Information maximising non-linear data compression

Tom Charnock

About me

Current work

- Started at the IAP in May
- Statistical problems
- Novel uses of neural networks
- IAP machine learning journal club

Previous work

- Theoretical particle physics PhD at Nottingham
- Statistical properties of the CMB
- Cosmic strings
- Classification of supernovae

Information maximising non-linear data compression

Fisher information

Fisher information matrix

$$\mathbf{F}_{\alpha\beta}(\theta) = -\left\langle \frac{\partial^2 \ln \mathcal{L}(\mathbf{d}|\theta)}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle \Bigg|_{\theta=\theta_{\rm fid}}$$

Fisher information of Gaussian likelihood

$$-2\ln\mathcal{L}\left(\mathbf{d}|\theta\right) = \left(\mathbf{d}-\mu(\theta)\right)^{\mathsf{T}}\mathbf{C}^{-1}\left(\mathbf{d}-\mu(\theta)\right) + n_{\mathbf{d}}\ln|2\pi\mathbf{C}|$$

$$\mathbf{F}_{\alpha\beta}(\theta) = \frac{1}{2}\mathrm{Tr}\left[\frac{\partial\mu(\theta)}{\partial\theta_\alpha}^{\mathsf{T}}\mathbf{C}^{-1}\frac{\partial\mu(\theta)}{\partial\theta_\beta} + \frac{\partial\mu(\theta)}{\partial\theta_\beta}^{\mathsf{T}}\mathbf{C}^{-1}\frac{\partial\mu(\theta)}{\partial\theta_\alpha}\right]$$

Assuming the covariance is independent of the parameters

Information conserving linear data compression

$$n_{\mathbf{d}} \rightarrow n_{\boldsymbol{\theta}}$$

Linear summaries

$$\tilde{d}_\alpha = \mathbf{r}_\alpha^{\mathsf{T}} \mathbf{d}$$

with each $\mathbf{r}_\alpha$ orthogonal such that $\tilde{d}_\alpha$ are uncorrelated.

$$\mathbf{r}_1 = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}^{\mathsf{T}}{}_{,1}\,\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}}$$

$$\mathbf{r}_\alpha = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}\left(\boldsymbol{\mu}^{\mathsf{T}}{}_{,\alpha}\,\mathbf{r}_\beta\right)\mathbf{r}_\beta}{\sqrt{\boldsymbol{\mu}^{\mathsf{T}}{}_{,1}\,\mathbf{C}^{-1}\boldsymbol{\mu}_{,1} - \sum_{\beta=1}^{\alpha-1}\left(\boldsymbol{\mu}^{\mathsf{T}}{}_{,\alpha}\,\mathbf{r}_\beta\right)^2}}$$

with $,_\alpha \equiv \partial/\partial\theta_\alpha$, $\boldsymbol{\mu} = \mu(\boldsymbol{\theta})$ and $\mathbf{C}$ and $\boldsymbol{\mu}$ evaluated at fiducial parameter values.

Less pleasant likelihoods

Unspecified function of the data $\mathrm{f}(\mathbf{d})$

$$-2\ln\mathcal{L}\left(\mathbf{d}|\boldsymbol{\theta}\right) = \left(\mathrm{f}(\mathbf{d}) - \boldsymbol{\mu}_\mathrm{f}\right)^\mathsf{T}\mathbf{C}_\mathrm{f}^{-1}\left(\mathrm{f}(\mathbf{d}) - \boldsymbol{\mu}_\mathrm{f}\right)$$

with mean and covariance over some simulations $\mathbf{s}_\mathrm{i}$

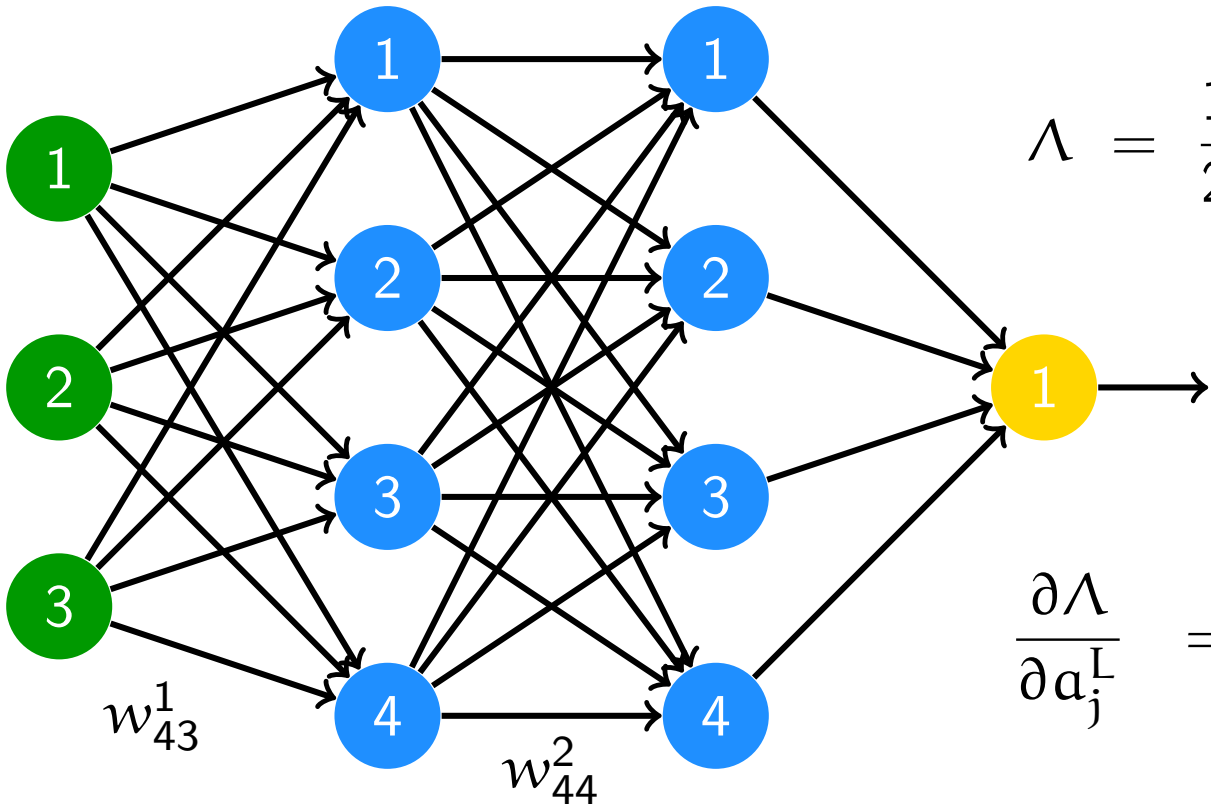
$$\boldsymbol{\mu}_\mathrm{f} = \frac{1}{\mathrm{n_s}}\sum_{\mathrm{i}=1}^{\mathrm{n_s}}\mathrm{f}(\mathbf{s}_\mathrm{i})$$

$$\mathbf{C}_\mathrm{f} = \frac{1}{\mathrm{n_s}-1}\sum_{\mathrm{i}=1}^{\mathrm{n_s}}\left(\mathrm{f}(\mathbf{s}_\mathrm{i}) - \boldsymbol{\mu}_\mathrm{f}\right)\left(\mathrm{f}(\mathbf{s}_\mathrm{i}) - \boldsymbol{\mu}_\mathrm{f}\right)$$

New Fisher information matrix

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\mu}_{\mathrm{f},\alpha}^\mathsf{T}\mathbf{C}_\mathrm{f}^{-1}\boldsymbol{\mu}_{\mathrm{f},\beta} + \boldsymbol{\mu}_{\mathrm{f},\beta}^\mathsf{T}\mathbf{C}_\mathrm{f}^{-1}\boldsymbol{\mu}_{\mathrm{f},\alpha}\right]$$

Neural networks - the unspecified function

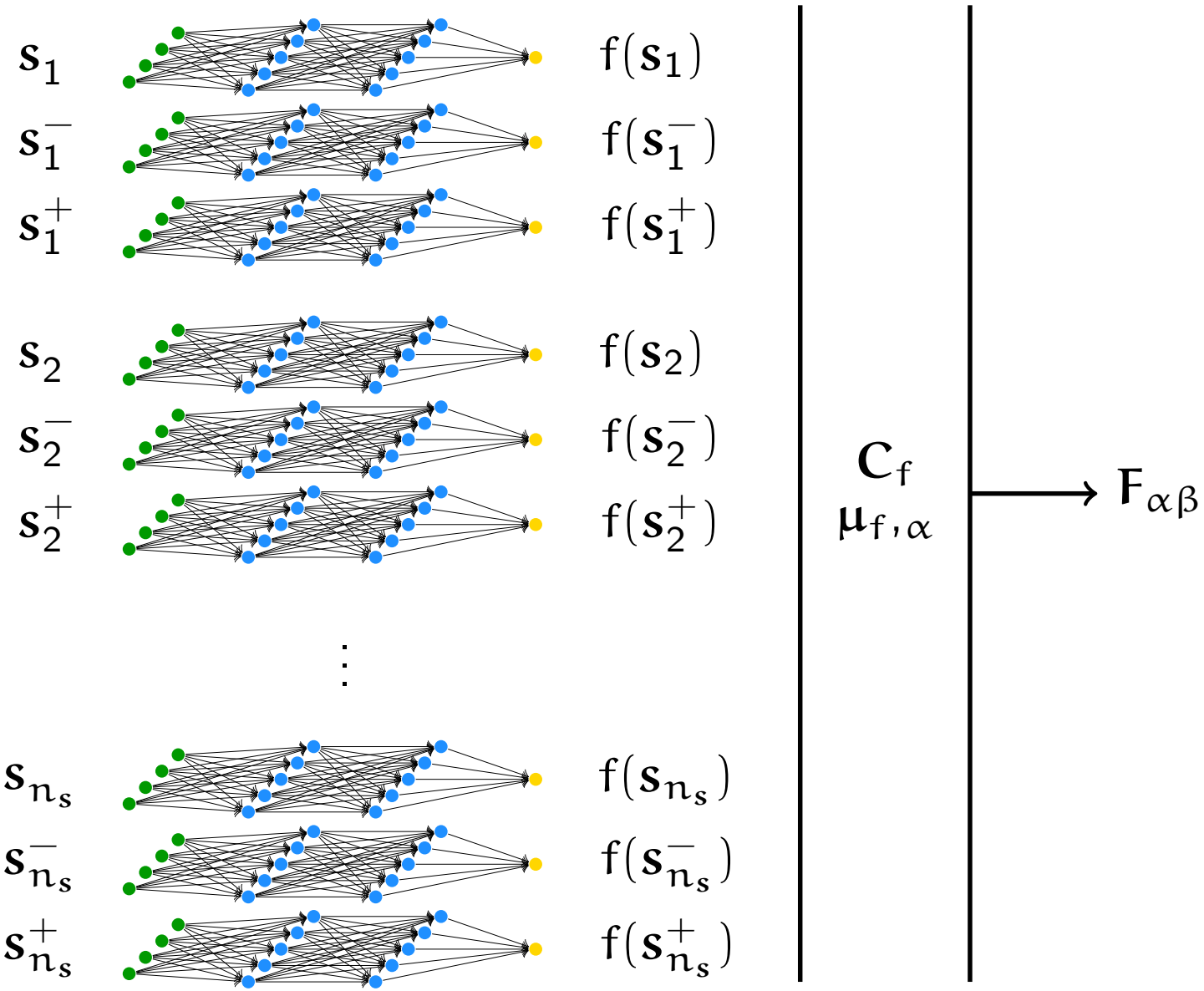


$$\Lambda = \frac{1}{2}\|t - a^L_j\|^2$$

$$\frac{\partial \Lambda}{\partial a^L_j} = t - a^L_j$$

$$\mathbf{x} = \phi(\mathbf{v}^0) \qquad \phi(\mathbf{v}^1) \qquad \phi(\mathbf{v}^2) \qquad \phi(\mathbf{v}^3) = \mathbf{a}^L$$

Information maximising neural network



Test model - the problem

Unknown variance Gaussian noise

Real data is $\mathbf{d} = \left\{ d_i \curvearrowleft \mathcal{N}(0, \theta) \middle| \; i = 1 \rightarrow n_{\mathbf{d}} \right\}$ where θ is unknown and $n_{\mathbf{d}} = 10$.

Analytically the likelihood is

$$-2 \ln \mathcal{L}(\mathbf{d}|\theta) = \frac{1}{\theta} \sum_{i=1}^{n_{\mathbf{d}}} d_i^2 + n_{\mathbf{d}} \ln [2\pi\theta]$$

Single sufficient statistic

$$T = \sum_{i=1}^{n_{\mathbf{d}}} d_i^2$$

The Fisher information is

$$\mathbf{F} = \frac{n_{\mathbf{d}}}{2\theta_{\text{fid}}}$$

Choosing $\theta_{\text{fid}} = 1$ gives $\mathbf{F} = 5$.

Test model - data compression

Linear data compression

The sufficient statistic for this problem is non-linear, i.e.

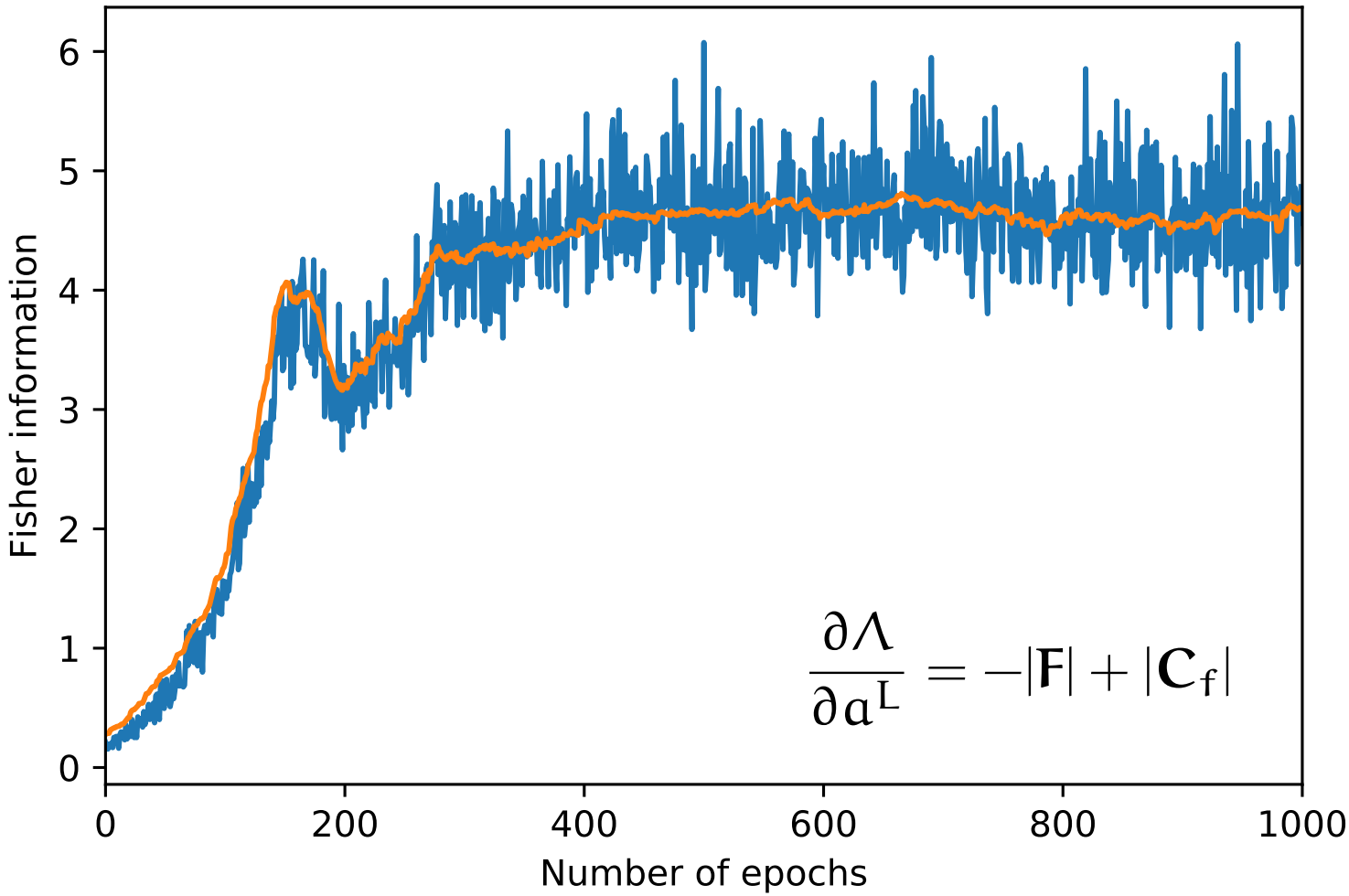
$$\mathrm{T} = \sum_{\mathrm{i}=1}^{\mathrm{n_d}} \mathrm{d_i^2}$$

Any linear combination of $\mathbf{d}$ will necessarily give $\mathbf{F}$ less than optimal.

Non-linear data compression

Can train the network to become the function f which fully specifies the sufficient statistic?

Test model - training



[96, 96], 0.5 dropout, leaky ReLU with $\alpha = 0.1$, $\eta = 0.01$, $\mathbf{b}^l_{\rm init} = 0.1$ $\boldsymbol{w}^l_{\rm init} = \mathcal{N}(0, \sqrt{2}/\kappa^{l-1})$,
1000 simulations, 200 derivatives, 2 batches, 1000 epochs

Artificial approximate Bayesian computation (AABC)

Get summary of real data $\mathrm{f}(\mathbf{d})$

Create simulation $\mathbf{s}_\mathrm{i}(\theta)$ within prior range of θ

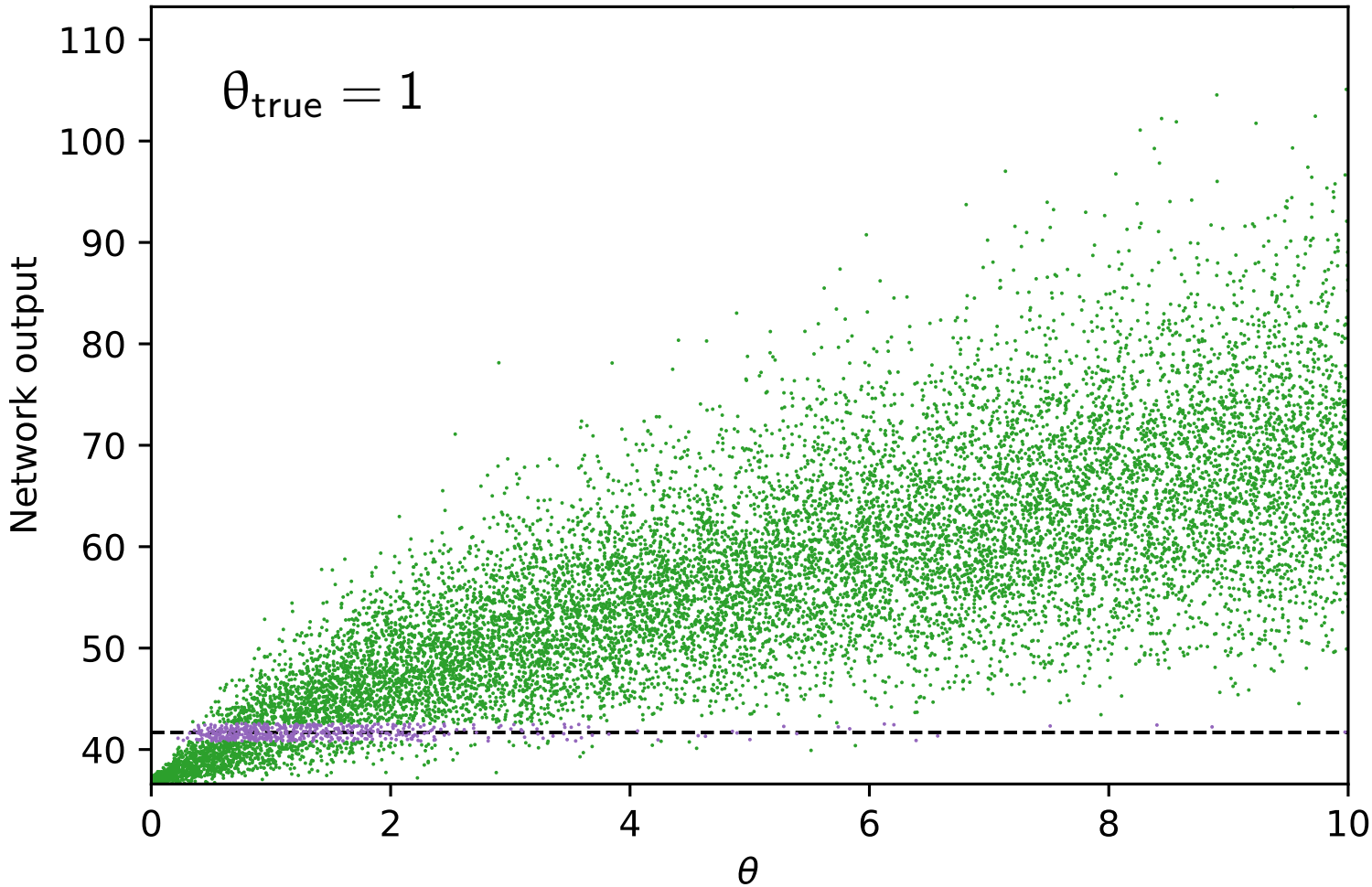
Calculate $\mathrm{f}(\mathbf{s}_\mathrm{i}(\theta))$

Accept point if

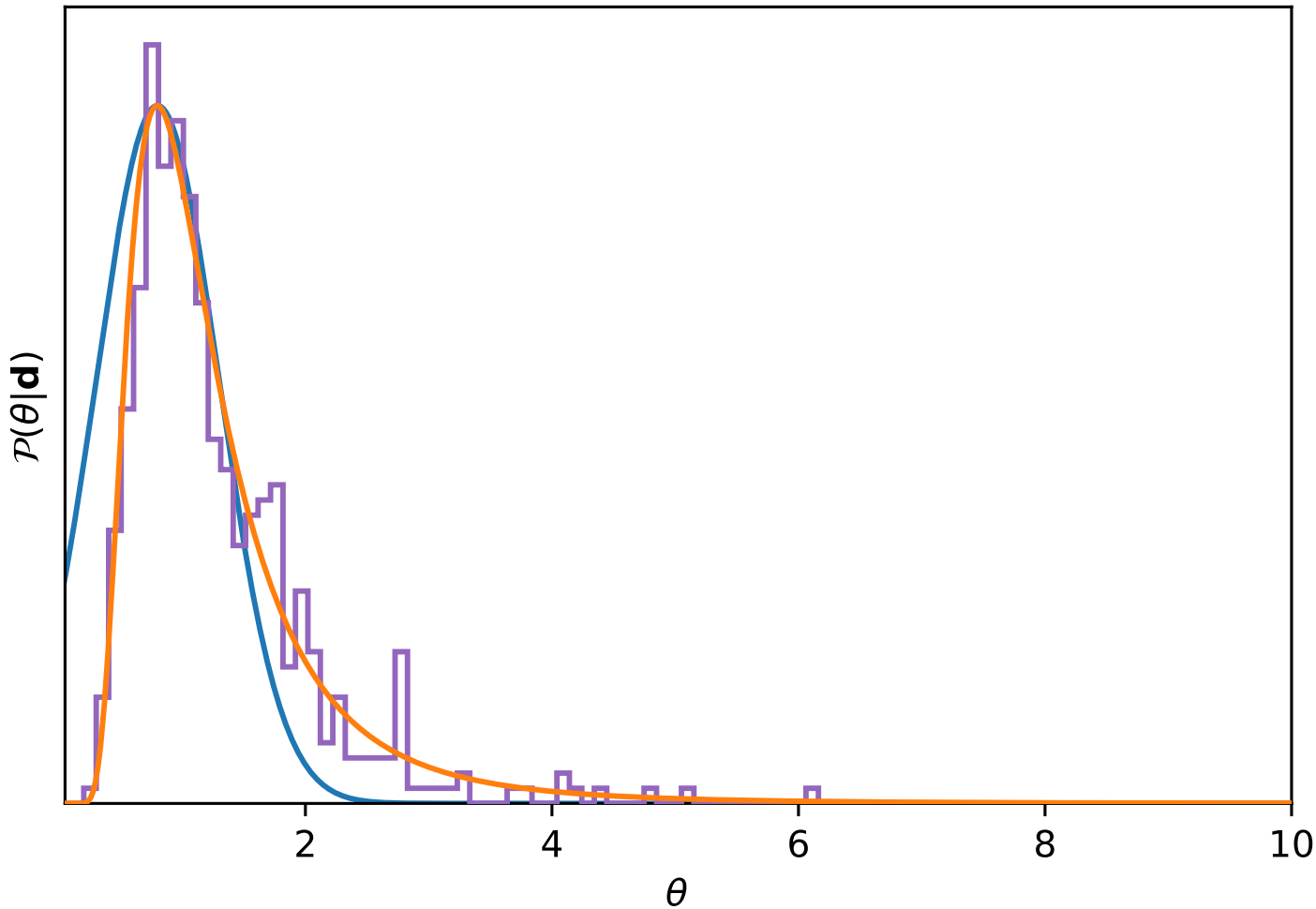
$$\rho = \sqrt{(\mathrm{f}(\mathbf{s}_\mathrm{i}(\theta)) - \mathrm{f}(\mathbf{d}))^2} < \varepsilon$$

Actually use PMC rather than random draws from the prior

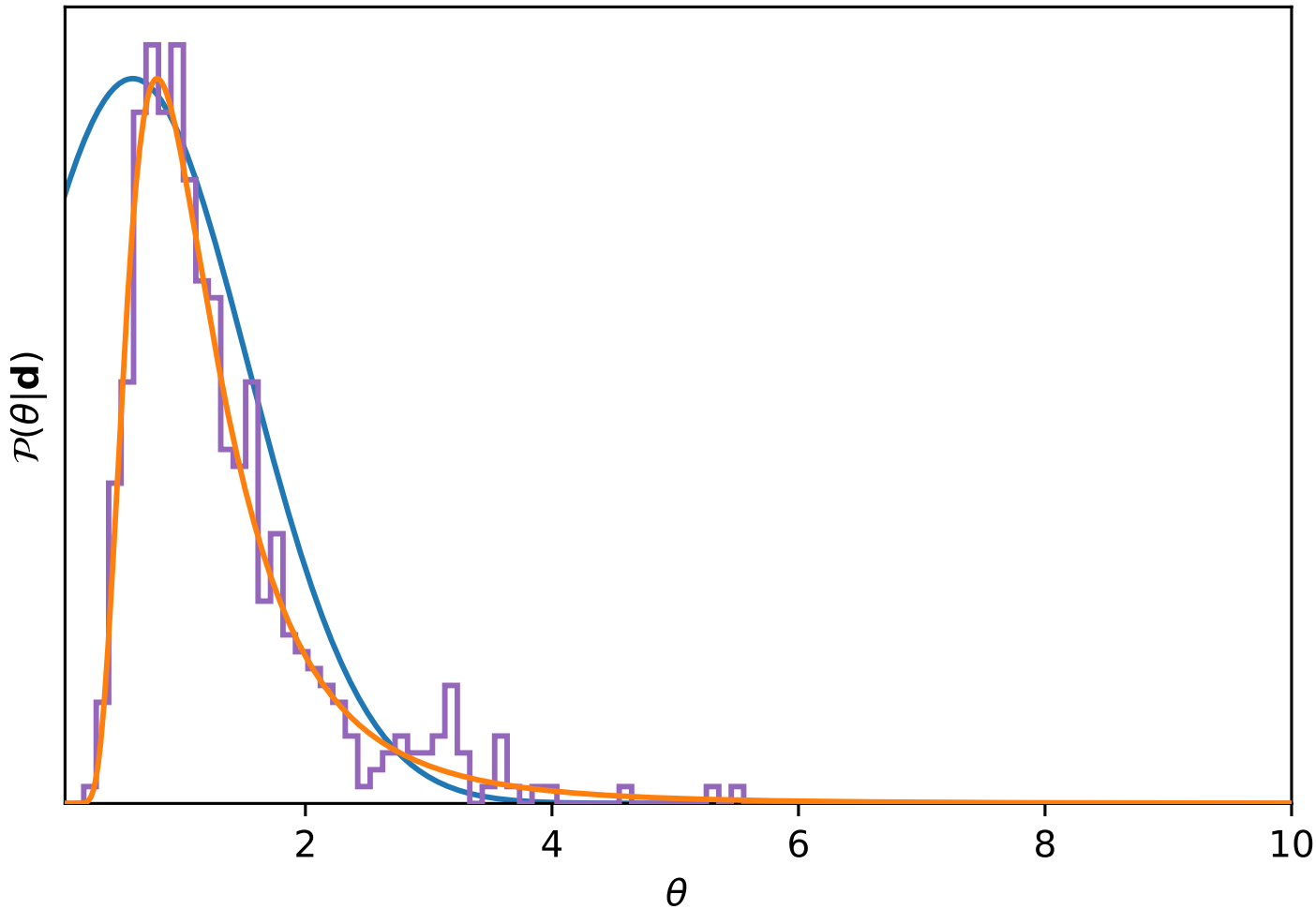
Network output



$P(\theta|\mathbf{d})$



$P(\theta|\mathbf{d})$ when $\theta_{\rm fid} = 2$



Information maximising non-linear data compression

- Linear data compression can find summaries of Gaussian likelihood (and approximations thereof).
- Non-linear data compression can find summaries of any data where a function of the data can be written like a Gaussian.
- The function of the data which maximises the Fisher information is the function which best summarises the data - although this function is unknown.
- A neural network can be trained (using relatively few training samples) to find the function by maximising the Fisher information as the loss function.
- This function can now be used for AABC to find $\mathrm{P}(\theta|\mathbf{d})$.

About me (non-work-related ice breakers)

Music

- Avid fan of traditional folk music from all over the world
- Played in several bands with styles from metal to blues to folk
- Melodeon
- Bouzouki
- Electric guitar

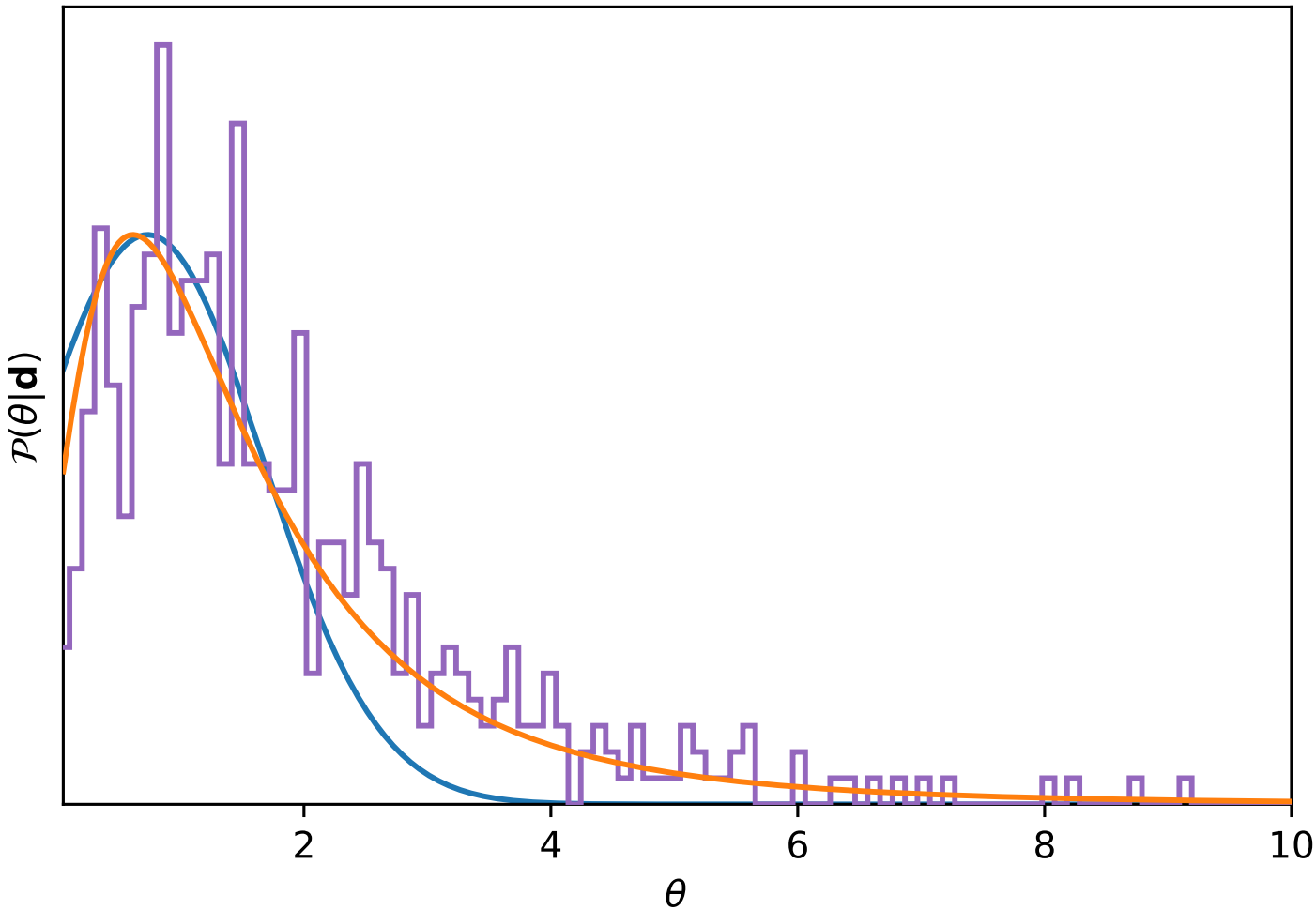


The finer things...

- A well-made (hipster?) coffee
- Craft ale
- Bread and cheese (no wonder I moved to France)
- Listening to music on vinyl

Etc...

$P(\theta|\mathbf{d})$ when $\theta_{\rm fid} = 1$ with known noise $\sigma^2_{\rm noise} = 1$



$P(\theta|\mathbf{d})$ when $\theta_{\rm fid} = 1$ with unknown noise $\sigma^2_{\rm noise} = [0, 2]$