-In-Time (JIT) compilation acceleration
  - Automatic vectorization/parallelization
  - Automatic differentiation

- **NumPyro**
  - Probabilistic Programming Language (PPL)
  - Powered by JAX
  - Integrated samplers

So JAX in practice?

- GPU accelerate

```
import jax.numpy as np
# then enjoy
```

- JIT compile

```
function = jax.jit(function)
# function is so fast now!
```

- Vectorize/Parallelize

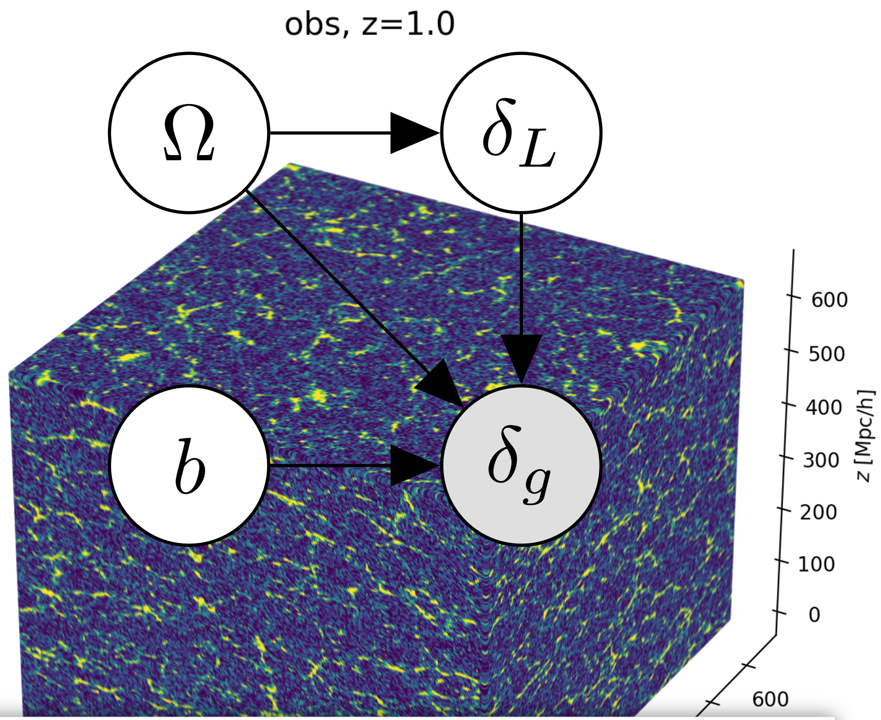
```
vfunction = jax.vmap(function)
pfunction = jax.pmap(function)
# for-loops are for-loosers
```

- Auto-diff

```
gradient = jax.grad(function)
# too bad if you love chain ruling by hand
```

Now let's build a cosmological model

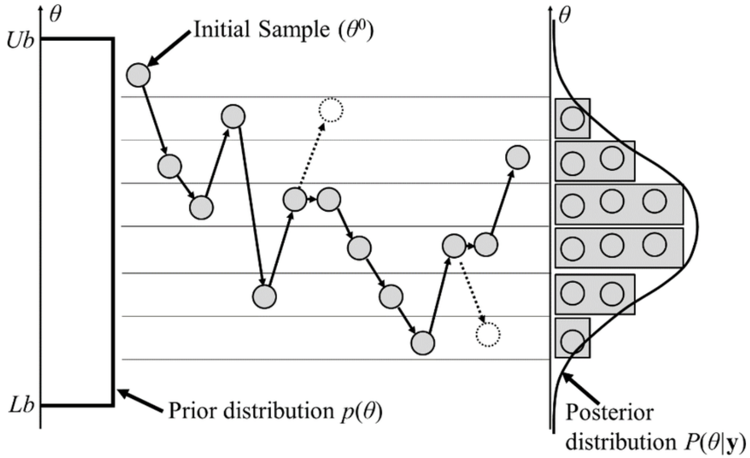
1. Prior on
   - Cosmology Ω
   - Initial field δ_L
   - Dark matter-galaxy connection (Lagrangian galaxy biases) b
2. Initialize matter particles
3. LSS formation (LPT+PM)
4. Populate matter field with galaxies
5. Galaxy peculiar velocities (RSD)
6. Observational noise



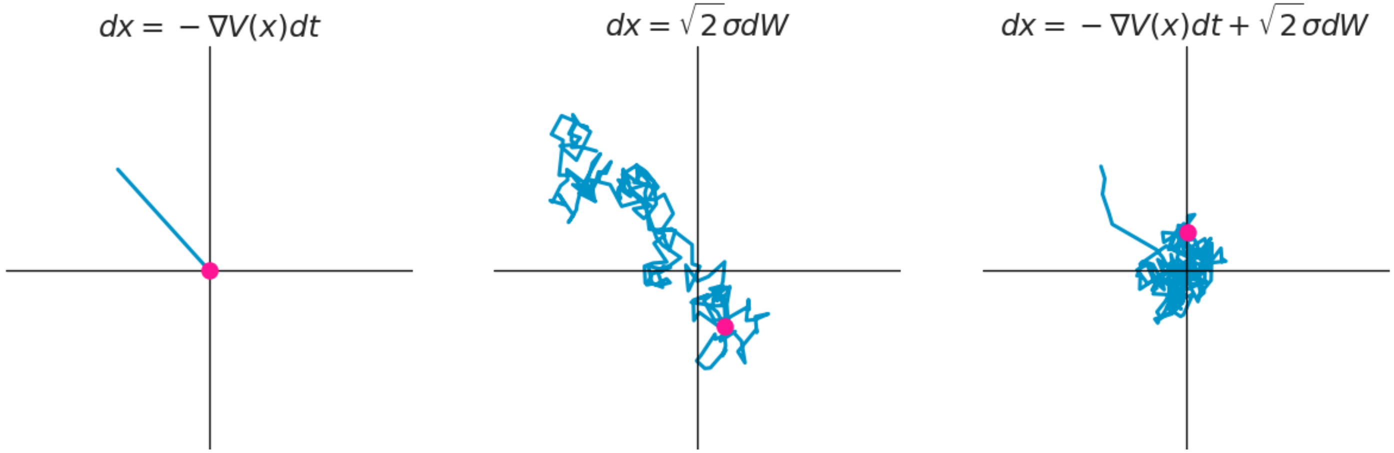
- **Fast** and **differentiable** model
- $\simeq 1024^3$ parameters is huge!
- Need inference methods that **scale to high dimensions**
- Some proposed by Lavaux+2018, Bayer+2023

Why care about differentiable model?

- **Classical MCMCs**
  - agnostic random moves
    + MH acceptance step
    = **blinded natural selection**
  - small moves yield **correlated samples**.



- **s.o.t.a. MCMCs** rely on the **gradient of the model log-proba**, to drive dynamic towards highest density regions.



gradient descent + Brownian = Langevin

posterior mode — *exploding Gaussian* — *posterior*

Hamiltonian Monte Carlo (HMC)

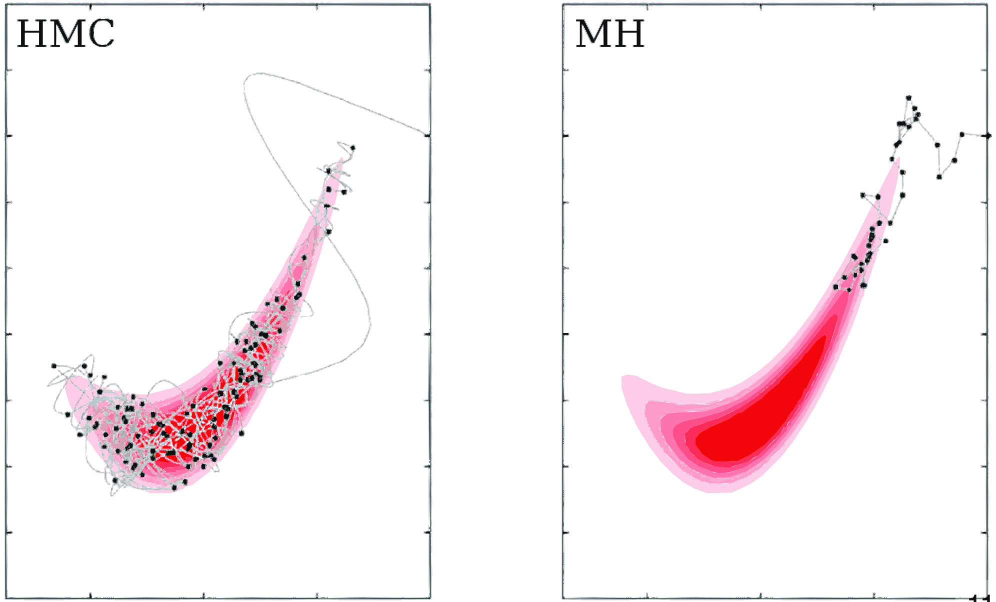
- To travel farther, add inertia.
  - sample particle at **position** q now have **momentum** p and **mass matrix** M
  - target $\boldsymbol{p}(q)$ becomes $\boldsymbol{p}(q,p) := e^{-\mathcal{H}(q,p)}$, with **Hamiltonian**

$$\mathcal{H}(q,p) := -\log \boldsymbol{p}(q) + \frac{1}{2} p^\top M^{-1} p$$

  - at each step, **resample momentum** $p \sim \mathcal{N}(0, M)$
  - let (q,p) follow the **Hamiltonian dynamic** during time length L, then arrival becomes new **MH proposal**.

Variations around HMC

- **No U-Turn Sampler (NUTS)**
  - trajectory length L auto-tuned
  - samples drawn along trajectory
- **NUTSGibbs** i.e. alternating sampling over parameter subsets.

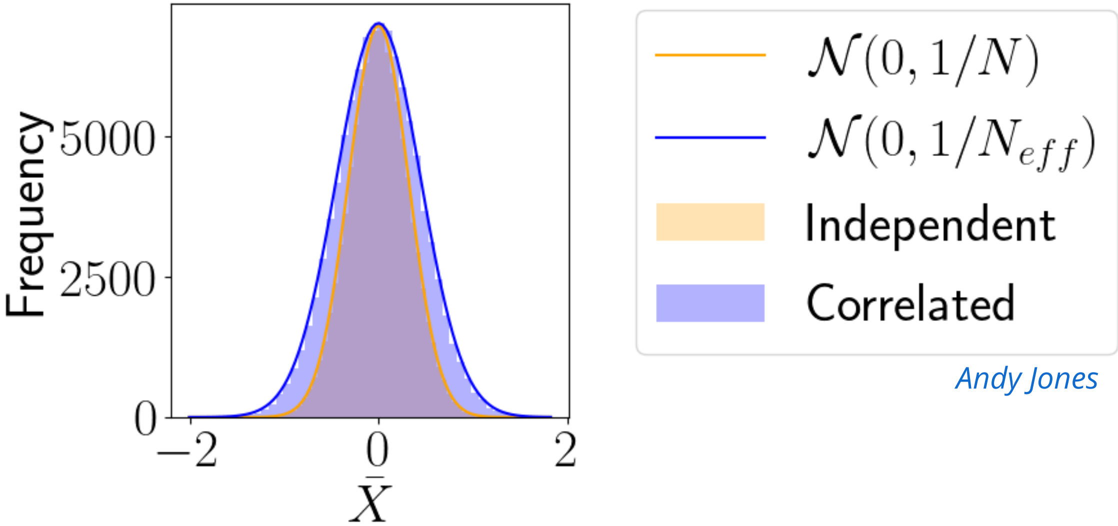


How to compare samplers?

- **Effective Sample Size (ESS)**
  - number of i.i.d. samples that yield same statistical power.
  - For sample sequence of size N and autocorrelation ρ

$$N_{\mathrm{eff}} = \frac{N}{1 + 2\sum_{t=1}^{+\infty} \rho_t}$$

  so aim for as less correlated sample as possible.

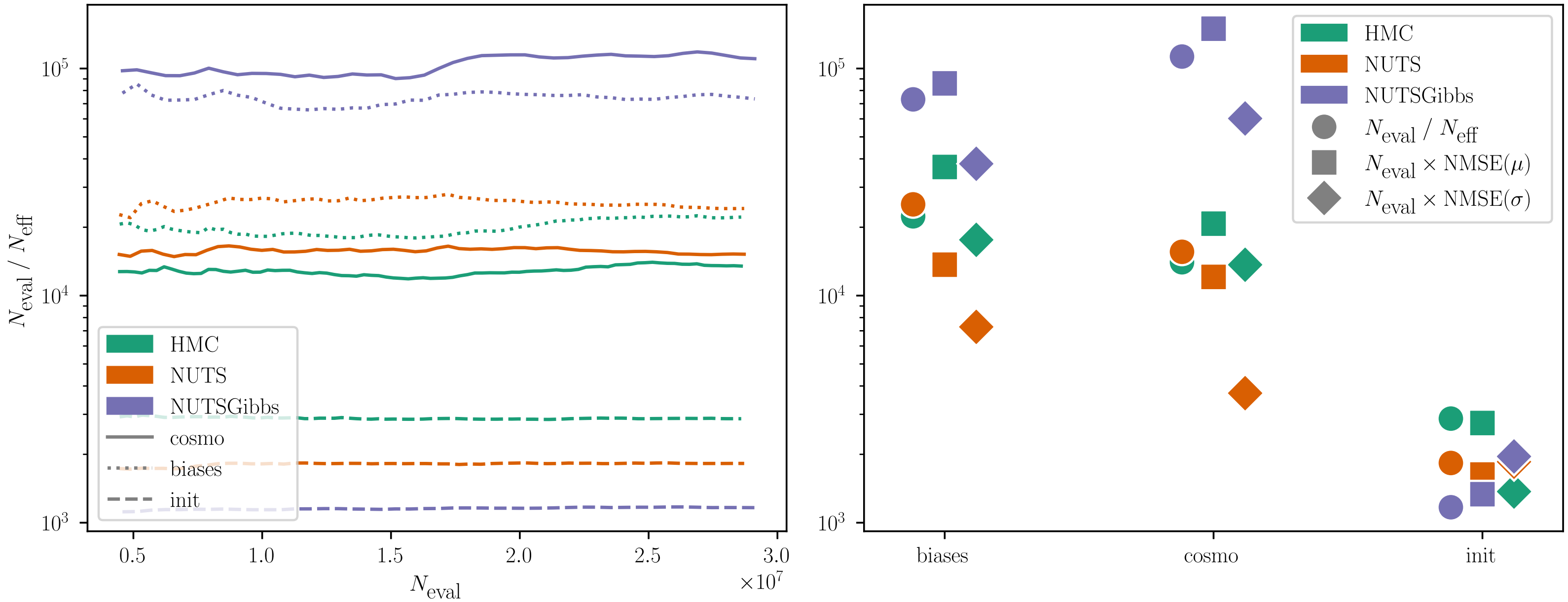


Andy Jones

- Main limiting computational factor is **model evaluation** (e.g. N-body), so characterize MCMC efficiency by $N_{\mathrm{eval}}/N_{\mathrm{eff}}$

Benchmarking

- **model setting**: 64^3 mesh, $(640\ \mathrm{Mpc/h})^3$ box, 1LPT, second order Lagrangian bias expansion, RSD and Gaussian observational noise.
- **parameter space**: initial field δ_L, cosmology $\Omega = \{\Omega_m, \sigma_8\}$, and galaxy biases $b = \{b_1, b_2, b_{s^2}, b_{\nabla^2}\}$. Total of $64^3 + 2 + 4$ parameters.
- For NUTSGibbs: split sampling between δ_L and the rest (common in lit.)



- Results suggest no particular advantage to splitting sampling between initial field and rest.

Recap...

- **Field-level inference** may be relevant to fully capture cosmological information in data.
- Leverage modern computational tools to build **fast** and **differentiable** cosmological model.
- Performing field-level inference becomes tractable. Many proposed methods in literature.
- Standardized benchmark for **comparing s.o.t.a. MCMC samplers** on field-level inference tasks, selecting proposed methods for **Stage-IV galaxy surveys**.

...and what's next

- Include more proposed samplers.
- Compare to SBI approaches.
- Move towards real applications on DESI data.