Nested Sampling

Jonathan Pritchard

ICIC
Imperial Centre
for Inference & Cosmology

Bayesian Inference

- Two stages:

  1) Parameter estimation - Posterior

  2) Model selection - Evidence

Inputs | Outputs

$$P(\mathbf{d}|\theta, M) \times P(\theta|M) = P(\theta|\mathbf{d}, M) \times P(\mathbf{d}|M)$$

Likelihood | Prior | Posterior | Evidence

ICIC

Model Selection

- Apply Bayes theorem to models rather than parameters

$$P(\mathcal{M}_i|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathcal{D})},$$

- The normalisation here can be written

$$P(\mathcal{D}) = \sum_j \mathcal{Z}_j \pi_j \qquad \mathcal{Z}_i \equiv P(\mathcal{D}|\mathcal{M}_i) = \int P(\mathcal{D}|\theta, \mathcal{M}_i)P(\theta|\mathcal{M}_i)\mathrm{d}\theta.$$

- So that the posterior of the model can be written in terms of the evidences and priors for the models

$$P(\mathcal{M}_i|\mathcal{D}) = \frac{\mathcal{Z}_i \pi_i}{\sum_j \mathcal{Z}_j \pi_j}.$$

ICIC

Model Selection

- For uniform priors on the models, we prefer a model with a larger Evidence

$$\frac{P(\mathcal{M}_1|\mathcal{D})}{P(\mathcal{M}_2|\mathcal{D})} = \frac{\mathcal{Z}_1}{\mathcal{Z}_2}\frac{\pi_1}{\pi_2}$$

- Evidence is key for Bayesian model selection!
- How can we calculate the evidence?

The evidence

- Evidence is integral over likelihood and prior

$$P(\mathcal{D}|\mathcal{M}) \equiv \mathcal{Z} = \int P(\mathcal{D}|\theta, \mathcal{M}) P(\theta|\mathcal{M}) \mathrm{d}\theta.$$

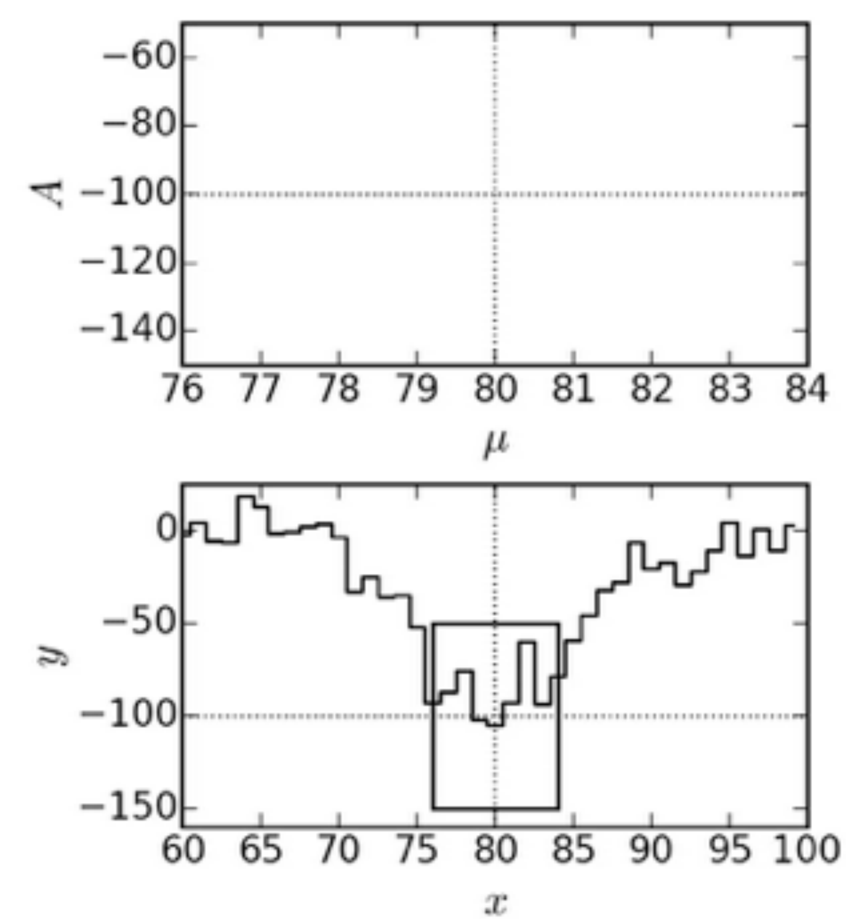
$$\mathcal{Z} = \int \mathcal{L}(\theta) \pi(\theta) \mathrm{d}\theta.$$

- Typically the integral is in a high-dimensional space, but only a small region contributes significantly to integral. Need to find it!

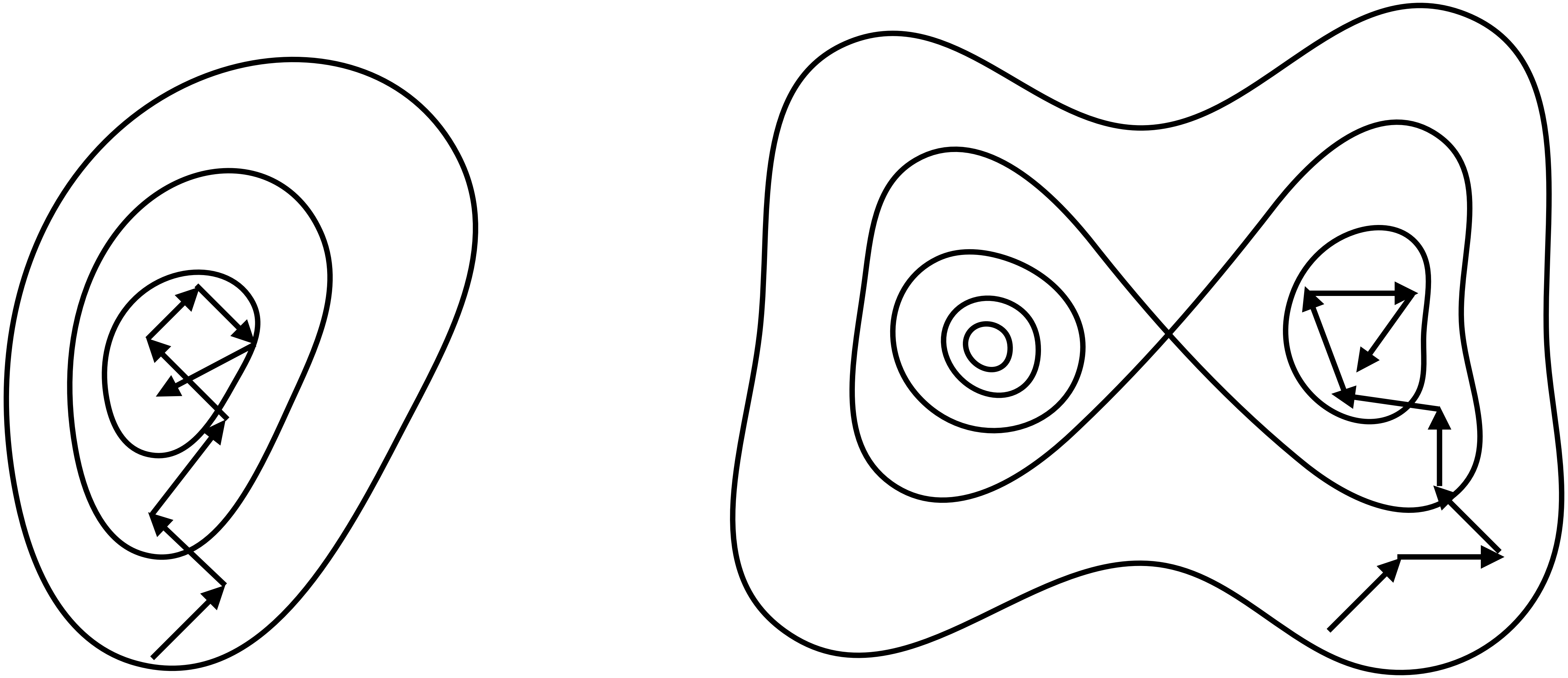
ICIC

Limitations of MCMC

- MCMC with Metropolis-Hastings typically focusses in on peak of posterior and explores in that vicinity
- Low sampling in tails of distribution. Not a problem for parameter estimation, but can be when calculating evidence.
- Difficult to handle multimodal posterior distributions
  - may get trapped



Multimodal likelihood



- MH may get trapped in local maximum without exploring full likelihood shape

ICIC

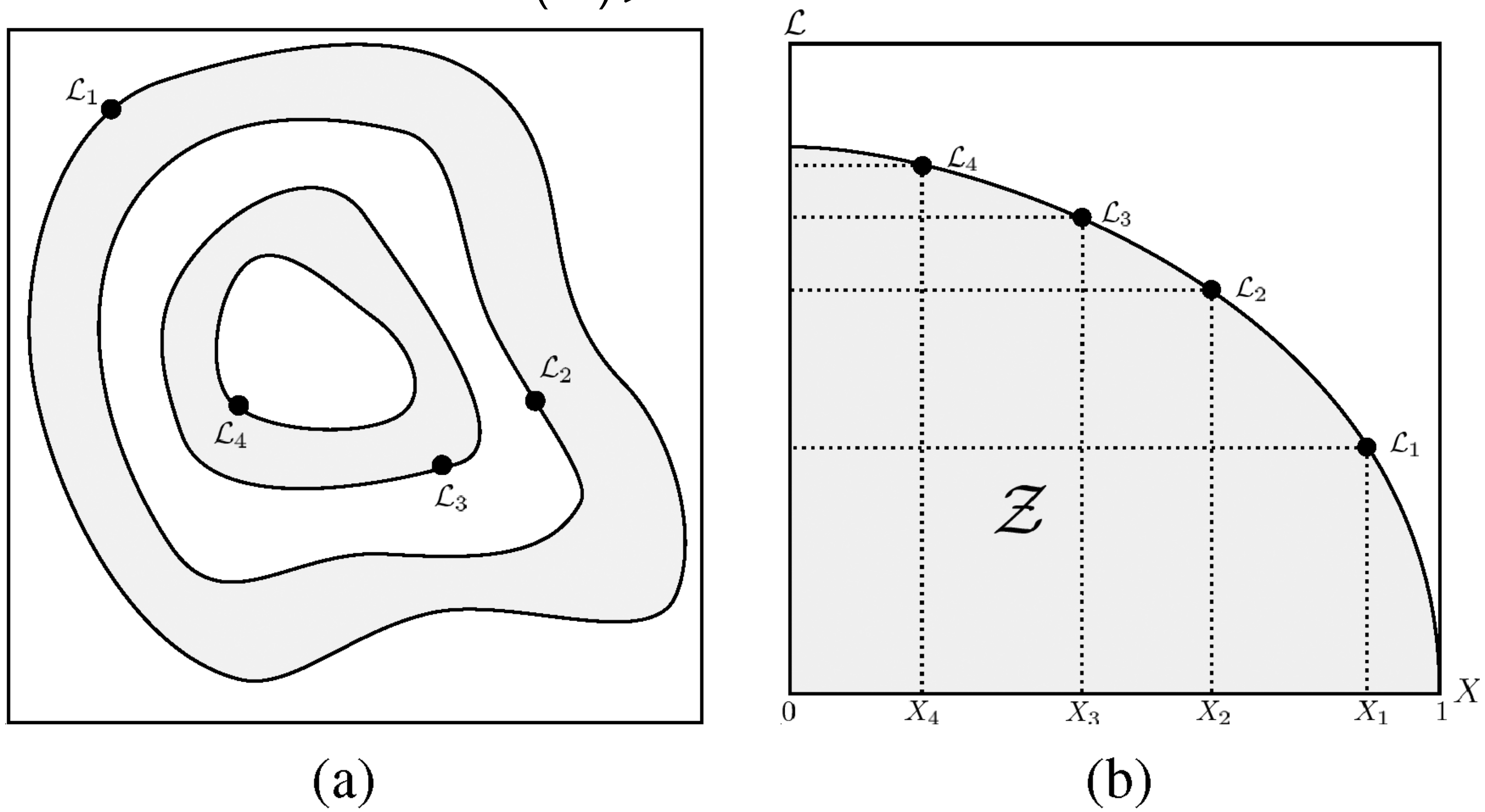
Nested Sampling

- Goal of efficiently evaluating evidence and returning posterior estimate.

ICIC

Nested Sampling

- Imagine ordering set of likelihood points
- Introduces *prior volume:* fraction of prior contained within an iso-likelihood contour
- $$X(\mathcal{L}) = \int_{\mathcal{L}(\theta)>\mathcal{L}} \pi(\theta)\mathrm{d}\theta$$



L(X)

ICIC

- Use to transform evidence calculation from multidimensional integral to a 1D integral

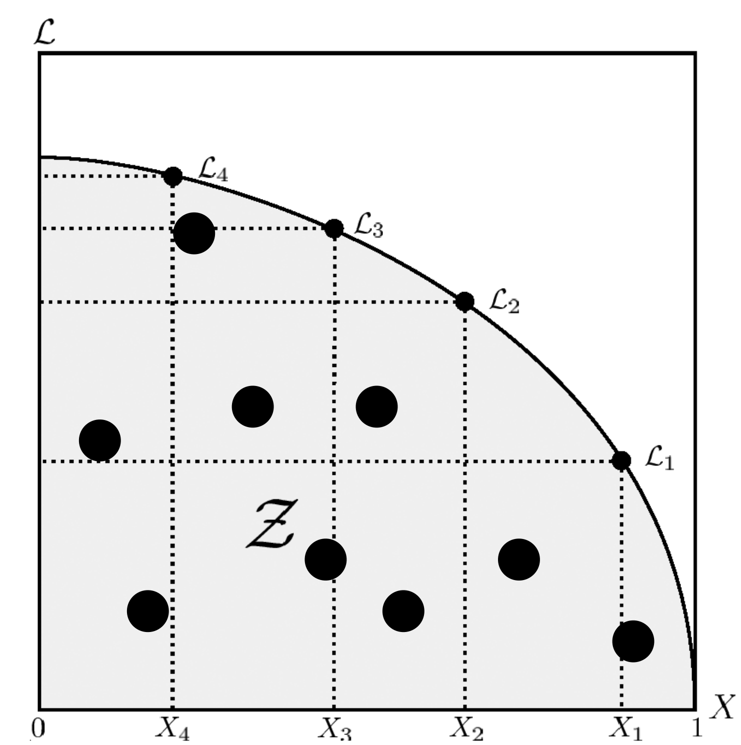
$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)\mathrm{d}\theta. \longrightarrow \mathcal{Z} = \int_0^1 \mathcal{L}(X)\mathrm{d}X.$$

- Ordered L(X) then gives evidence via 1D integration e.g. via quadrature

$$\mathcal{Z} = \sum_{i=1}^{M} \mathcal{L}_i w_i \qquad w_i = \tfrac{1}{2}(X_{i-1} - X_{i+1})$$

- Points chosen randomly from region L(X) are representative of posterior

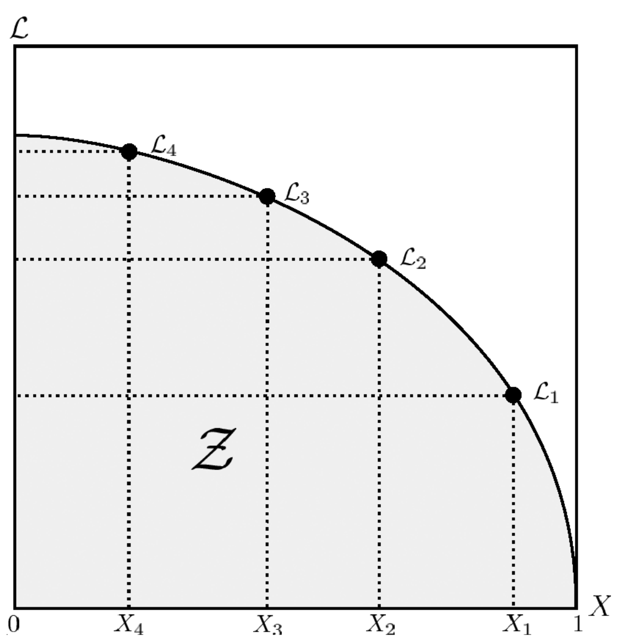
$$P(X_i) = \frac{\mathcal{L}(X_i) w_i}{\mathcal{Z}}$$

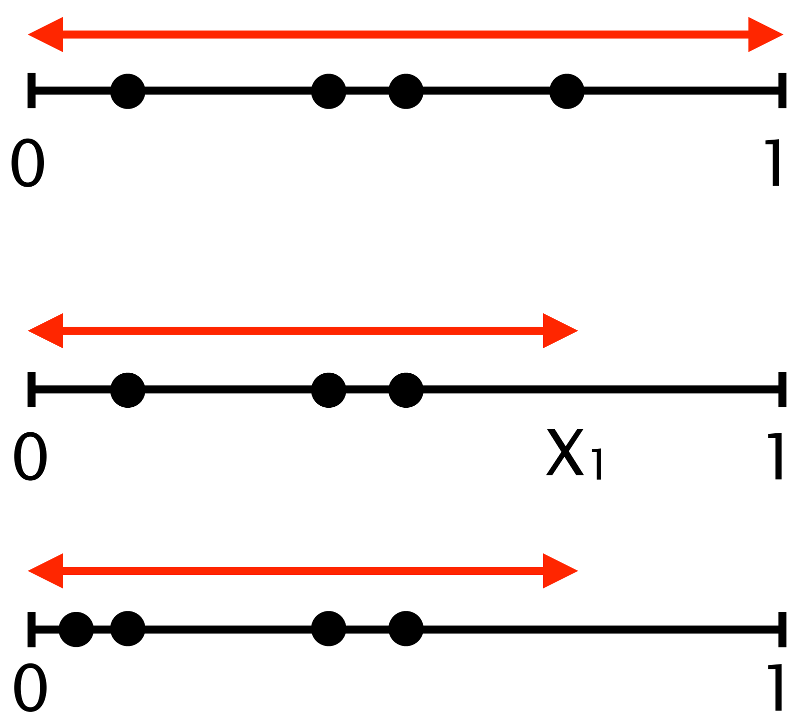


ICIC

Nested Sampling

- Uniformly sample from prior maintaining a population of live points that is updated so that they contract around the peak(s) of posterior





4 points uniformly sampled from prior (equivalent to X)

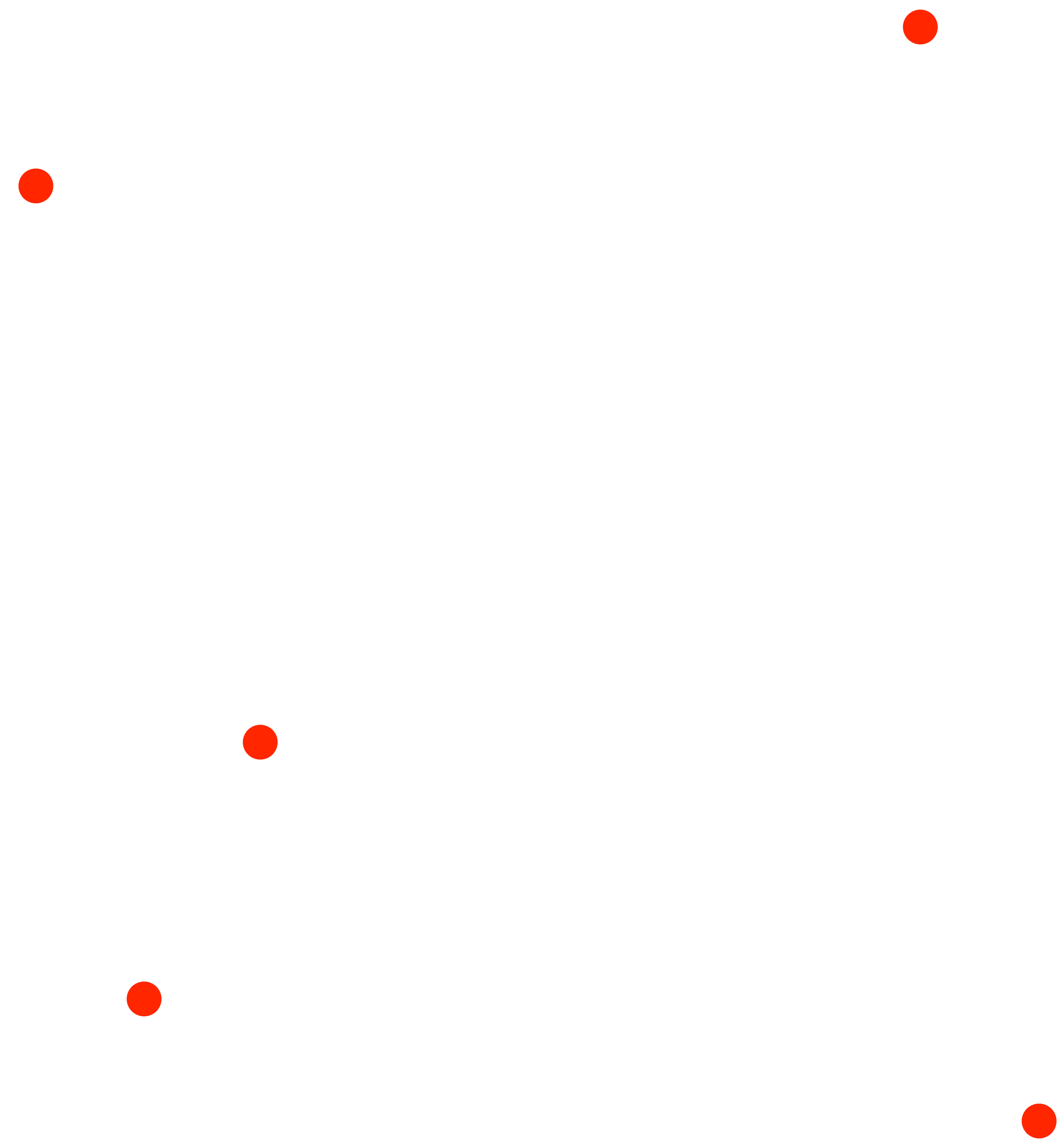
Store worst point X_1 i.e. lowest likelihood

Generate a new point from uniform dist on $[0, X_1]$

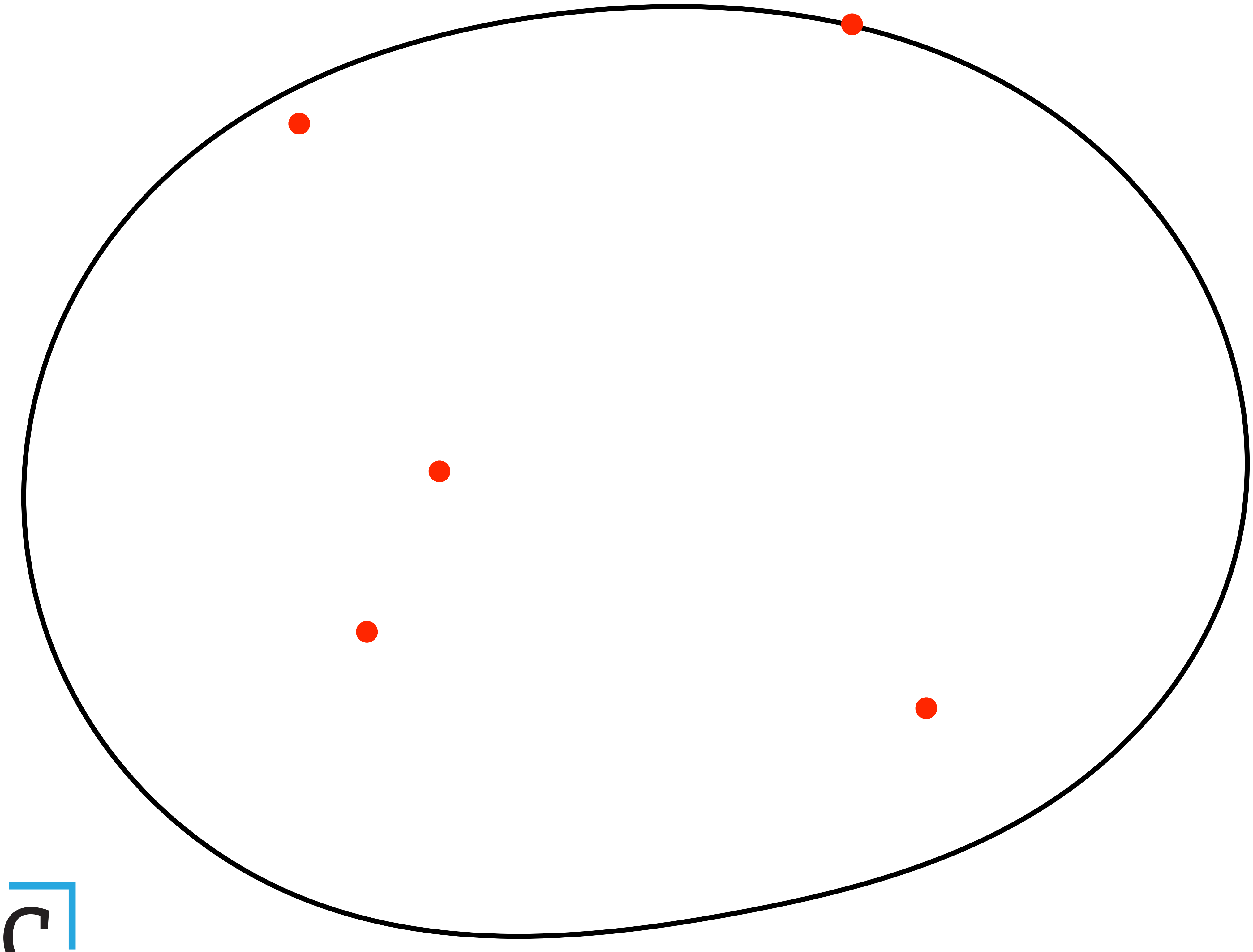
- Assign X values on basis of statistics of uniform dist

$$\log X_i \approx -(i \pm \sqrt{i})/N$$

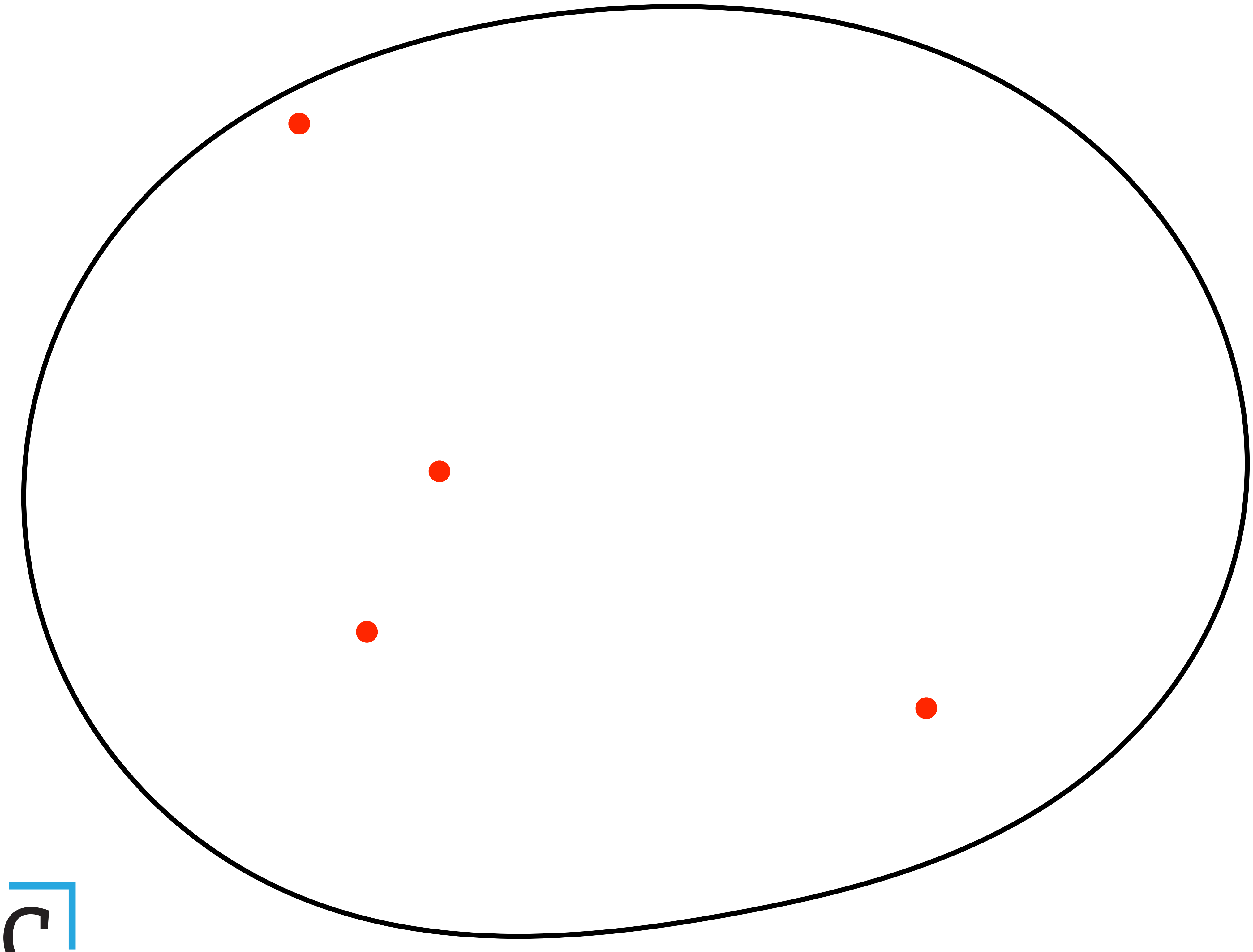
Points exponentially hone in on high L(X) as $X_k \sim \exp(-k/n)$ for n points five live points chosen uniformly from prior



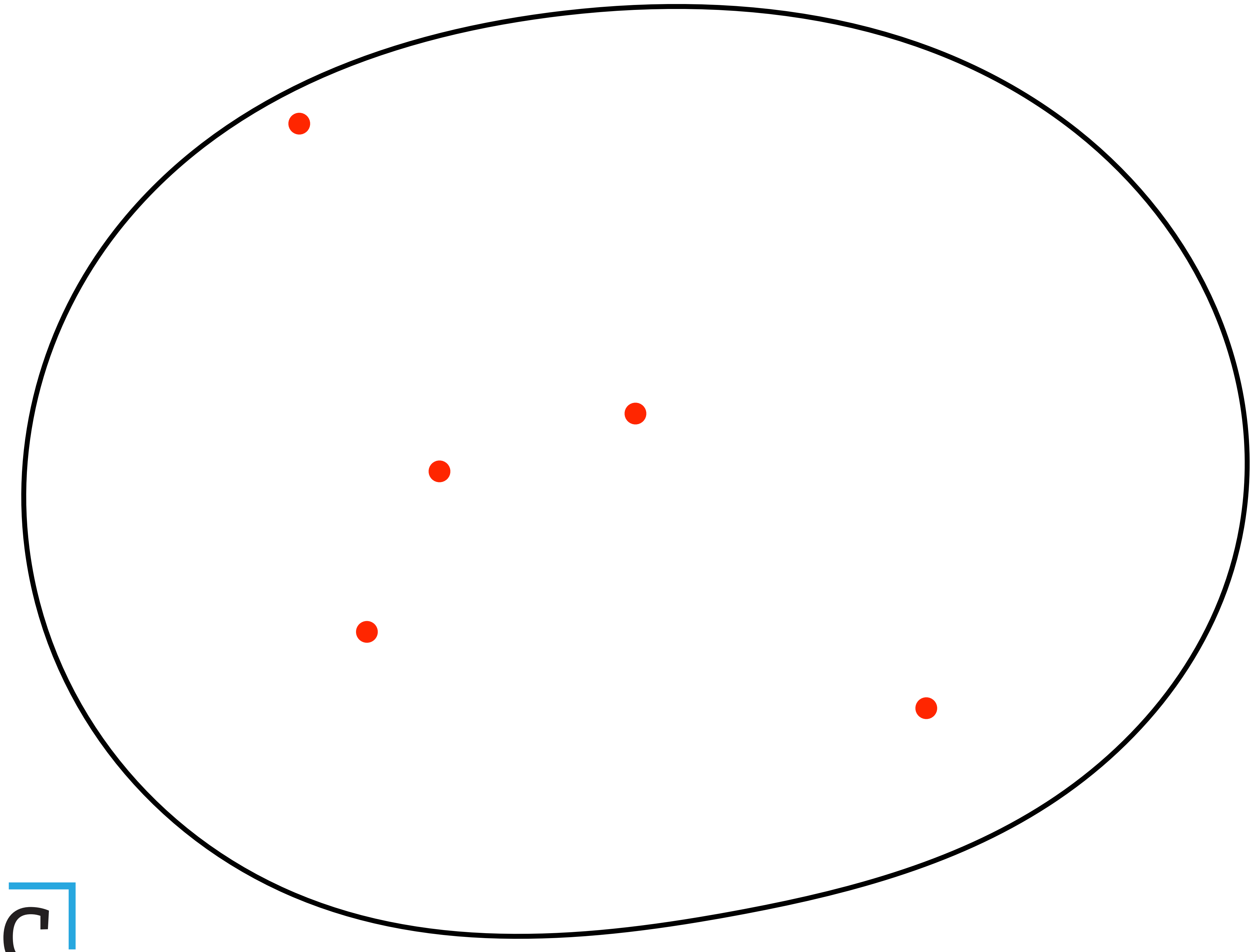
Define likelihood contour from lowest point



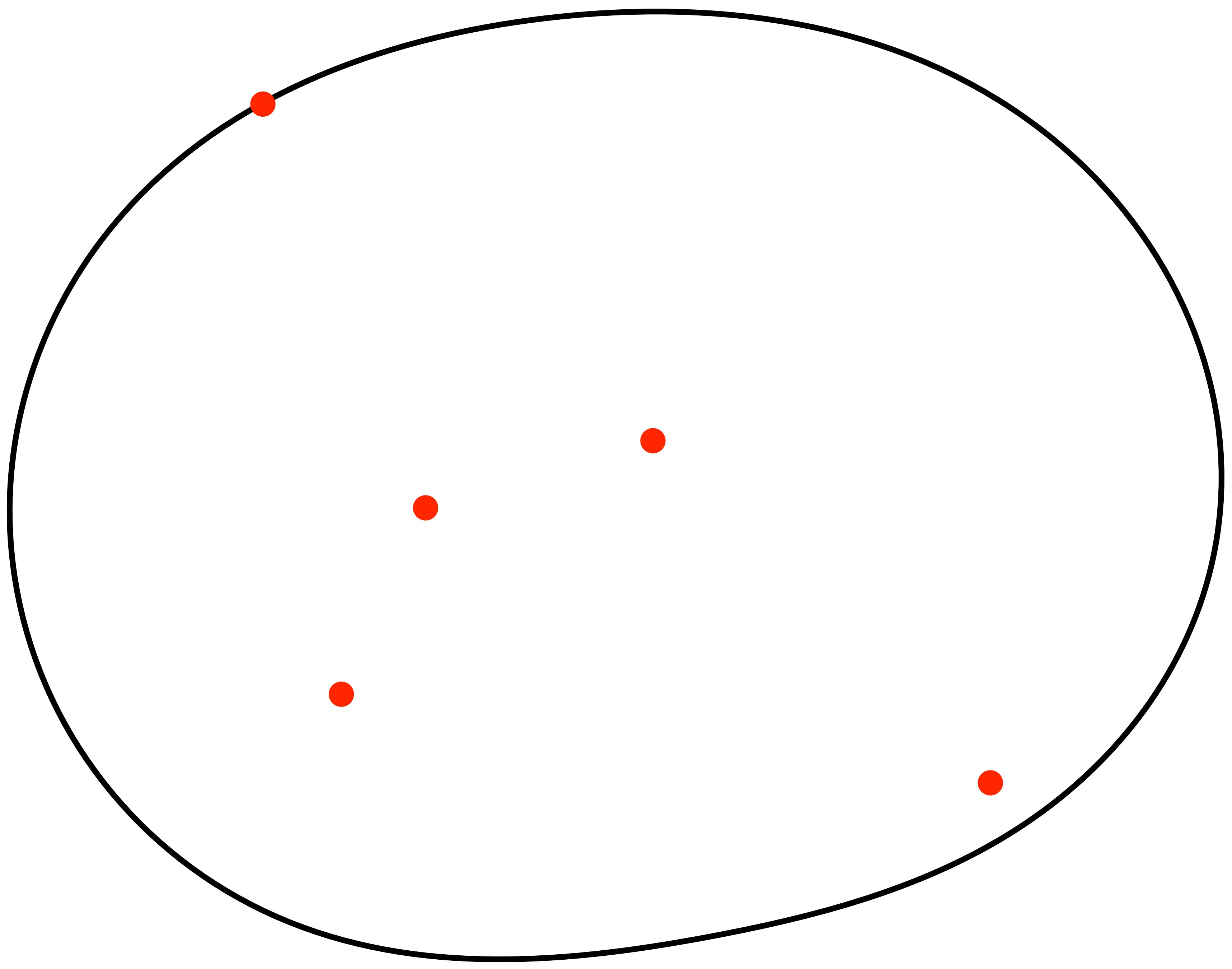
Delete that point (storing it's value)

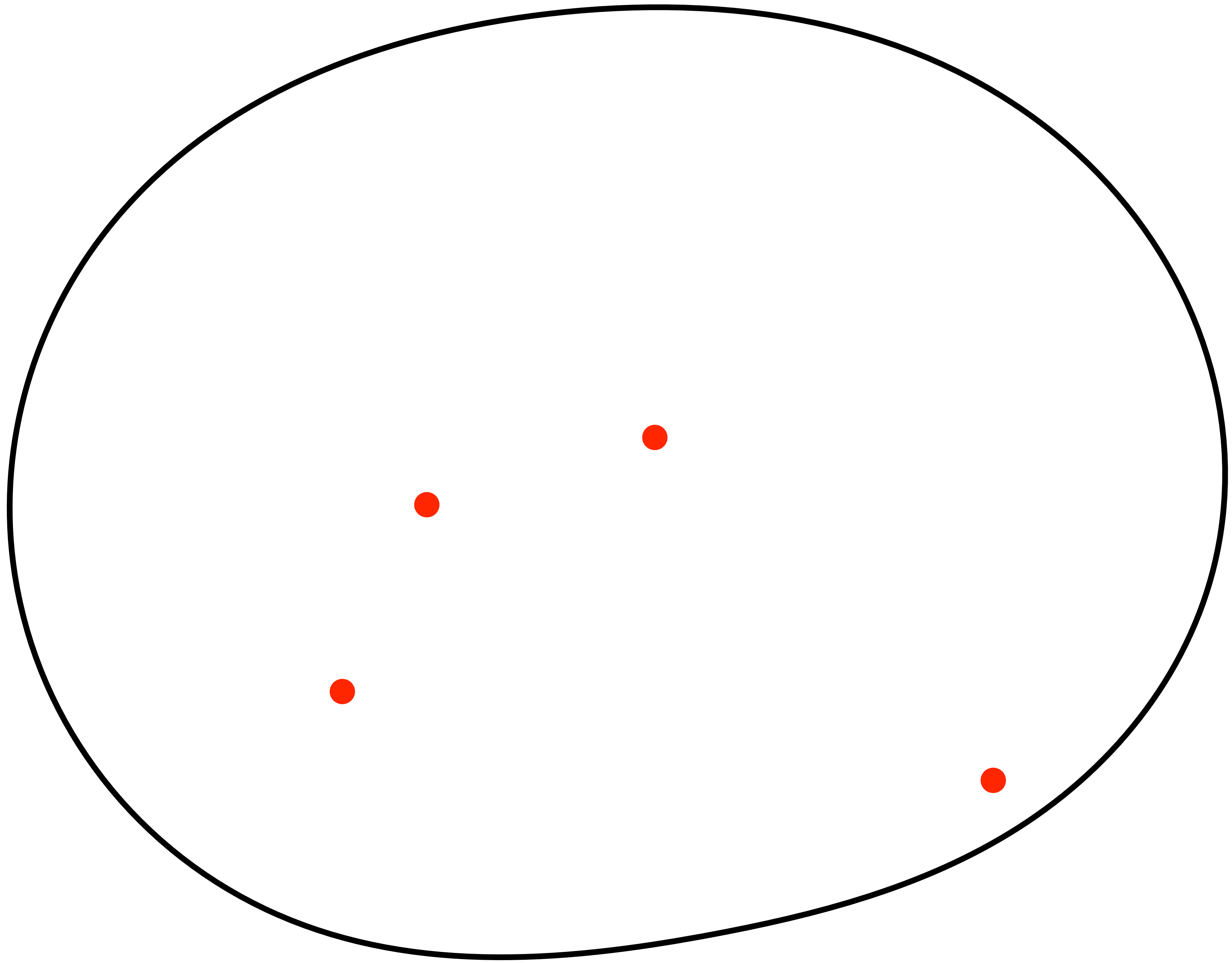


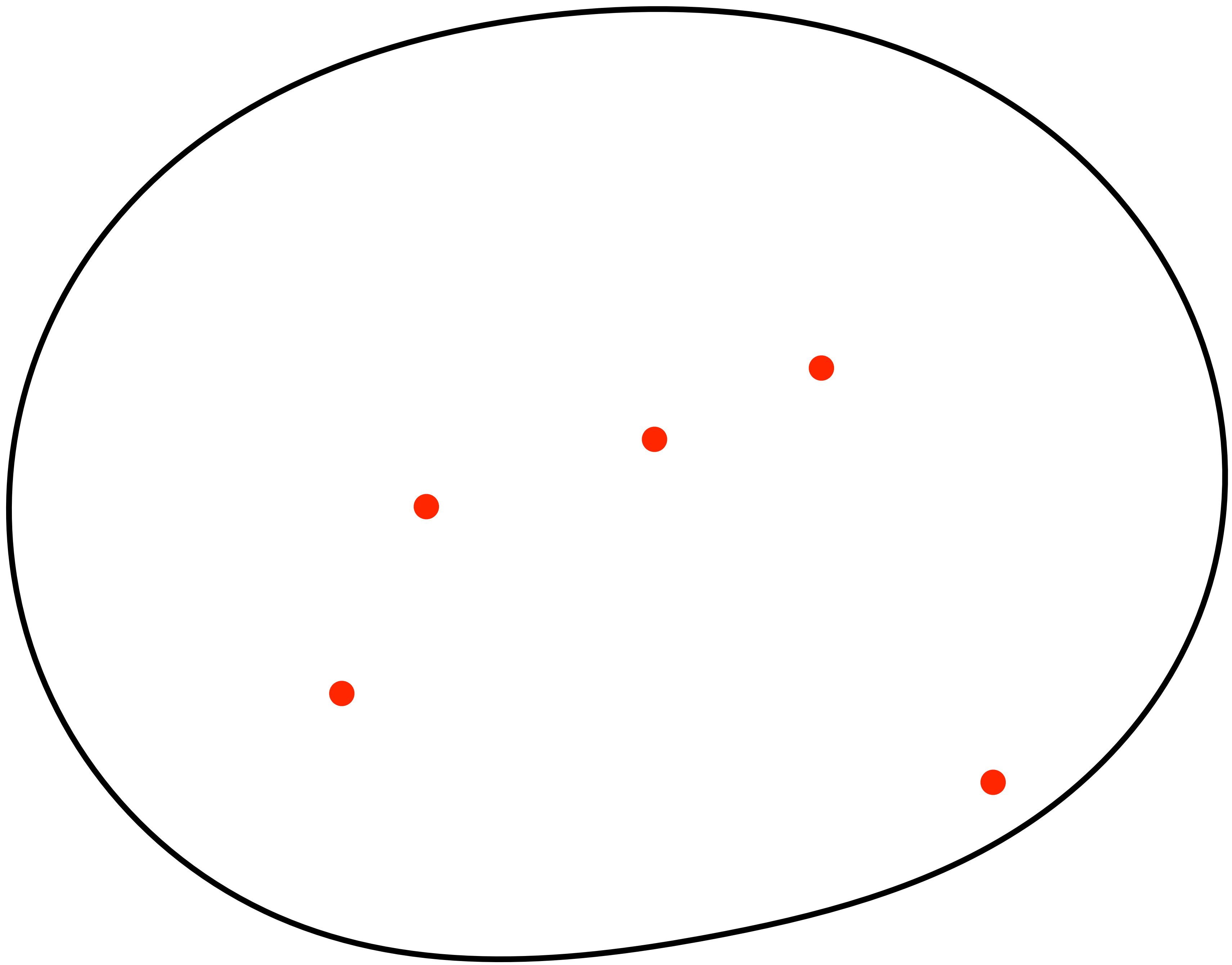
Select a new point uniformly sampled subject to requirement $L(X_{new}) > L(X_{old})$



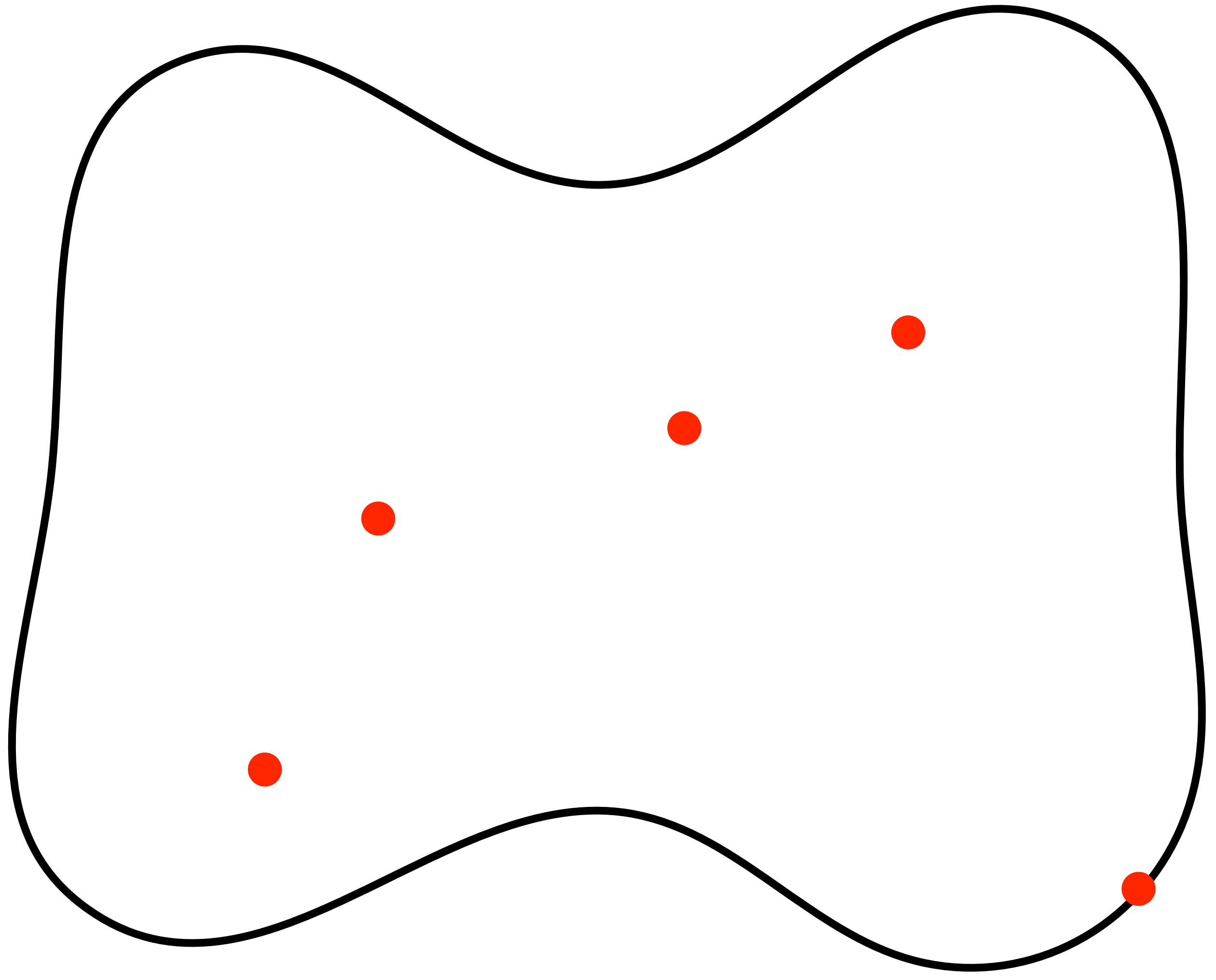
Iterate - contour shrinks by X~exp(1/n)



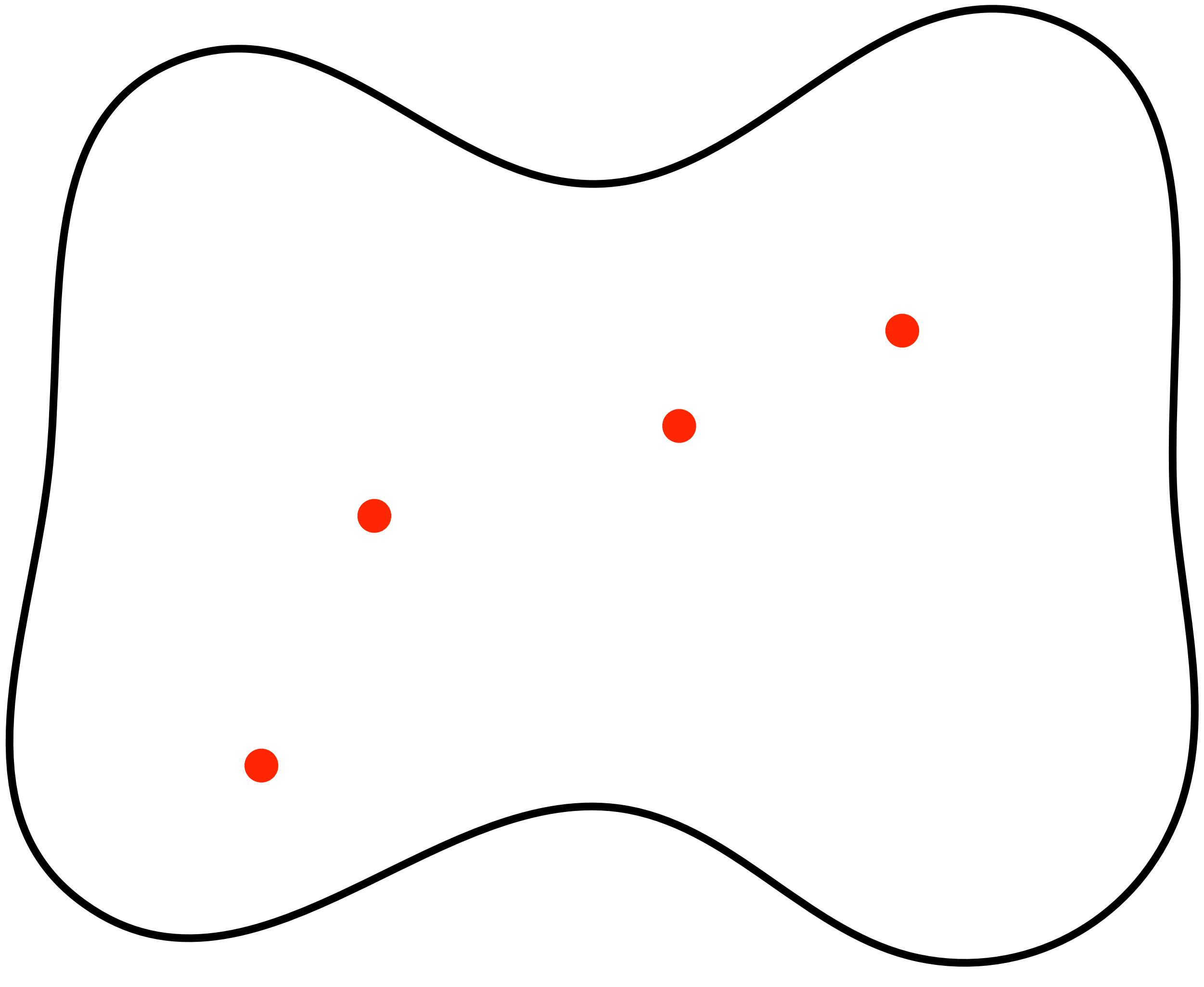


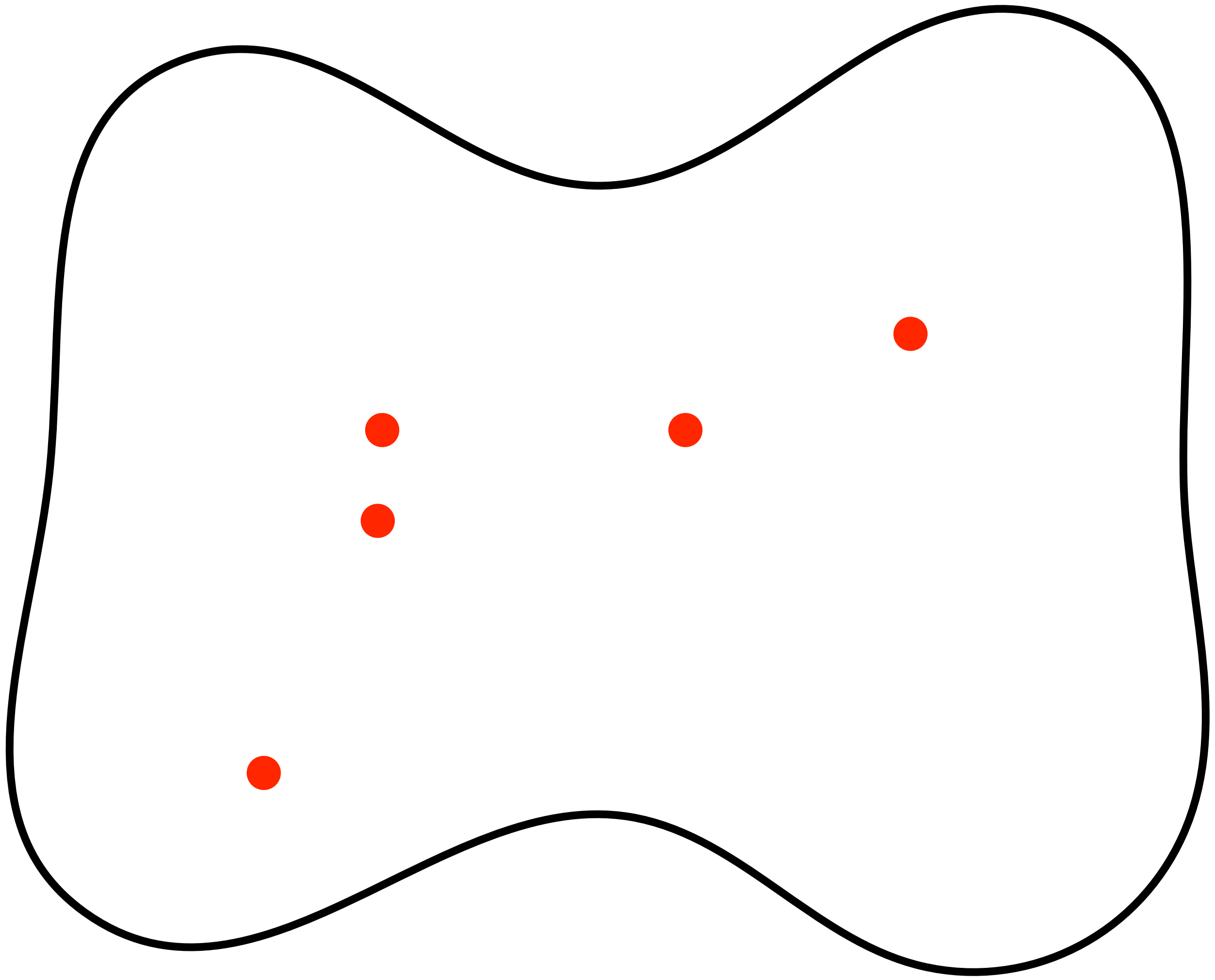


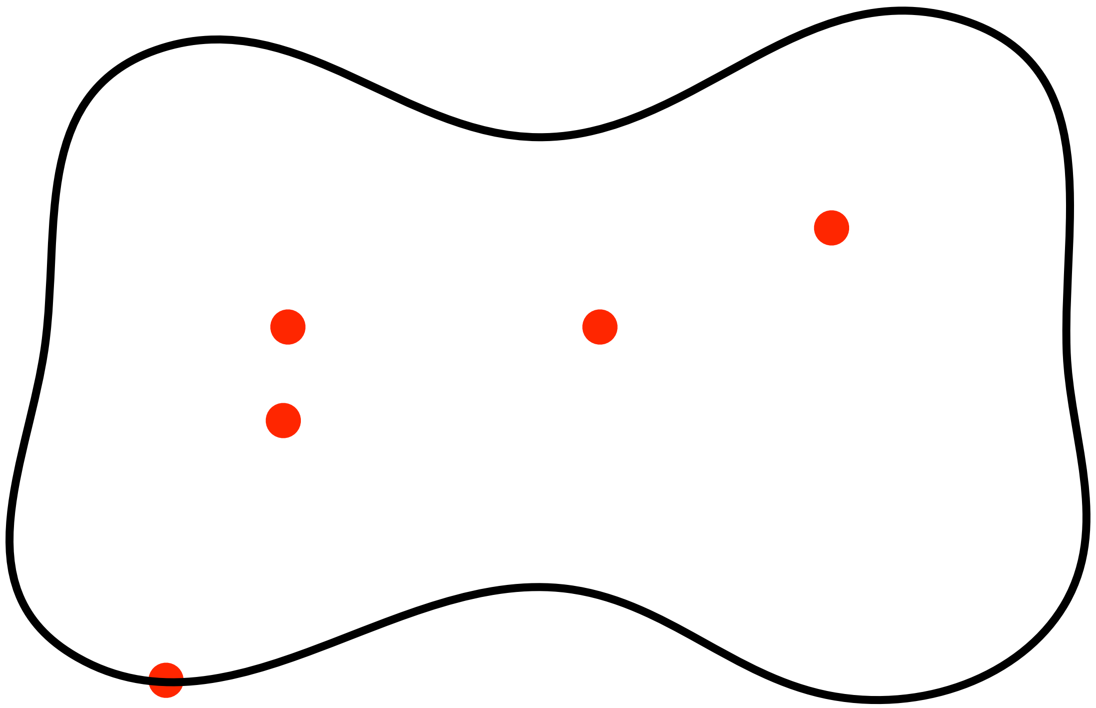
ICIC



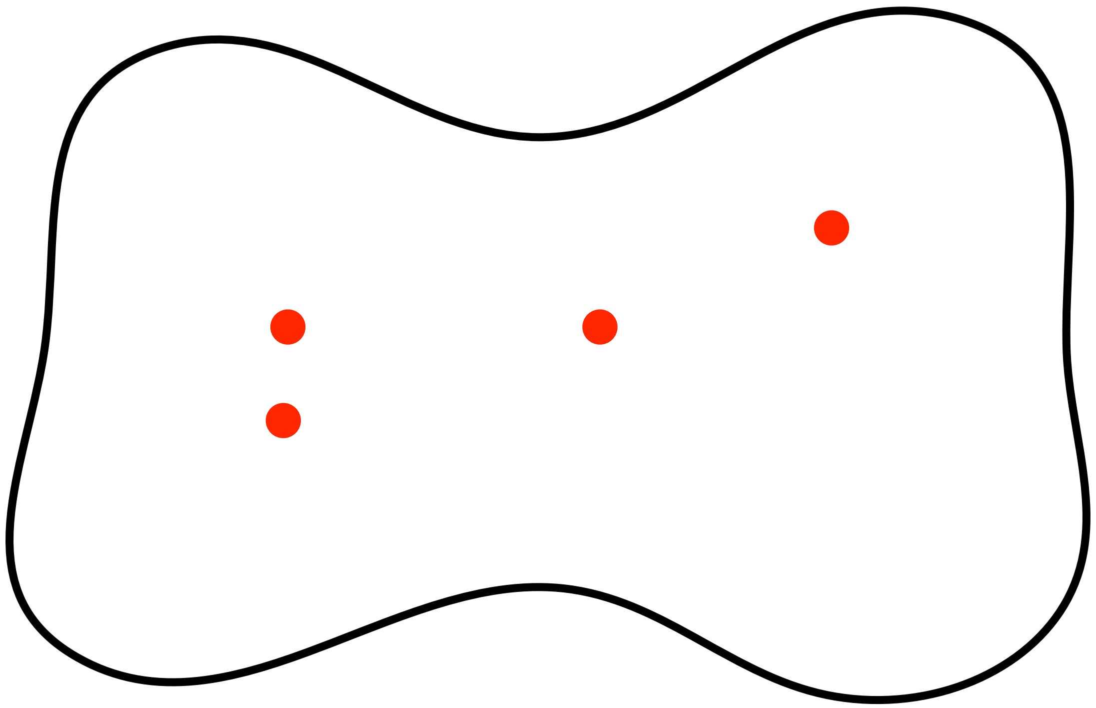
ICIC

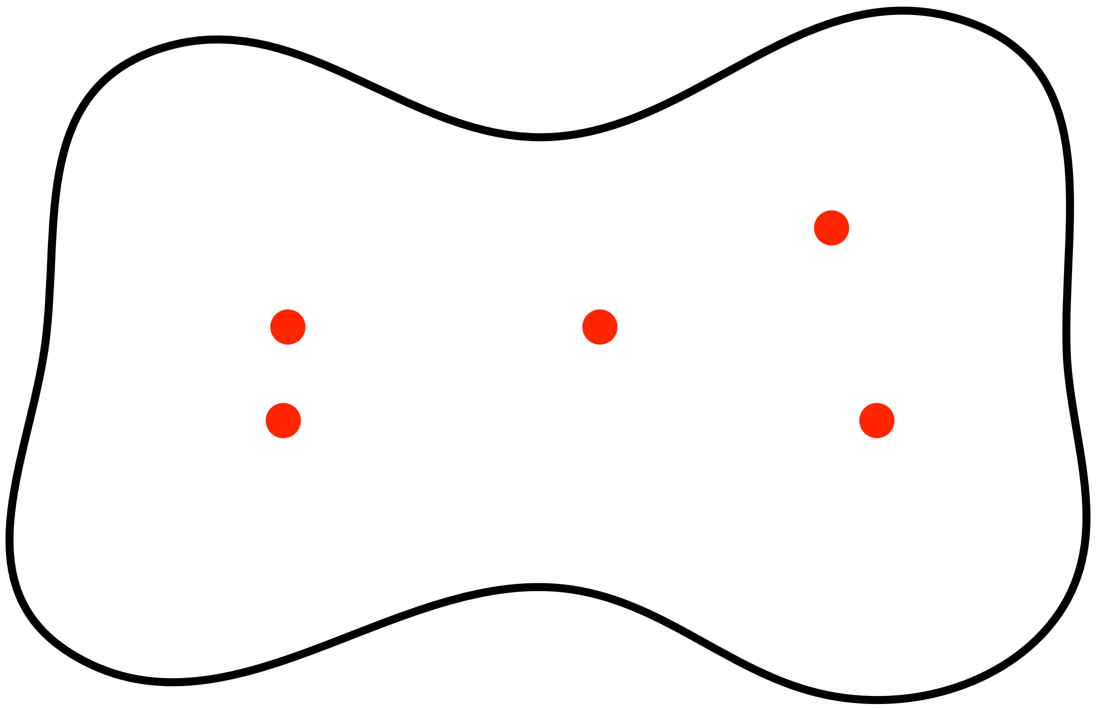


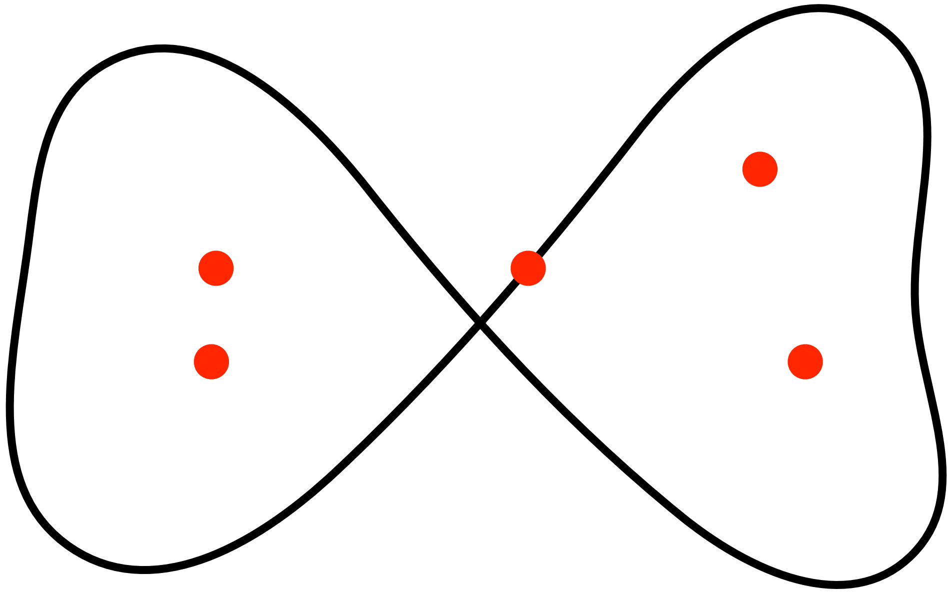


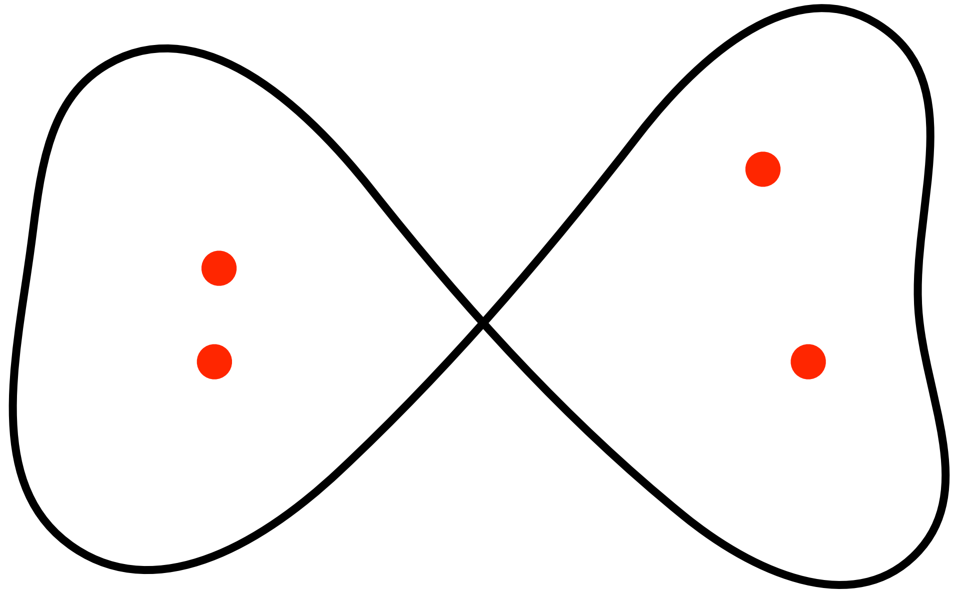


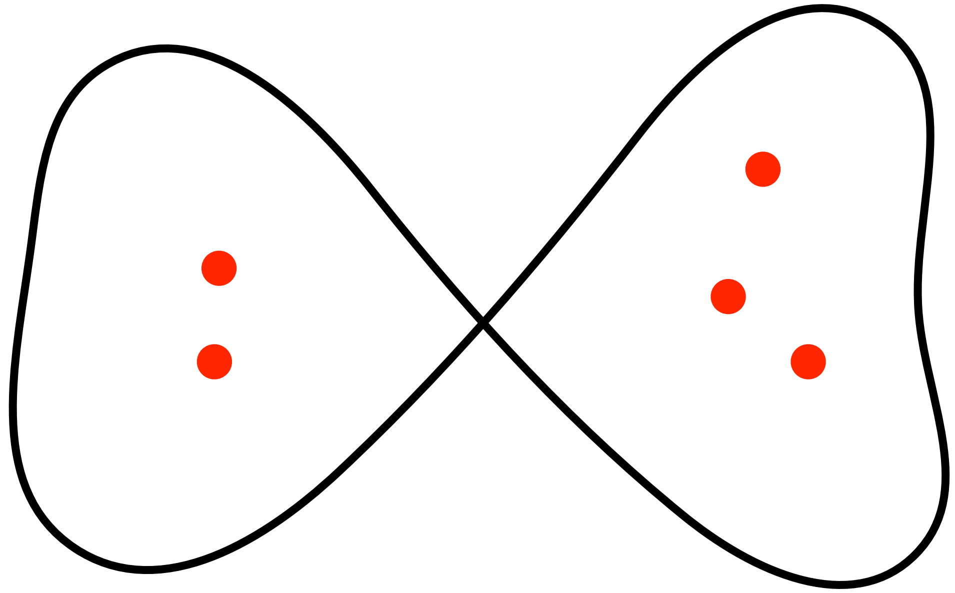
ICIC

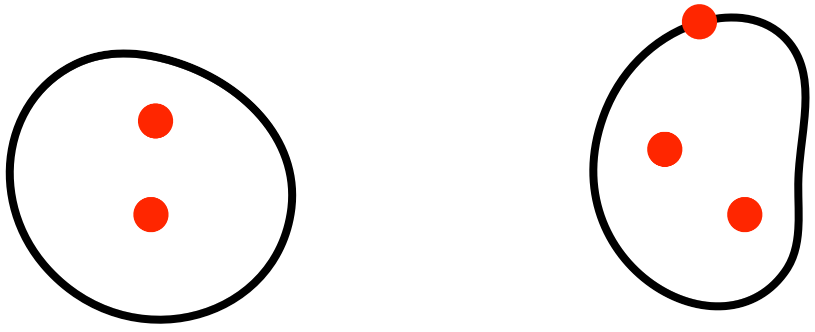


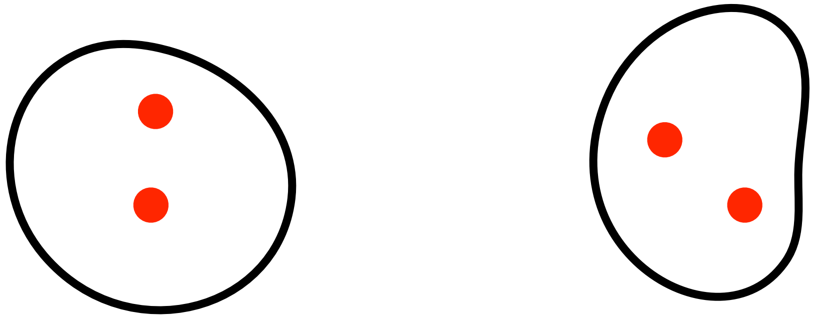




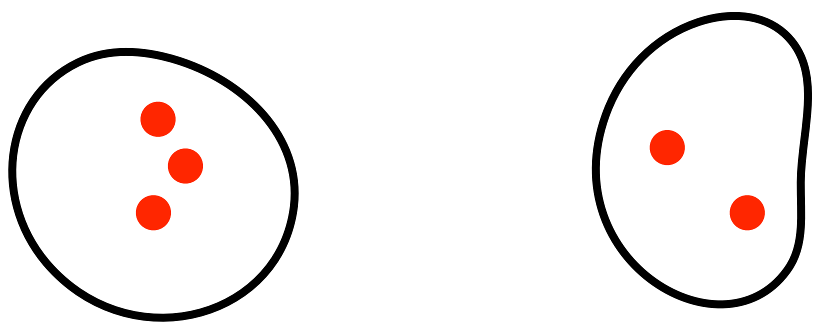




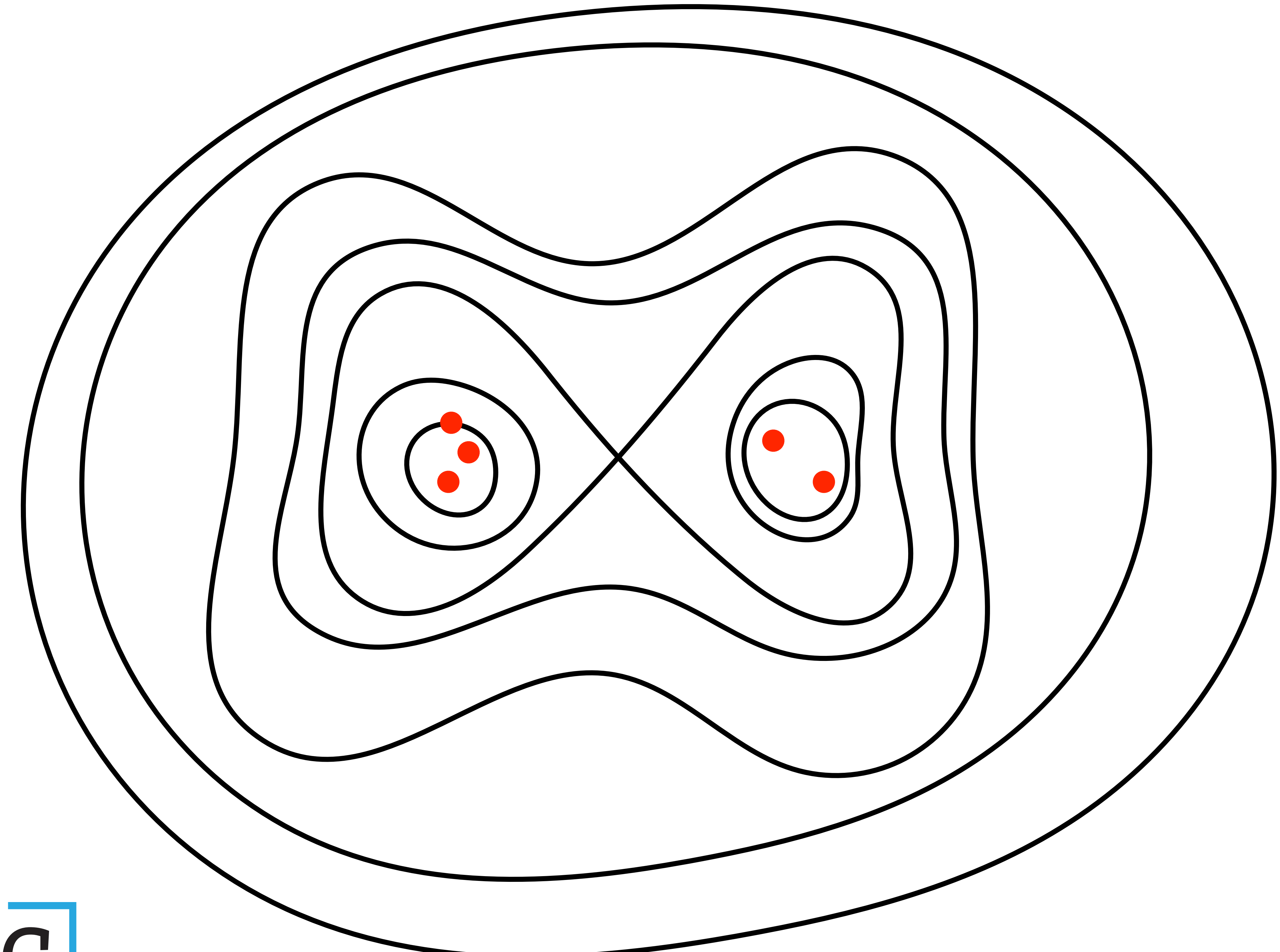




ICIC



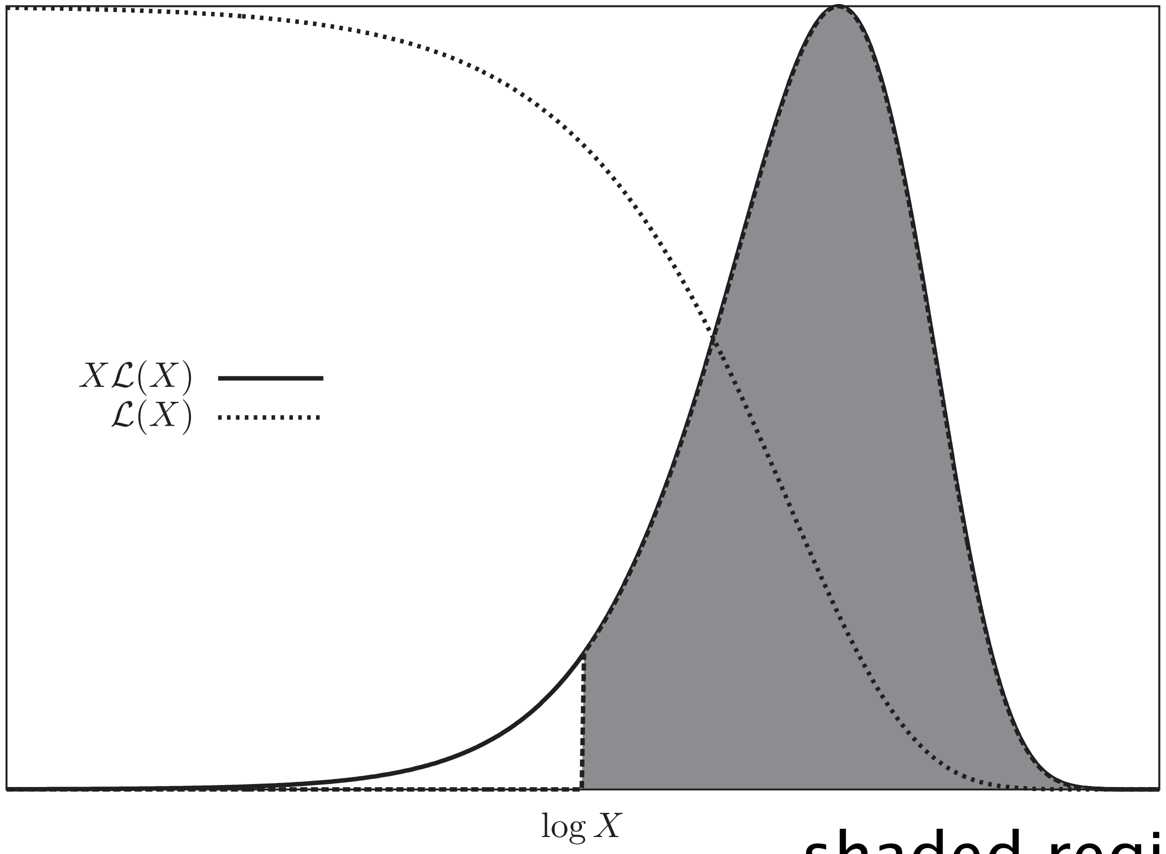
Successive points map likelihood in ordered way



Stopping criteria

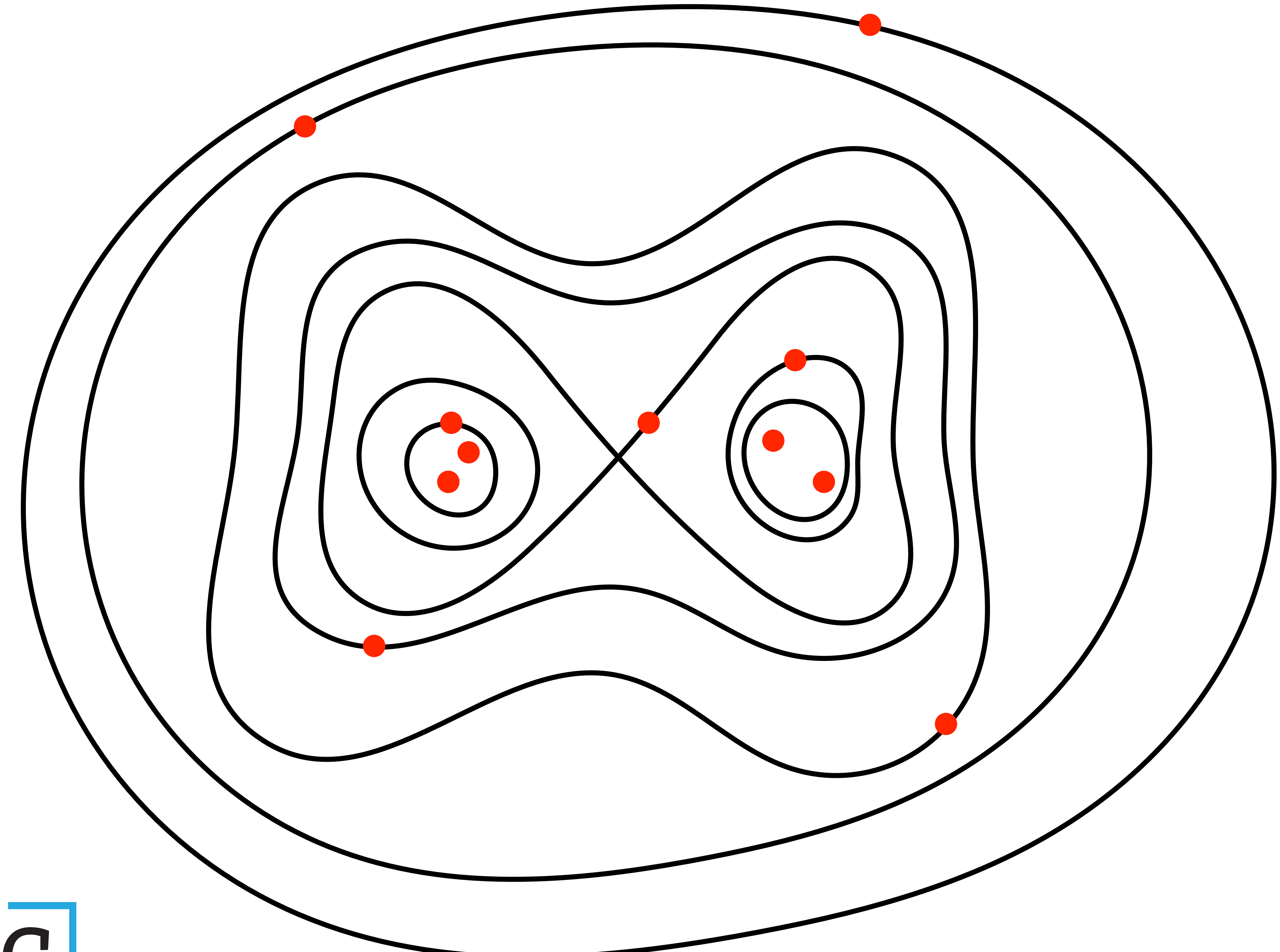
- Can decide to stop based on error estimate on evidence
- Likelihood increases, but separation of points decreases, so contribution to integral converges

$$\Delta \mathcal{Z}_{\mathrm{i}} = \mathcal{L}_{\max} X_{\mathrm{i}}$$



shaded region is running contribution to evidence

Ultimately left with a set of points with known $\{\theta_i, L_i\}$, inferred X_i and estimate of evidence Z



Incomplete algorithm

- Uniform sampling from prior subject to L(X)>λ is not straightforward.
- e.g. Ellipsoidal rejection sampling (MultiNest)

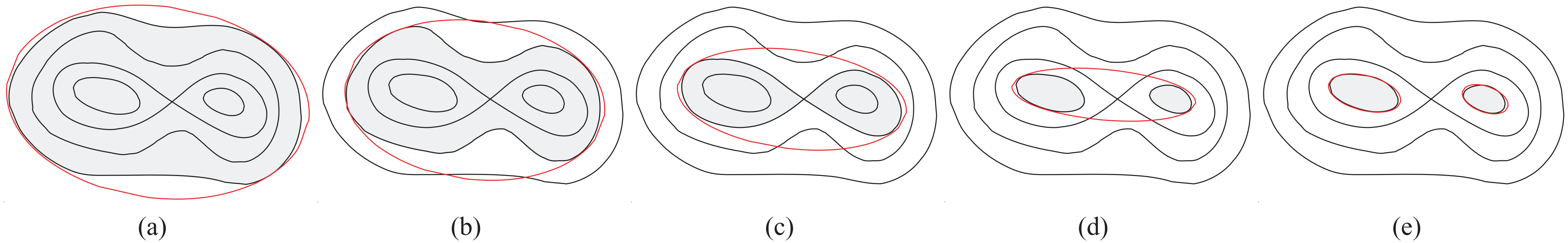
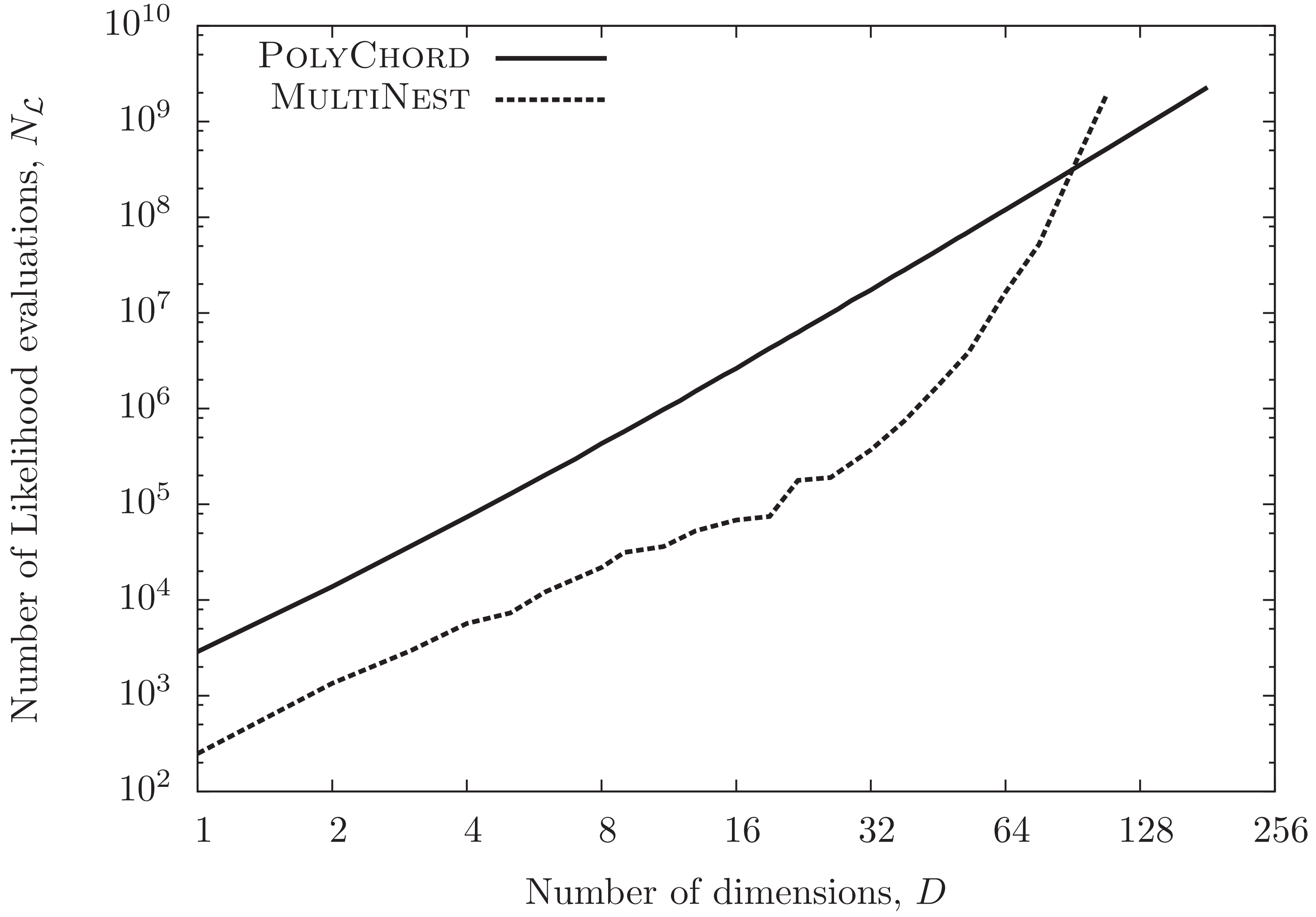


Figure 2. Cartoon of ellipsoidal nested sampling from a simple bimodal distribution. In (a) we see that the ellipsoid represents a good bound to the active region. In (b)–(d), as we nest inwards we can see that the acceptance rate will rapidly decrease as the bound steadily worsens. (e) illustrates the increase in efficiency obtained by sampling from each clustered region separately.

ICIC

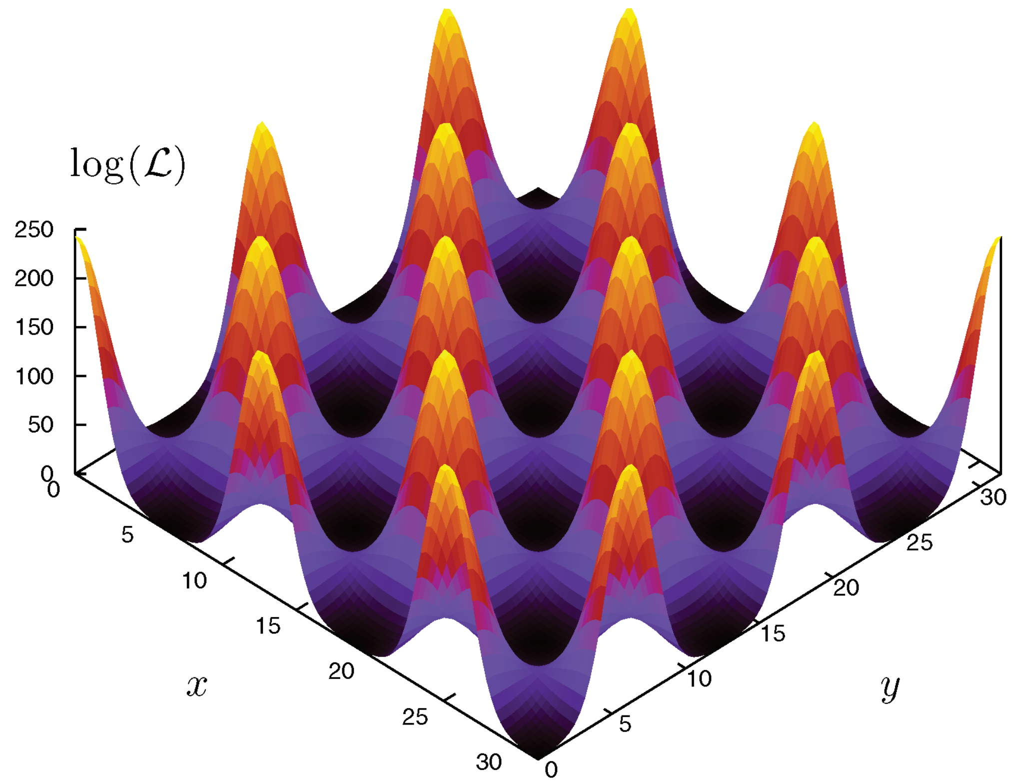
Codes

- Two examples:
  MultiNest - ellipsoidal rejection sampling
  PolyChord - slice sampling

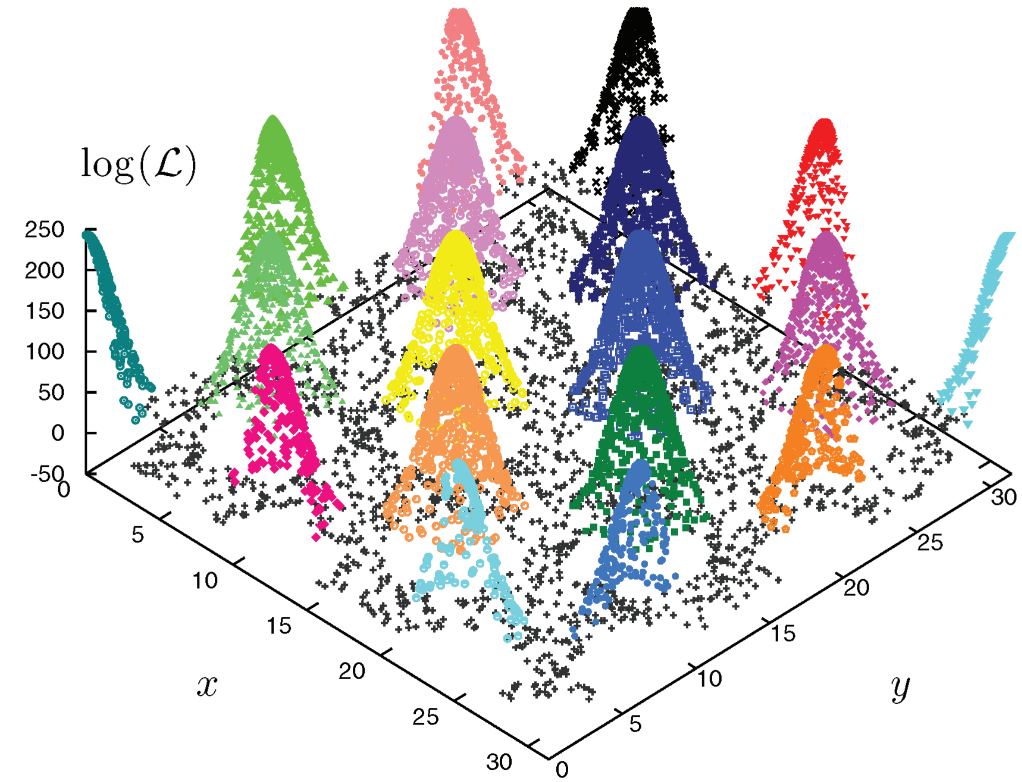


(python wrapper - pymultinest)

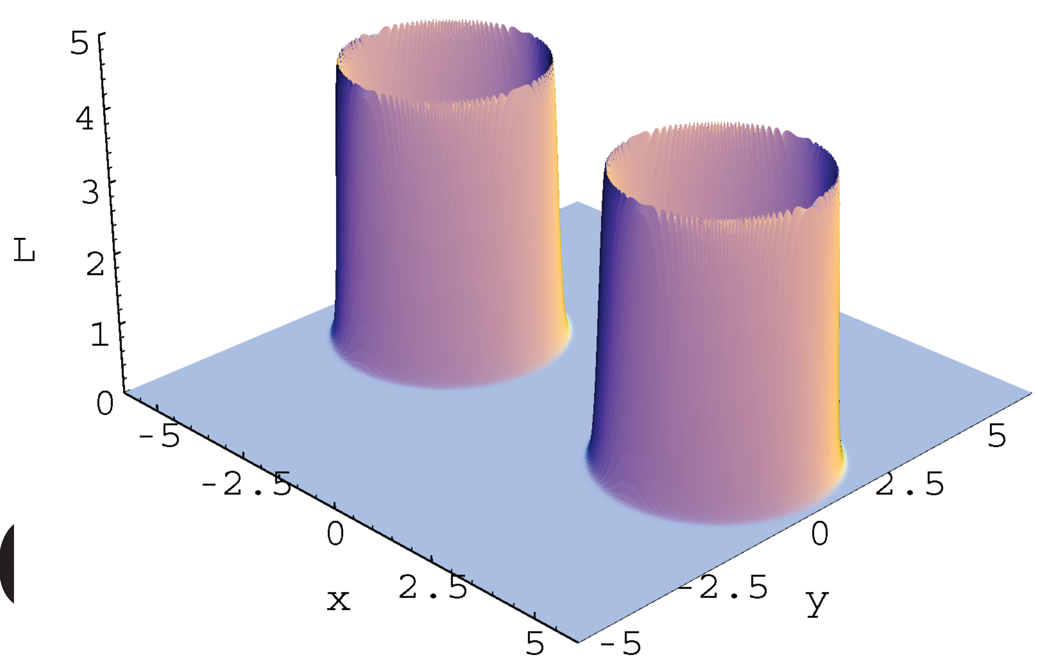
MultiNest



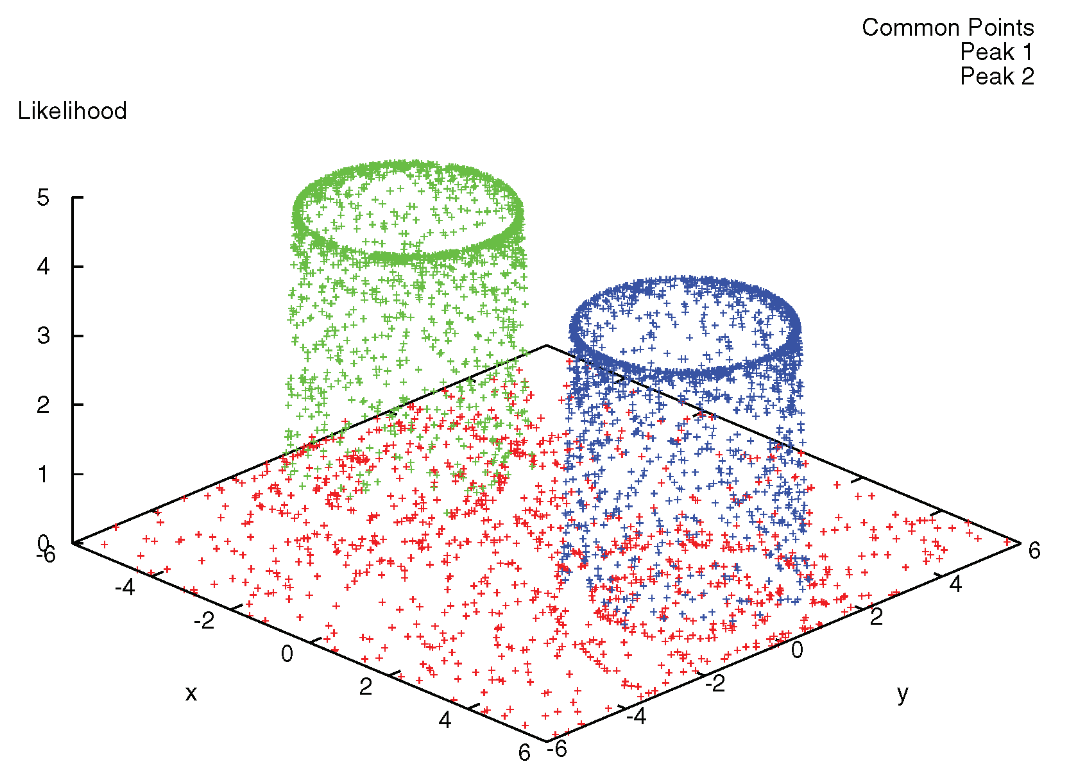
(a)



(b)



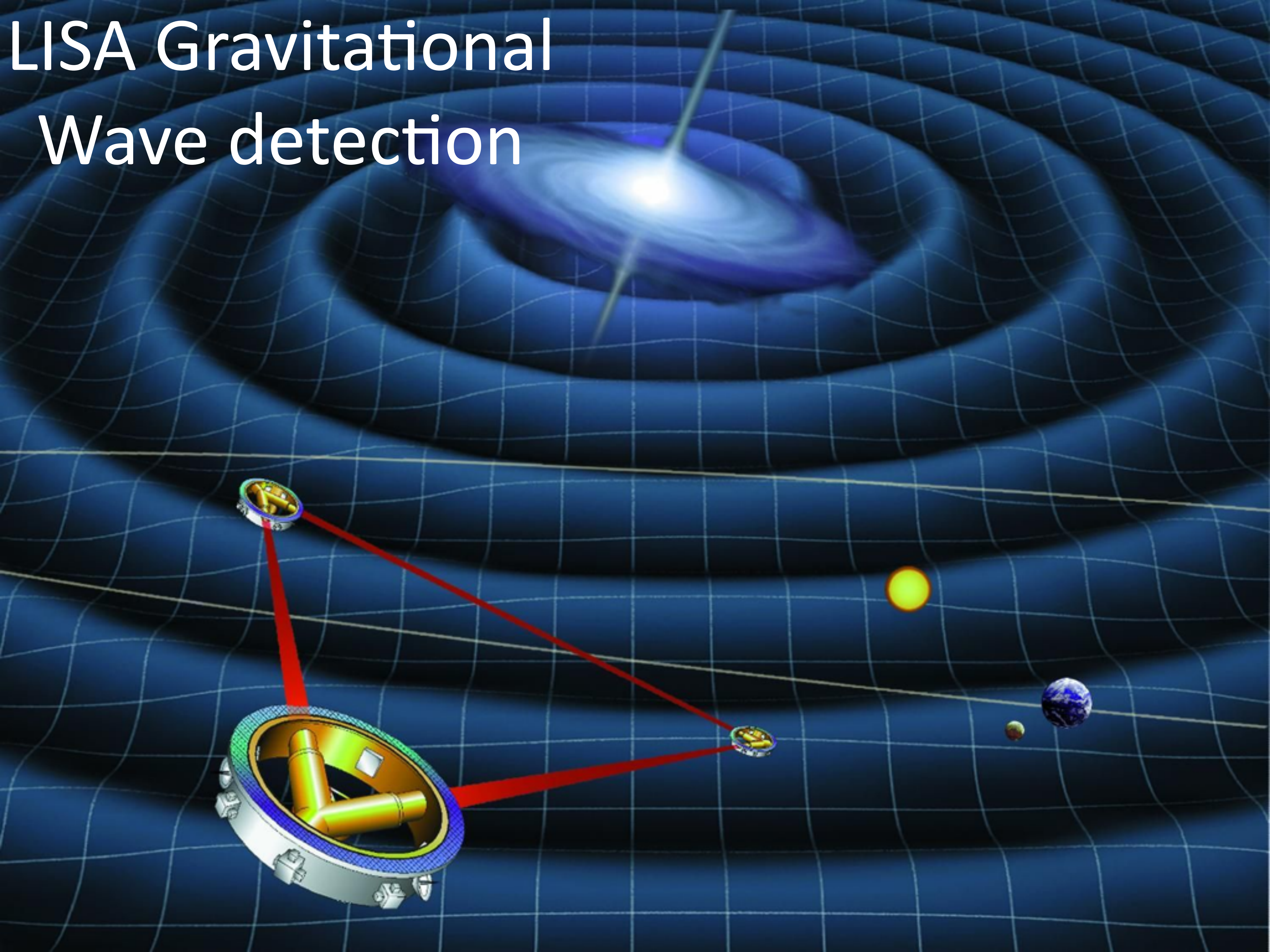
(a)



(b)

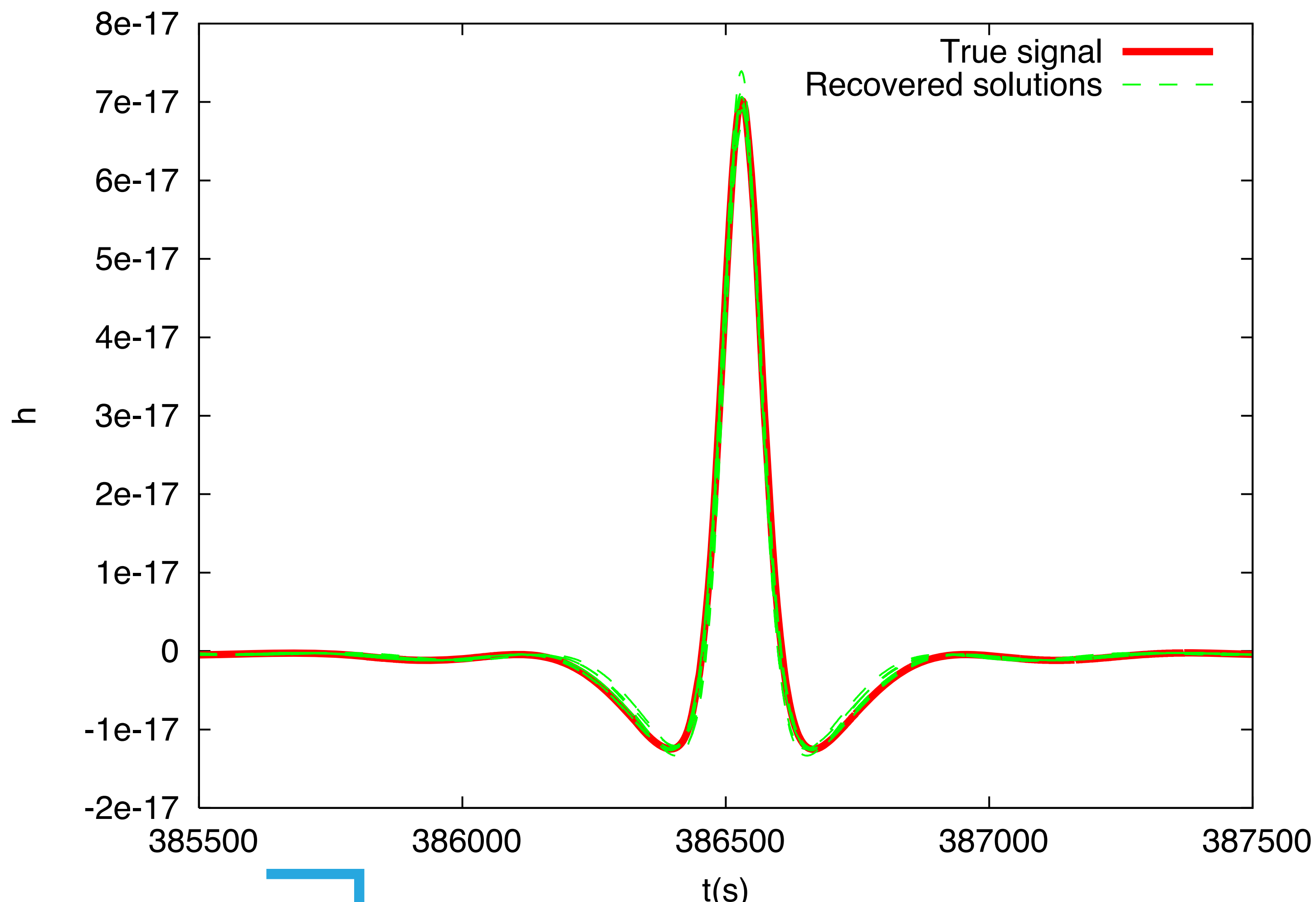
Examples

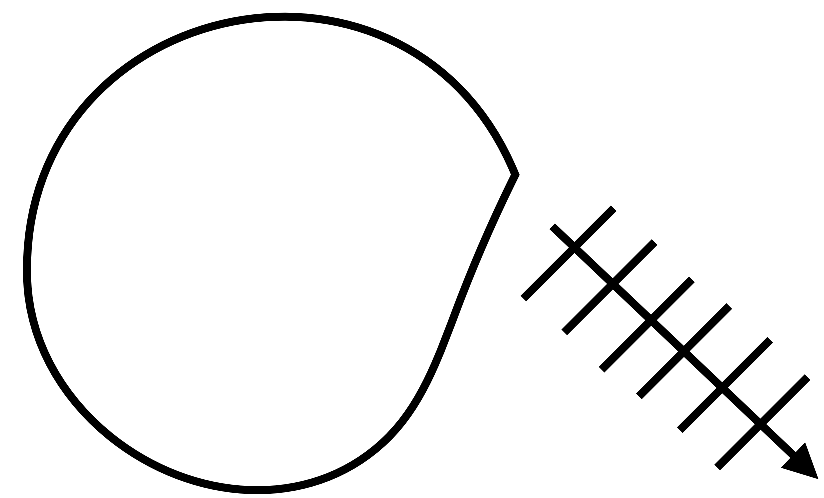
LISA Gravitational Wave detection



LISA time series analysis

- Cosmic strings produce beamed burst of gravitational waves via cusp formation





Target for LISA

ICIC

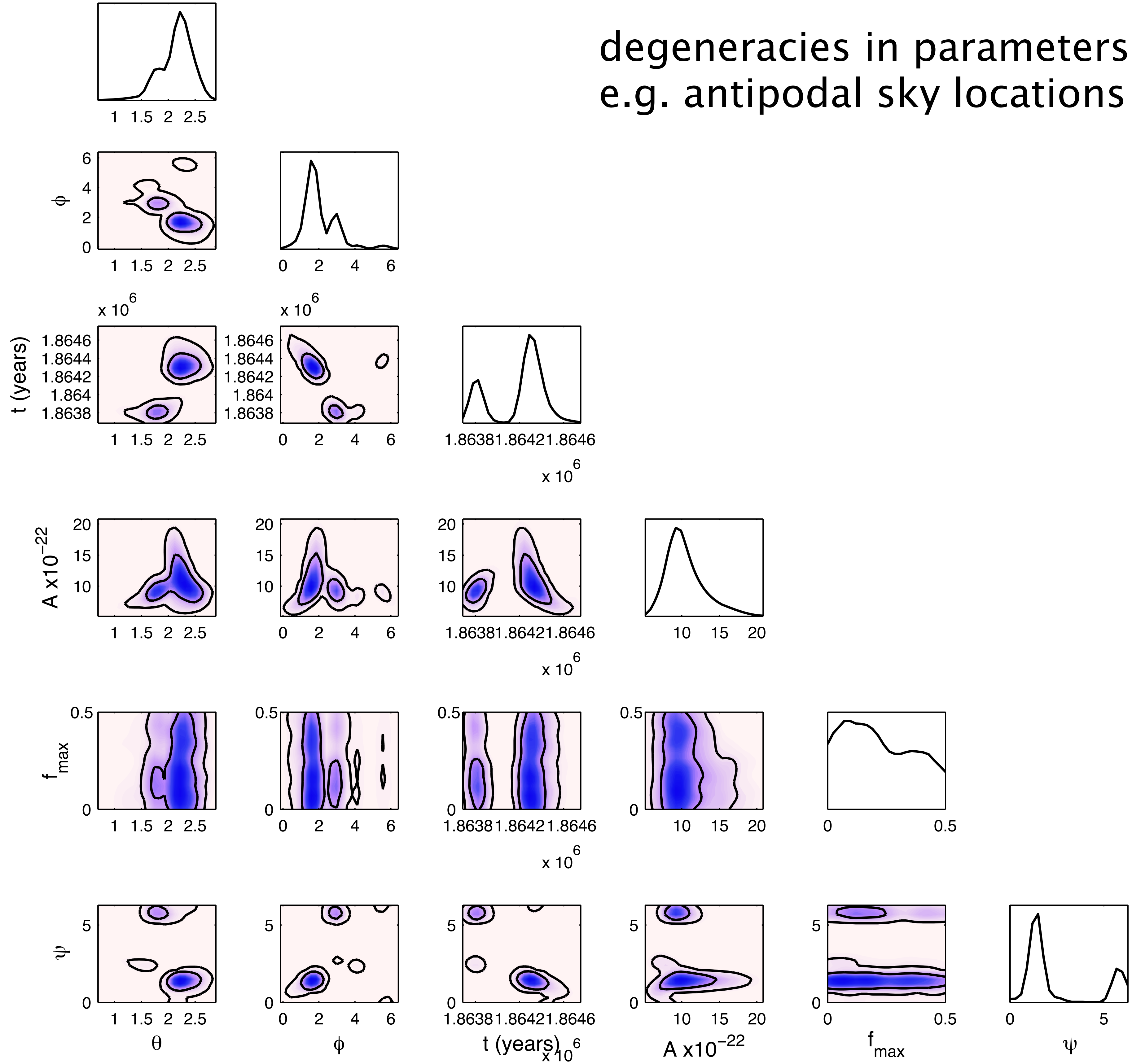
Ferhoz, Gair + 2010

$$|h(f)| = \begin{cases} \mathcal{A} f^{-4/3} & f < f_{\max} \\ \mathcal{A} f^{-4/3} \exp\left(1 - \frac{f}{f_{\max}}\right) & f > f_{\max}. \end{cases}$$

phase by $\exp(2\pi \imath f t_c)$,

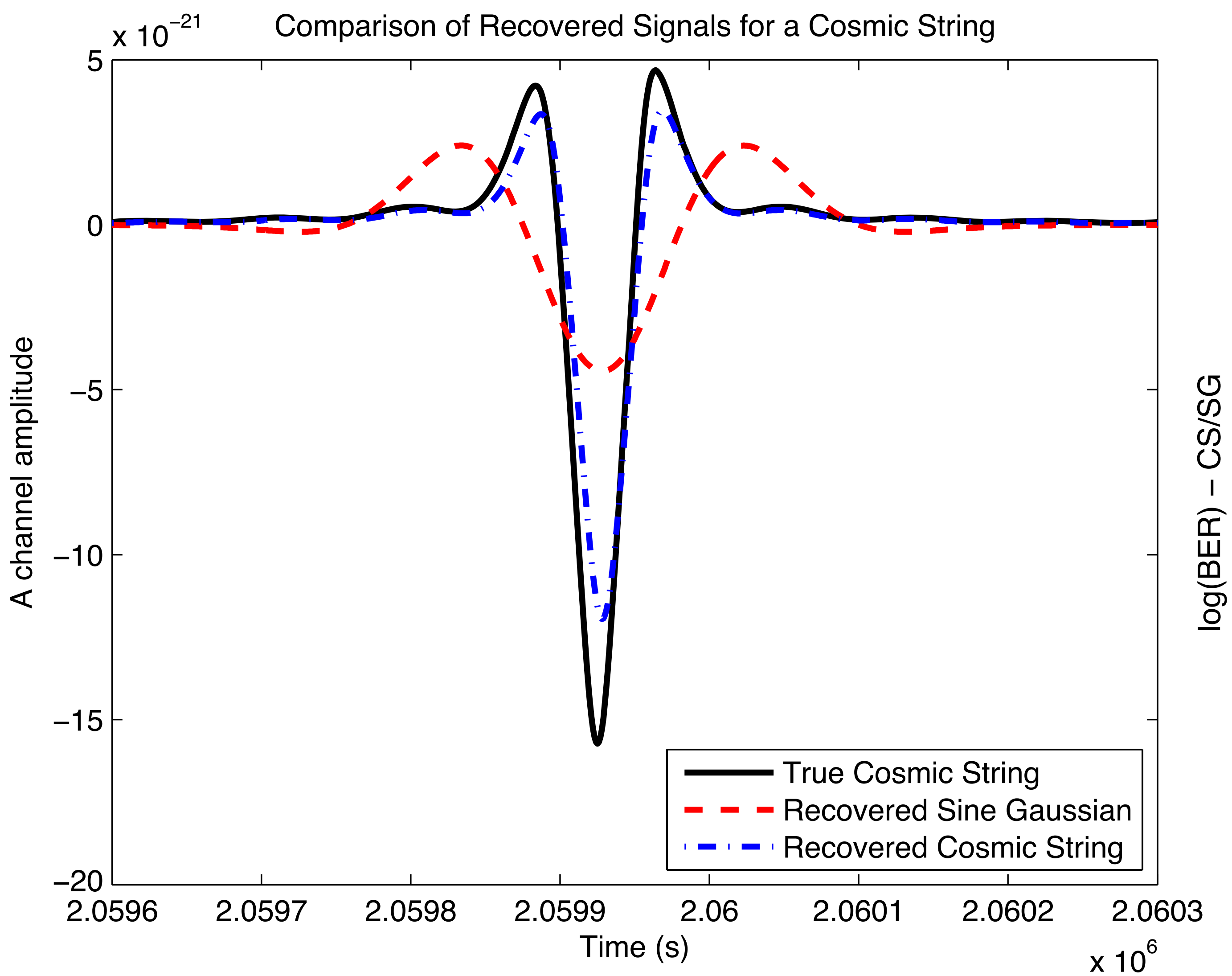
MultiNest used to search mock LISA timestream for cosmic string signal

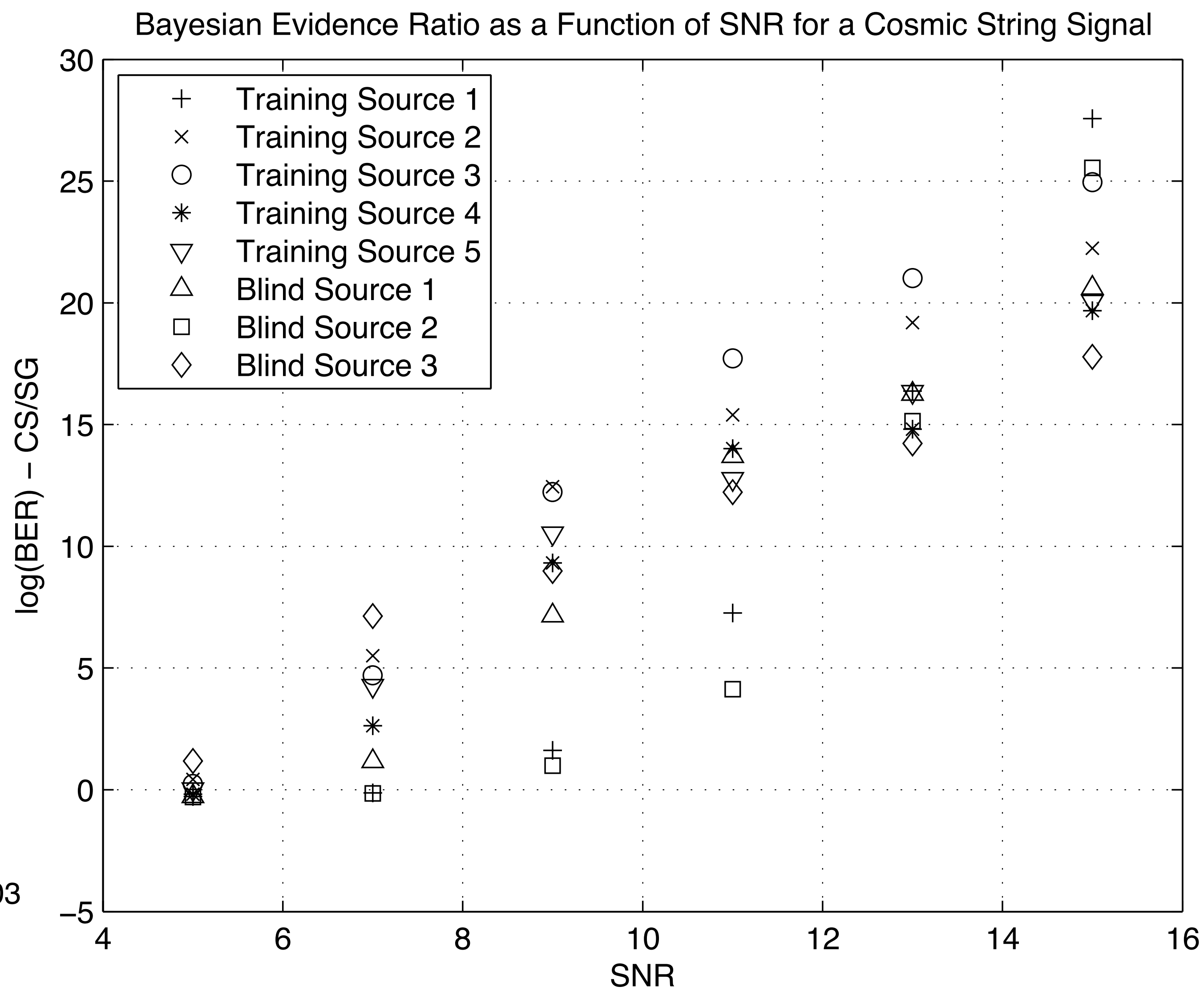
degeneracies in parameters
e.g. antipodal sky locations



colatitude longitude burst time burst amplitude break freq. waveform polarisation

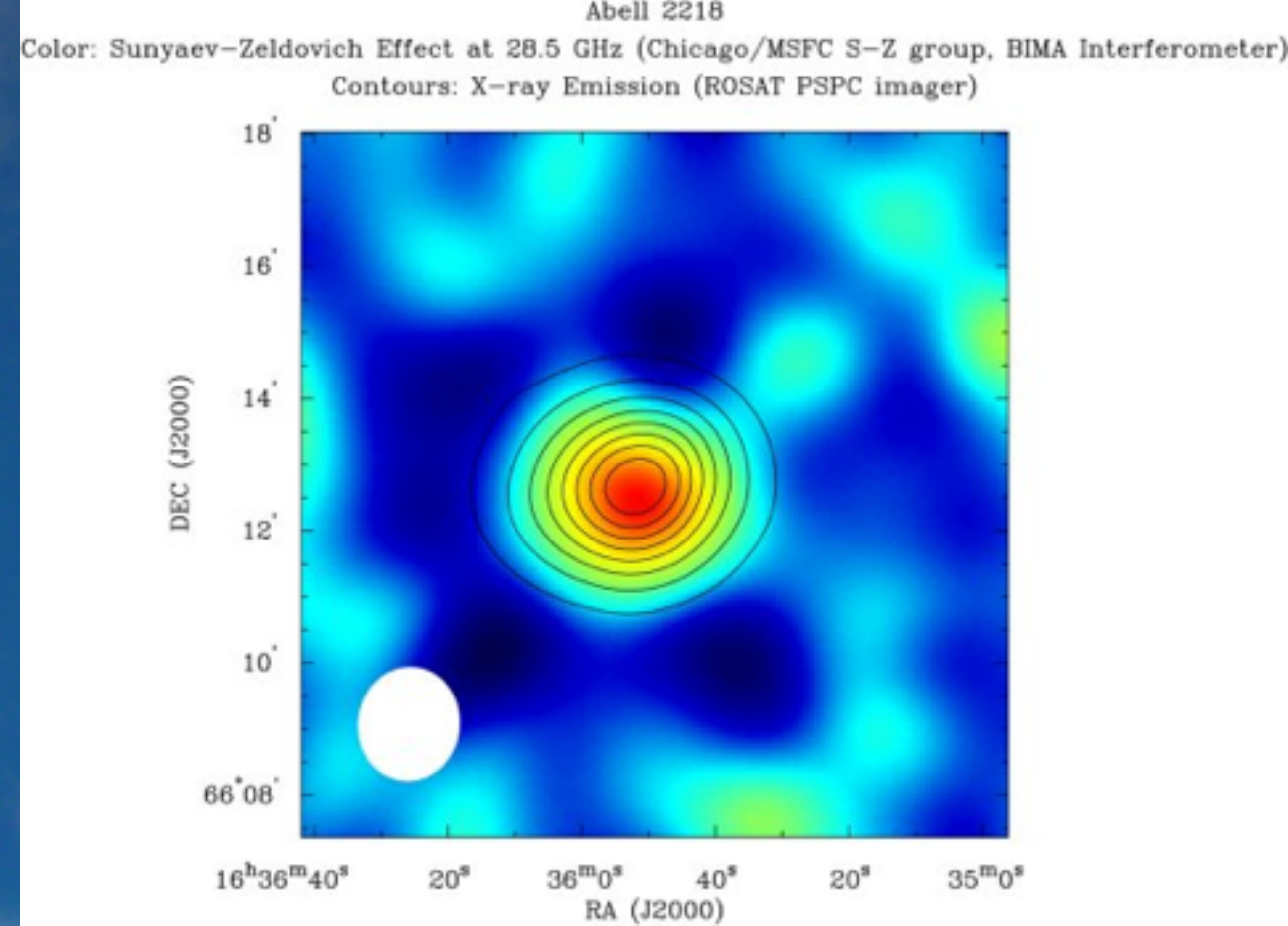
- Model selection to determine type of burst: cosmic string versus Sine-Gaussian model





ICIC

Bayesian Image detection



Bayesian Image detection

- e.g. finding circular objects in noisy data
- $$\tau(\boldsymbol{x};\boldsymbol{a}) = A \exp\left[-\frac{(x-X)^2+(y-Y)^2}{2R^2}\right] \qquad \boldsymbol{a} = \{X, Y, A, R\}$$

Hobson & McLachlan (2003)

- Sky model will be sum of such objects

$$\boldsymbol{D} = \boldsymbol{n} + \sum_{k=1}^{N_{\rm obj}} \boldsymbol{s}(\boldsymbol{a}_k),$$

- Gaussian noise determines likelihood

$$\Pr(\boldsymbol{D}\mid\boldsymbol{\theta}) = \frac{\exp\left\{-\frac{1}{2}[\boldsymbol{D}-\boldsymbol{s}(\boldsymbol{a})]^{\rm t}\mathbf{N}^{-1}[\boldsymbol{D}-\boldsymbol{s}(\boldsymbol{a})]\right\}}{(2\pi)^{N_{\rm pix}/2}\,|\mathbf{N}|^{1/2}},$$

ICIC

- Assume prior is separable by object

$$\Pr(\boldsymbol{\theta}) = \Pr(N_{\rm obj})\Pr(\boldsymbol{a}) = \Pr(N_{\rm obj})\Pr(\boldsymbol{a}_1)\Pr(\boldsymbol{a}_2)\cdots\Pr\left(\boldsymbol{a}_{N_{\rm obj}}\right)$$

Toy model

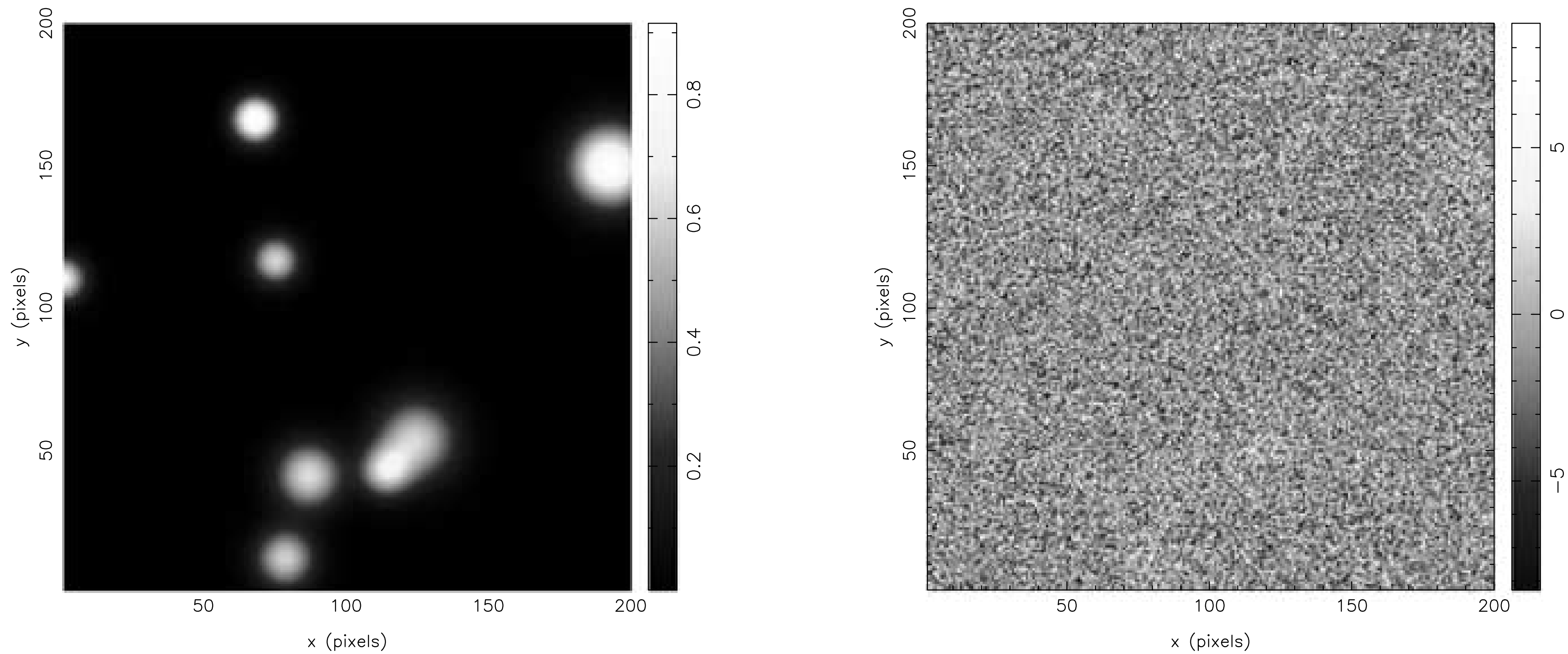


Figure 1. The toy problem discussed in Section 4.3. The 200 × 200 pixel test image (left panel) contains eight discrete Gaussian-shaped objects of varying widths and amplitudes; the parameters X_k, Y_k, A_k and r_k for each object are listed in Table 1. The corresponding data map (right panel) has independent Gaussian pixel noise added with an rms of 2 units. This figure is available in colour in the on-line version of the journal on *Synergy*.

ICIC

N=1 fitting

- Simplified analysis with just one object: leads to multimodal posterior with peaks at object locations

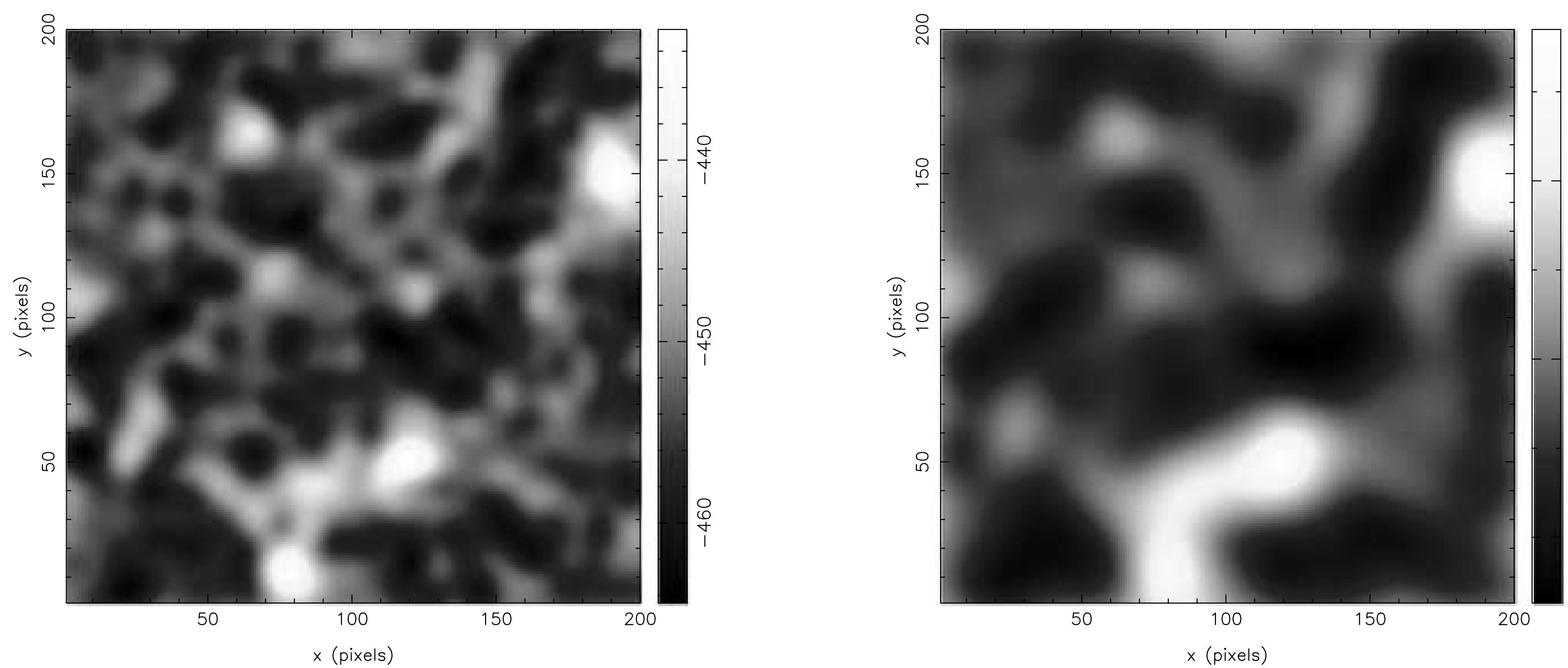


Figure 3. The two-dimensional conditional log-posterior distributions in the (X, Y)-subspace for the toy problem illustrated in Fig. 1, where the model contains a single object parametrized by $\boldsymbol{a} = \{X, Y, A, R\}$. The values of the amplitude A and size R are conditioned at $A = 0.75$, $R = 5$ (left panel) and $A = 0.75$, $R = 10$ (right panel). This figure is available in colour in the on-line version of the journal on *Synergy*.

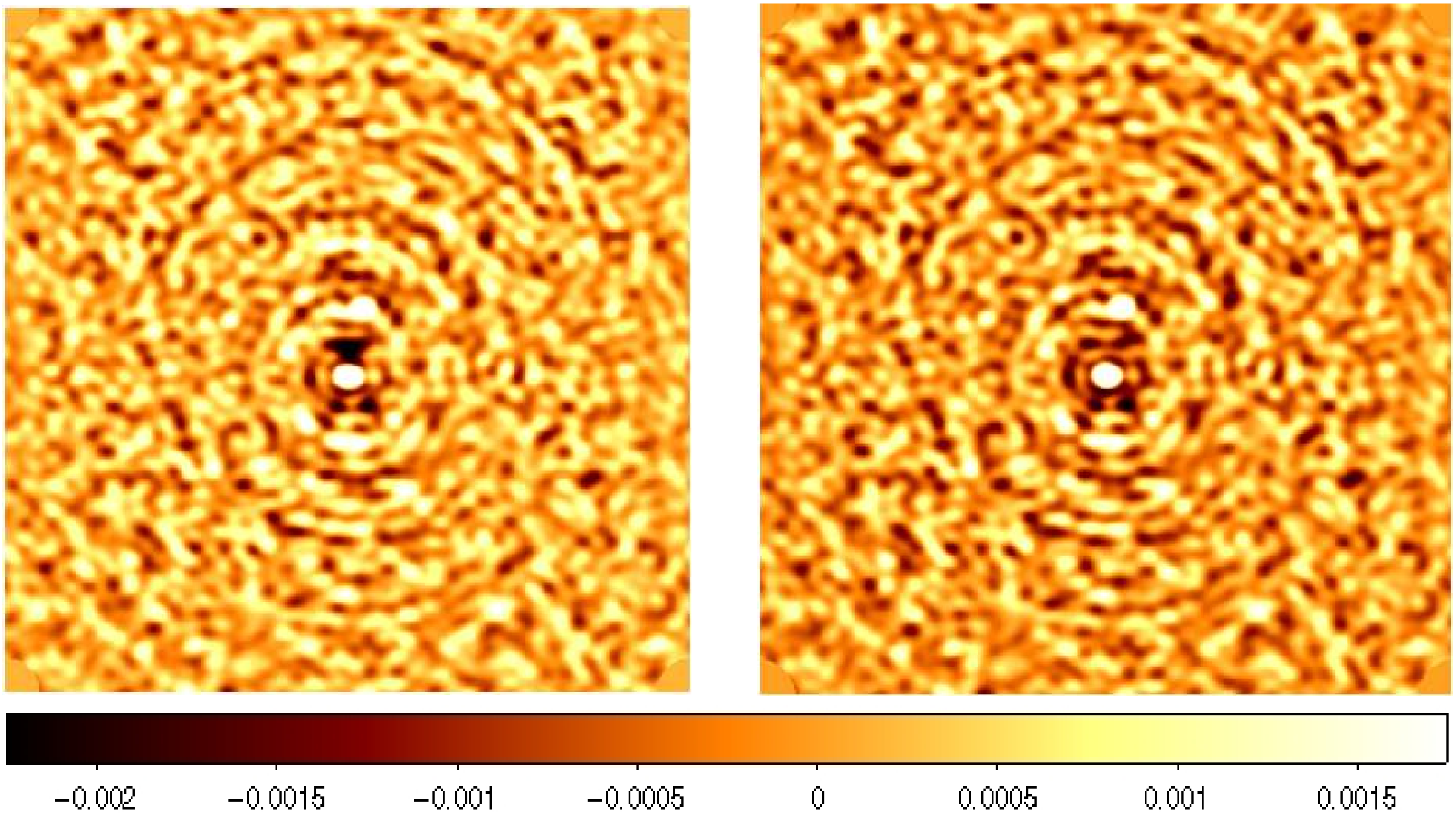
Identifying real sources

- Not all peaks in posterior will be sources - apply model selection to distinguish

$$H_0 = \text{`a cluster with } M_{\rm g,min} < M_{\rm g} \leqslant M_{\rm g,lim} \text{ is centred in } S\text{'},$$ no-cluster

$$H_1 = \text{`a cluster with } M_{\rm g,lim} < M_{\rm g} < M_{\rm g,max} \text{ is centred in } S\text{'},$$ cluster

SZ cluster detection with AMI
Feroz+ 2009



Odds ratio $e^{12.2\pm0.2}$ 0.32 ± 0.03

Global particle physics analysis

- Combine particle physics and astrophysics constraints to learn about beyond the Standard Model physics

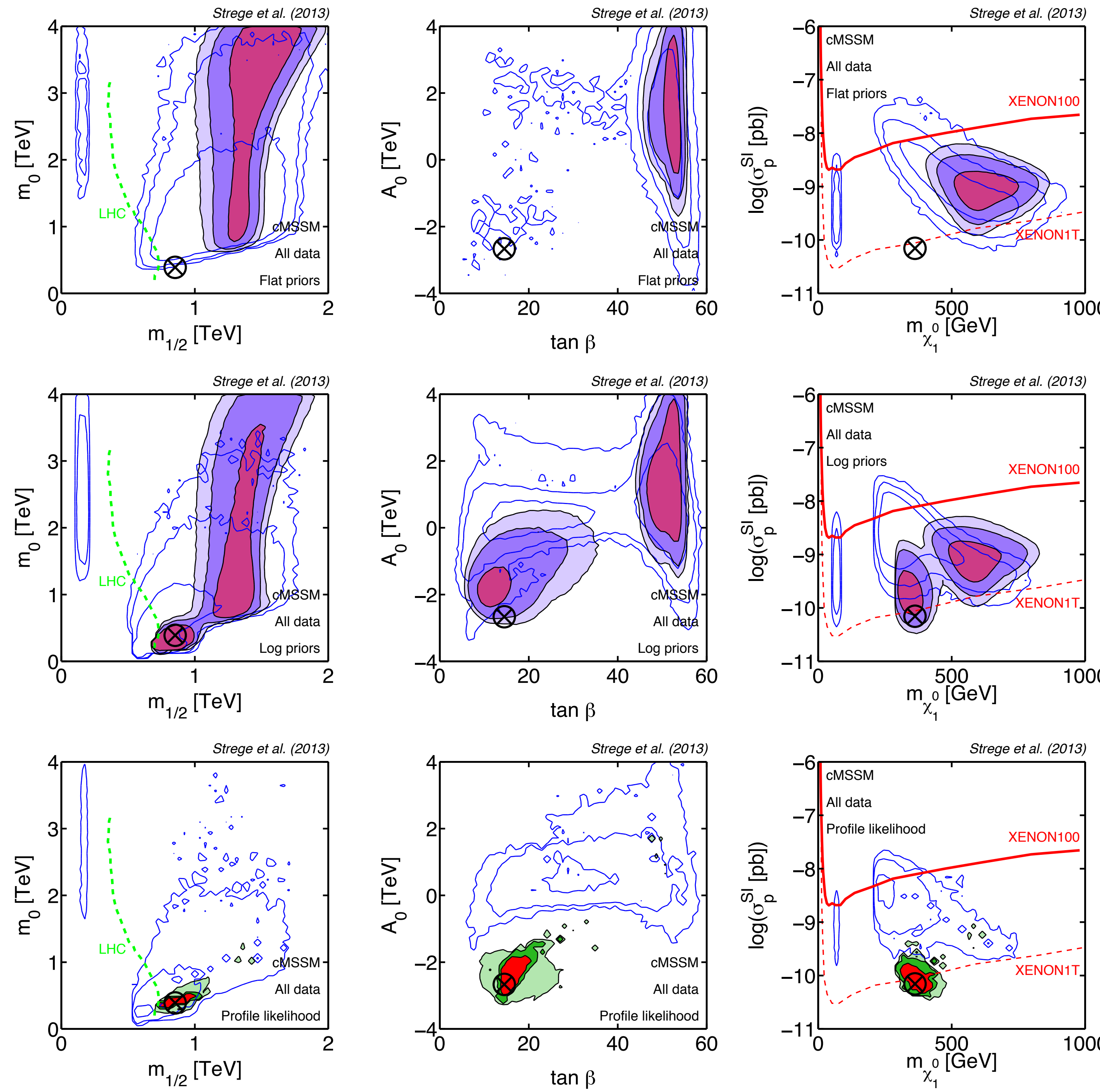
Observable	Mean value μ	Uncertainties σ (exper.)	τ (theor.)	Ref.
M_W [GeV]	80.399	0.023	0.015	[34]
$\sin^2\theta_{eff}$	0.23153	0.00016	0.00015	[34]
$\delta a_\mu^{\rm SUSY} \times 10^{10}$	28.7	8.0	2.0	[35]
$BR(\bar{B} \to X_s\gamma) \times 10^4$	3.55	0.26	0.30	[36]
$R_{\Delta M_{B_s}}$	1.04	0.11	-	[37]
$\frac{BR(B_u\to\tau\nu)}{BR(B_u\to\tau\nu)_{SM}}$	1.63	0.54	-	[36]
$\Delta_{0-} \times 10^2$	3.1	2.3	-	[38]
$\frac{BR(B\to D\tau\nu)}{BR(B\to De\nu)} \times 10^2$	41.6	12.8	3.5	[39]
R_{l23}	0.999	0.007	-	[40]
$BR(D_s \to \tau\nu) \times 10^2$	5.38	0.32	0.2	[36]
$BR(D_s \to \mu\nu) \times 10^3$	5.81	0.43	0.2	[36]
$BR(D \to \mu\nu) \times 10^4$	3.82	0.33	0.2	[36]
$\Omega_\chi h^2$	0.1109	0.0056	0.012	[41]
m_h [GeV]	125.8	0.6	2.0	[19]
$BR(\overline{B}_s \to \mu^+\mu^-)$	3.2×10^{-9}	1.5×10^{-9}	10%	[20]
	Limit (95% CL)		τ (theor.)	Ref.
Sparticle masses	As in table 4 of Ref. [42].			
$m_0, m_{1/2}$	ATLAS, $\sqrt{s} = 8$ TeV, 5.8 fb^{-1} 2012 limits			[17]
$m_A, \tan\beta$	CMS, $\sqrt{s} = 7$ TeV, 4.7 fb^{-1} 2012 limits			[18]
$m_\chi - \sigma^{\rm SI}_{\tilde{\chi}^0_1-p}$	XENON100 2012 limits (224.6×34 kg days)			[21]

ICIC

Constrained MSSM

- WMAP7+ATLAS+CMS Higgs mass + XENON100+…

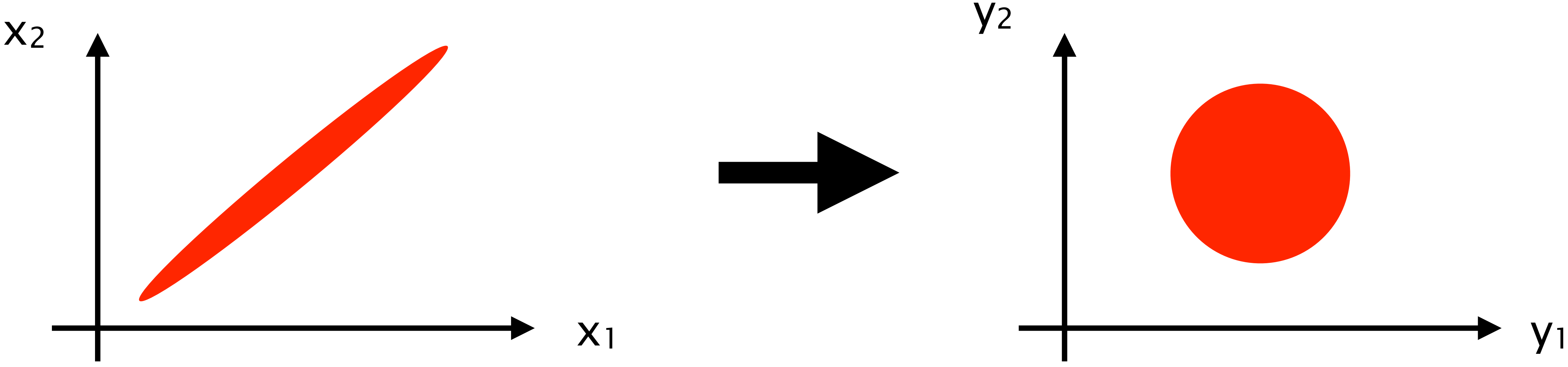
Strege+ (2013)
arXiv:1212.2636



Other useful packages

Affine invariance

- Common problem for MCMC is need to sample likelihood with narrow degeneracies
- Requires carefully chosen proposal distribution
- Would like to exploit an affine transformation to a more symmetric space



$$p(x) \propto f\left(-\frac{(x_1 - x_2)^2}{2\epsilon} - \frac{(x_1 + x_2)^2}{2}\right)$$

$$y_1 = \frac{x_1 - x_2}{\sqrt{\epsilon}}, \qquad y_2 = x_1 + x_2.$$

Emcee - "The MC Hammer"

- Emcee - affine invariant ensemble sampler (available as python module) Foreman-Mackey+ (2012)
- Evolve ensemble of walkers (c.f. Metropolis-Hastings)
- Stretch-move:

For walker at X_k, chose another walker Xj and propose move

$$X_k(t) \to Y = X_j + Z[X_k(t) - X_j],$$

Z random variable drawn from distribution chosen to ensure detailed balance

$$g(z) \propto \begin{cases} \frac{1}{\sqrt{Z}} & \text{if } z \in \left[\frac{1}{a}, a\right] \\ 0 & \text{otherwise} \end{cases}$$

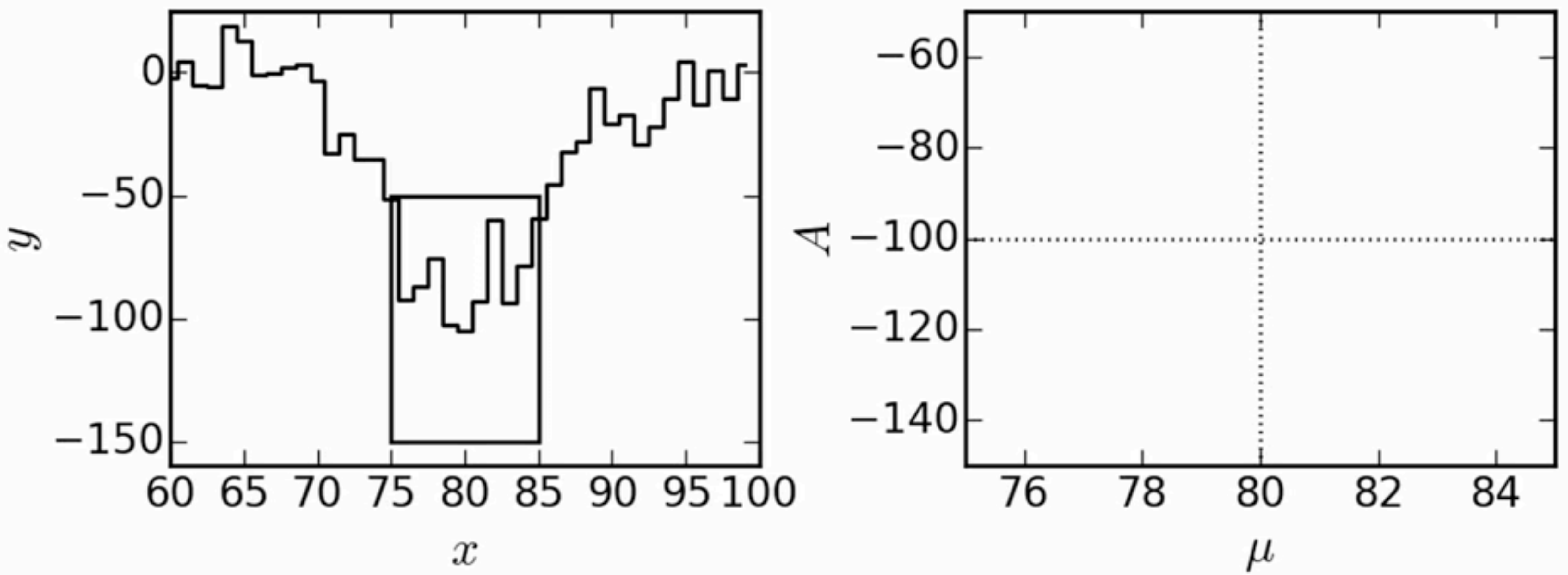
Accept proposal according to

$$q = \min\left(1, Z^{N-1} \frac{p(Y)}{p(X_k(t))}\right),$$

Repeat for all walkers in series.

ICIC

- Versitile module for general MCMC problems



Movie credit: Jordan Mirocha (UCLA)

COSMOSIS

- Modular Cosmology analysis code
  Zuntz et al: http://arxiv.org/abs/1409.3409
  - includes MultiNest, CosmoMC
  - integrates likelihoods from Planck, WMAP, DES,…
  - easy to switch in and out different samplers/
  datasets

ICIC

Conclusions

- Nested sampling offers alternative to MCMC to sampling posterior and provides evidence
- Nested sampling is well suited to problems with multi-modal posteriors
- Codes like MultiNest and PolyChord are freely available and should be in your toolkit
- Many other useful packages: COSMOSIS, EMCEE,SuperBayes, …

Some notation

- Likelihood

$$P(\mathcal{D}|\theta, \mathcal{M}) \equiv \mathcal{L},$$

$$P(\theta|\mathcal{M}) \equiv \pi,$$

$$P(\mathcal{D}|\mathcal{M}) \equiv \mathcal{Z} = \int P(\mathcal{D}|\theta, \mathcal{M}) P(\theta|\mathcal{M}) \mathrm{d}\theta.$$

$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)\mathrm{d}\theta.$$

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}.$$