Field-level inference lecture 2: Monte Carlo techniques

Cosmology Beyond the Analytic Lamppost course (2025)

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

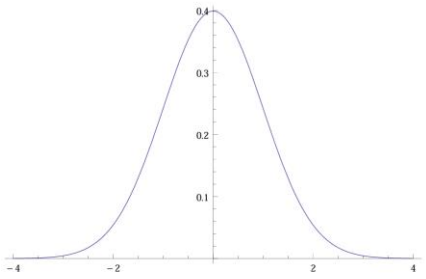
16 June 2025

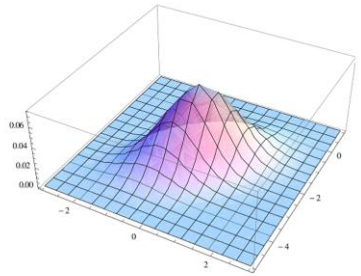
Parc de Sceaux, France

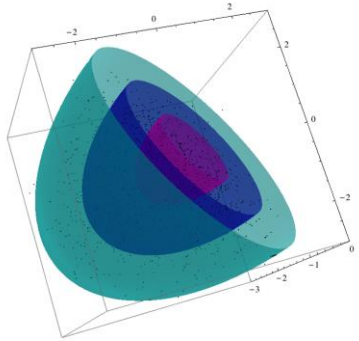
02 MONTE CARLO TECHNIQUES

Exploration of the posterior – Reporting inferences

- The output of a Bayesian analysis is a pdf: the posterior.
- The posterior cannot always be easily represented. Communication can take various forms:
  - Direct visualisation if the parameter space has sufficiently small dimension,



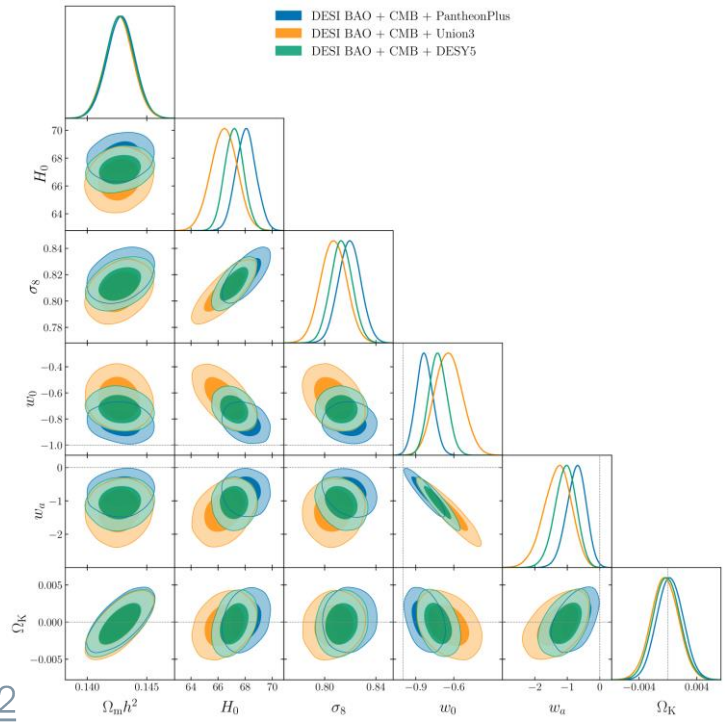




  - Credible regions (e.g. shortest interval containing 68% of the posterior probability),
    Warning: this has not the same meaning as a frequentist confidence interval (although the 2 might be formally identical)

Leclercq (2015), chap. 3

- Statistical summaries of the posterior, e.g.
  - the mean, the median, or the mode of the distribution of each parameter, marginalising over all others;
  - its standard deviation;
  - the means and covariance matrices of some groups of parameters,
- "Corner plots"

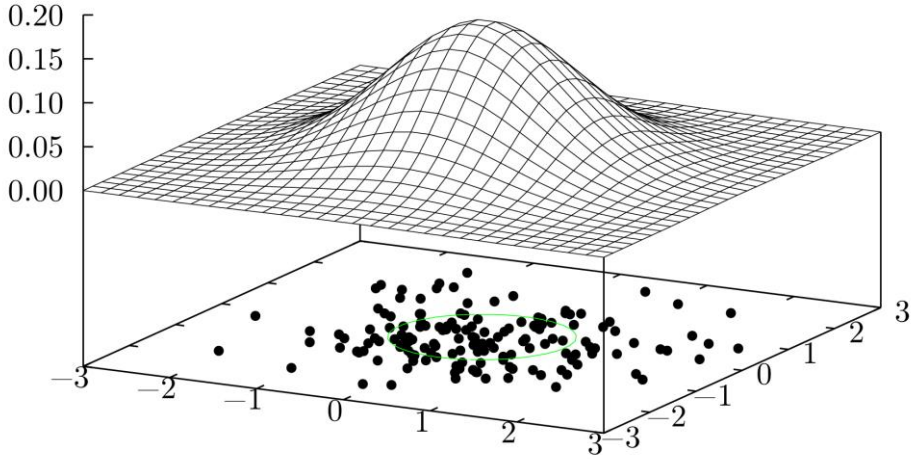


DESI 2024 VI, 2404.03002

Exploration of the posterior: challenges

- The number of grid points increases exponentially with D: direct mapping of the posterior density is practically impossible for $D \geq 5$.
- Computing statistical summaries by marginalisation means integrating out other parameters:
  - Analytical integration is rarely possible (except for GRFs)
  - Even numerical integration is basically hopeless for $D > 5$.

- In high dimension, direct evaluation of the posterior is impossible and one has to rely on a numerical approximation: representing the posterior distribution by a set of samples.



$$p(\theta|d) \approx p_N(\theta|d) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathrm{D}}(\theta - \theta_i) \quad \text{with} \quad \theta_i \curvearrowleft p(\theta|d)$$

Leclercq (2015), chap. 3

Living in a world made of samples

- Each sample is one “possible version of the truth”.
- The variation among different samples quantifies the uncertainty.
- In a world made of samples:
  - Multiplication is hard…

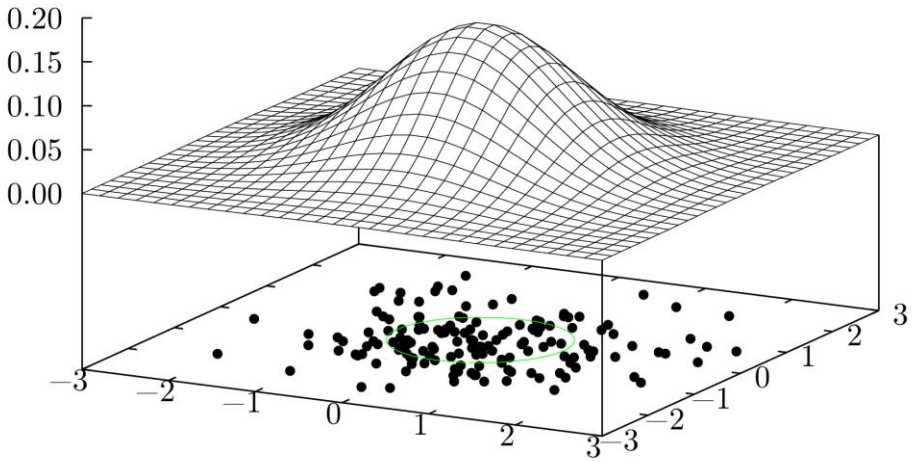
$$\left[\sum_{i=1}^{N_\mathrm{A}} \delta_\mathrm{D}(\theta-\theta_i)\right] \times \left[\sum_{j=1}^{N_\mathrm{B}} \delta_\mathrm{D}(\theta-\theta_j)\right]$$ gives (almost certainly) zero!

  - But integration is easy!

$$\langle\theta\rangle_{p(\theta|d)} = \int \theta\, p(\theta|d)\,\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N}\theta_i \quad \text{with} \quad \theta_i \curvearrowleft p(\theta|d)$$

More generally:

$$\langle f(\theta)\rangle_{p(\theta|d)} = \int f(\theta)\, p(\theta|d)\,\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \quad \text{with} \quad \theta_i \curvearrowleft p(\theta|d)$$



- In particular, marginalisation is trivial:

$$\langle\theta\rangle_{p(\theta,\varphi|d)} = \iint \theta\, p(\theta,\varphi|d)\,\mathrm{d}\theta\mathrm{d}\varphi = \int \theta\, p(\theta|d)\,\mathrm{d}\theta$$
$$\approx \frac{1}{N}\sum_{i=1}^{N}\theta_i \quad \text{with} \quad (\theta_i,\varphi_i) \curvearrowleft p(\theta,\varphi|d)$$

  Marginalisation is achieved simply by “not looking” at the values of φ_i in the samples (θ_i,φ_i)

- This procedure works in arbitrarily high dimension.

SUFFICIENT SUMMARY STATISTICS AND RAO-BLACKWELL ESTIMATORS

Sufficient summary statistics

- A summary statistic $t = t(d)$ of the dataset d is sufficient for underlying parameter θ if and only if (equivalent definitions):
  1. It contains all of the information that the dataset d provides about the model parameters θ:
     $$p(\theta|t) = p(\theta|d)$$
  2. The conditional distribution of the data d does not further depend on the parameters θ once the statistic t is known:
     $$p(d|\theta, t) = p(d|t)$$
- Proof: using twice the product rule: $p(d, \theta|t) = p(d|\theta, t)p(\theta|t) = p(\theta|d, t)p(d|t)$
  1. Assuming 2., $p(d|t)p(\theta|t) = p(\theta|d, t)p(d|t)$, so we have $p(\theta|t) = p(\theta|d, t) = p(\theta|d)$ (unless $p(d|t) = 0$)
  2. Assuming 1., $p(d|\theta, t)p(\theta|d) = p(\theta|d, t)p(d|t) = p(\theta|d)p(d|t)$, so we have $p(d|\theta, t) = p(d|t)$ (unless $p(\theta|d) = 0$)
- There exists a third equivalent definition, using the notion of mutual information.

Sufficient summary statistics

- Sufficiency is closely related to the concepts of:
  - Ancillarity: an ancillary statistic contains no information about the model parameters (it has the same distribution regardless of the value of the parameters)
  - Completeness: a complete statistic only contains information about the parameters and no ancillary information
- Neyman-Fisher factorisation theorem: $t(d)$ is sufficient for θ if and only if nonnegative functions g and h can be found such that $p(d|\theta) = g(t(d), \theta)h(d)$ where the function $h(d)$ does not depend on the parameters.

Rao-Blackwell estimators

- Frequentist version: if $g(d)$ is any kind of estimator of a parameter θ, then the conditional expectation of $g(d)$ given $t(d)$, where $t(d)$ is a sufficient statistic, is typically a "better" estimator of θ, and is never "worse".
- More precisely, define:
  - $\delta(d)$ an "original estimator" of θ
  - $\delta_1(d) = E[\delta(d) \mid t(d)]$ the Rao-Blackwell "improved estimator" (it shall be observable, i.e. not depend on θ)
- Then: the mean squared error of the Rao-Blackwell estimator does not exceed that of the original estimator, i.e. $E[(\delta_1(d) - \theta)^2] \leq E[(\delta(d) - \theta)^2]$.
- Demonstration: the mean square error of the Rao-Blackwell estimator has the following decomposition:

$$E[(\delta_1(d) - \theta)^2] = E[(\delta(d) - \theta)^2] - E[\mathrm{Var}(\delta(d) \mid t(d))],$$
$$\text{and } E[\mathrm{Var}(\delta(d) \mid t(d))] \geq 0.$$

- Improving an estimator based on this procedure is called "Rao-Blackwellisation".

Rao-Blackwell estimators

- Bayesian version: suppose we have data d, an underlying signal s, and a property x which does not depend on the data when the signal is known, i.e. $p(x|s,d) = p(x|s)$ (s is a sufficient summary statistic of d).
- Further, suppose we have a way to generate samples of the joint posterior, $(x_i, s_i) \curvearrowleft p(x,s|d)$.
- The naïve estimator of the marginal pdf $p(x|d)$ is $p(x|d) \approx \frac{1}{N}\sum_{i=1}^{N} x_i \equiv \hat{x}$
- But we also have

$$p(x|d) = \int p(x,s|d)\,\mathrm{d}s = \int p(x|s,d)p(s|d)\,\mathrm{d}s = \int p(x|s)p(s|d)\,\mathrm{d}s$$
$$\approx \frac{1}{N}\sum_{i=1}^{N} p(x|s_i) \equiv \hat{x}_1 \quad \text{with} \quad s_i \curvearrowleft p(s|d) \text{ : the Rao-Blackwell estimator of } p(x|d)$$

- Then one can show that $\hat{x}_1$ is a "better" estimator of $p(x|d)$ than $\hat{x}$, in any reasonable sense.

Robert & Roberts, 2101.01011

Rao-Blackwell estimators

- In particular if $p(x|s_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}\right]$
- The original estimator $p(x|d) \approx \frac{1}{N}\sum_{i=1}^{N} x_i \equiv \hat{x}$ is the sum of N independent Gaussian variables. It is a Gaussian with

  mean: $E(\hat{x}) = \frac{1}{N}\sum_{i=1}^{N}\mu_i$ variance: $E\left[(\hat{x} - E(\hat{x}))^2\right] = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2$

- The Rao-Blackwell estimator is $p(x|d) \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}\right] \equiv \hat{x}_1$. It is non-Gaussian but has:

  mean: $E(\hat{x}_1) = \frac{1}{N}\sum_{i=1}^{N}\mu_i = E(\hat{x})$

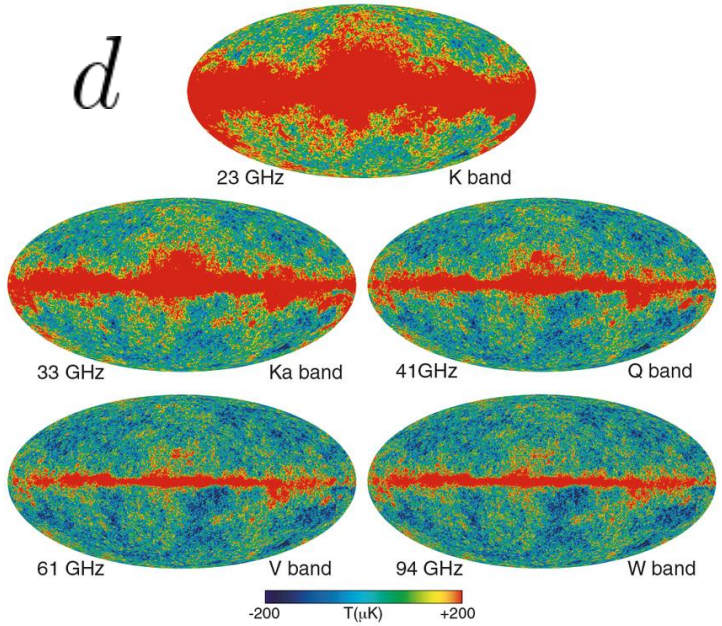
  variance: $E\left[(\hat{x}_1 - E(\hat{x}_1))^2\right] \approx \frac{1}{N}\sum_{i=1}^{N}(\mu_i^2 + \sigma_i^2) - \left(\frac{1}{N}\sum_{i=1}^{N}\mu_i\right)^2$

  $\leq \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 = E\left[(\hat{x} - E(\hat{x}))^2\right]$

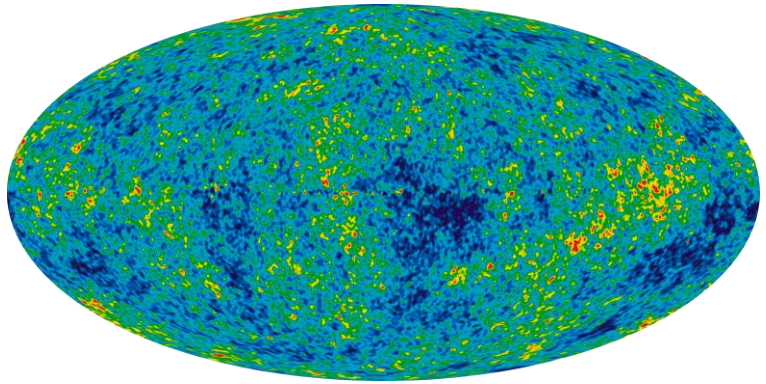
Rao-Blackwell estimators: example

- For analysis of the cosmic microwave background, we want to know the distribution of the power spectrum coefficients C_ℓ given the data d (frequency maps). The signal s is the cosmic microwave background map.

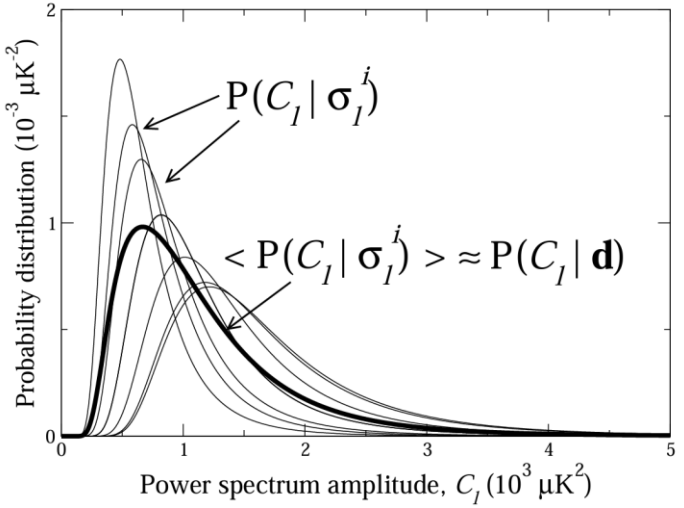
d



s



C_ℓ



Assumption: $p(C_\ell|s) = p(C_\ell|\sigma_\ell)$ where $\sigma_\ell = \sum_m s_{\ell m}$

$$p(C_\ell|d) = \int p(C_\ell, s|d)\,\mathrm{d}s = \int p(C_\ell|s)p(s|d)\,\mathrm{d}s = \int p(C_\ell|\sigma_\ell)p(\sigma_\ell|d)\,\mathrm{d}\sigma_\ell \approx \frac{1}{N}\sum_{i=1}^{N} p(C_\ell|\sigma_\ell^i)$$

Wandelt, Larson & Lakshminarayanan (2003), astro-ph/0310080

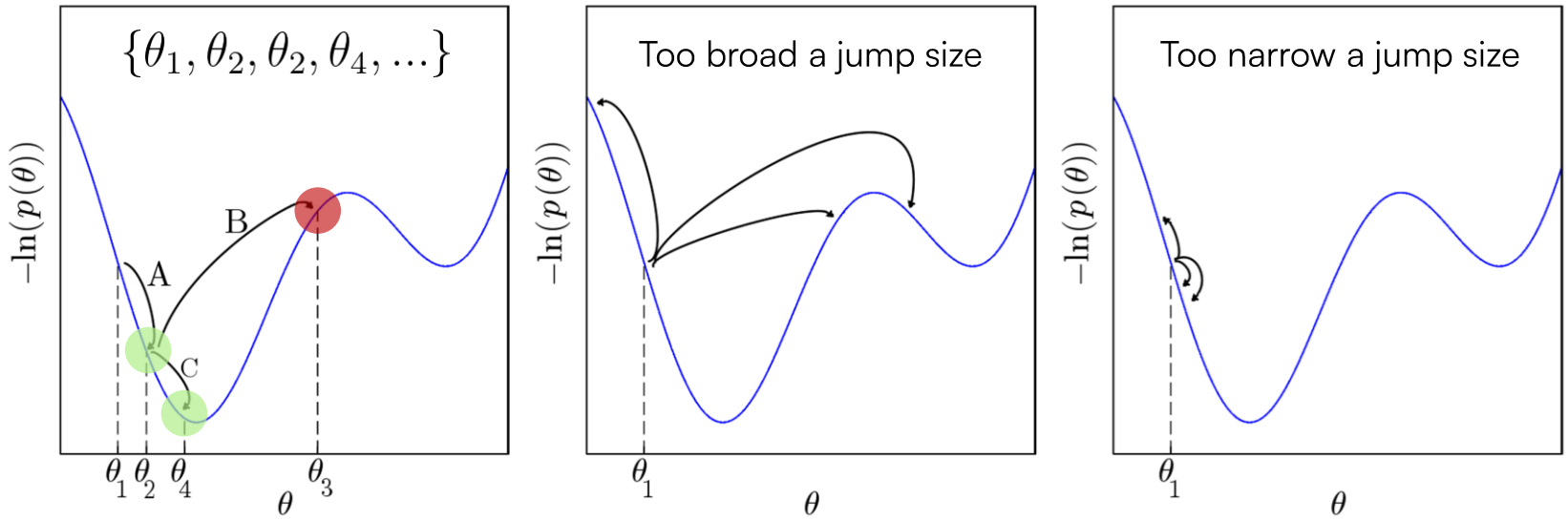
MARKOV CHAIN MONTE CARLO

Markov Chain Monte Carlo (MCMC)

- The Markov property:
  - The useful information for predicting the future is entirely contained in the present state of the process and does not depend on past states (the system has no “memory”).
  - Mathematically: The conditional probability distribution of future states, given past states and the present state, depends only on the present state and not on past states:

$$p(\boldsymbol{x}_{n+1} | \{\boldsymbol{x}_i\}_{1 \leq i \leq n}) = p(\boldsymbol{x}_{n+1} | \boldsymbol{x}_n)$$

- Constructing Markov Chains: accepting and rejecting moves



Leclercq (2015), chap. 3

Markov Chain Monte Carlo (MCMC): the Metropolis-Hastings algorithm

- Metropolis-Hastings algorithm:

begin
  initialise $\boldsymbol{x}_{(0)}$;
  for $i = 1$ *to* n **do**
    $\boldsymbol{x}^* \curvearrowleft q(\boldsymbol{x}^*|\boldsymbol{x})$ (proposal distribution);
    $\alpha \curvearrowleft \mathcal{U}(0,1)$ (uniform distribution);
    if $\alpha < \min\left[1, r(\boldsymbol{x}, \boldsymbol{x}^*)\right]$ **then**
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}^*$;
    else
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}_{(i-1)}$;
    end
  end
  return $(\boldsymbol{x}_{(0)}, \ldots, \boldsymbol{x}_{(n)})$;
end

General case*: $r(\boldsymbol{x}, \boldsymbol{x}^*) \equiv \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{q(\boldsymbol{x}|\boldsymbol{x}^*)}{q(\boldsymbol{x}^*|\boldsymbol{x})}$ (Hastings ratio)

Particular case: $r(\boldsymbol{x}, \boldsymbol{x}^*) = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})}$ (Metropolis update)

for a symmetric proposal pdf: $q(\boldsymbol{x}^*|\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{x}^*)$

* This is only a very simplified and practical guide to MCMC. A more extensive treatment requires introducing the notions *stationarity* and *global/detailed balance*.

- It is possible to prove that the chain has the target distribution as its stationary distribution, i.e. elements of the chain (asymptotically) become correlated samples of $p(\boldsymbol{x})$.
- A frustrating property: the optimal proposal distribution $q(\boldsymbol{x}^*|\boldsymbol{x})$ to sample from the target distribution $p(\boldsymbol{x})$ is... the target distribution $p(\boldsymbol{x})$ itself!

Metropolis-Hasting algorithm: implementation

- Metropolis-Hastings algorithm:

begin
  initialise $\boldsymbol{x}_{(0)}$;
  for $i = 1$ *to* n **do**
    $\boldsymbol{x}^* \curvearrowleft q(\boldsymbol{x}^*|\boldsymbol{x})$ (proposal distribution);
    $\alpha \curvearrowleft \mathcal{U}(0,1)$ (uniform distribution);
    if $\alpha < \min[1, r(\boldsymbol{x}, \boldsymbol{x}^*)]$ **then**
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}^*$;
    else
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}_{(i-1)}$;
    end
  end
  return $(\boldsymbol{x}_{(0)}, \ldots, \boldsymbol{x}_{(n)})$;
end

```python
def MH_sampler(Ntries,theta_start,Ntrials,Nsuccesses,lh,prior,proposal_sigma):
    Naccepted=0
    samples=np.zeros(Ntries+1)
    samples[0]=theta_start
    theta=theta_start
    for i in range(Ntries):
        theta_p = theta + proposal_pdf(proposal_sigma).rvs()
        # the Gaussian proposal pdf satisfies the detailed balance equation, so the
        # acceptance ratio simplifies to the Metropolis ratio
        # for numerical reasons, it is better to use the log of the ratio
        log_a = min(0, target_logpdf(theta_p,Ntrials,Nsuccesses,lh,prior) - target_logpdf(theta,Ntrials,Nsuccesses,lh,prior))
        u = np.random.uniform()
        if np.log(u) < log_a:
            Naccepted+=1
            theta=theta_p
        samples[i+1] = theta
    return Naccepted, samples
```

General case: $r(\boldsymbol{x}, \boldsymbol{x}^*) \equiv \dfrac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \dfrac{q(\boldsymbol{x}|\boldsymbol{x}^*)}{q(\boldsymbol{x}^*|\boldsymbol{x})}$ (Hastings ratio)

Particular case: $r(\boldsymbol{x}, \boldsymbol{x}^*) = \dfrac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})}$ (Metropolis update) for a symmetric proposal pdf: $q(\boldsymbol{x}^*|\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{x}^*)$

A toy Bayesian problem

- Considered problem: a Bernoulli experiment (N_{trials} independent trials each with a probability of success θ)
- The likelihood for this problem is a binomial distribution:

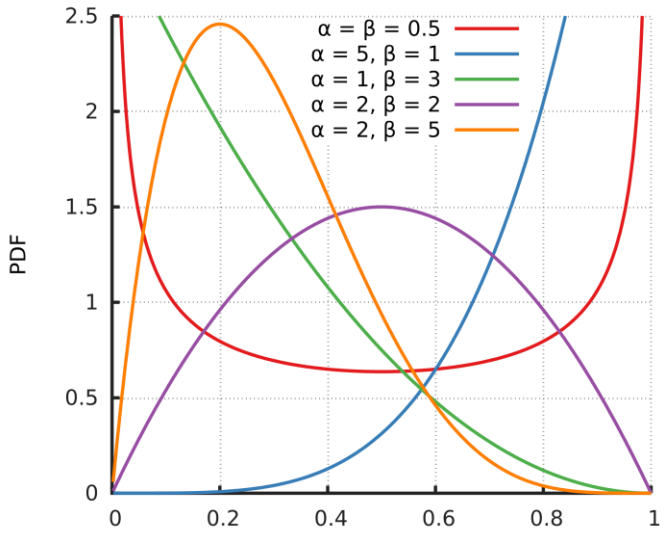
$$p(N_{\text{successes}}, N_{\text{trials}}, \theta) = \binom{N_{\text{trials}}}{N_{\text{successes}}} \theta^{N_{\text{successes}}} (1-\theta)^{N_{\text{trials}} - N_{\text{successes}}}$$

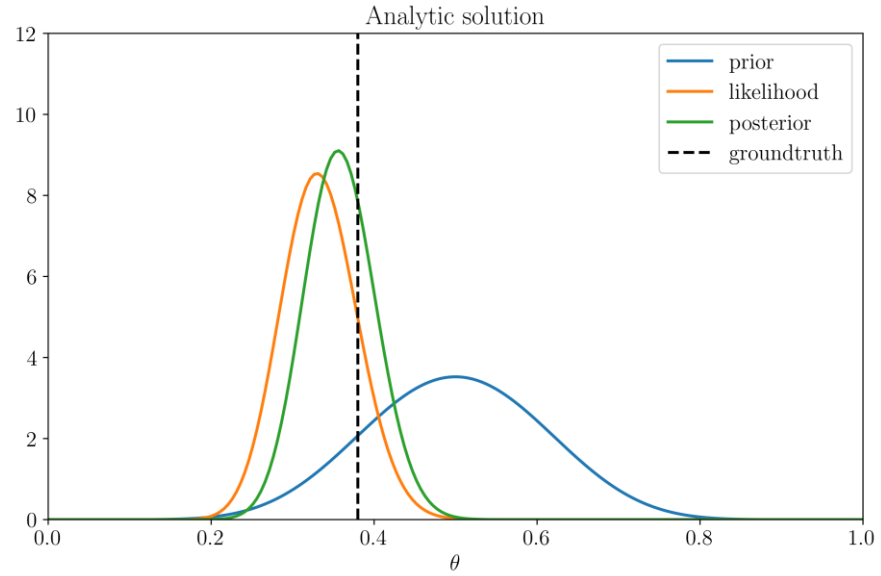
- Analytic result: the beta distribution gives a family of conjugate priors, meaning that if the prior is $\mathcal{B}(\alpha, \beta)$, then the posterior is $\mathcal{B}(\alpha', \beta')$ with $\alpha' = \alpha + N_{\text{sucesses}}$

$$\beta' = \beta + N_{\text{trials}} - N_{\text{sucesses}}$$

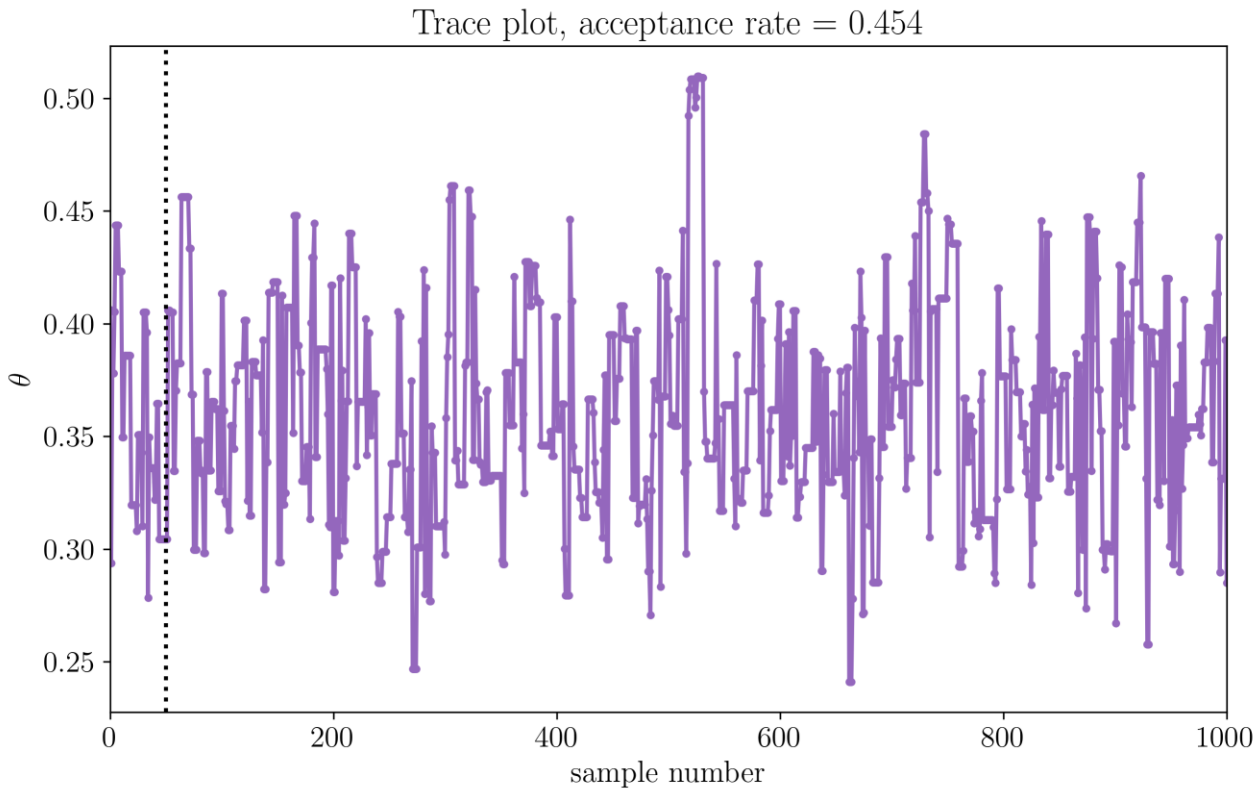
$$\mathcal{B}(\alpha, \beta)(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

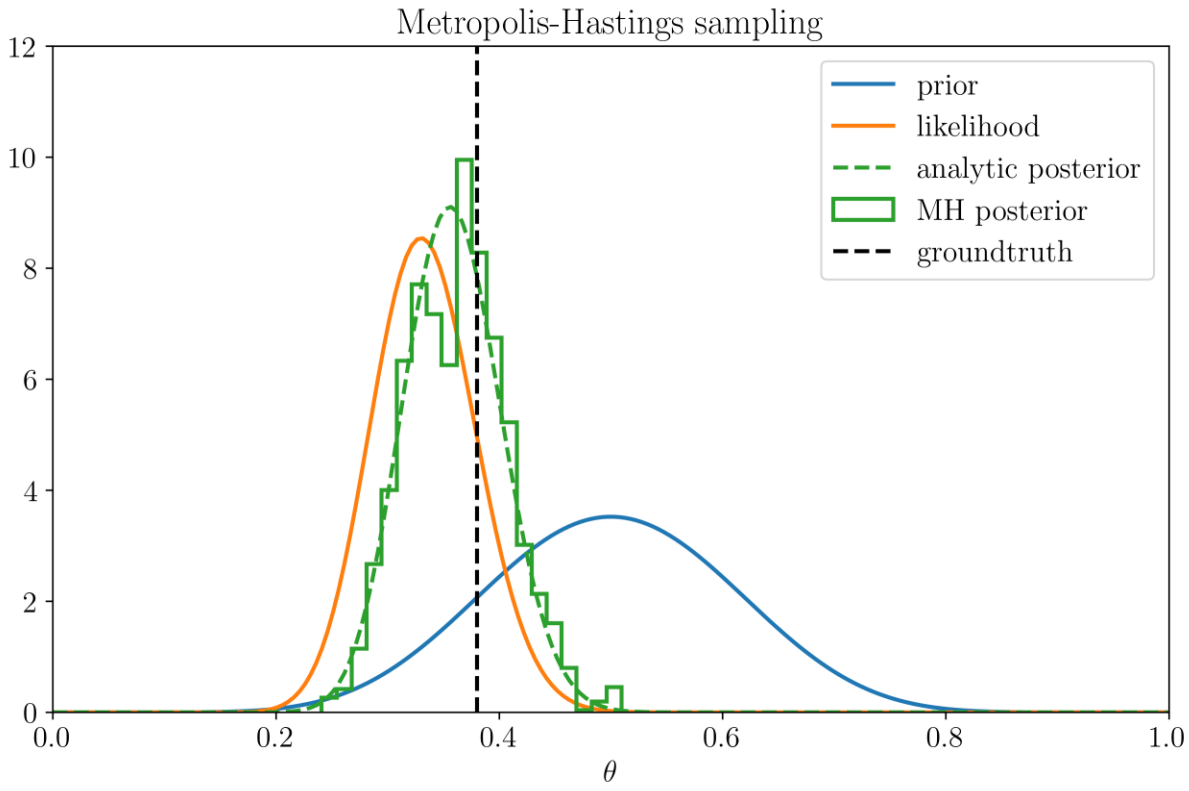
with $\mathrm{B}(\alpha, \beta) \equiv \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$





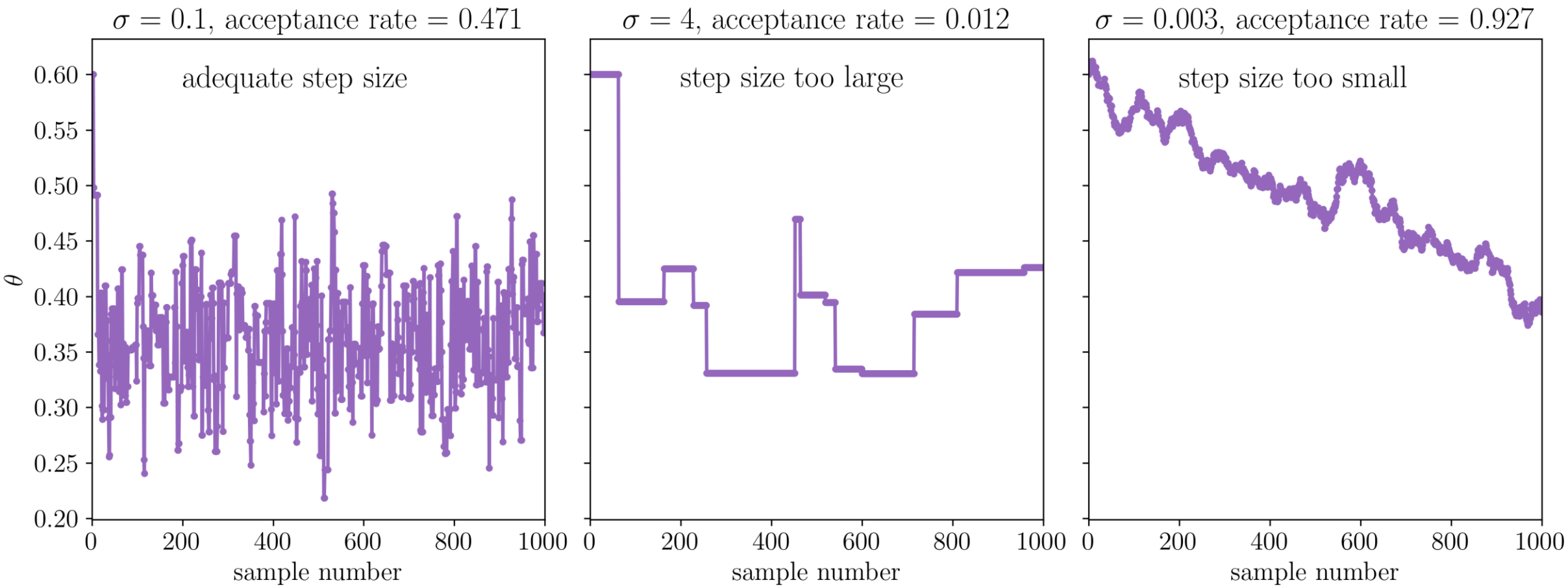
Diagnostics of Markov chains: burn-in





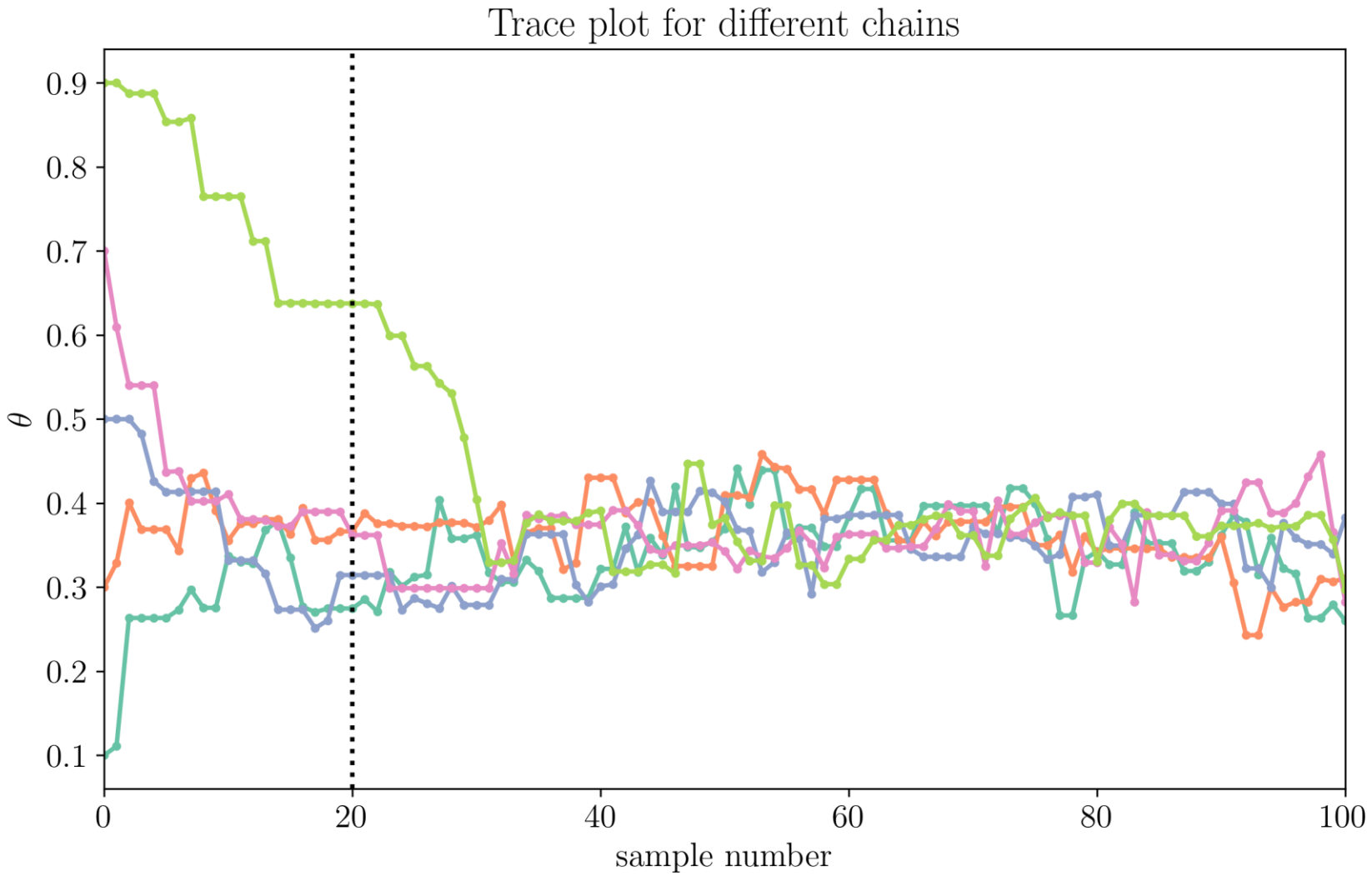
Diagnostics of Markov chains: trace plots

- Adjusting the proposal distribution (here by changing the step size):



Diagnostics of Markov chains: mixing

- Several independent chains, different starting points:



Diagnostics of Markov chains: convergence – the Gelman-Rubin test

- Parameters:
  - m : number of chains
  - n: length of chains
- Definitions:
  - "between"-chains variance: $B \equiv \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\psi}_{.j} - \bar{\psi}_{..}\right)^2$
  - "within"-chains variance: $W \equiv \frac{1}{m} \sum_{j=1}^{m} s_j^2$

$$\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$$

$$\bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$$

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\psi_{ij} - \bar{\psi}_{.j}\right)^2$$

- Estimators of the marginal posterior variance of the estimand (for each parameter):
  - $\widehat{\mathrm{var}}^- \equiv W$: underestimates the variance
  - $\widehat{\mathrm{var}}^+ \equiv \frac{n}{n-1} W + \frac{1}{n} B$: overestimates the variance
- Gelman-Rubin test:
  - Potential scale reduction factor: $\widehat{R} \equiv \sqrt{\frac{\widehat{\mathrm{var}}^+}{\widehat{\mathrm{var}}^-}}$
  - Test: $\widehat{R} \to 1$ as $n \to \infty$
  - Typically, one aims for $\widehat{R} - 1 \lesssim 10^{-2}$ for all parameters.

Gelman *et al.*, Bayesian Data Analysis (third edition), p. 284-285

Diagnostics of Markov chains: correlation length and effective sample size

- Correlation length:
  - For a Markov chain $\{\theta_n\}_{n=1}^{N}$, the integrated autocorrelation time is defined as $\tau_\theta \equiv \sum_{\tau=-\infty}^{+\infty} \rho_\theta(\tau)$ where $\rho_\theta(\tau)$ is the normalised autocorrelation function of the stochastic process that generated the chain for θ.
  - We can estimate $\rho_\theta(\tau)$ using a finite chain as $\hat{\rho}_\theta(\tau) = \frac{\hat{c}_\theta(\tau)}{\hat{c}_\theta(0)}$ where
$$\hat{c}_\theta(\tau) \equiv \frac{1}{N-\tau} \sum_{n=1}^{N-\tau} (\theta_n - \mu_\theta)(\theta_{n+\tau} - \mu_\theta) \quad \text{and} \quad \mu_\theta \equiv \frac{1}{N} \sum_{n=1}^{N} \theta_n$$
  - In practice, it is more computationally efficient to compute $\hat{c}_\theta(\tau)$ using fast Fourier transforms than summing it directly. The Wiener-Khinchin theorem states that the autocorrelation function of a signal is equal to the inverse Fourier transform of its power spectral density: $\hat{c}_\theta(\tau) \propto \mathcal{F}^{-1}\left\{|\mathcal{F}\{\theta - \mu_\theta\}|^2\right\}$
  - A good estimator for τ_θ is: $\hat{\tau}_\theta = 1 + 2\sum_{\tau=1}^{M} \hat{\rho}_\theta(\tau)$ for some $M \ll N$ (it is possible to use an automated windowing procedure).
- Effective sample size:
  - The (estimated) effective sample size is $N_{\text{eff}} = \frac{\hat{\tau}_\theta}{N}$.
  - The figure of merit for the sampling efficiency of any MCMC algorithm is the effective sample size per function (model, likelihood, gradient) evaluation.

emcee tutorial: https://emcee.readthedocs.io/en/stable/tutorials/autocorr/

MARKOV CHAIN MONTE CARLO BEYOND METROPOLIS-HASTINGS

MCMC beyond Metropolis-Hastings: Gibbs sampling

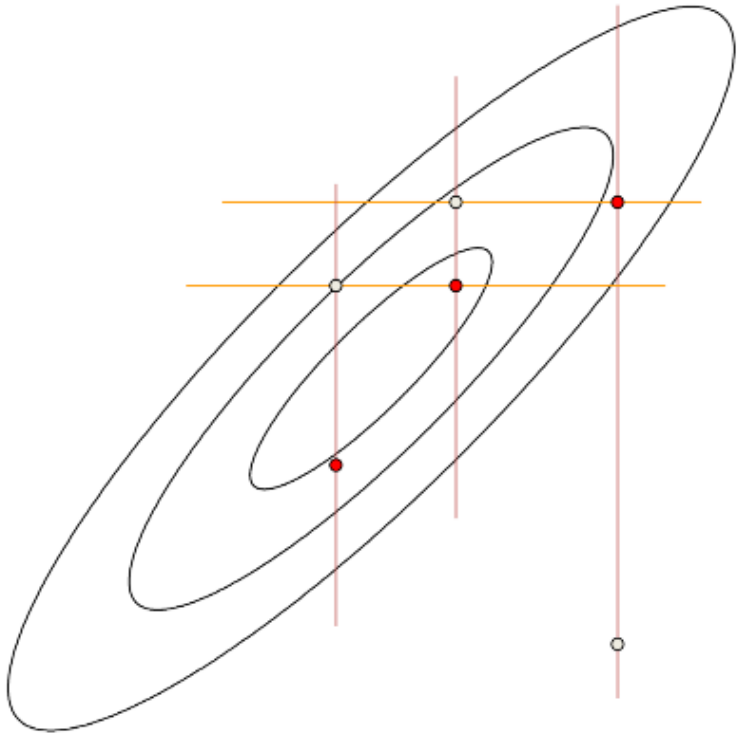
- Shortcomings of standard Metropolis-Hastings:
  - Tuning of proposal distributions
  - Curse of dimensionality

- Gibbs sampling:
  - Uses conditionals of the target pdf: we need to be able to sample $p(x|y)$ and $p(y|x)$
  - Is a special case of Metropolis-Hastings with acceptance ratio unity:
    - To update x to x^* given y: $q(x^*|x,y) = p(x^*|y)$

$$r(x, x^*|y) = \frac{p(x^*, y)}{p(x, y)} \frac{q(x|x^*, y)}{q(x^*|x, y)} = \frac{p(x^*|y)p(y)}{p(x|y)p(y)} \frac{p(x|y)}{p(x^*|y)} = 1$$

    - To update y to y^* given x: $q(y^*|x,y) = p(y^*|x)$

$$r(y, y^*|x) = \frac{p(x, y^*)}{p(x, y)} \frac{q(y|x, y^*)}{q(y^*|x, y)} = \frac{p(y^*|x)p(x)}{p(y|x)p(x)} \frac{p(y|x)}{p(y^*|x)} = 1$$



A two-dimensional test pdf

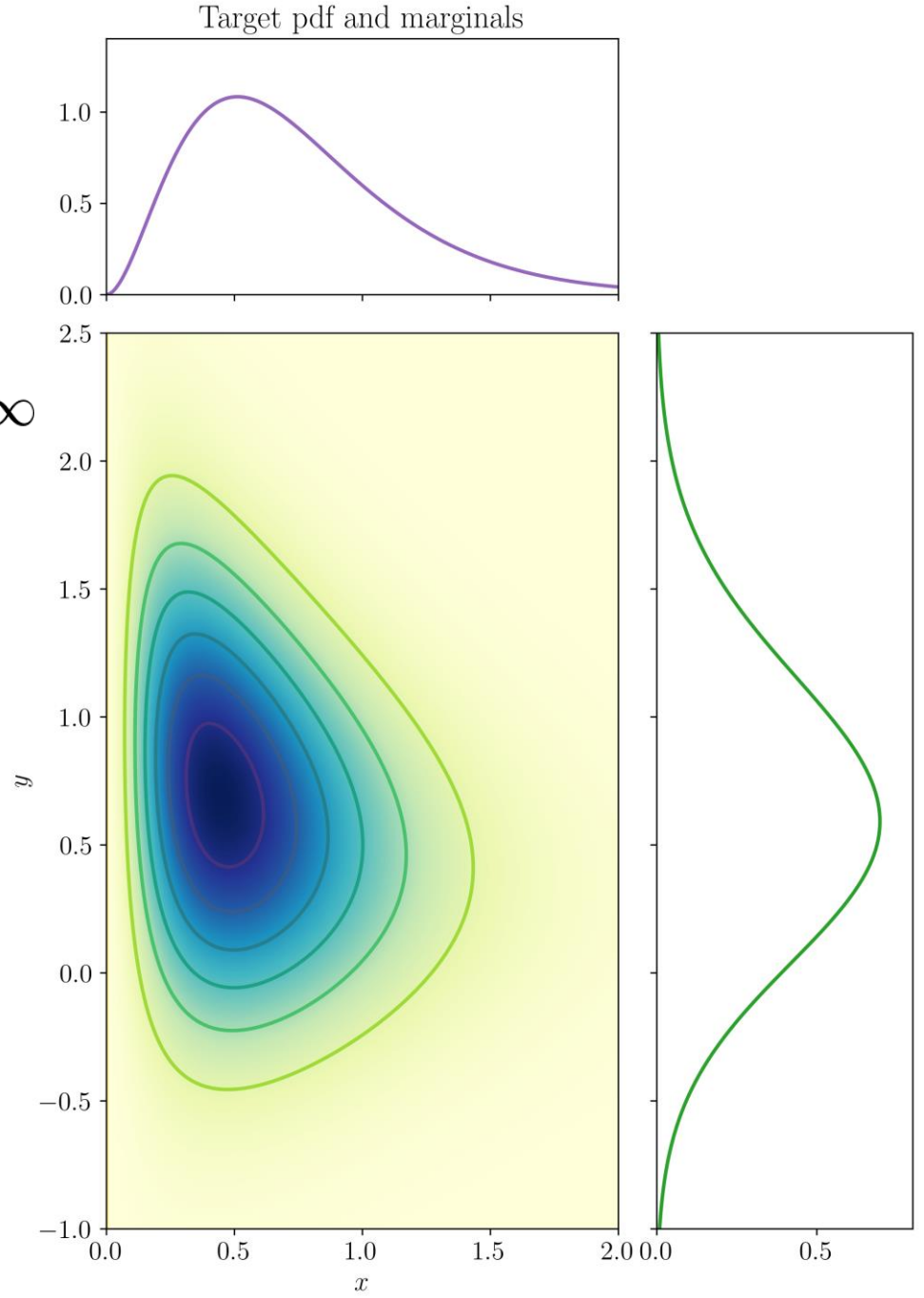
- Joint probability distribution:

$$p(x,y) \propto x^2 \exp\left(-xy^2 - y^2 + 2y - 4x\right) \quad \begin{cases} 0 < x < +\infty \\ -\infty < y < +\infty \end{cases}$$

- Marginals:

$$p(x) \propto x^2 \mathrm{e}^{-4x} \frac{1}{\sqrt{x+1}} \mathrm{e}^{\frac{1}{x+1}}$$

$$p(y) \propto \frac{\mathrm{e}^{-y^2+2y}}{(y^2+4)^3}$$



A two-dimensional test pdf

- Joint probability distribution:

$$p(x,y) \propto x^2 \exp\left(-xy^2 - y^2 + 2y - 4x\right) \quad \begin{cases} 0 < x < +\infty \\ -\infty < y < +\infty \end{cases}$$

- Marginals:

$$p(x) \propto x^2 \mathrm{e}^{-4x} \frac{1}{\sqrt{x+1}} \mathrm{e}^{\frac{1}{x+1}}$$

$$p(y) \propto \frac{\mathrm{e}^{-y^2+2y}}{(y^2+4)^3}$$

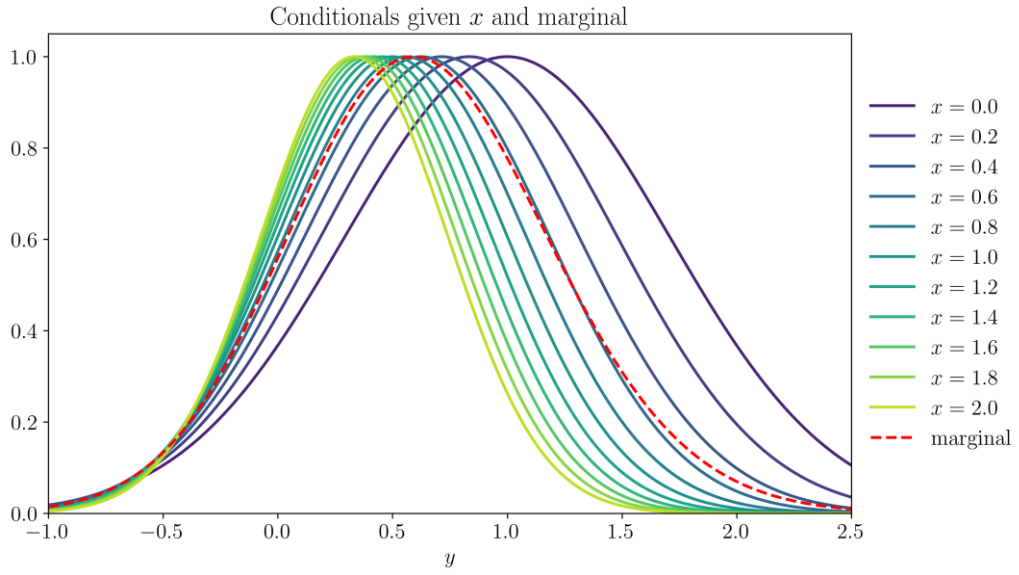
- Conditionals:

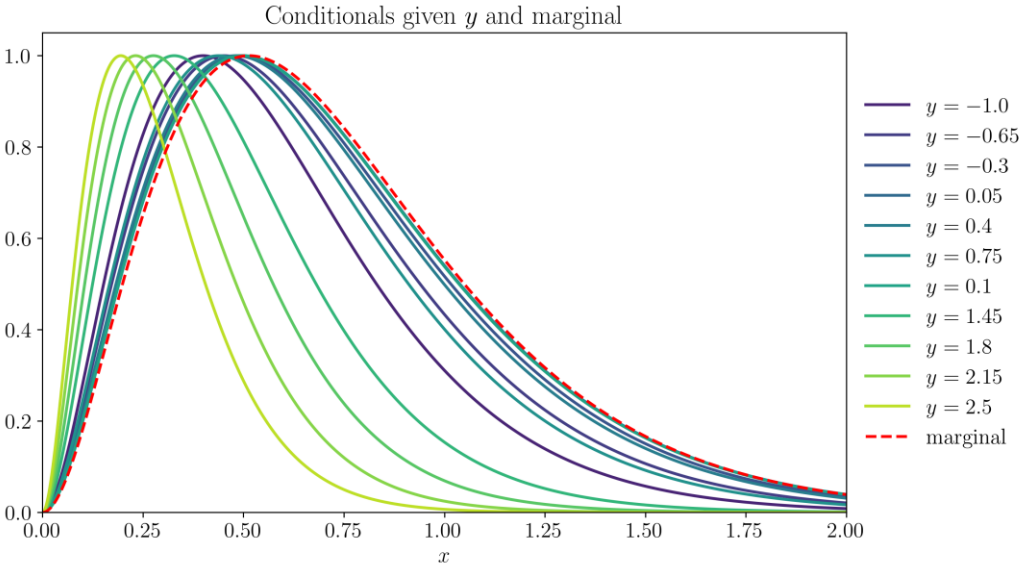
$$p(y|x) = \mathcal{G}\left(\mu = \frac{1}{1+x}, \sigma^2 = \frac{1}{2(1+x)}\right)(y)$$

with $\mathcal{G}\left(\mu, \sigma^2\right)(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(t-\mu)^2}{\sigma^2}\right]$

$$p(x|y) = \Gamma(k=3, \theta = y^2+4)(x)$$

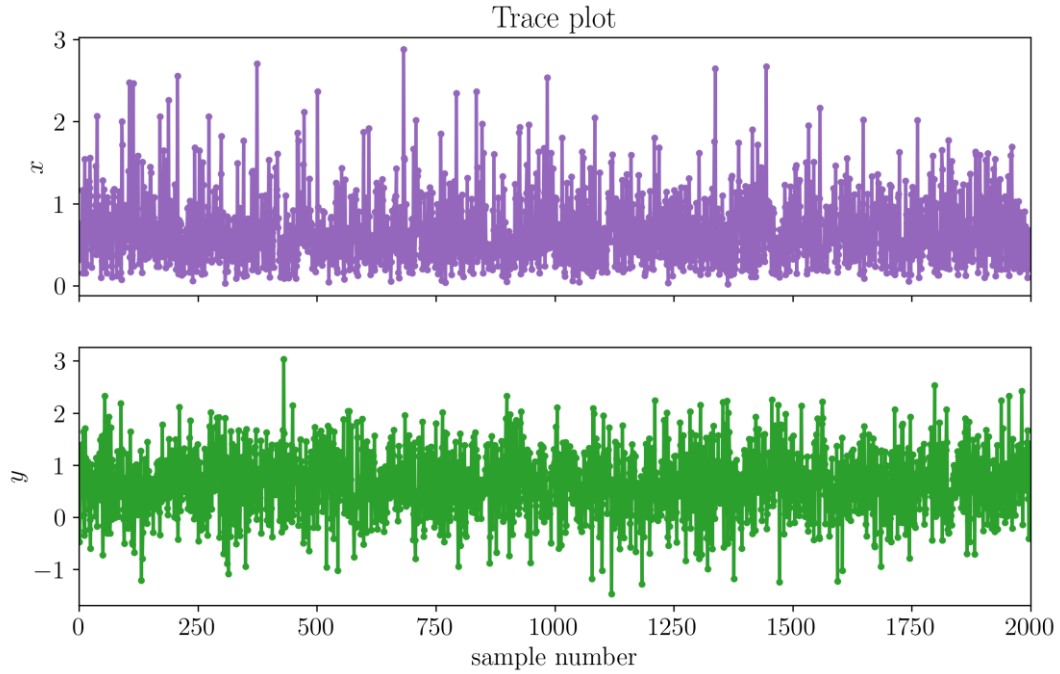
with $\Gamma(k,\theta)(t) = \frac{1}{\Gamma(k)\theta^k} t^{k-1} \mathrm{e}^{-t/\theta}$

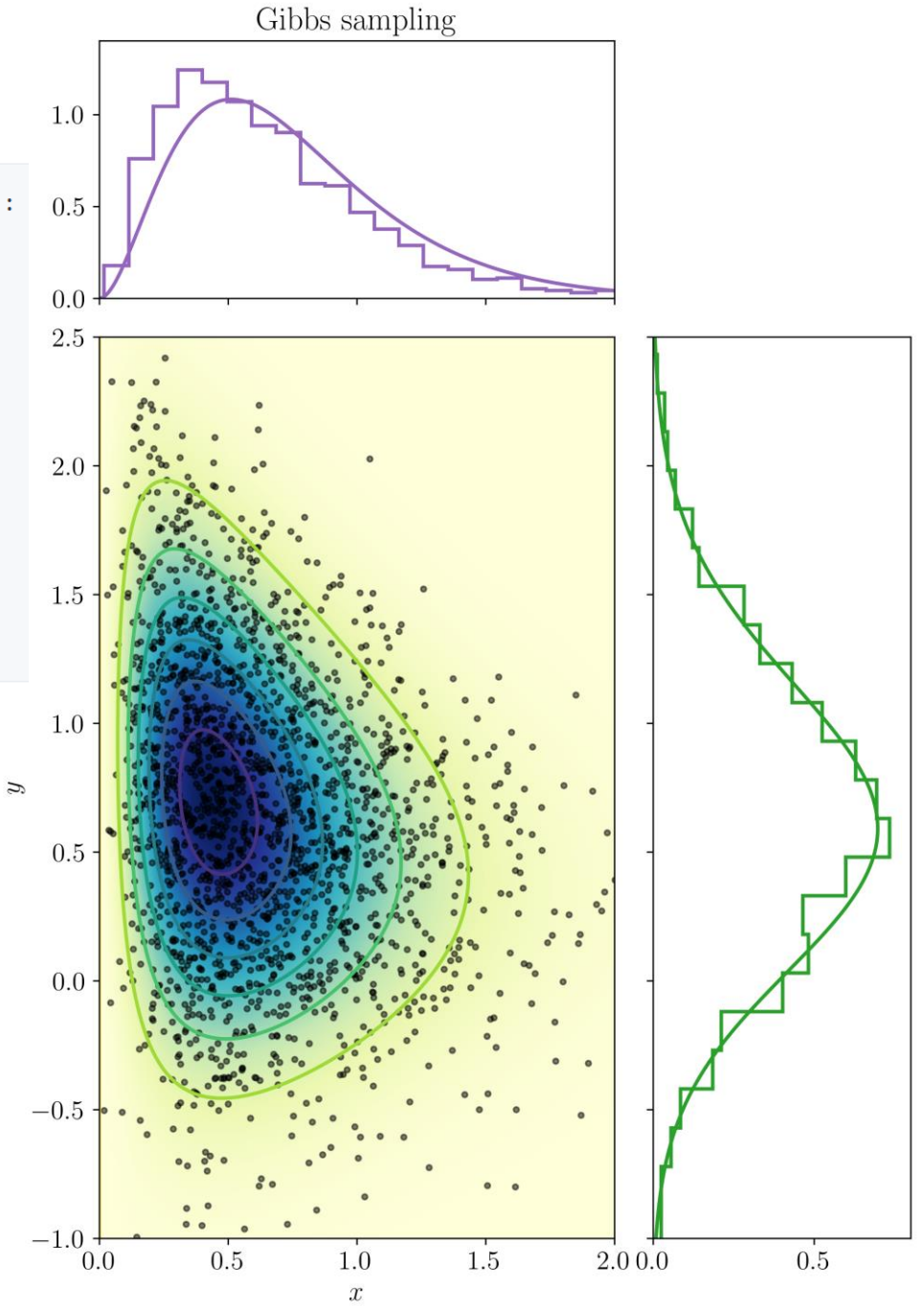




Gibbs sampler

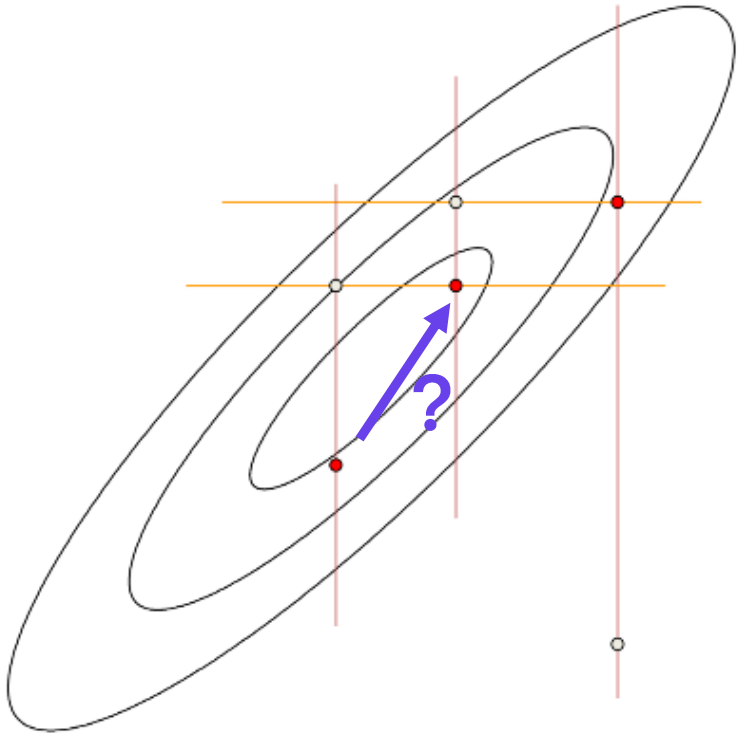
```python
def Gibbs_sampler(target_conditional_given_x,target_conditional_given_y,Nsamp,x_start,y_start):
    samples = np.zeros((Nsamp, 2))
    x=x_start
    y=y_start
    for i in range(Nsamp):
        # first sample x given y
        x = target_conditional_given_y(y).rvs()
        # then sample y given x
        y = target_conditional_given_x(x).rvs()
        # since Gibbs sampling is rejection-free,
        # here we don't even check for acceptance
        samples[i, :] = (x, y)
    return samples
```





MCMC beyond Metropolis-Hastings

- Shortcomings of standard Metropolis-Hastings:
  - Tuning of proposal distributions
  - Curse of dimensionality
- Shortcomings of Gibbs sampling:
  - Needs conditionals of the target pdf
  - Inefficient if parameters are strongly correlated
  - How does one take diagonal steps in parameter space?



Hamiltonian (Hybrid) Monte Carlo

- Use classical mechanics to solve statistical problems!
  - The potential: $\psi(\mathbf{x}) \equiv -\ln p(\mathbf{x})$
  - The Hamiltonian: $H(\mathbf{x}, \mathbf{p}) \equiv \frac{1}{2}\mathbf{p}^\intercal \mathbf{M}^{-1}\mathbf{p} + \psi(\mathbf{x})$

$$(\mathbf{x}, \mathbf{p}) \Rightarrow \left\{ \begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\frac{\partial H}{\partial \mathbf{x}} = -\frac{\mathrm{d}\psi(\mathbf{x})}{\mathrm{d}\mathbf{x}} \end{aligned} \right\} \Rightarrow (\mathbf{x}^*, \mathbf{p}^*)$$

gradients of the log-pdf

$$r(\mathbf{x}, \mathbf{x}^*) = \mathrm{e}^{-(H^* - H)} = 1 \Leftarrow$$ acceptance ratio unity

- HMC beats the curse of dimensionality by:
  - Exploiting gradients
  - Using conservation of the Hamiltonian

Duane *et al.* (1987), Phys. Lett. B 195, 2; Neal, 1206.1901

Hamiltonian Monte Carlo: implementation

- Usual implementation, including a Metropolis-Hastings test for numerical errors:

begin
  initialise $\boldsymbol{x}_{(0)}$;
  for $i = 1$ *to* n **do**
    $\boldsymbol{p} \curvearrowleft \mathcal{N}(0,1)$ (normal distribution);
    $(\boldsymbol{x}^*_{(0)}, \boldsymbol{p}^*_{(0)}) = (\boldsymbol{x}_{(i-1)}, \boldsymbol{p})$;
    for $j = 1$ *to* N_{steps} **do**
      make a leapfrog move: $(\boldsymbol{x}^*_{(j-1)}, \boldsymbol{p}^*_{(j-1)}) \to (\boldsymbol{x}^*_{(j)}, \boldsymbol{p}^*_{(j)})$;
    end
    $(\boldsymbol{x}^*, \boldsymbol{p}^*) = (\boldsymbol{x}_{(N_{\text{steps}})}, \boldsymbol{p}_{(N_{\text{steps}})})$;
    $\alpha \curvearrowleft \mathcal{U}(0,1)$ (uniform distribution);
    if $\alpha < \min\left(1, \exp\left\{-\left[H(\boldsymbol{x}^*, \boldsymbol{p}^*) - H(\boldsymbol{x}^*_{(0)}, \boldsymbol{p}^*_{(0)})\right]\right\}\right)$ **then**
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}^*$;
    else
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}_{(i-1)}$;
    end
  end
  return $(\boldsymbol{x}_{(0)}, \ldots, \boldsymbol{x}_{(n)})$;
end

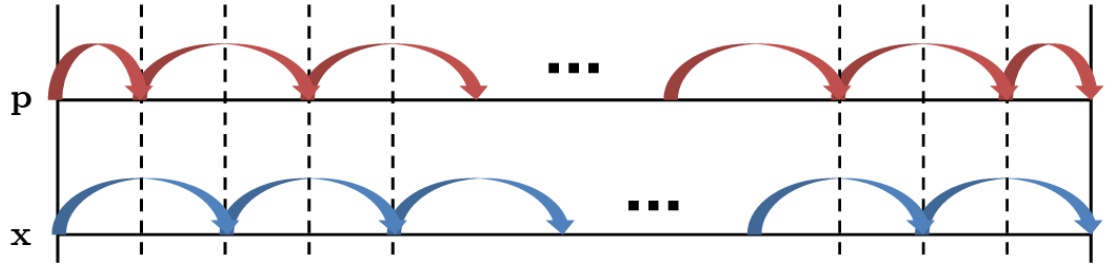
- Setting up the Hamiltoninan

$$H(\mathbf{x}, \mathbf{p}) \equiv \frac{1}{2}\mathbf{p}^\intercal \mathbf{M}^{-1}\mathbf{p} + \psi(\mathbf{x})$$

requires choosing (and tuning) a mass matrix $\mathbf{M}$.

- Equations of motion need to be discretised. The leapfrog or kick-drift-kick integrator is the usual choice because it is symplectic:

$$\begin{cases} p_k\left(t + \frac{\epsilon}{2}\right) = p_k(t) - \frac{\epsilon}{2}\frac{\partial \psi}{\partial x_k}\Big|_{\boldsymbol{x}(t)} \\ x_k\left(t + \epsilon\right) = x_k(t) + \epsilon\, p_k\left(t + \frac{\epsilon}{2}\right) \\ p_k\left(t + \epsilon\right) = p_k\left(t + \frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{\partial \psi}{\partial x_k}\Big|_{\boldsymbol{x}(t+\epsilon)} \end{cases}$$



Hamiltonian sampler

```
def Hamiltonian_sampler(psi,dpsi_dx,dpsi_dy,MassMatrix,stepstd,Nsamp,x_start,y_start):
    InvMassMatrix=np.linalg.inv(MassMatrix)
    Naccepted=0
    x=x_start
    y=y_start
    samples = np.zeros((Nsamp,2))
    samples[0,0]=x
    samples[0,1]=y
    for i in range(Nsamp-1):
        # compute potential energy and gradient
        old_x = x
        old_y = y
        old_psi = psi(old_x,old_y)
        dpsidx = dpsi_dx(old_x,old_y)
        dpsidy = dpsi_dy(old_x,old_y)

        # randomly draw momenta
        p_x = norm(0.,1.).rvs()
        p_y = norm(0.,1.).rvs()
        p = np.array((p_x,p_y))

        # compute kinetic energy
        old_K = p.T.dot(InvMassMatrix).dot(p)/2.

        # compute Hamiltonian
        old_H = old_K + old_psi

        # do 3 leapfrog step
        for tau in range(3):
            # draw stepsize
            stepsize = norm(0.,stepstd).rvs()

            # Kick: make half step in p_x, p_y
            p_x -= stepsize*dpsidx/2.0
            p_y -= stepsize*dpsidy/2.0
            # compute velocities
            p = np.array((p_x,p_y))
            v_x,v_y = InvMassMatrix.dot(p)
            # Drift: make full step in (x,y)
            new_x = old_x+stepsize*v_x
            new_y = old_y+stepsize*v_y
            # compute new gradient
            dpsidx = dpsi_dx(new_x,new_y)
            dpsidy = dpsi_dy(new_x,new_y)
            # Kick: make half step in p_x, p_y
            p_x -= stepsize*dpsidx/2.0
            p_y -= stepsize*dpsidy/2.0
            p = np.array((p_x,p_y))

        # compute new energy and Hamiltonian
        new_psi = psi(new_x,new_y)
        new_K = p.T.dot(InvMassMatrix).dot(p)/2
        new_H = new_K + new_psi
        dH = new_H - old_H

        # accept/reject new candidate x,y using the standard Monte Carlo rule
        if(x<0.):
            accept=False
        else:
            if(dH<0.0):
                accept=True
            else:
                a = np.exp(-dH)
                u = np.random.uniform()
                if(u < a):
                    accept=True
                else:
                    accept=False

        if(accept):
            x=new_x
            y=new_y
            Naccepted+=1
        else:
            x=old_x
            y=old_y
        samples[i+1, :] = (x, y)

    return Naccepted, samples
```

A two-dimensional test pdf

- Joint probability distribution:

$$p(x,y) \propto x^2 \exp\left(-xy^2 - y^2 + 2y - 4x\right)$$

$$\begin{cases} 0 < x < +\infty \\ -\infty < y < +\infty \end{cases}$$

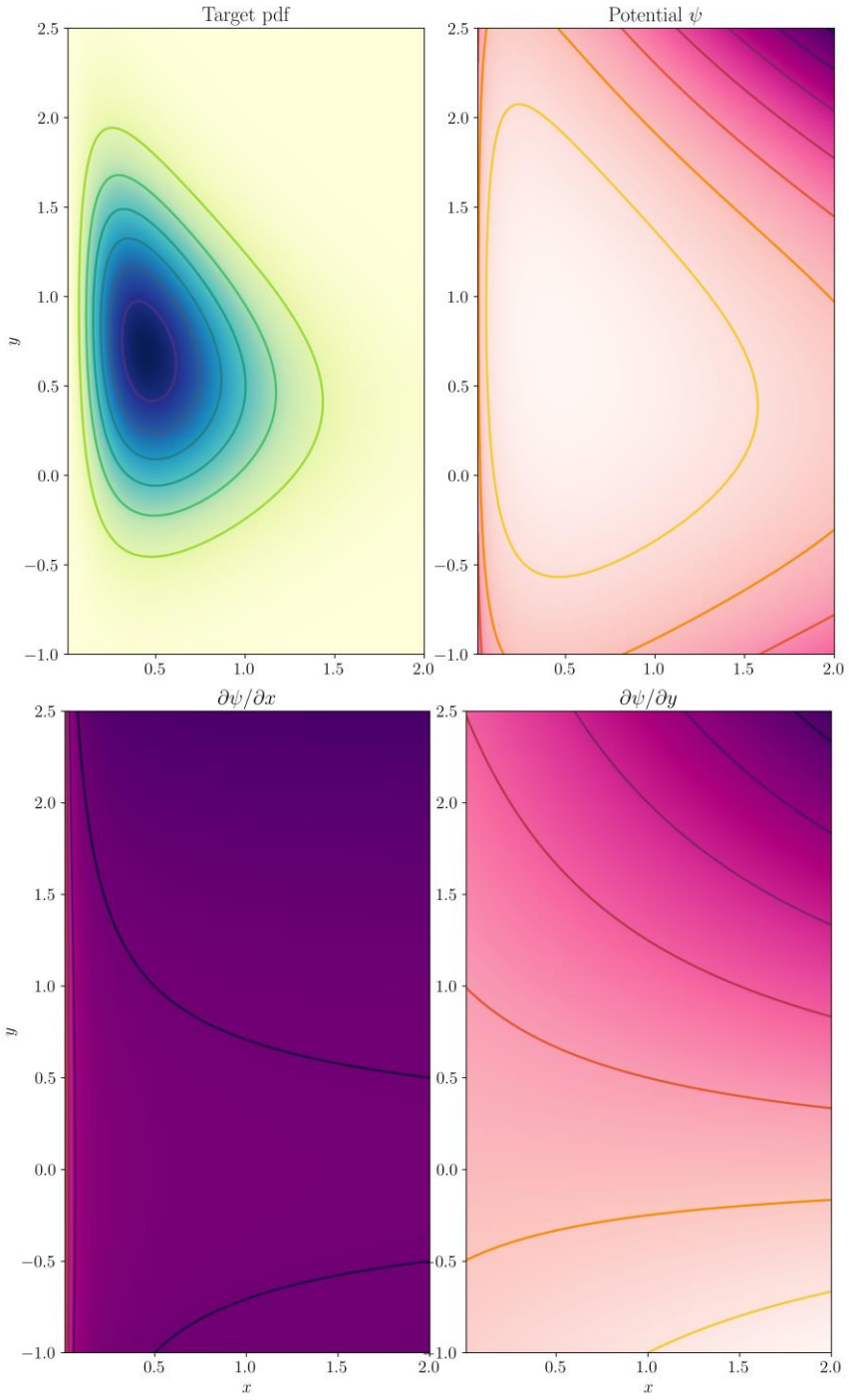
- Potential:

$$\psi(x,y) \equiv -\ln p(x,y) = -2\ln x + xy^2 + y^2 - 2y + 4x$$

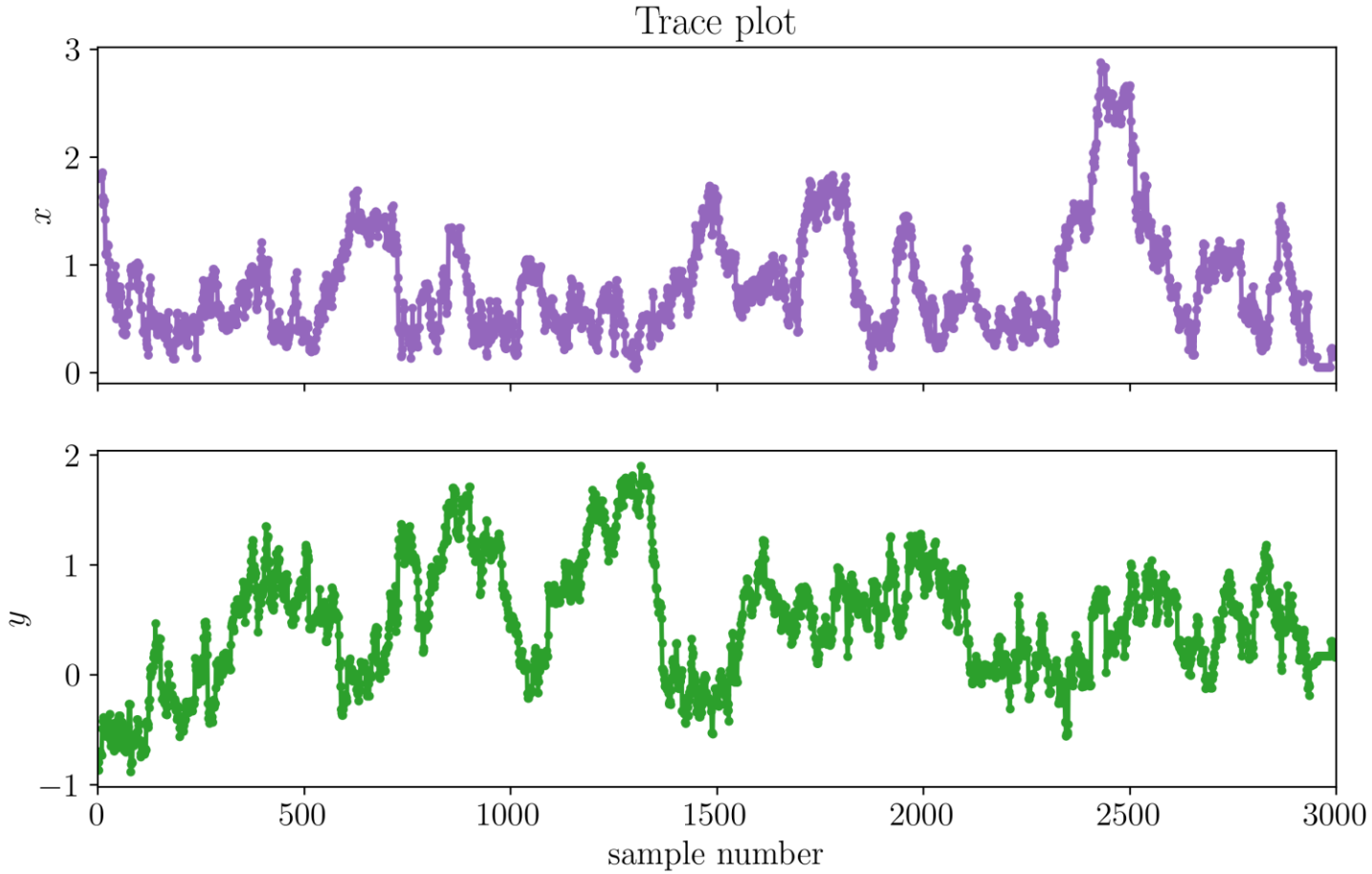
- Gradients of the potential:

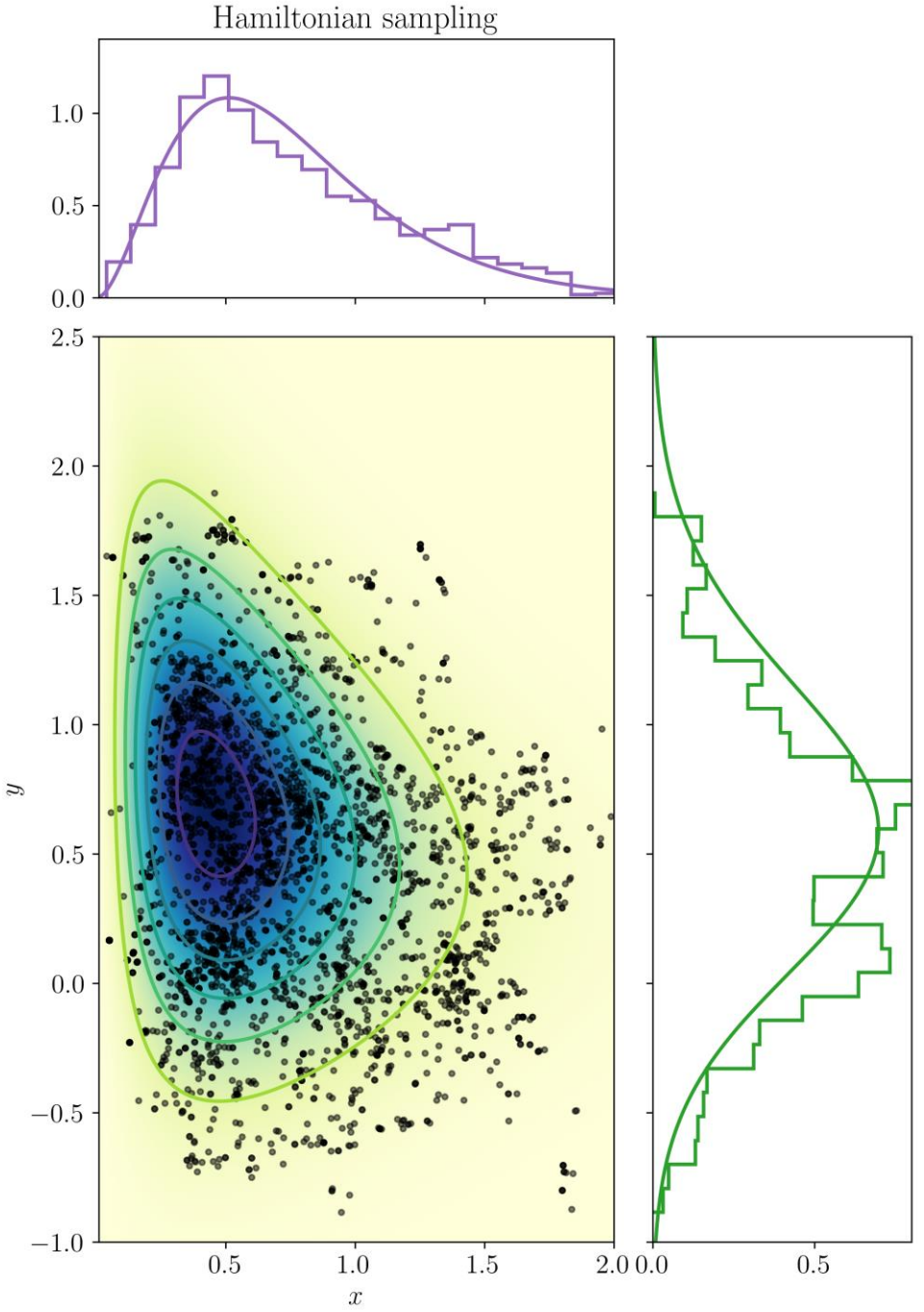
$$\frac{\partial \psi}{\partial x}(x,y) = -\frac{2}{x} + y^2 + 4$$

$$\frac{\partial \psi}{\partial y}(x,y) = 2xy + 2y - 2$$



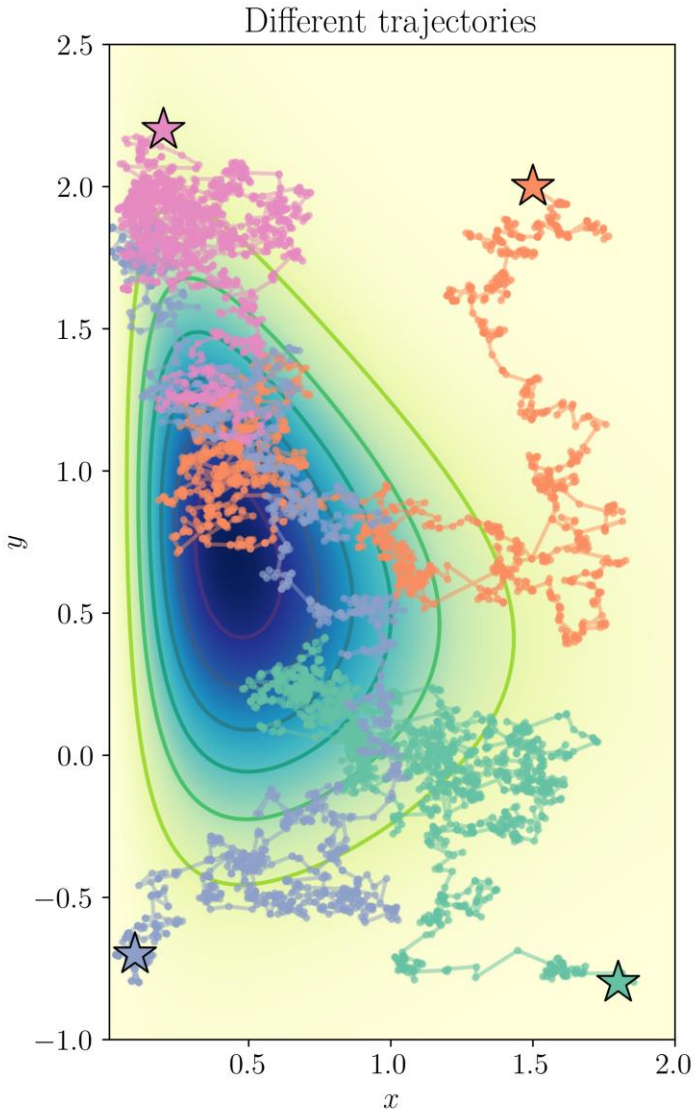
Hamiltonian sampling



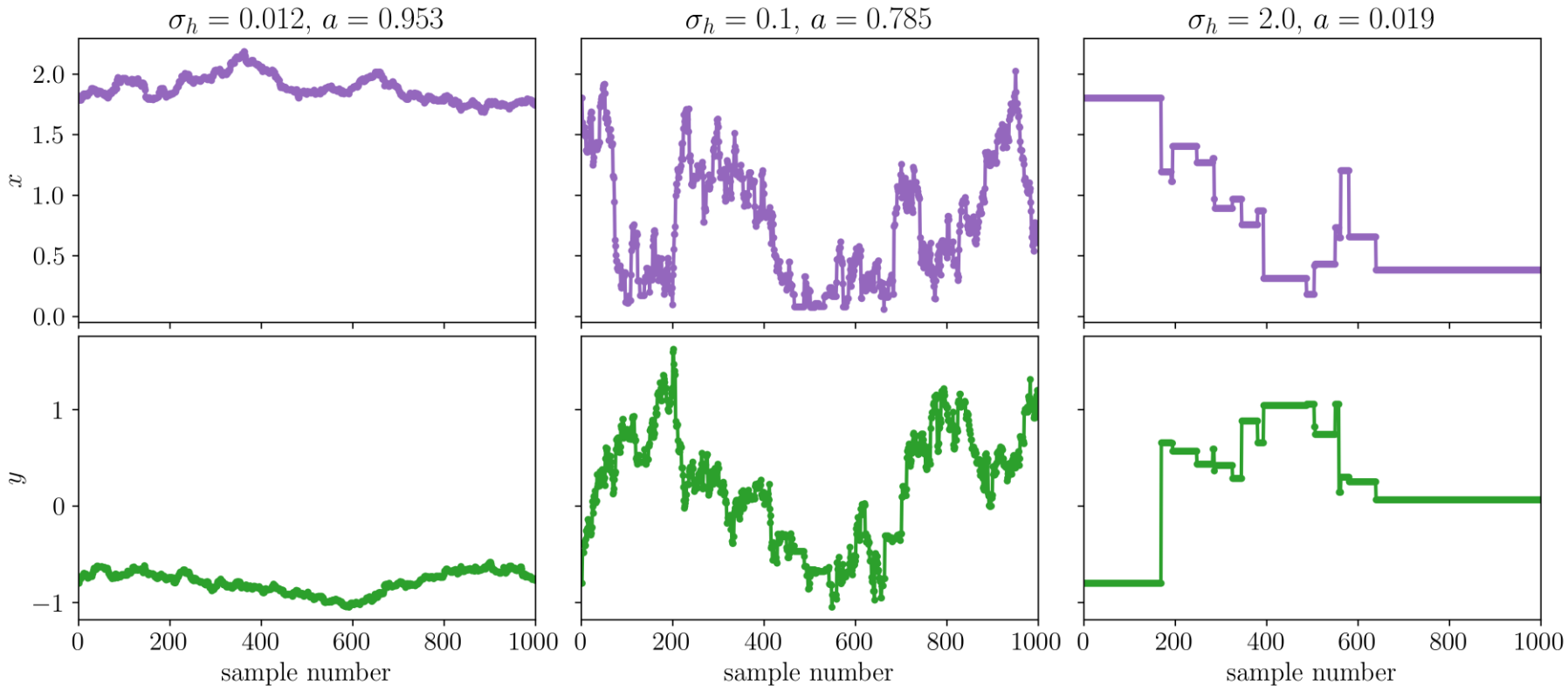


Tuning a Hamiltonian sampler

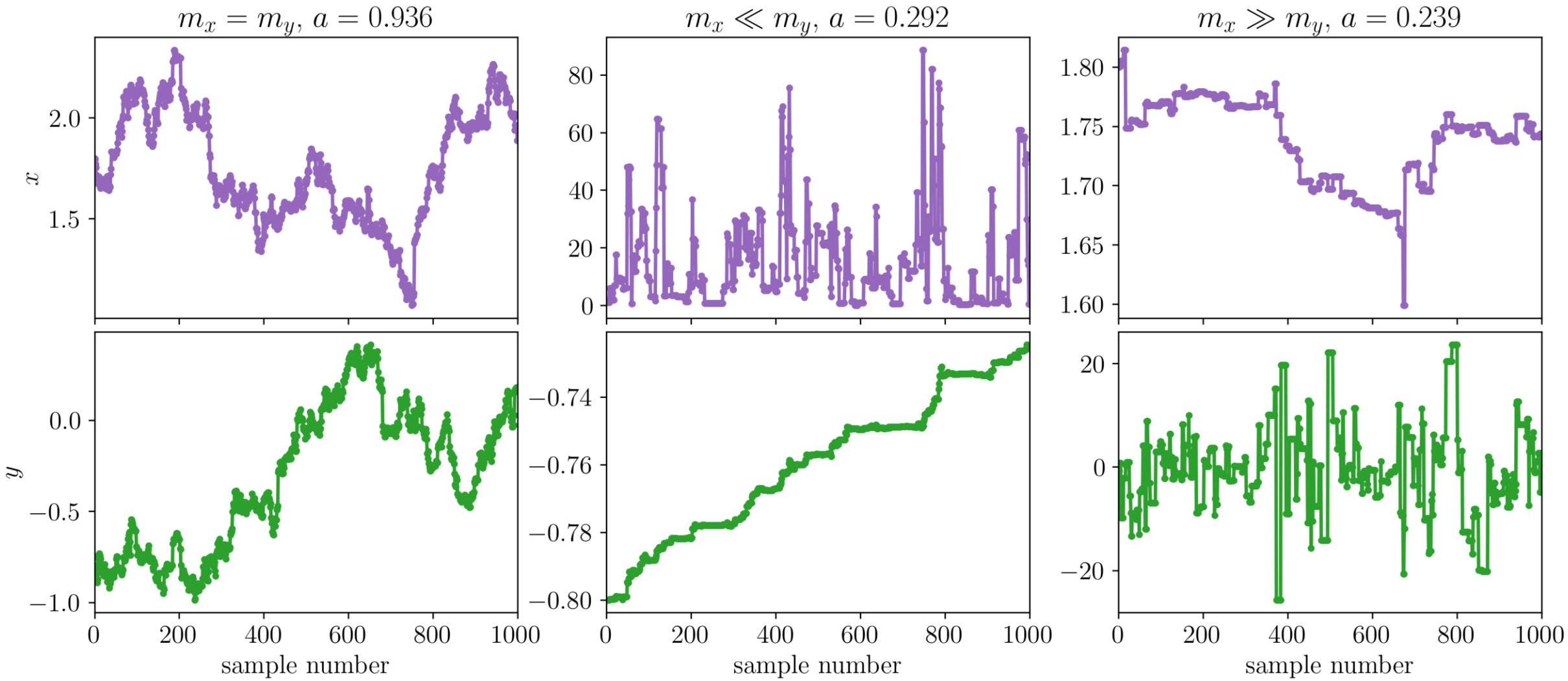
- Different trajectories



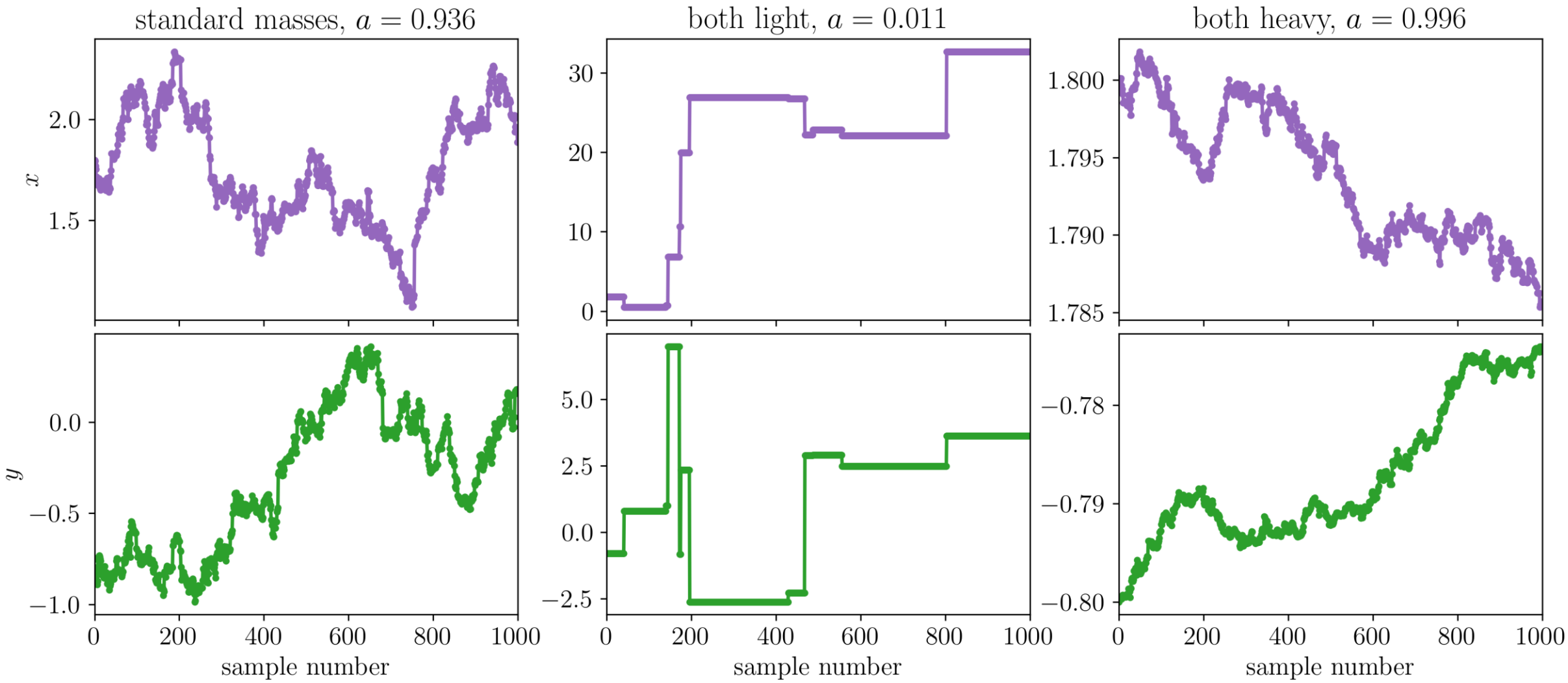
- Tuning the step size ϵ



Tuning the mass matrix



Tuning the mass matrix



FIELD-LEVEL INFERENCE WITH NON-LINEAR MODELS

A non-linear field-level model

Exercise: Non-linear models

- Model:
  - The signal s is a white noise field.
  - The primordial gravitational potential Φ_{L} is a Gaussian random field with phases given by s, zero mean and power spectrum

    $$P(k) = A_{\mathrm{s}} k^{n_{\mathrm{s}}-1}$$

    (i.e. a diagonal covariance matrix in Fourier space), where A_{s} and n_{s} are fixed cosmological parameters.
  - The non-linear gravitational potential Φ_{NL} follows

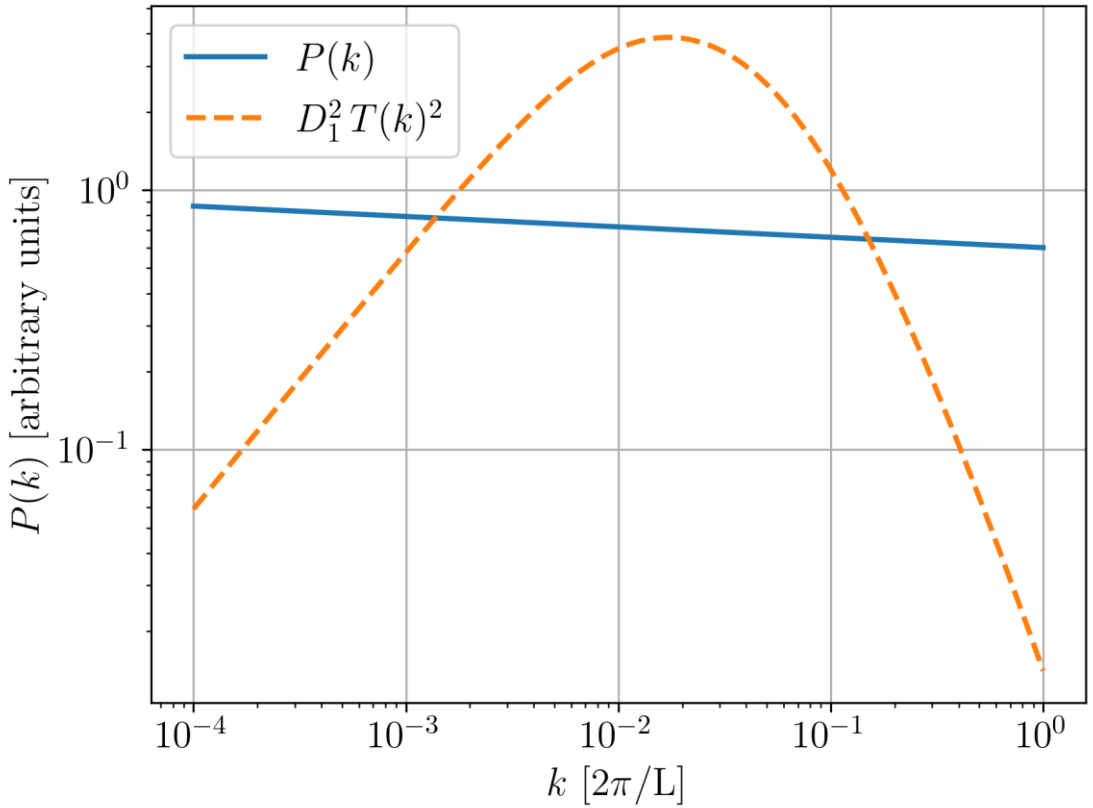
    $$\Phi_{\mathrm{NL}} = \Phi_{\mathrm{L}} + f_{\mathrm{NL}} \Phi_{\mathrm{L}}^2$$

    where f_{NL} is a fixed parameter.
  - The density contrast δ follows (in Fourier space)

    $$\delta(k) = D_1 \, T(k) \, \Phi_{\mathrm{NL}}(k)$$

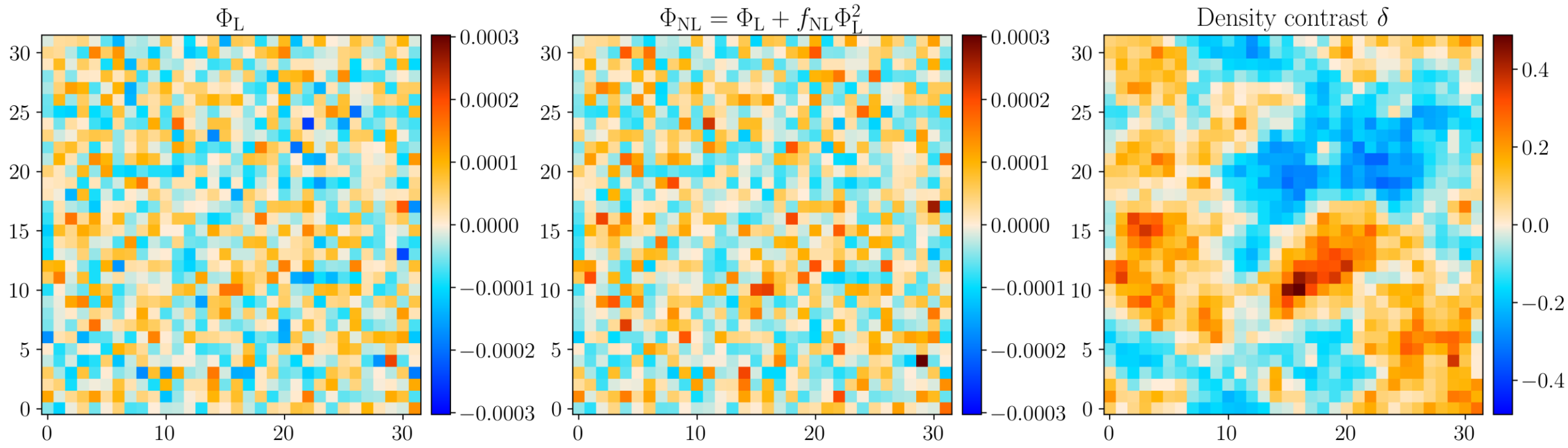
    where D_1 is a fixed (arbitrary) coefficient and $T(k)$ is a "BBKS" transfer function.
  - The noise is Gaussian and additive, to give the observed data:

    $$d = \delta(s) + n$$



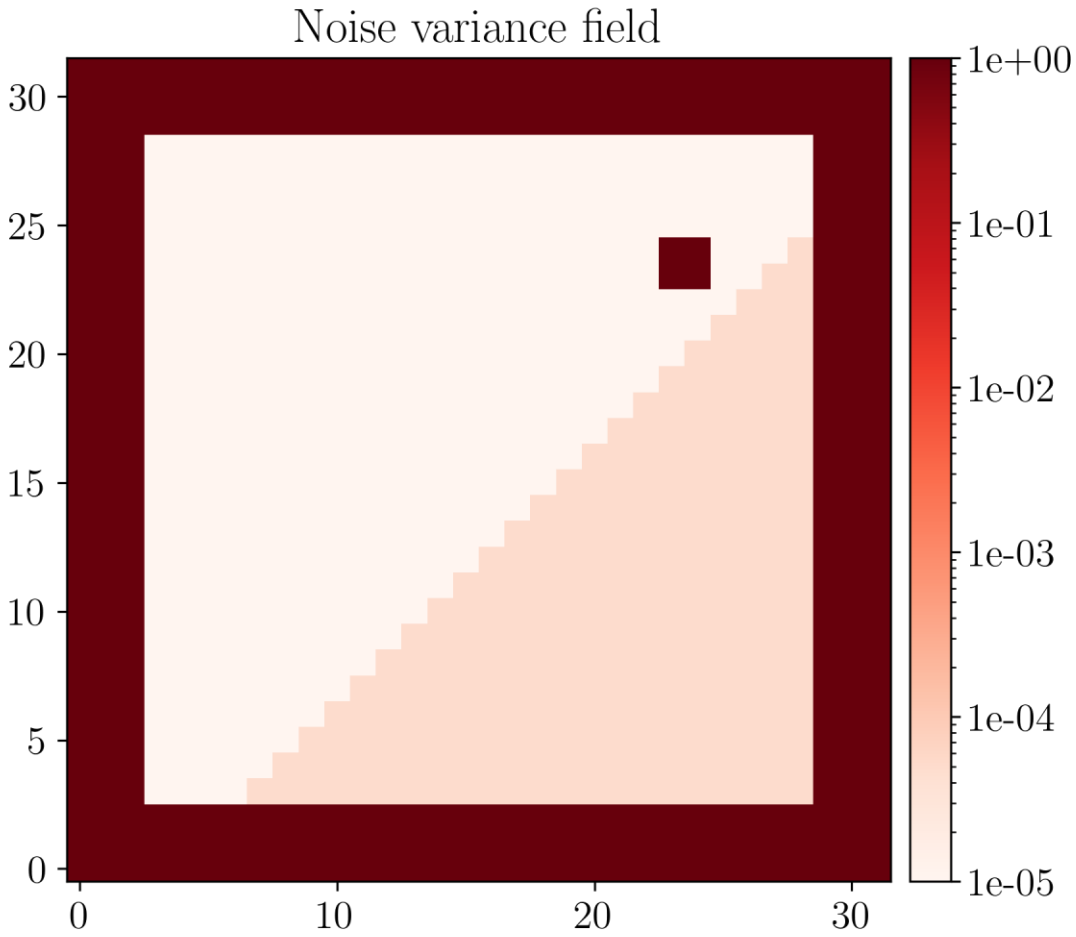
A non-linear field-level model

- Generate signal and ground truth latent fields

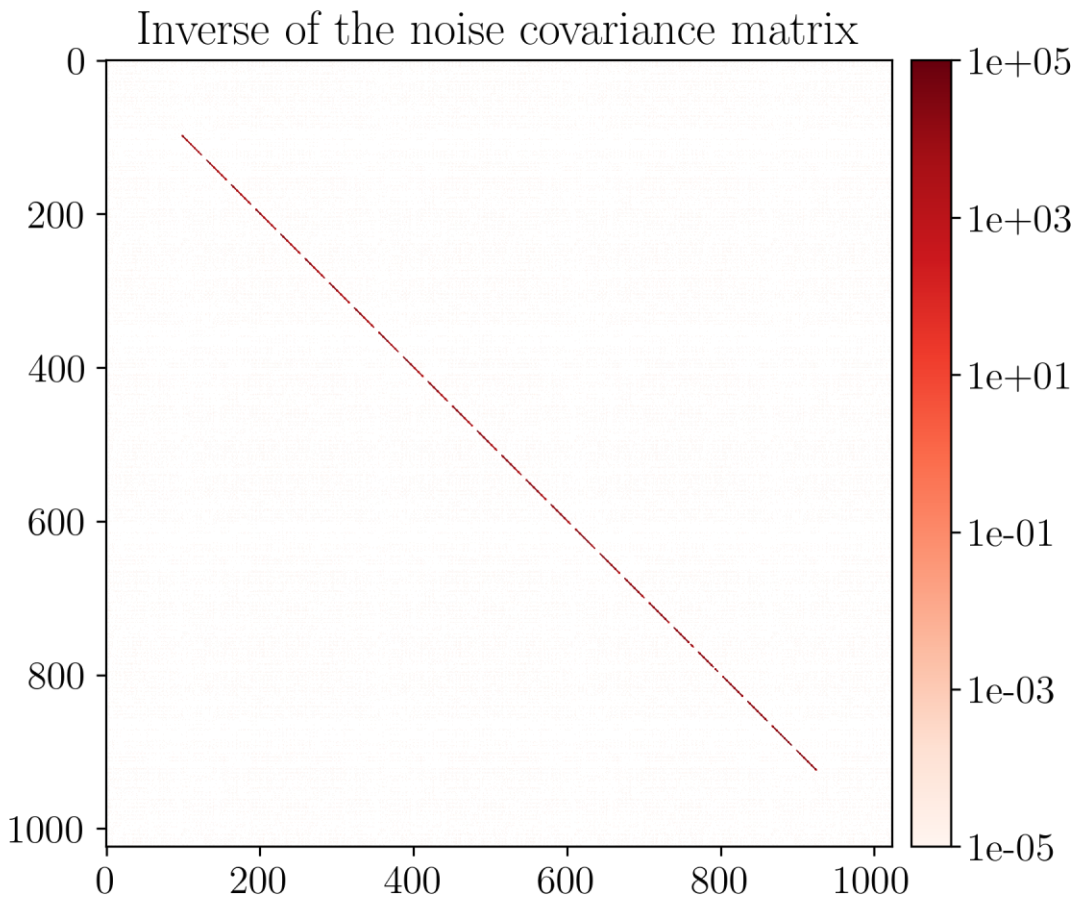


A non-linear field-level model

- Setup Gaussian noise with a covariance matrix as before (diagonal in pixel space)



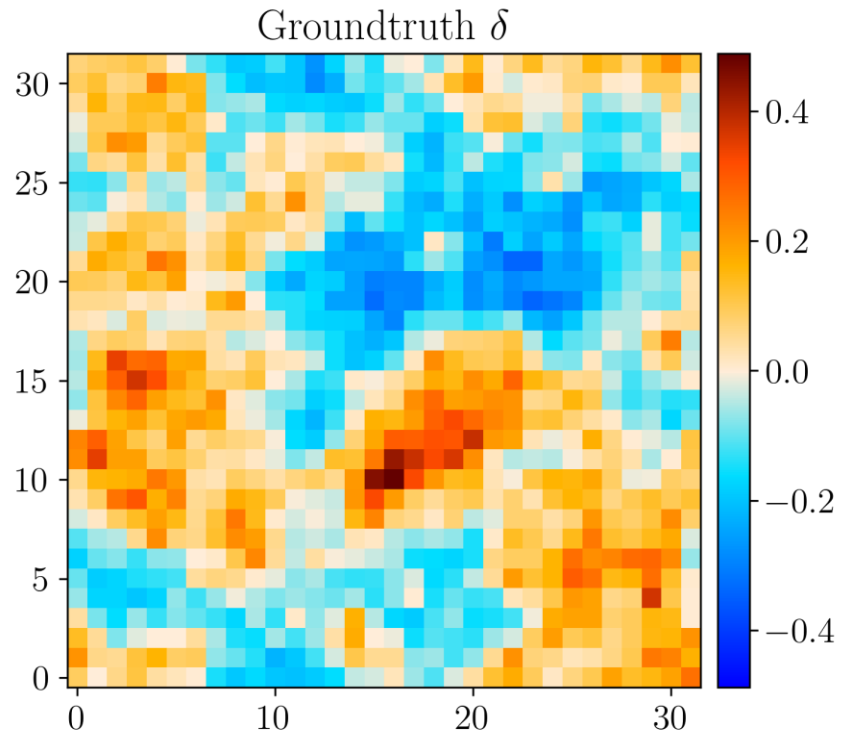
N^{-1}

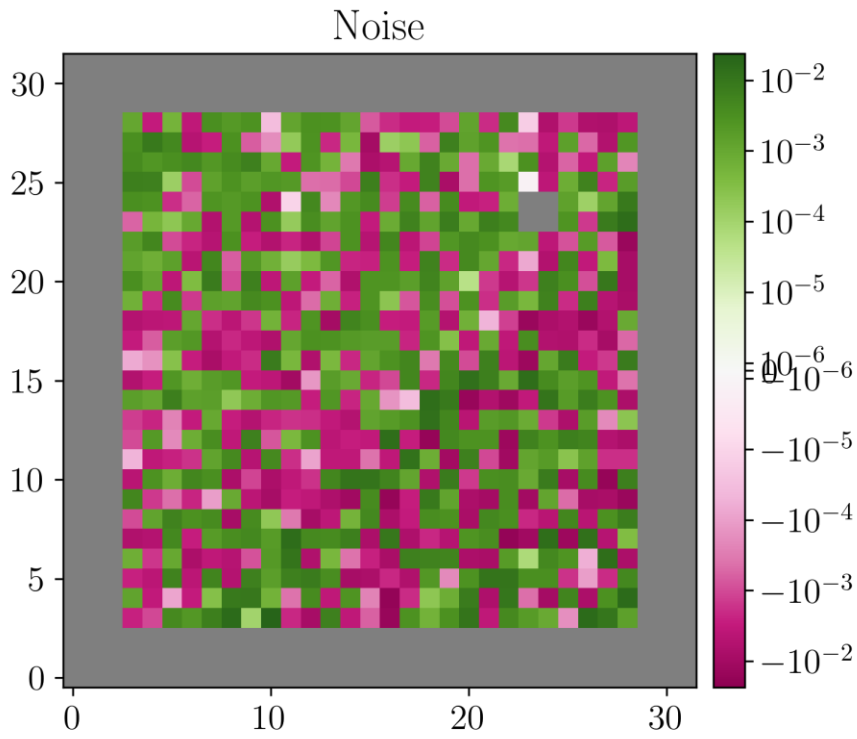


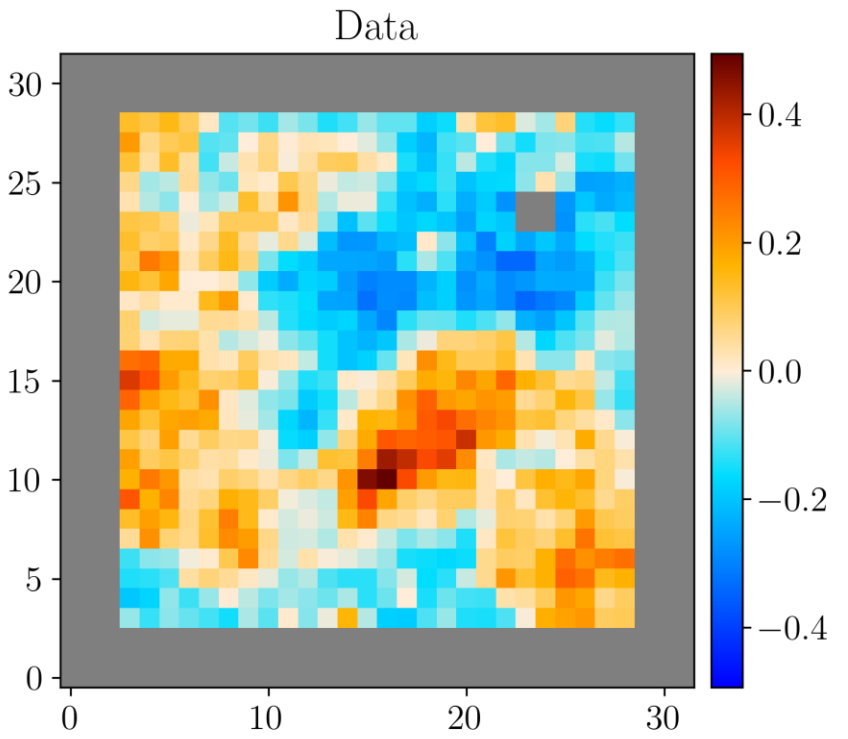
A non-linear field-level model

- Generate mock data

$$d = \delta(s) + n$$





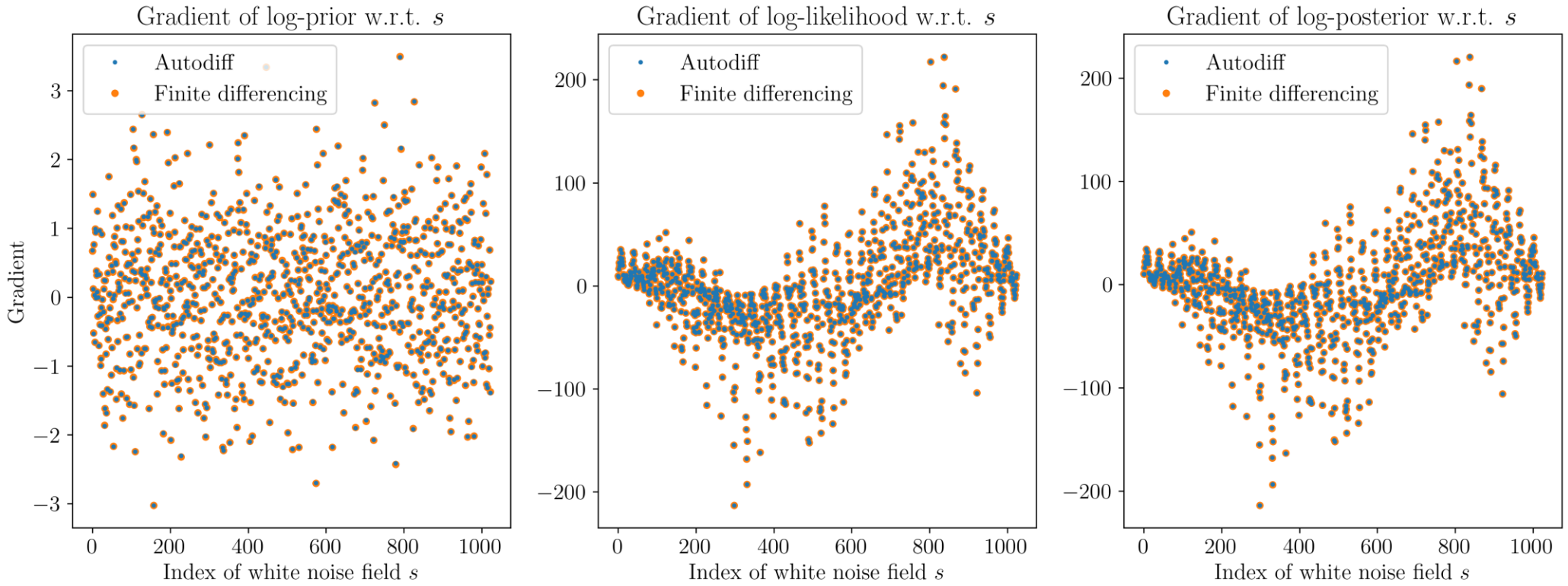


Gradients of the non-linear model and its posterior

- Define log-prior, log-likelihood and log-posterior:

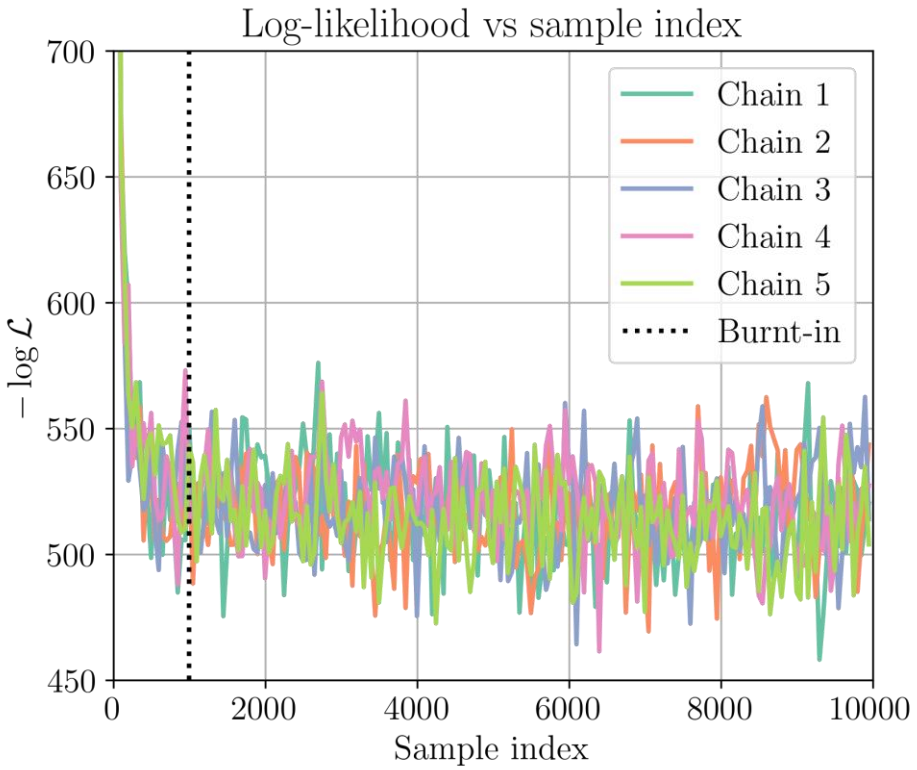
$$\ln p(s) = -\frac{1}{2} s^\intercal s + \text{const.}$$
$$\ln p(d|s) = -\frac{1}{2} \left[\delta(s) - d\right]^\intercal N^{-1} \left[\delta(s) - d\right] + \text{const.}$$
$$\ln p(s|d) = \ln p(s) + \ln p(d|s) + \text{const.}$$

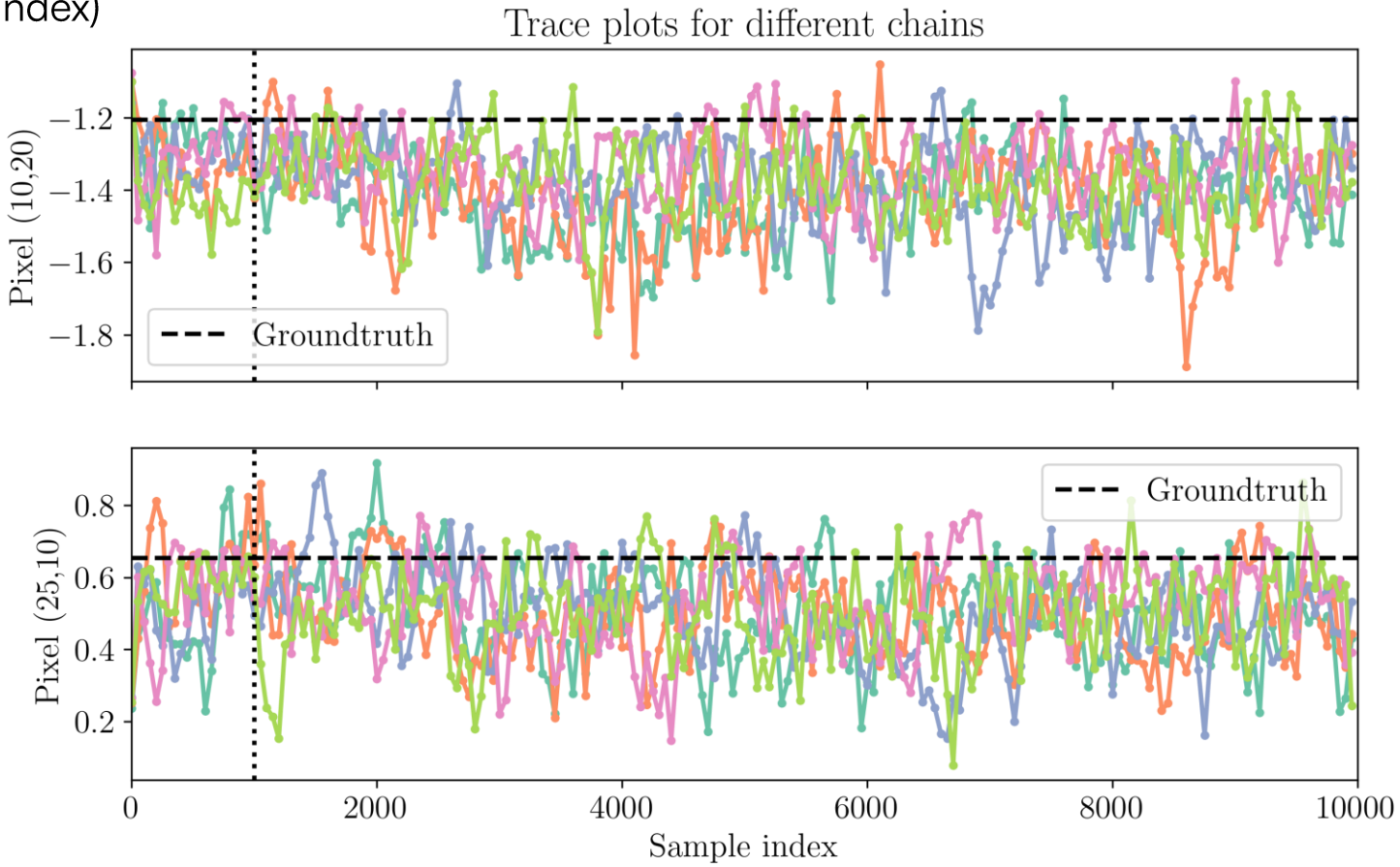
- Compute their gradients (analytically, "manual" differentiation, automatic differentiation).
- Checking the gradient code versus finite differencing is always a useful sanity test.



Field-level inference with non-linear models: sampler tuning and burn-in

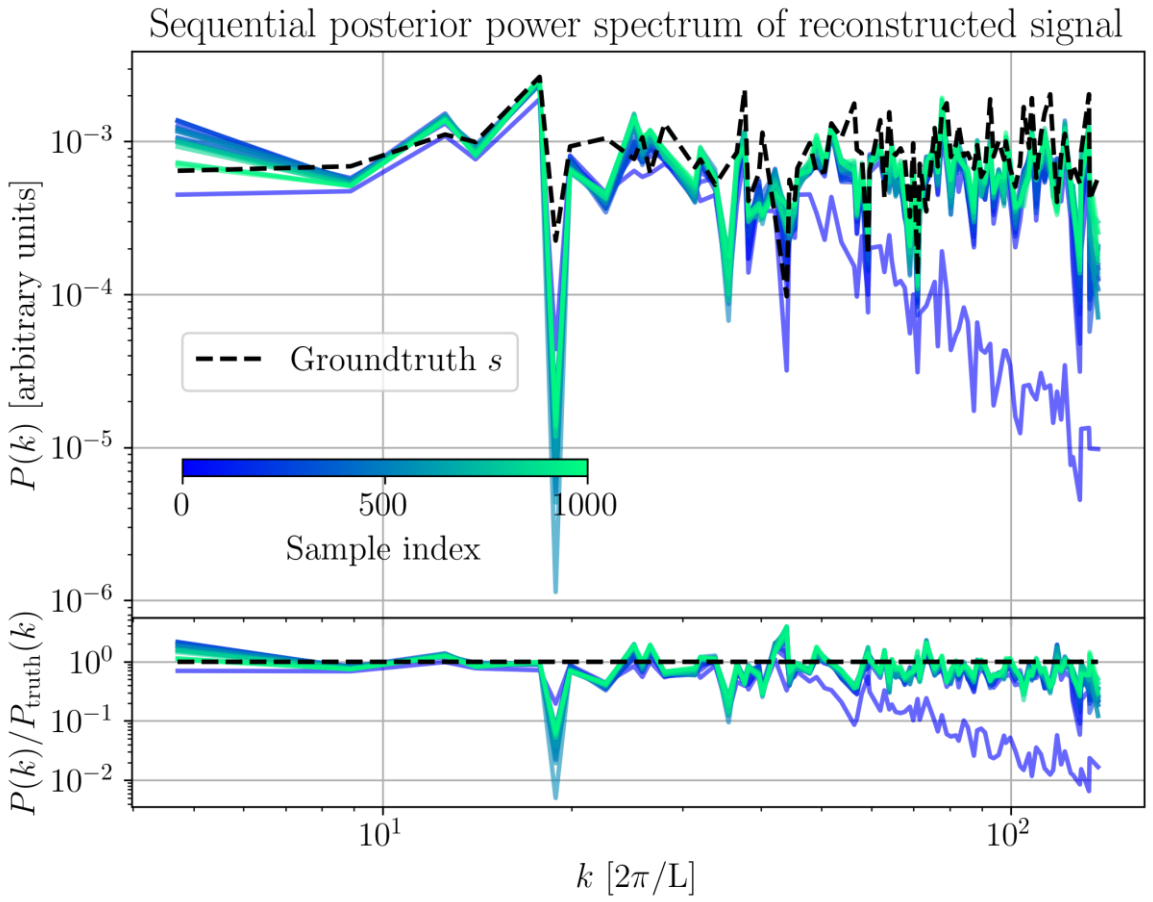
- Run HMC (or NUTS and variants), tune your sampler (ϵ, number of steps, mass matrix…)
- Check the length of the burn-in period: useful diagnostics:
  - log-likelihood vs sample index
  - trace plots (i.e. parameter value vs sample index)

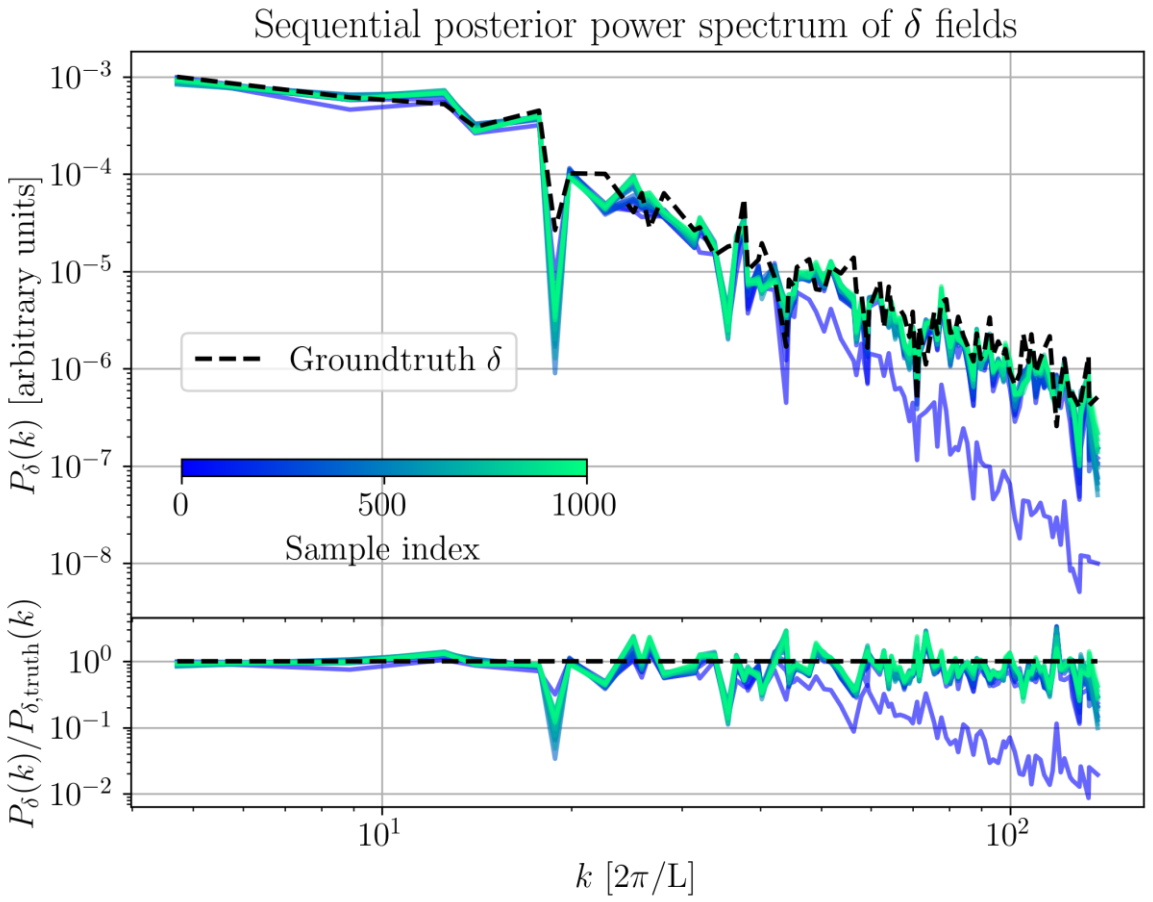




Field-level inference with non-linear models: sampler tuning and burn-in

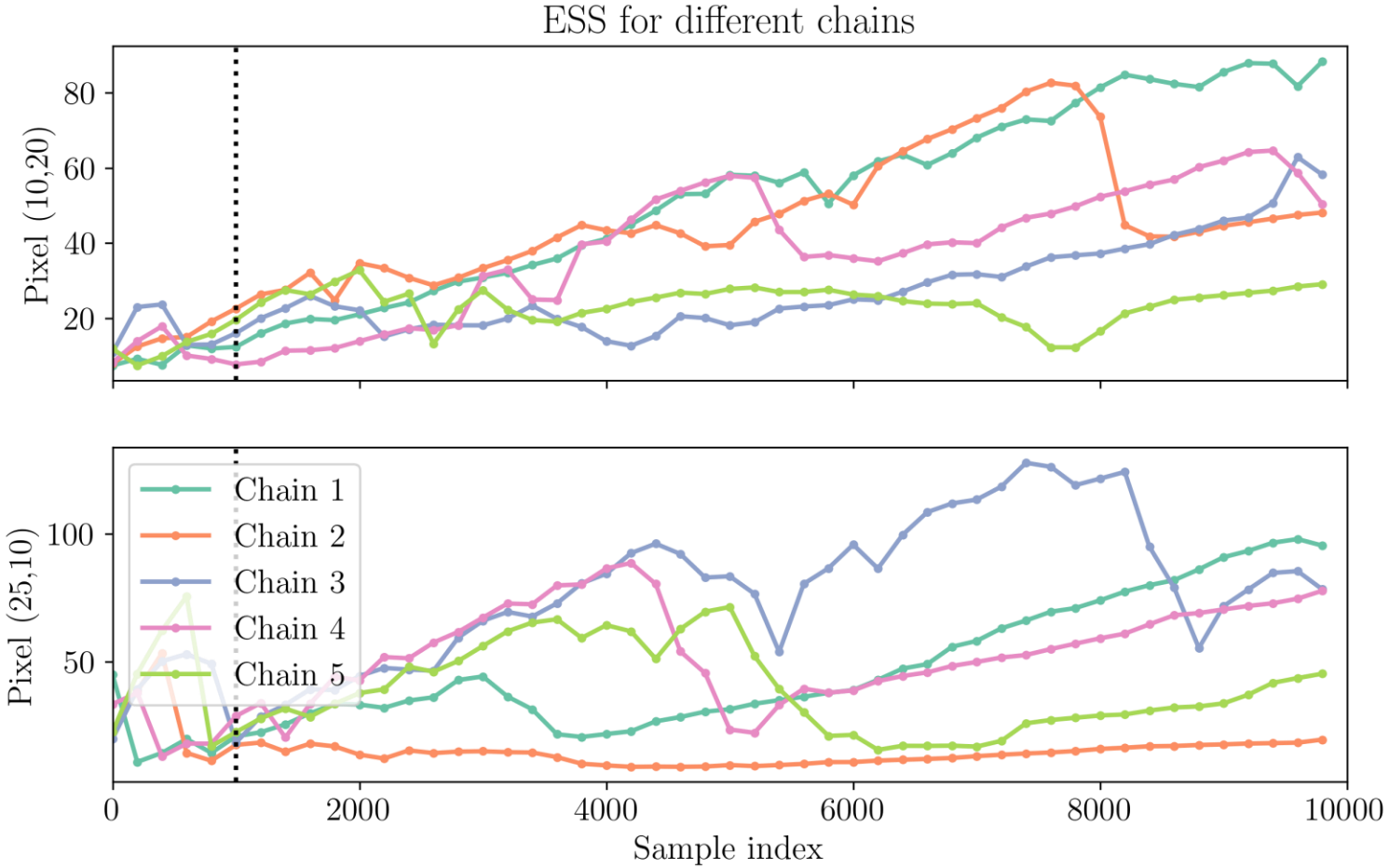
- Pro tip: start your chain from an over-dispersed state and check the sequential posterior power spectrum of samples (reconstructed signal or latent field) with respect to your expectation (ground truth or theory)

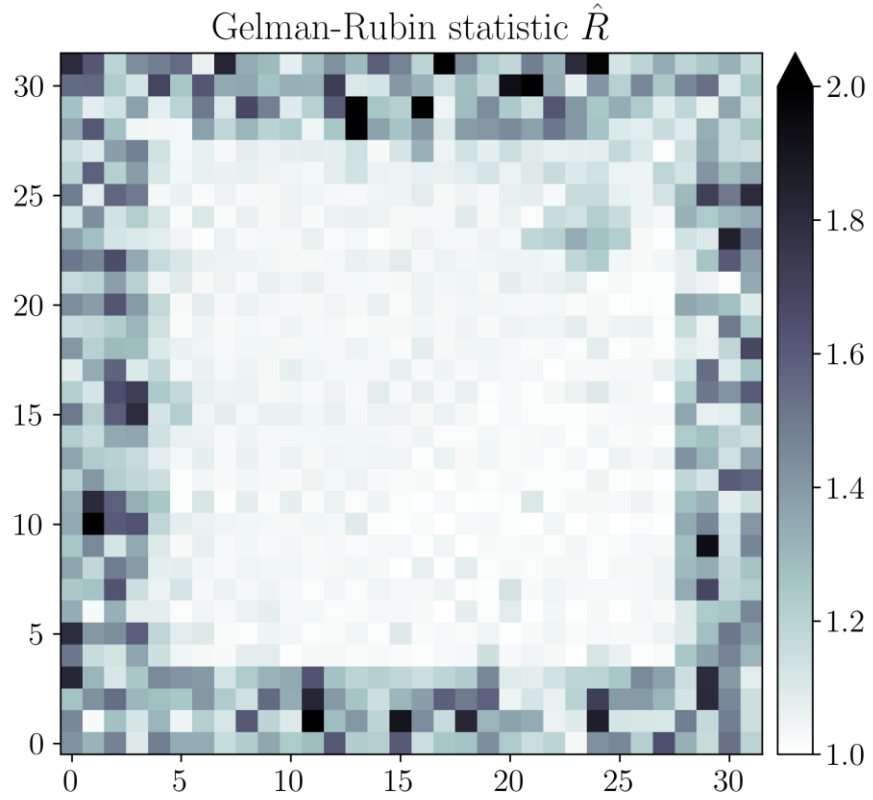




Field-level inference with non-linear models: autocorrelation and convergence

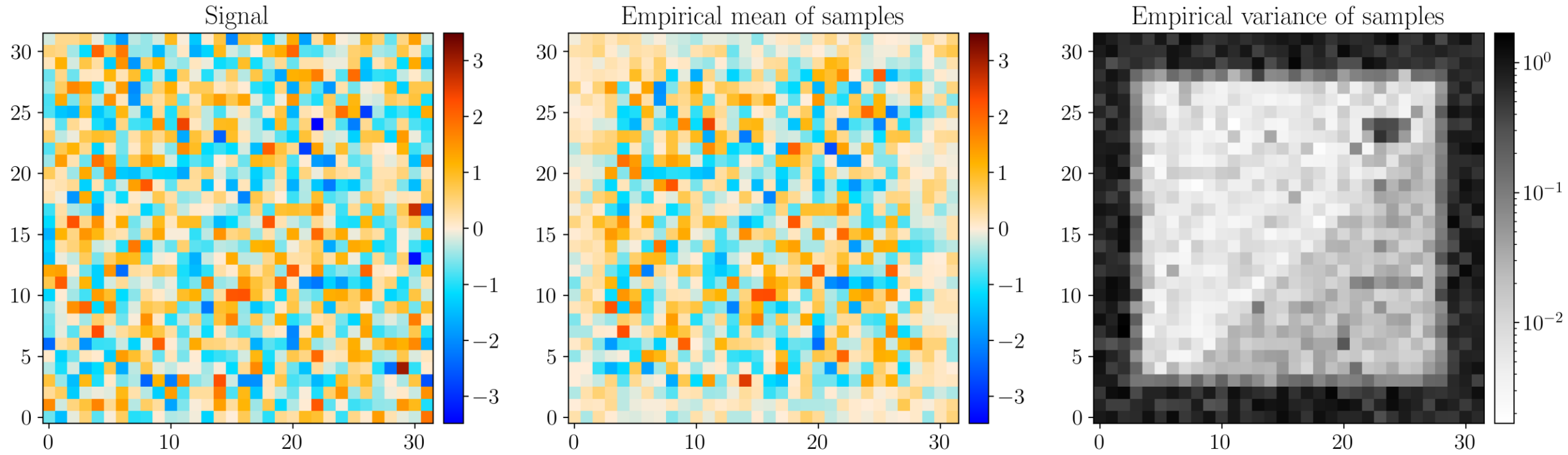
- The effective sample size (ESS) is useful to get a feeling for the number of independent samples (chains are rarely converged in the Gelman-Rubin sense).
- But of course, it is always possible to show the Gelman-Rubin statistic at the field level.



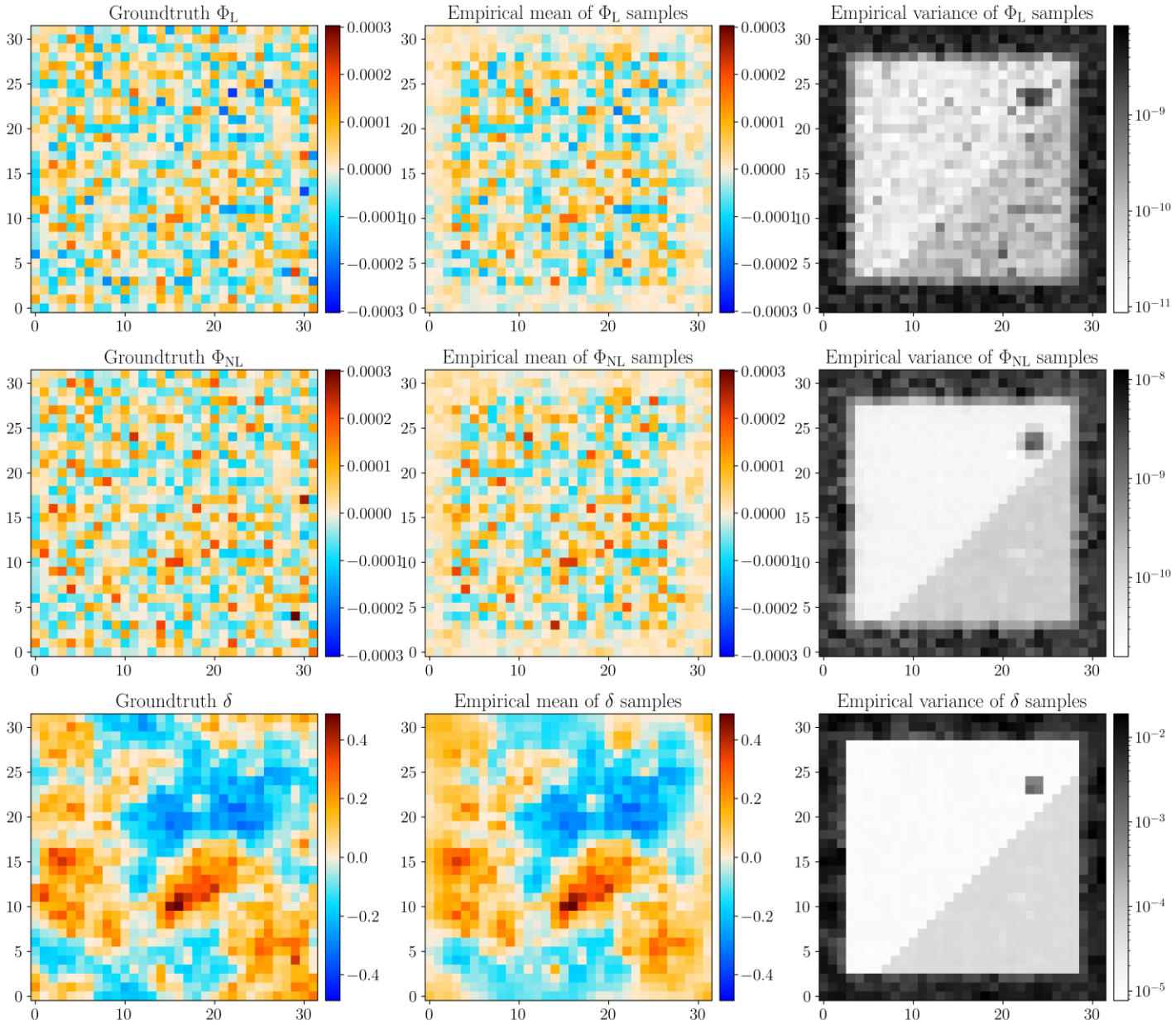


Field-level inference with non-linear models

- Visualise the empirical mean and empirical variance of samples for the reconstructed signal (these are the target parameters of the problem).



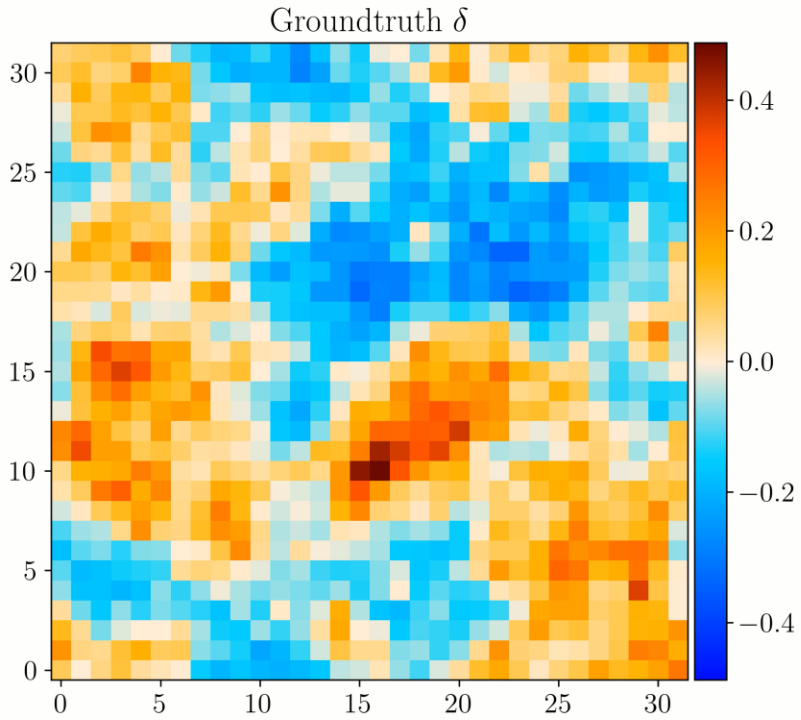
Field-level inference with non-linear models

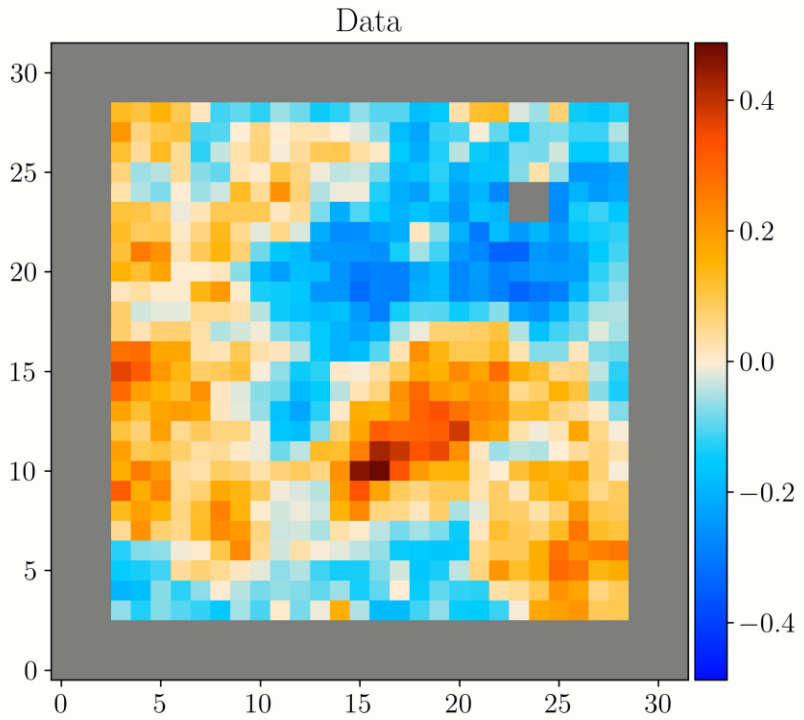


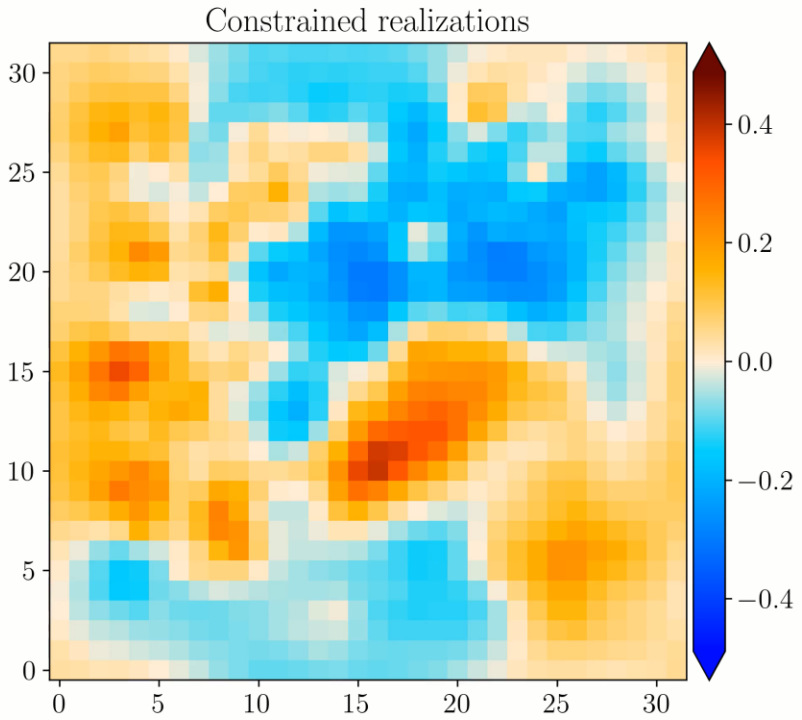
- It is also usual to show the empirical mean and empirical variance of samples for latent fields that have an interpretable physical meaning.

Field-level inference with non-linear models

- Each sample of the chain is a constrained realisation of the corresponding field, given the data.







References and further reading

References:

- A. Gelman *et al.* (2021), *Bayesian Data Analysis, Third edition*
- C. Geyer (2011), *Introduction to Markov Chain Monte Carlo*
- R. M. Neal (2011), 1206.1901, *MCMC using Hamiltonian Dynamics*

https://florent-leclercq.eu/teaching.php