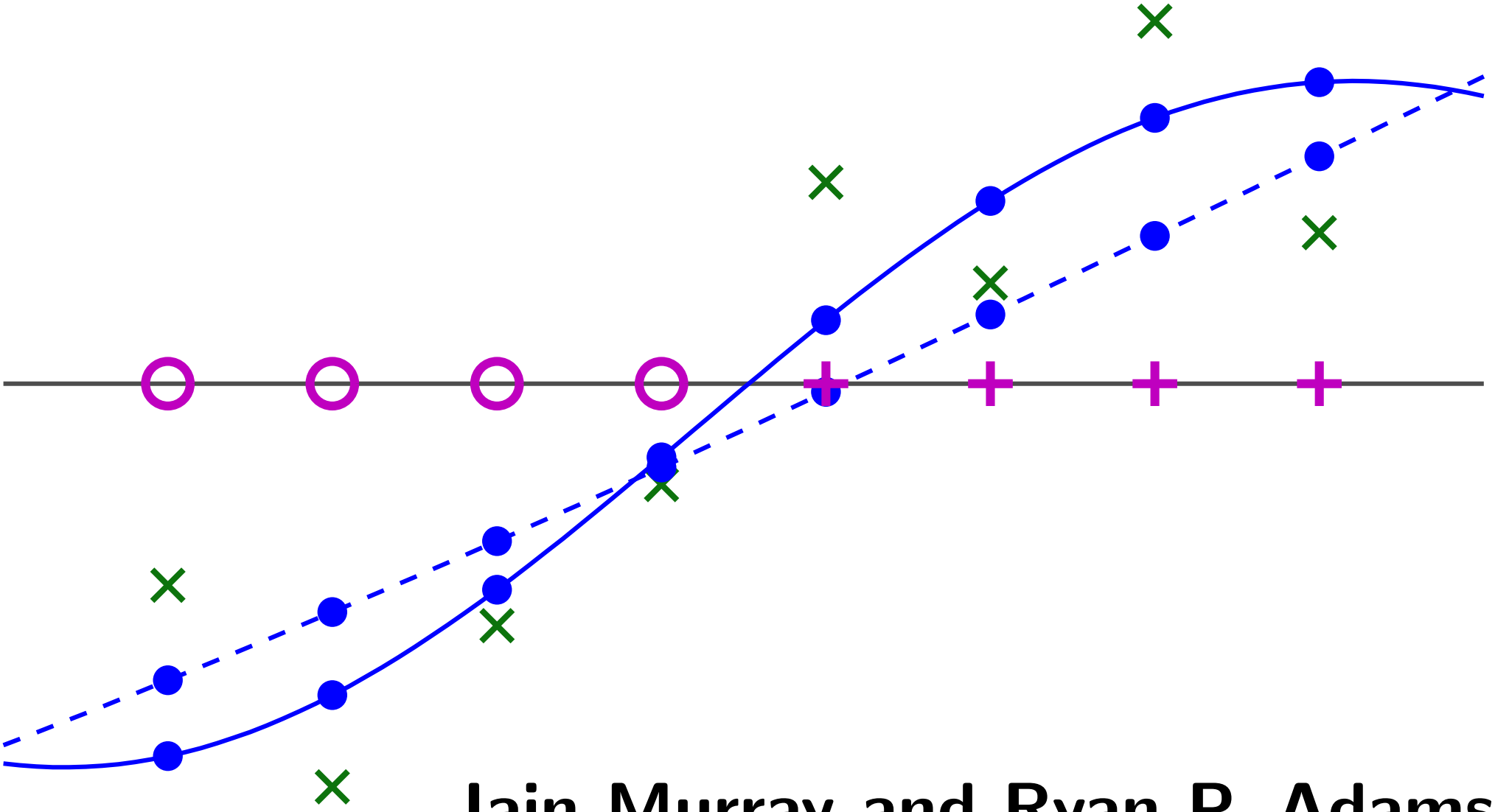
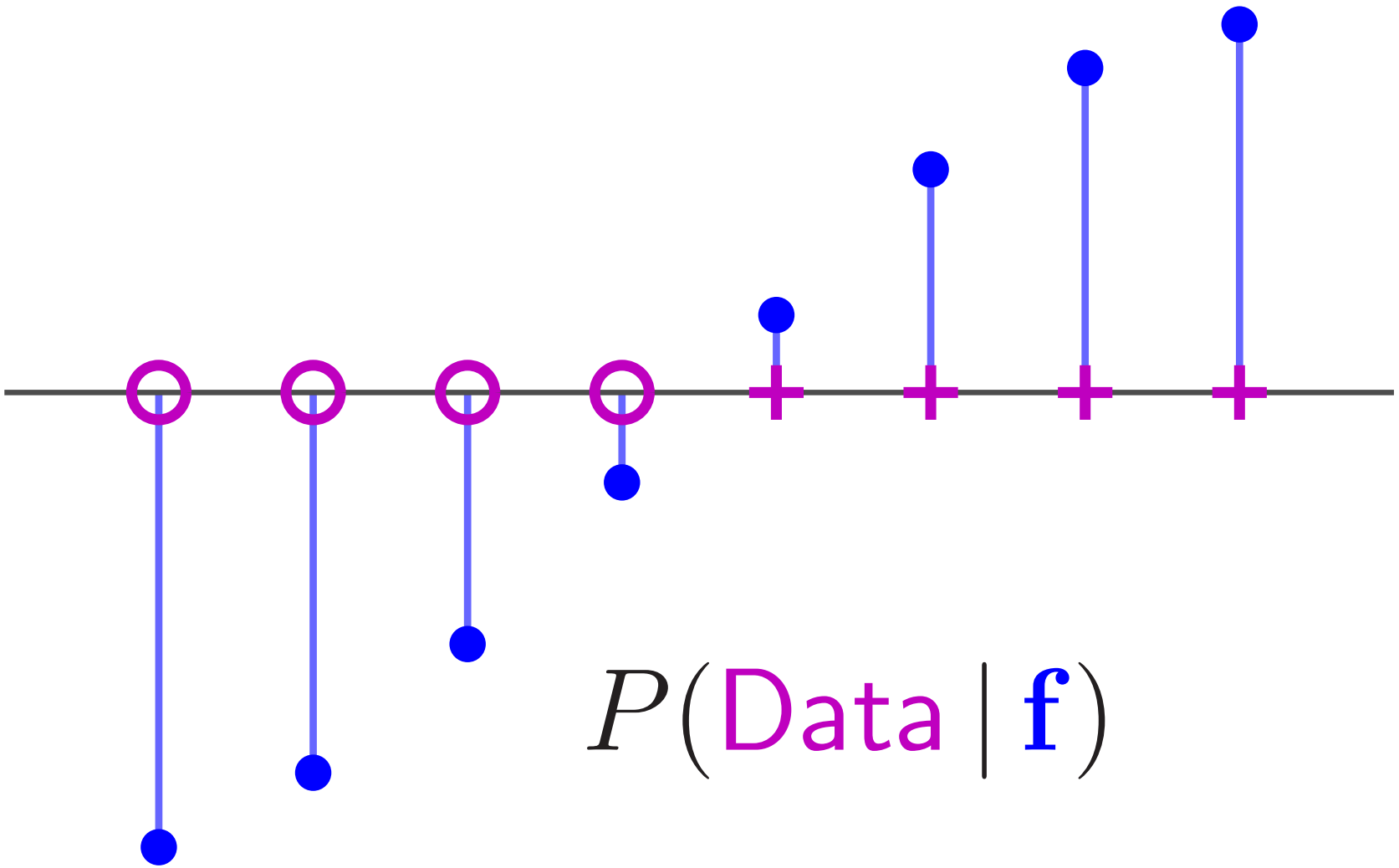
Slice sampling covariance hyperparameters of latent Gaussian models



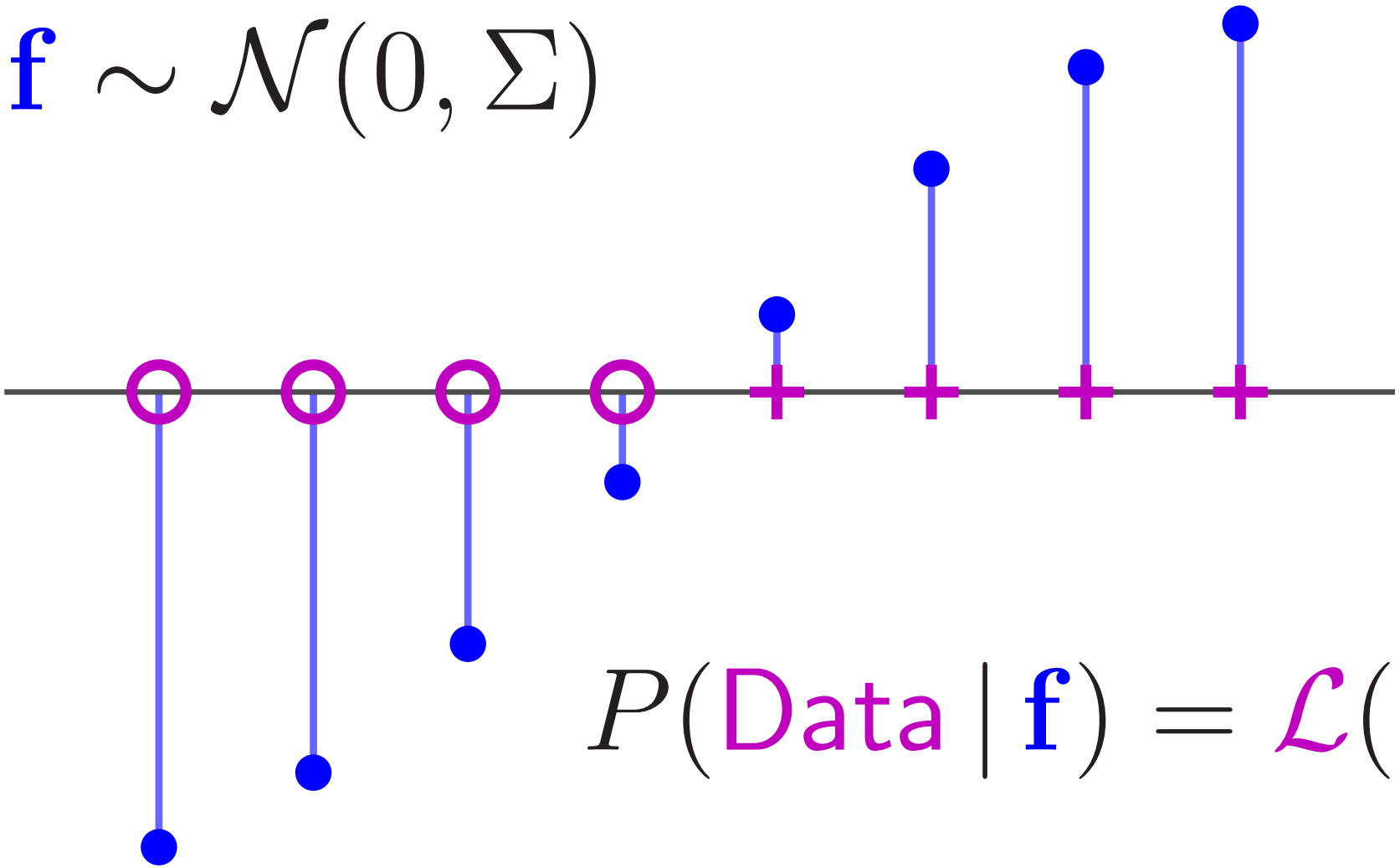
Iain Murray and Ryan P. Adams

University of Edinburgh and University of Toronto

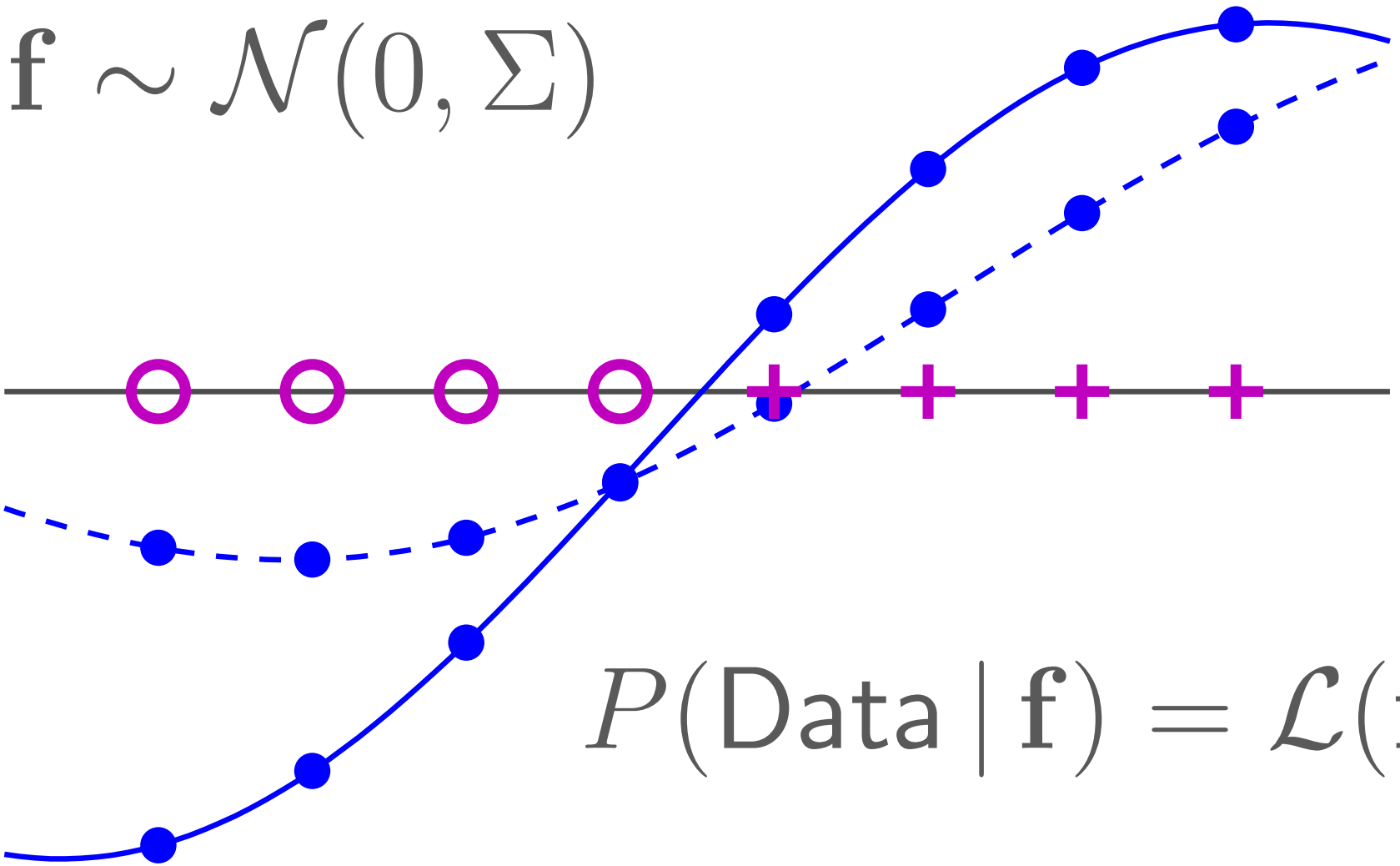
Data


$P(\text{Data} \mid \mathbf{f})$

$\mathbf{f} \sim \mathcal{N}(0, \Sigma)$



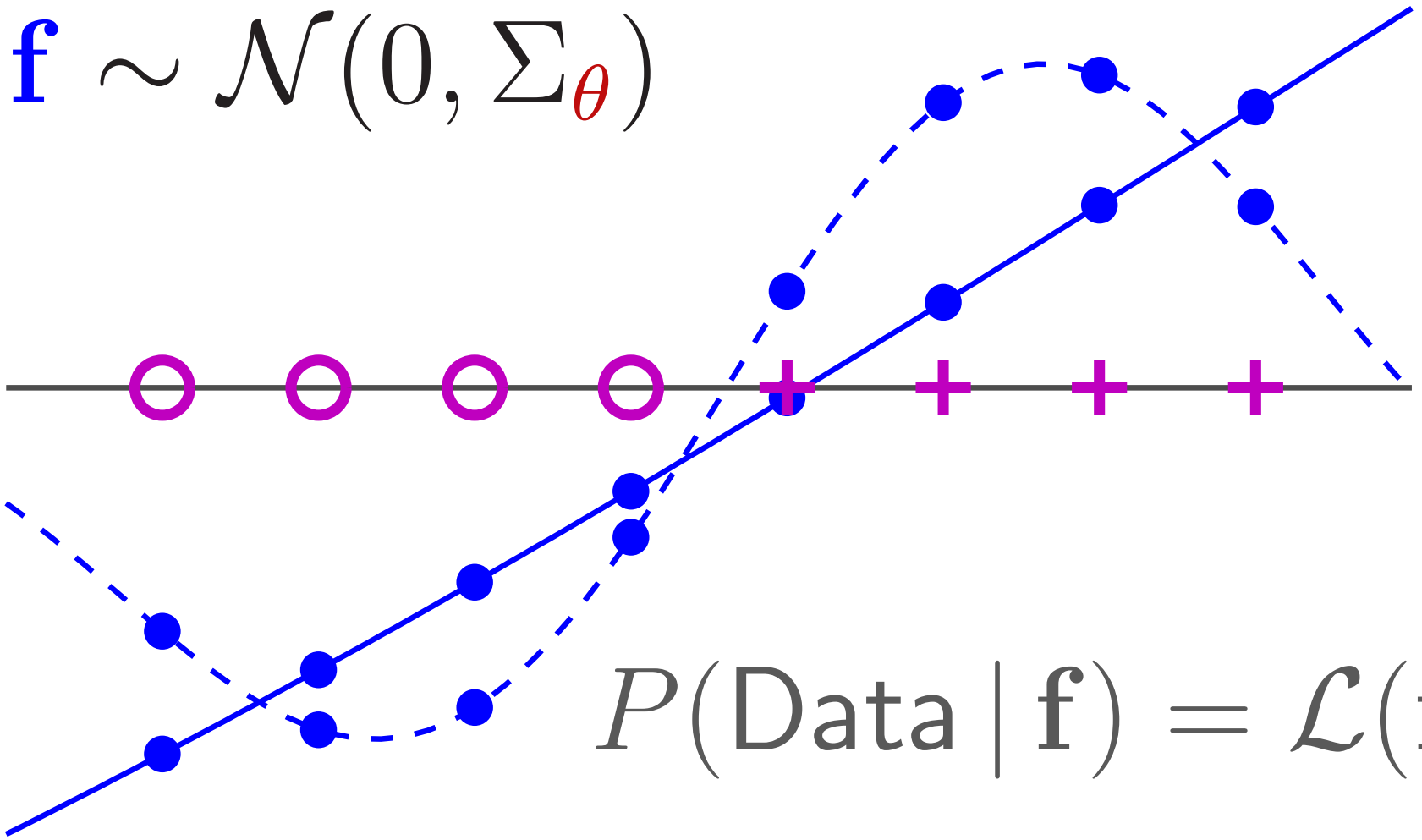
$P(\text{Data} \mid \mathbf{f}) = \mathcal{L}(\mathbf{f})$



$$P(\mathbf{f} \mid \text{Data}) \propto \mathcal{N}(\mathbf{f}; 0, \Sigma)\, \mathcal{L}(\mathbf{f})$$

$$\theta \sim p_h$$

$$\mathbf{f} \sim \mathcal{N}(0, \Sigma_\theta)$$



$$P(\text{Data} \mid \mathbf{f}) = \mathcal{L}(\mathbf{f})$$

$$P(\mathbf{f}, \theta \mid \text{D}) \propto p(\theta)\, \mathcal{N}(\mathbf{f}; 0, \Sigma_\theta)\, \mathcal{L}(\mathbf{f})$$

Beyond Classification

Other regression problems:

— heavy tailed noise; ordinal, multi-class, counts

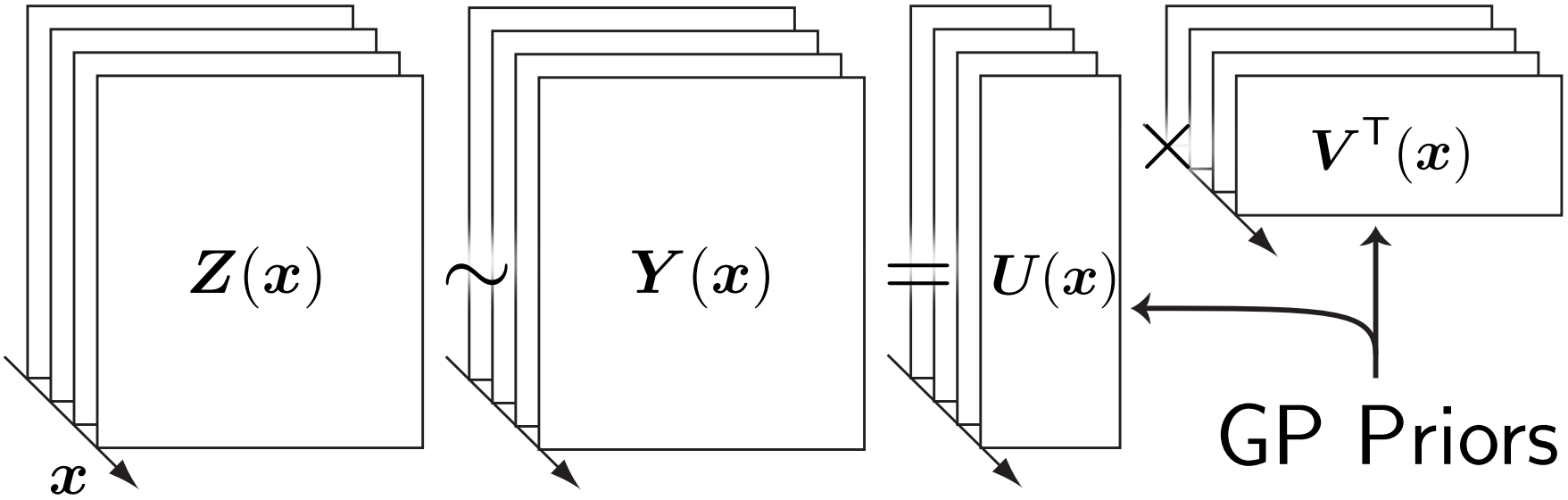
Component of *many* models, e.g.:

— Copula Processes (Wilson and Ghahramani, yesterday)

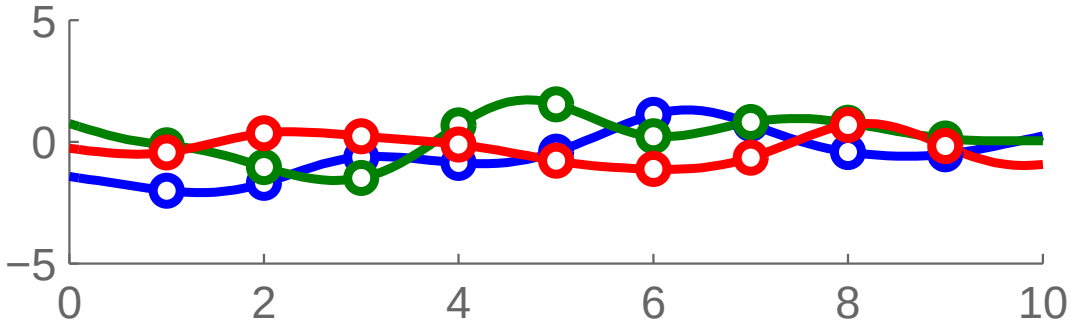
— (Sigmoidal) Gaussian Cox process (A, M, MacKay)

— Logistic GP; GP density sampler (A, M, MacKay)

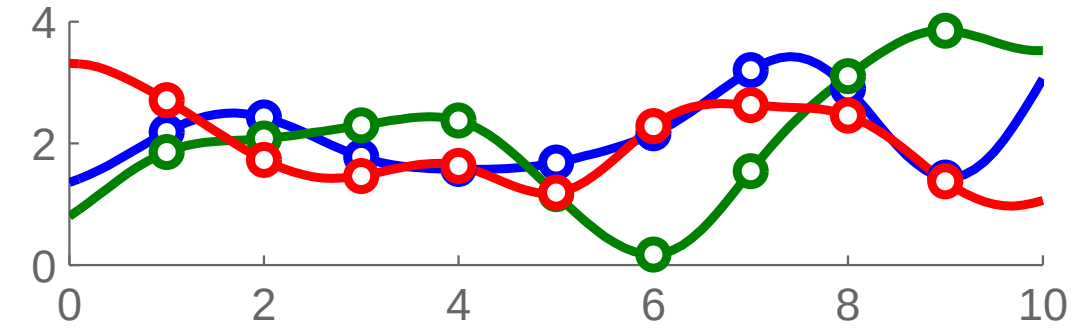
— Probabilistic matrix factorization (A, Dahl, M)



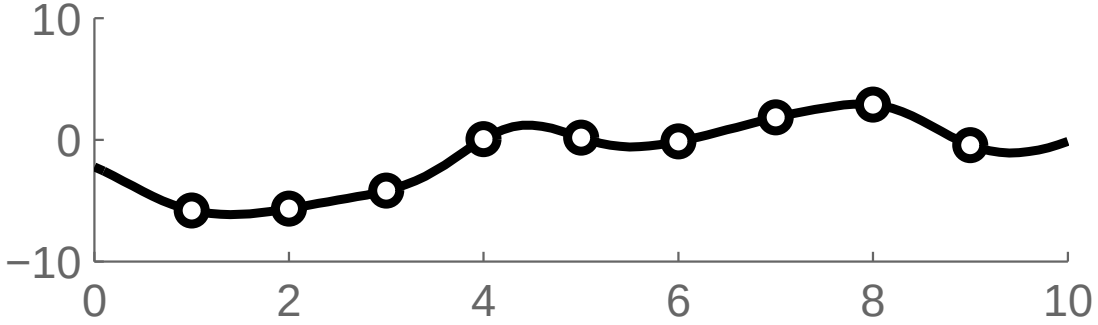
$$\mathbf{u}_m(\mathbf{x})$$



$$\mathbf{v}_m(\mathbf{x})$$



$$Y_{m,n}(\mathbf{x}) = \mathbf{u}_m^\top(\mathbf{x})\mathbf{v}_n(\mathbf{x})$$



Role of hyperparameters

Basic properties:

— Lengthscale, smoothness

— Amplitude

— Noise level

Structure:

— Soft feature selection (ARD)

— Inhomogeneities

— Periodicity

MCMC inference

Simulate Markov chain with equilibrium

$$P(\mathbf{f}, \theta \,|\, \text{Data}) \propto p(\theta)\, \mathcal{N}(\mathbf{f}; 0, \Sigma_\theta)\, \mathcal{L}(\mathbf{f})$$

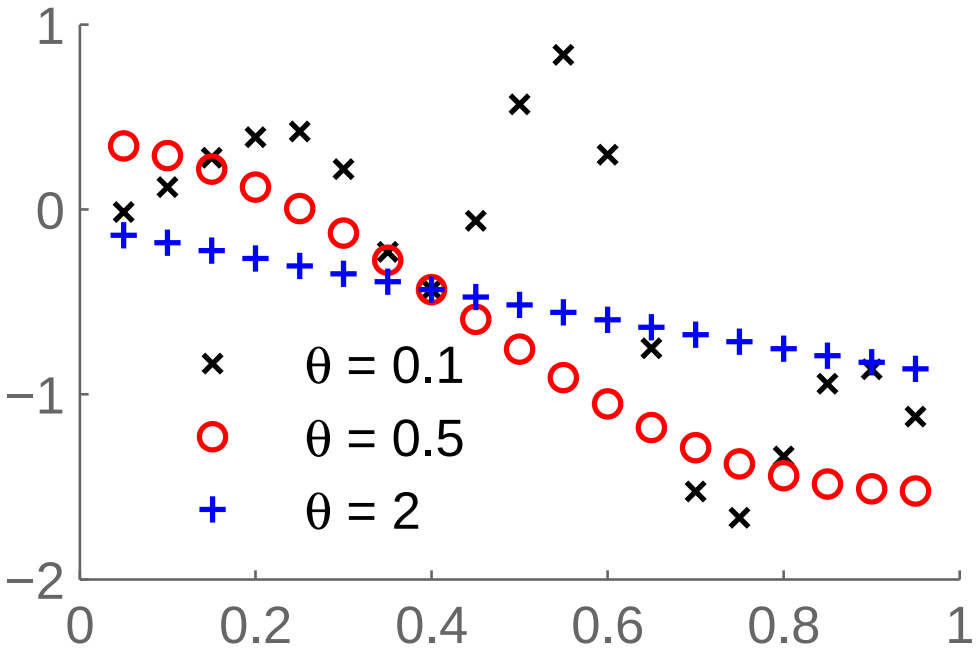
Straightforward approach:

— Update $\mathbf{f} \,|\, \theta, \text{Data}$

— Update $\theta \,|\, \mathbf{f}$

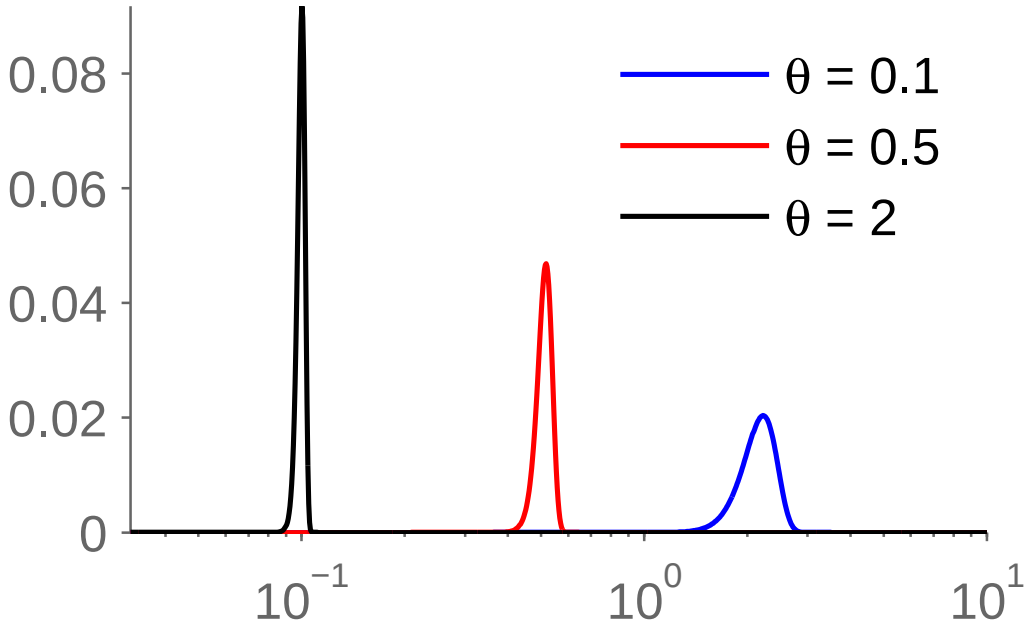
Prior mixing problem

Latent values, $\mathbf{f}$



Input Space

$P(\theta \mid \mathbf{f})$

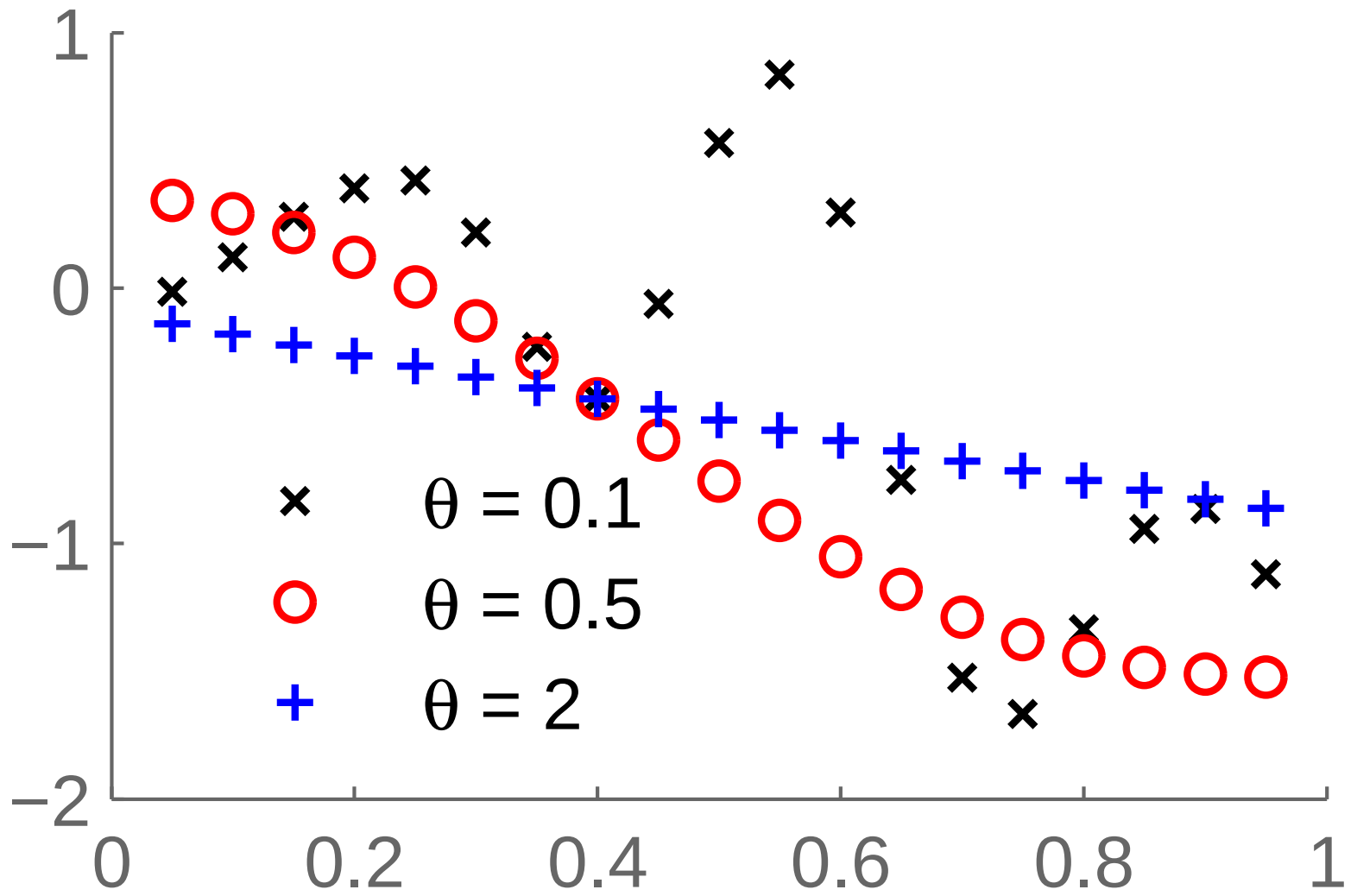


lengthscale, l

Any method updating $\theta \mid \mathbf{f}$ is slow

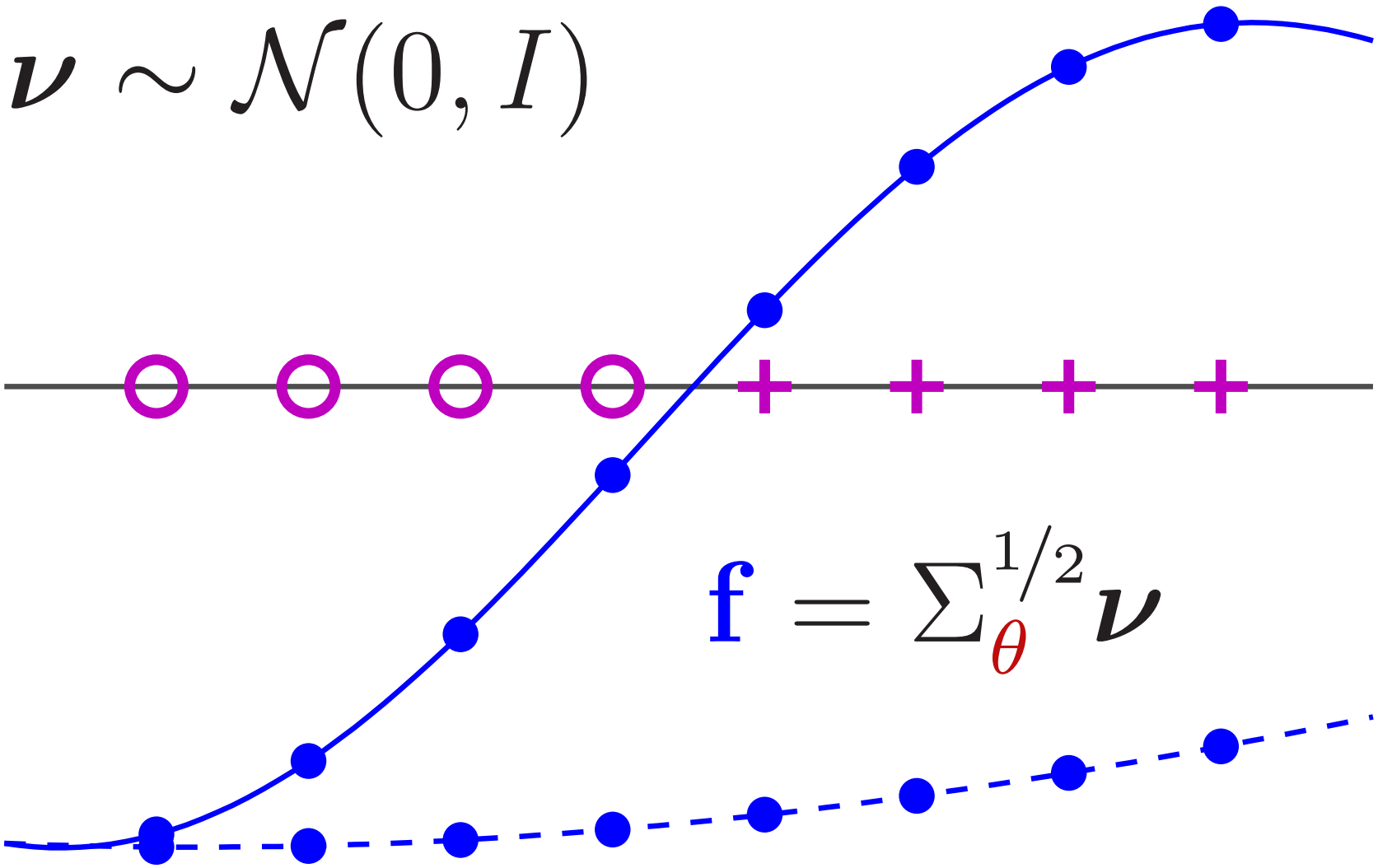
Whitening the prior

$$\theta \sim p_h, \quad \boldsymbol{\nu} \sim \mathcal{N}(0, I), \quad \mathbf{f} = \Sigma_\theta^{1/2} \boldsymbol{\nu}$$



$\theta \sim p_h$

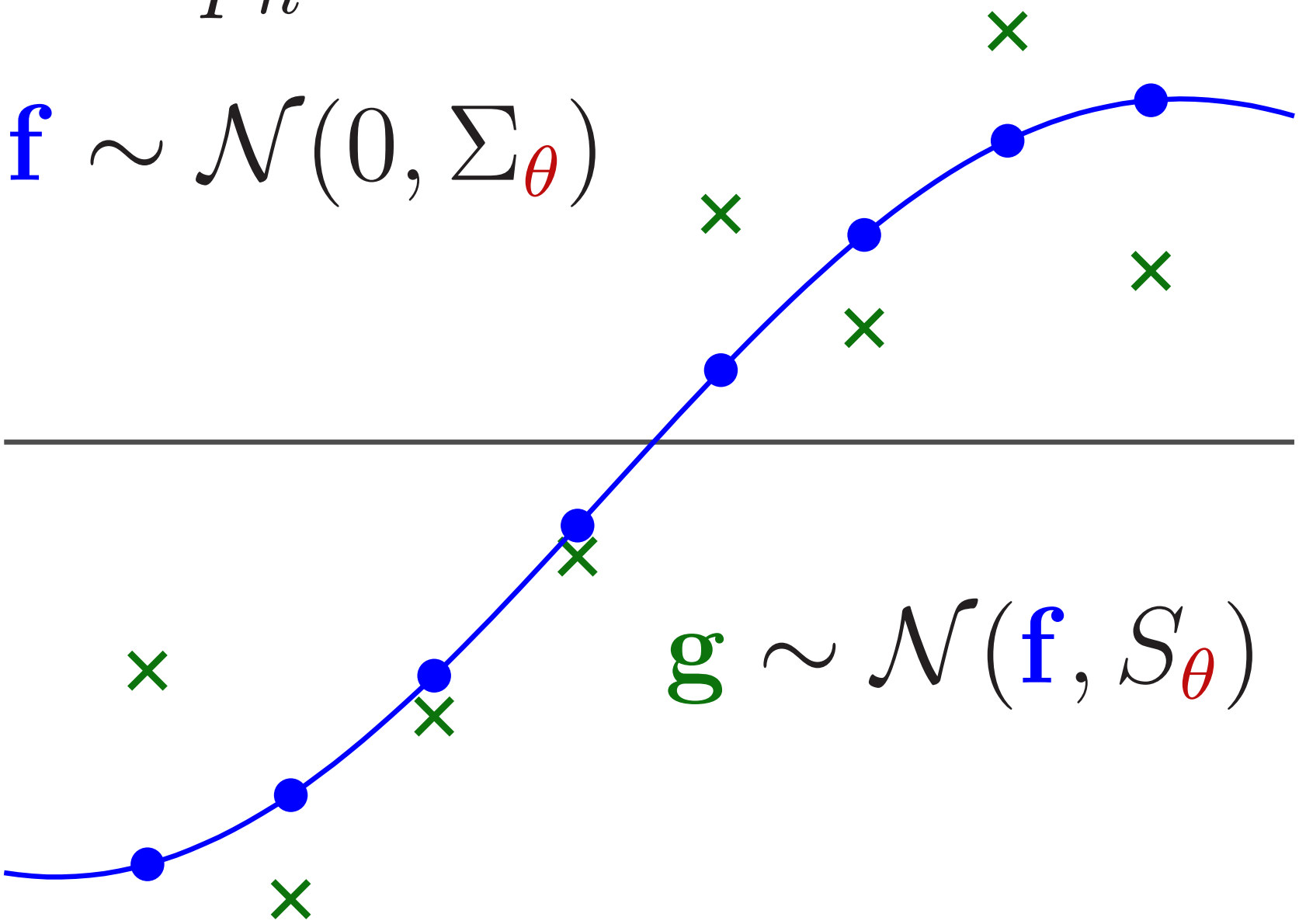
$\boldsymbol{\nu} \sim \mathcal{N}(0, I)$



Propose $\theta \longrightarrow$ long lengthscale

$\theta \sim p_h$

$\mathbf{f} \sim \mathcal{N}(0, \Sigma_\theta)$

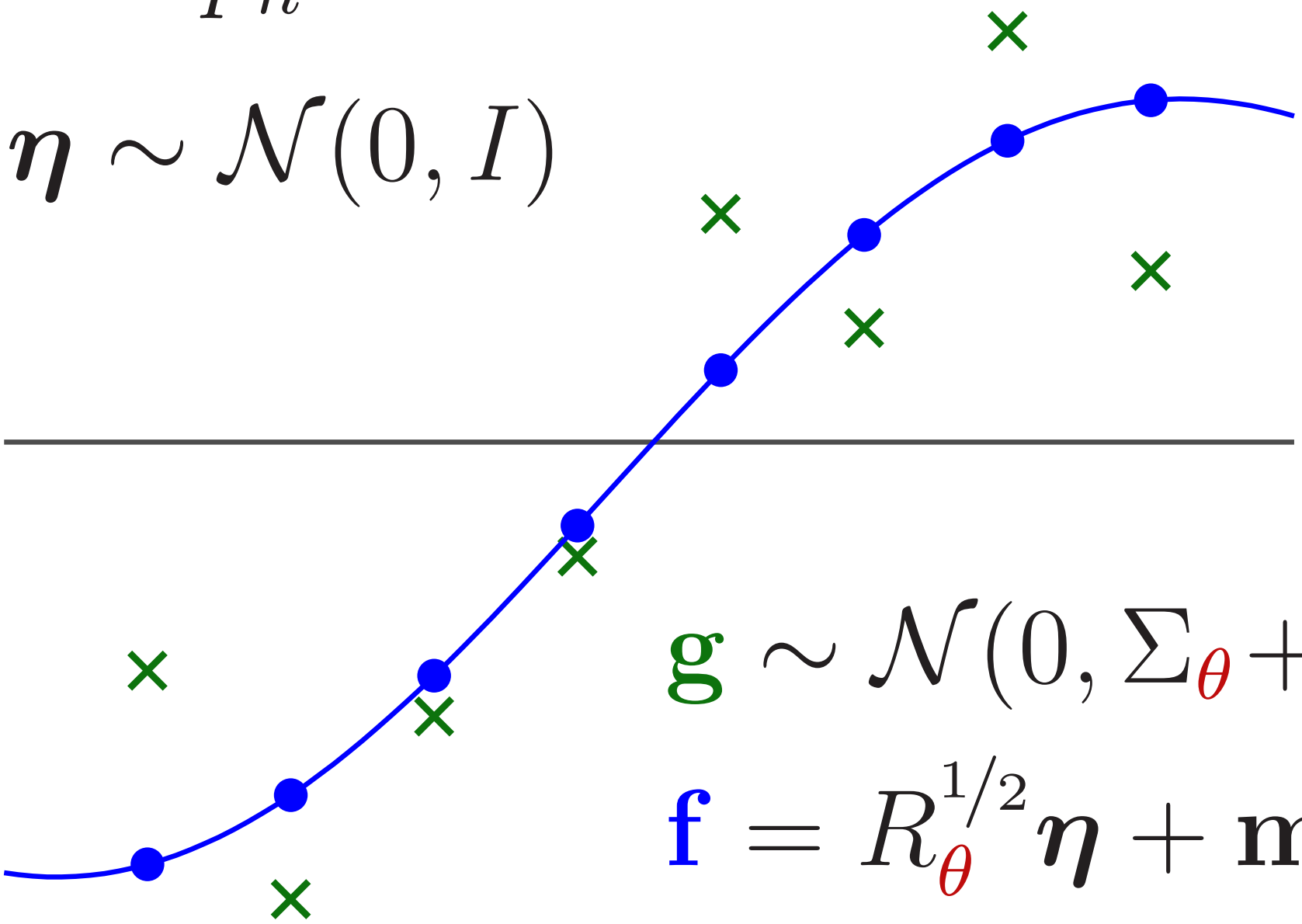


$\mathbf{g} \sim \mathcal{N}(\mathbf{f}, S_\theta)$

Marginally $\mathbf{g} \sim \mathcal{N}(0, \Sigma_\theta + S_\theta)$

$\theta \sim p_h$

$\boldsymbol{\eta} \sim \mathcal{N}(0, I)$



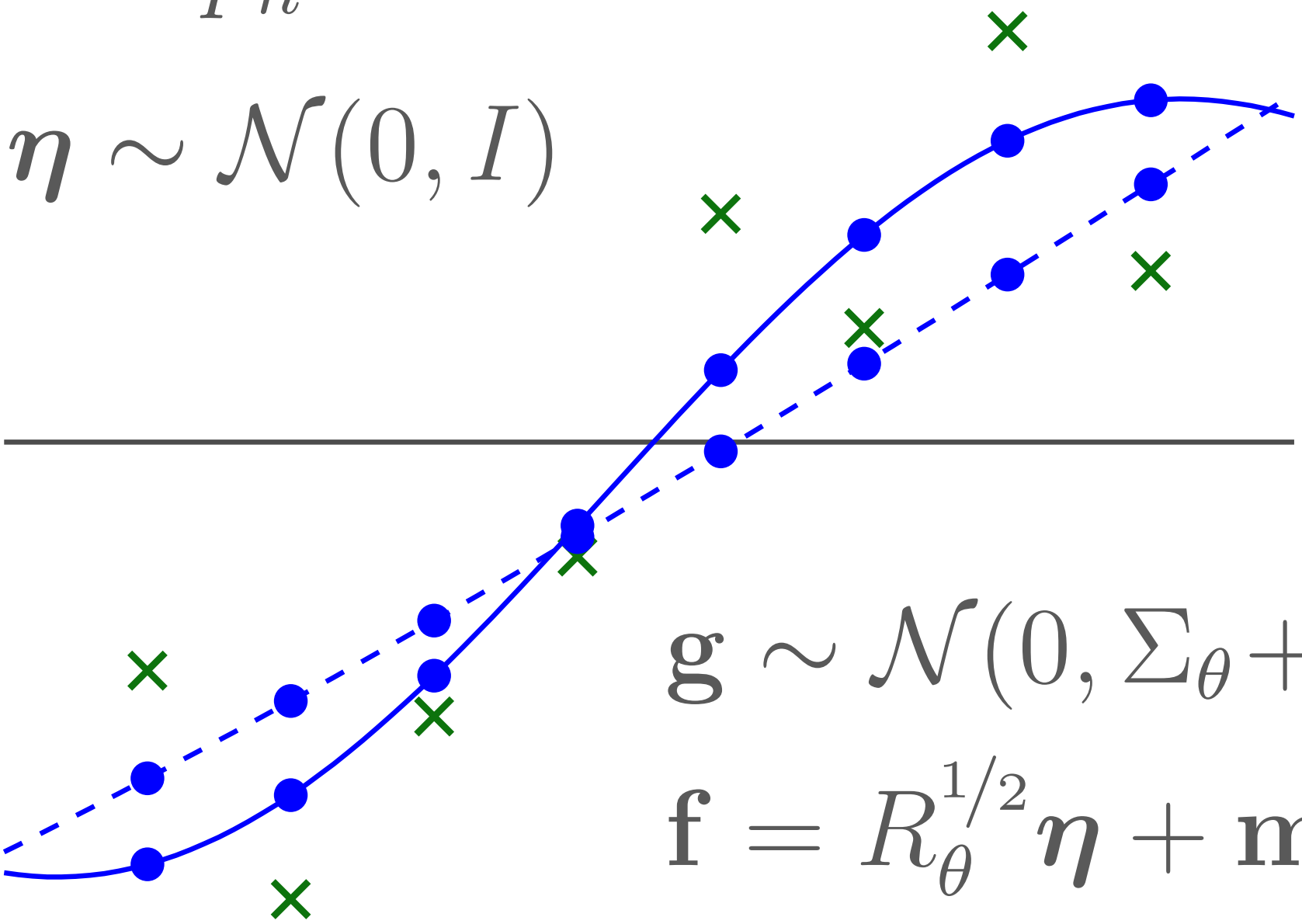
$\mathbf{g} \sim \mathcal{N}(0, \Sigma_\theta + S_\theta)$

$\mathbf{f} = R_\theta^{1/2} \boldsymbol{\eta} + \mathbf{m}_{\mathbf{g},\theta}$

Marginally $\mathbf{f} \sim \mathcal{N}(0, \Sigma_\theta)$

$\theta \sim p_h$

$\boldsymbol{\eta} \sim \mathcal{N}(0, I)$

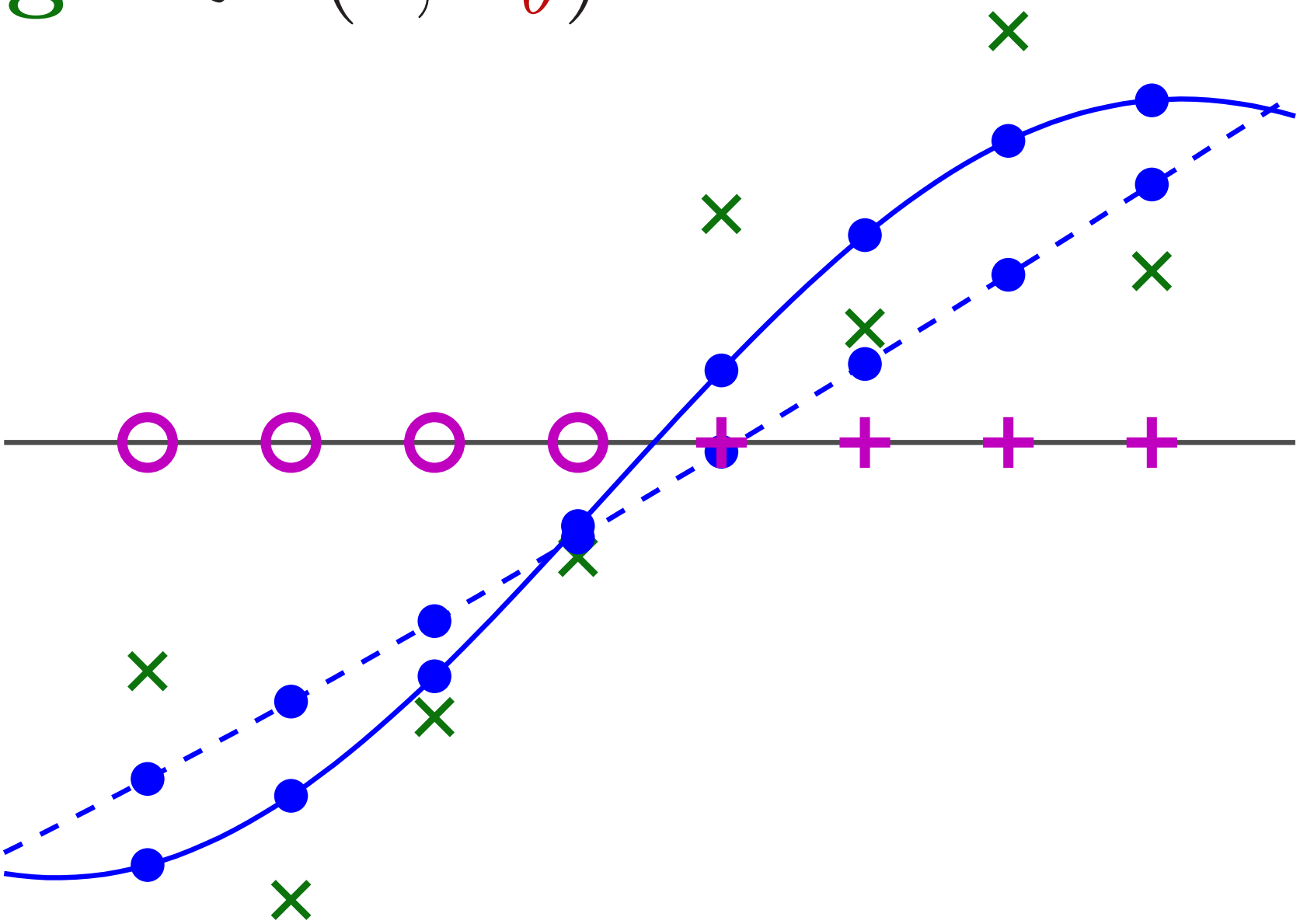


$\mathbf{g} \sim \mathcal{N}(0, \Sigma_\theta + S_\theta)$

$\mathbf{f} = R_\theta^{1/2} \boldsymbol{\eta} + \mathbf{m}_{\mathbf{g},\theta}$

Move $\theta \,|\, \mathbf{g}, \boldsymbol{\eta} \longrightarrow$ long lengthscale

$\mathbf{g} \sim \mathcal{N}(\mathbf{f}, S_\theta)$



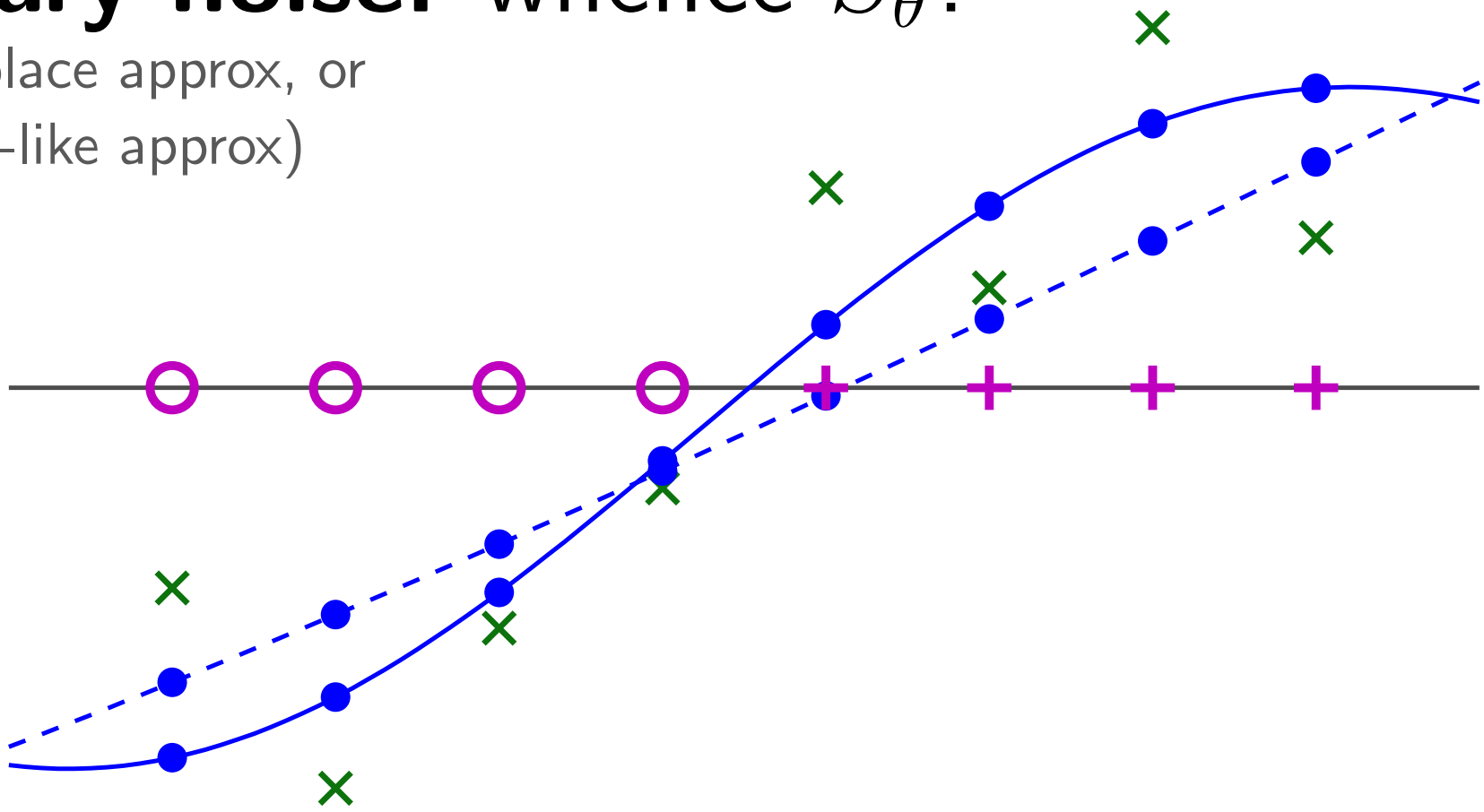
Propose move $\theta \,|\, \mathbf{g}, \boldsymbol{\eta}$

Setting everything

Step-sizes: how big are proposals?

Auxiliary noise: whence S_θ?

(Tune, Laplace approx, or
simple EP-like approx)



Slice sampling

Conservative steps: as $\theta' \to \theta$, $\mathbf{f}' \to \mathbf{f}$

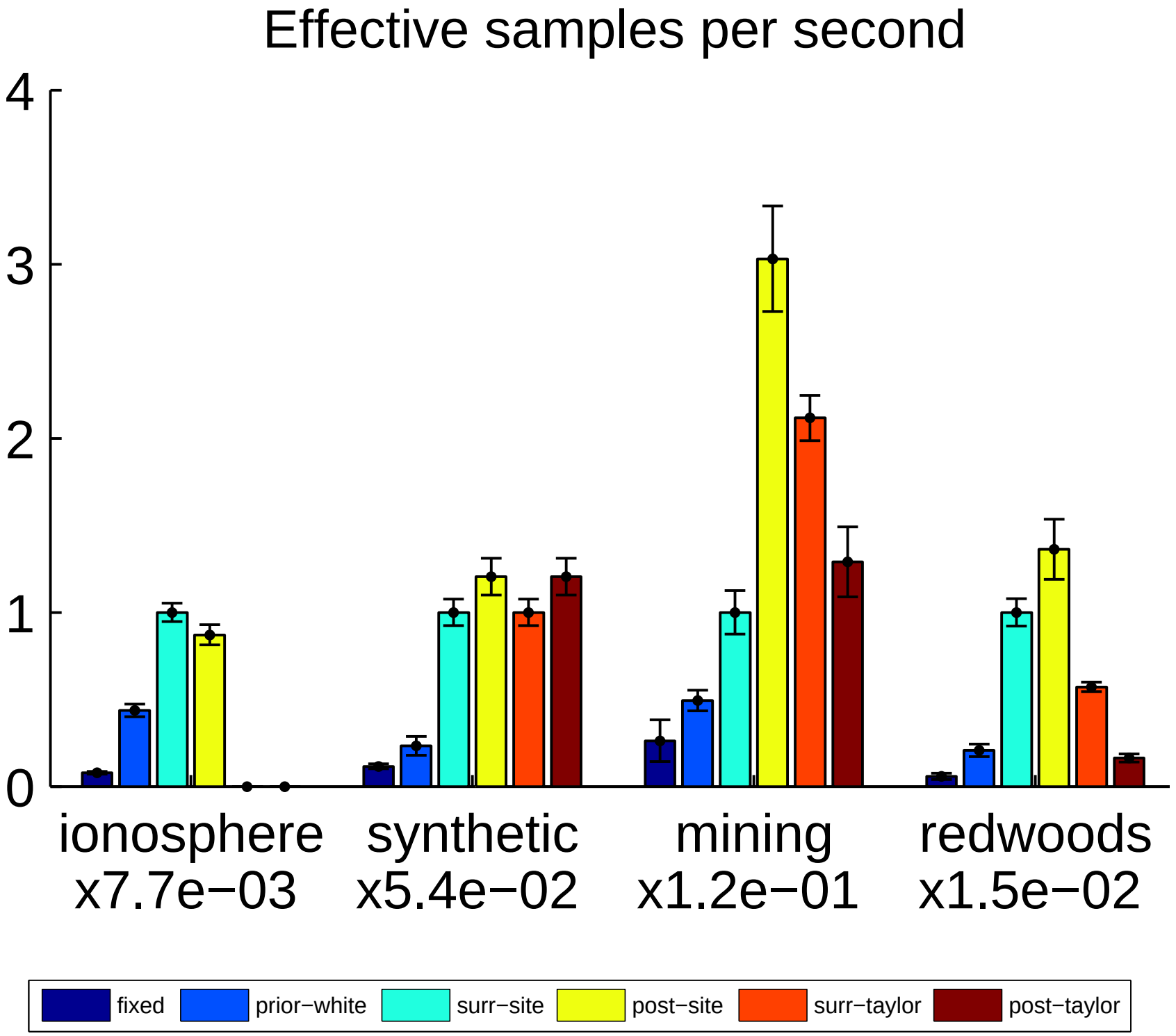
Slice Sampling (Neal, 2003)

— Always moves: shrinks steps as necessary

Elliptical Slice Sampling (M, A, MacKay, 2010)

— Optionally zero free parameters

— Especially suited to updating $\mathbf{f} \,|\, \theta$


Effective samples per second
4
3
2
1
0
ionosphere
x7.7e−03
synthetic
x5.4e−02
mining
x1.2e−01
redwoods
x1.5e−02
fixed
prior−white
surr−site
post−site
surr−taylor
post−taylor

Results: executive summary

Across a range of GP applications:

— prior whitening 2–10$\times$ faster than fixing $\mathbf{f}$

— simulations with surrogate data faster still

Some closely related work

Michalis Titsias (2010)
'Synthetic data' for MCMC on GPs

Christensen, Roberts and Skãld (2006)
Related 'robust' reparameterizations

Rue, Martino, Chopin; Cseke & Heskes
Recent deterministic developments

Summing up

Latents and hypers strongly coupled

(+ 'entropic' mixing problems)

Known problem:

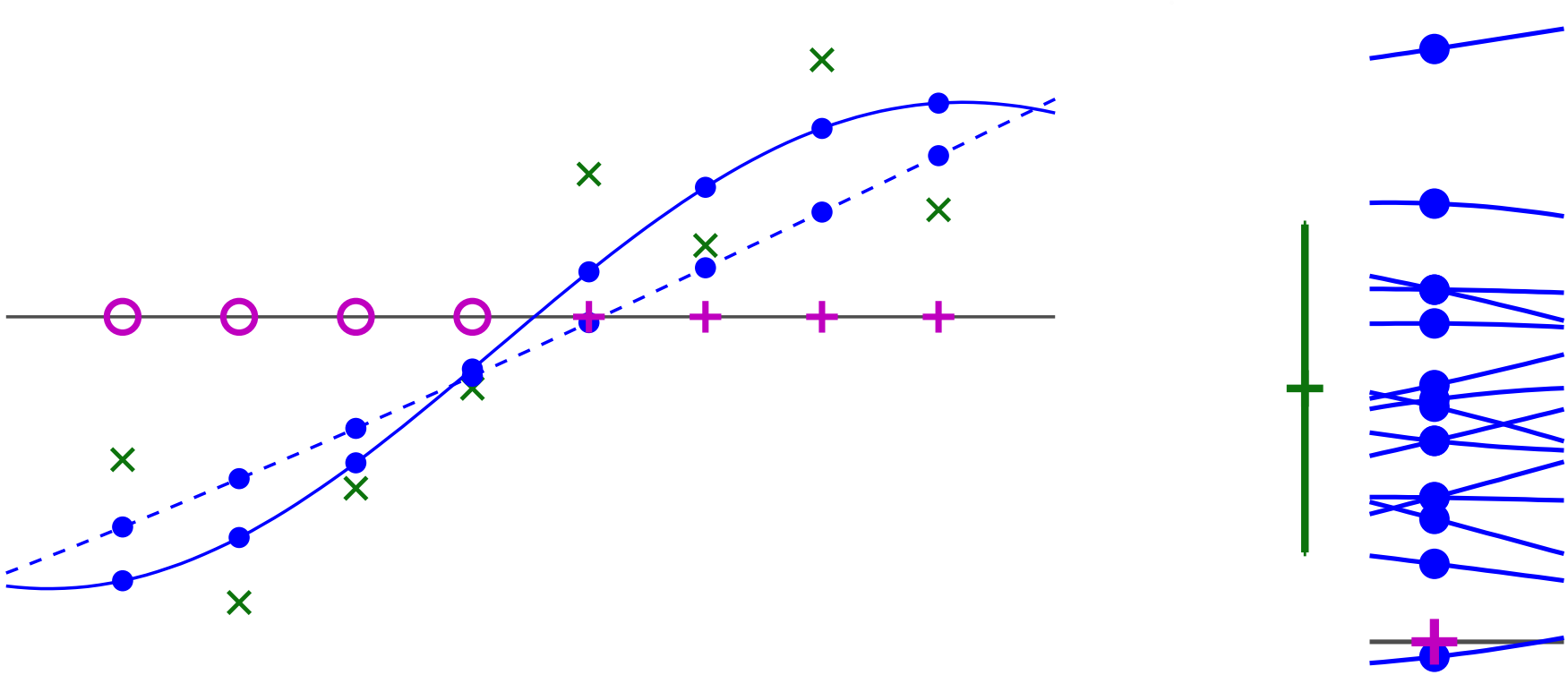
— we should do more automatically

— for non-Gaussian latents too!

Appendix Slides

Setting auxiliary noise

$\mathbf{g} \sim \mathcal{N}(\mathbf{f}, S_\theta)$ are 'surrogate data'

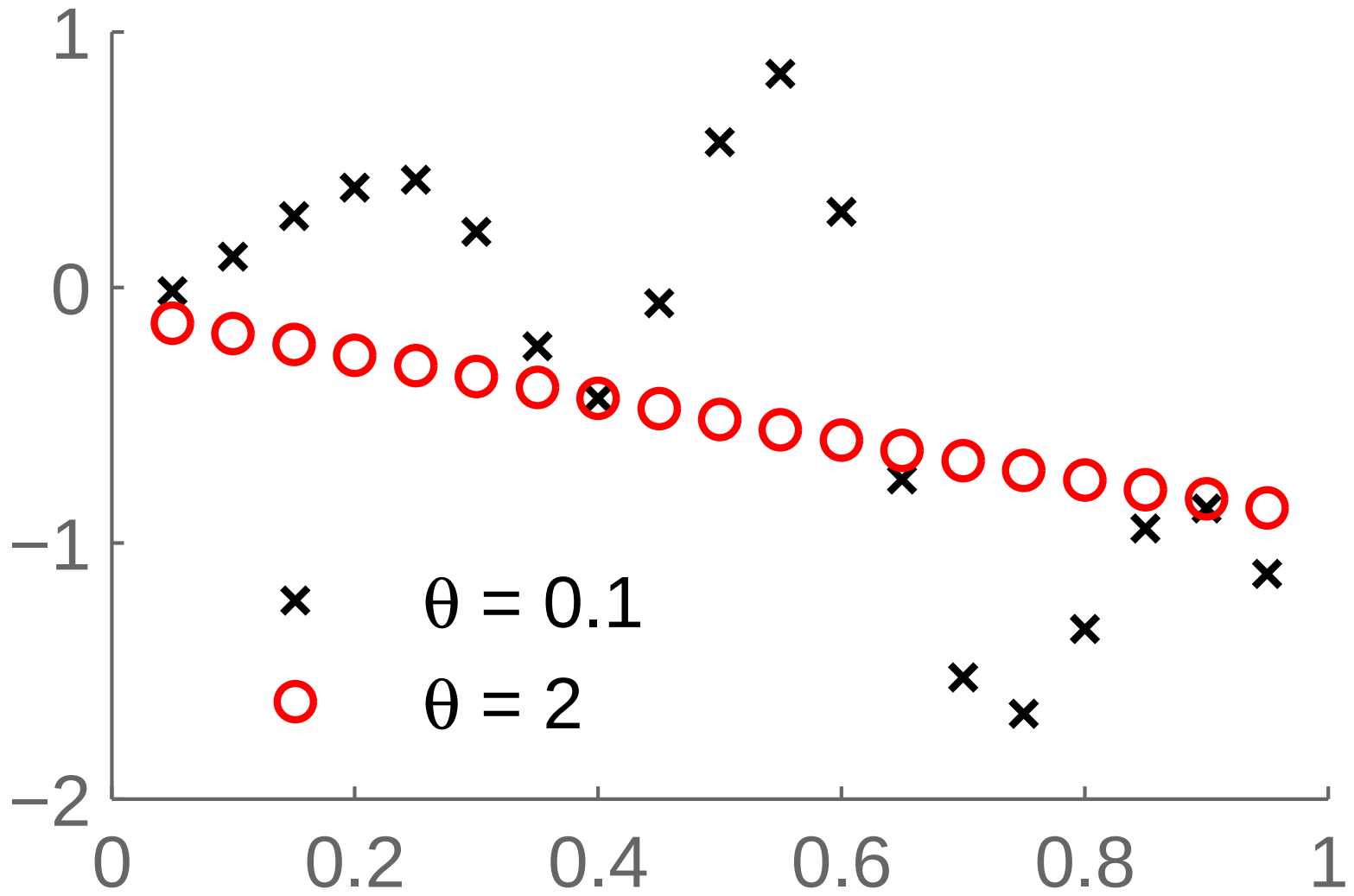


Consider isolated datapoints

Make site surrogate data posterior width $\approx$ site posterior

We're not mode-searching

Start at **Red** values. Propose short scale $\theta = 0.1$.



Red values are $> 500\times$ more probable than **Black**