Some Background in Probabilistic Modelling

Maneesh Sahani

maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

Simple Probabilistic Models

- beliefs about data represented by parameterised distribution

$$P(\mathcal{D} \mid \theta, m) = \prod_{i=1}^{n} P(x_i \mid \theta, m)$$

- learning involves estimating a value (or distribution) for θ

Bayes $$P(\theta \mid \mathcal{D}, m) = \frac{P(\mathcal{D} \mid \theta, m) P(\theta \mid m)}{P(\mathcal{D} \mid m)}$$

MAP $$\theta^* = \operatorname*{argmax}_{\theta} P(\theta \mid \mathcal{D}, m)$$

ML $$\theta^* = \operatorname*{argmax}_{\theta} P(\mathcal{D} \mid \theta, m)$$

and (possibly) model selection

$$P(m|\mathcal{D}) \propto \int d\theta \; P(\mathcal{D}|\theta, m) P(\theta|m)$$

Isn’t that it?

Isn’t that it?

No!

Isn’t that it?

No!

- direct application of Bayes’ rule is intractable for all but the simplest models.

Isn’t that it?

No!

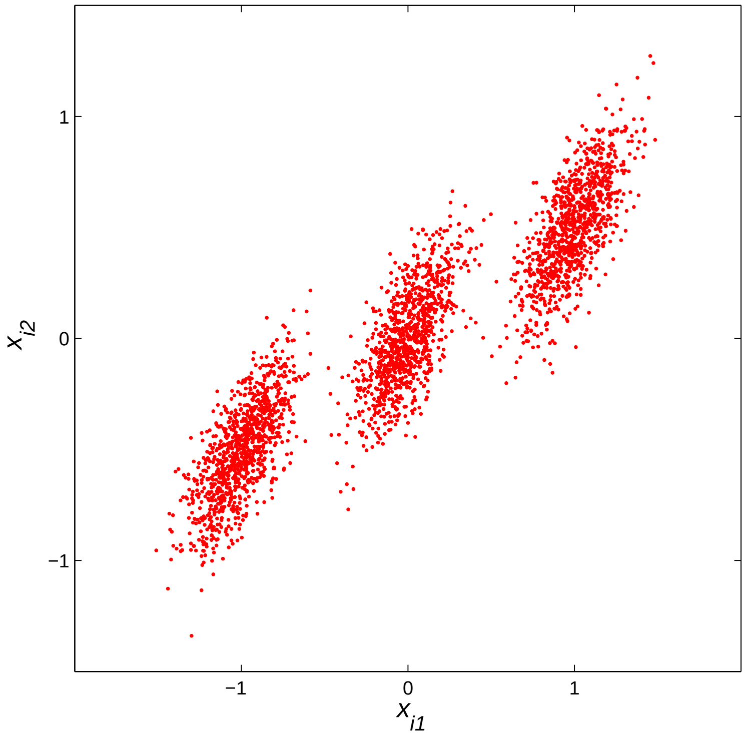
- direct application of Bayes’ rule is intractable for all but the simplest models.
- even maximum-likelihood (or similar) learning may be prohibitive.

Isn’t that it?

No!

- direct application of Bayes’ rule is intractable for all but the simplest models.
- even maximum-likelihood (or similar) learning may be prohibitive.
- algorithms (and approximations) are dictated by form of distribution.

What about these data?

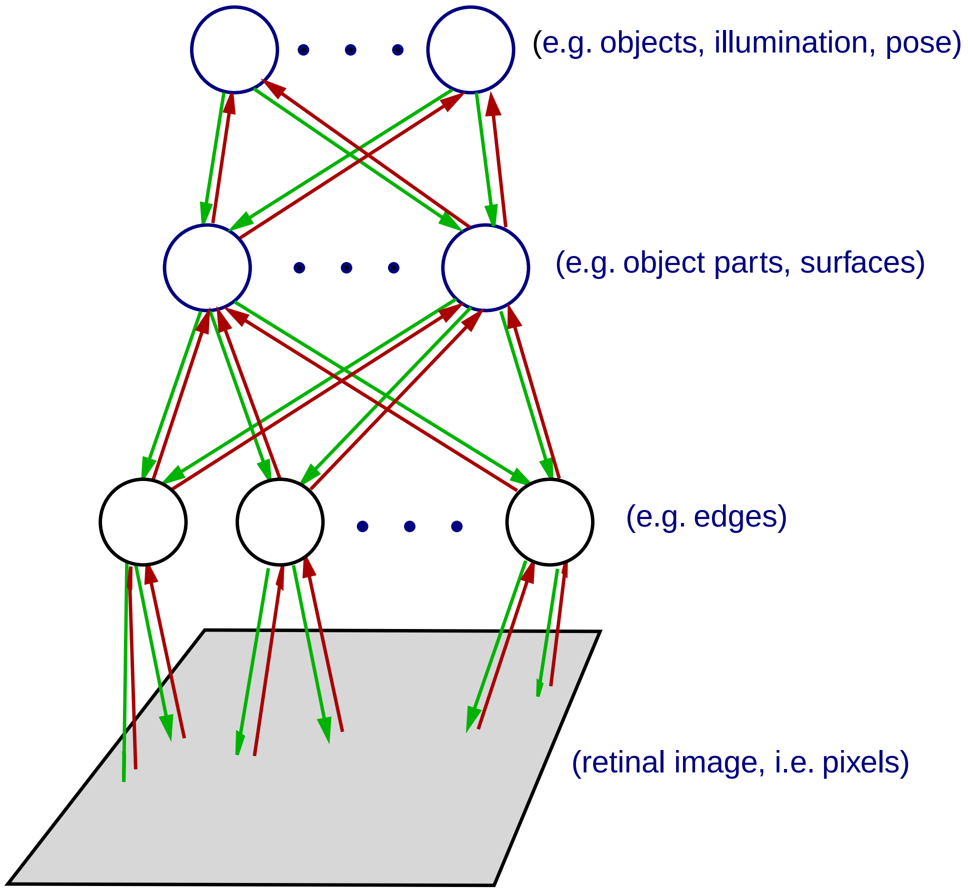


- joint distribution $p(x_{i1}, x_{i2})$ is not Normal.
- conditional distributions $p(x_{i1} \mid x_{i2})$ and $p(x_{i2} \mid x_{i1})$ vary, and are difficult to codify.
- easier to describe by an latent tri-valued discrete process

$$\begin{aligned} s_i &\sim Discrete[1/3, 1/3, 1/3] \\ \mathbf{x}_i &\sim \mathcal{N}\left(\mu_{s_i}, \Sigma\right) \end{aligned}$$

Latent Variable Models

Explain correlations in **x** by assuming some latent variables **y**


(e.g. objects, illumination, pose)
(e.g. object parts, surfaces)
(e.g. edges)
(retinal image, i.e. pixels)

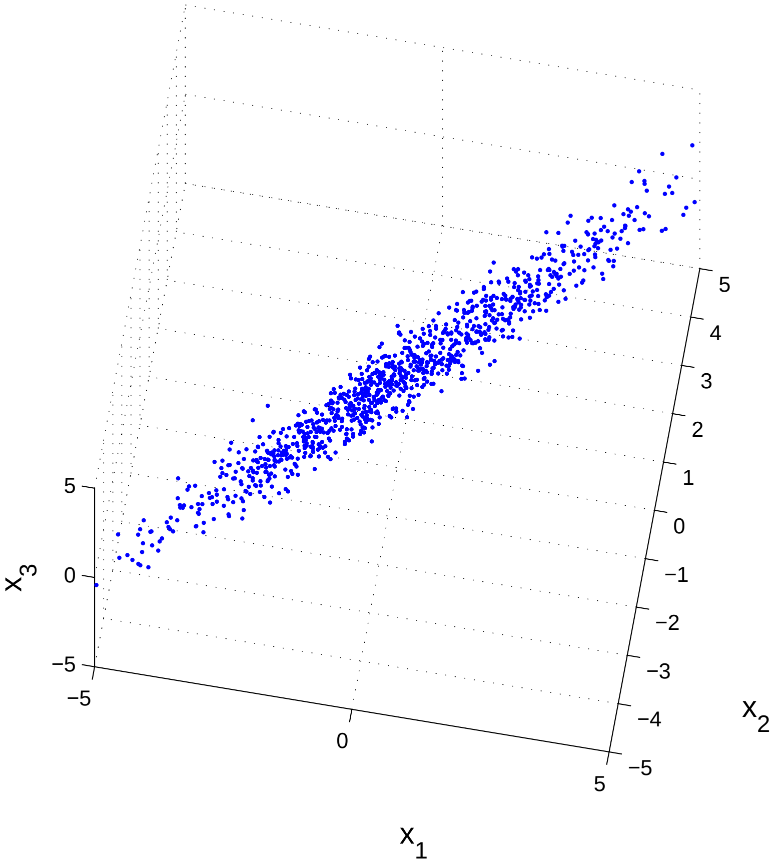
$$\mathbf{y} \sim \mathcal{P}[\theta_y]$$

$$\mathbf{x} \mid \mathbf{y} \sim \mathcal{P}[\theta_x]$$

$$p(\mathbf{x}, \mathbf{y}; \theta_x, \theta_y) = p(\mathbf{x} \mid \mathbf{y}; \theta_x) p(\mathbf{y}; \theta_y)$$

$$p(\mathbf{x}; \theta_x, \theta_y) = \int d\mathbf{y}\; p(\mathbf{x} \mid \mathbf{y}; \theta_x) p(\mathbf{y}; \theta_y)$$

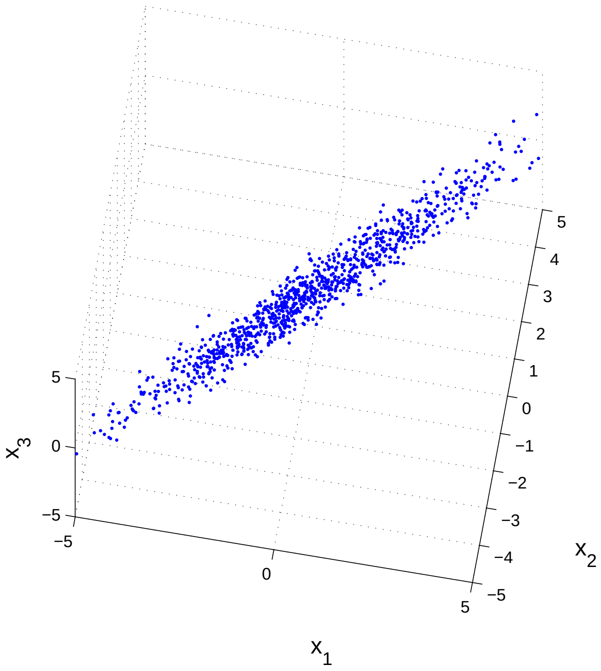
Latent variables and Gaussians



The data were measured in 3D, but are mostly specified by position on a 2D surface.

Principal Components Analysis

Data such as these are often modelled using Principal Components Analysis (PCA).



Assume data $\mathcal{D} = \{\mathbf{x}_i\}$ have zero mean (if not, subtract it!).

- Find direction of greatest variance – $\boldsymbol{\lambda}_{(1)}$.

$$\boldsymbol{\lambda}_{(1)} = \underset{\|\mathbf{v}\|=1}{\operatorname{argmax}} \sum_n (\mathbf{x}_n^{\mathsf{T}}\mathbf{v})^2$$

- Find direction orthogonal to $\boldsymbol{\lambda}_{(1)}$ with greatest variance – $\boldsymbol{\lambda}_{(2)}$

  ⋮

- Find direction orthogonal to $\{\boldsymbol{\lambda}_{(1)}, \boldsymbol{\lambda}_{(2)}, \ldots, \boldsymbol{\lambda}_{(n-1)}\}$ with greatest variance – $\boldsymbol{\lambda}_{(n)}$.
- Terminate when remaining variance drops below a threshold.

Eigendecomposition of a Covariance Matrix

The eigendecomposition of a covariance matrix makes finding the PCs easy.
Recall that $\mathbf{u}$ is an eigenvector, with scalar eigenvalue ω, of a matrix C if

$$C\mathbf{u} = \omega\mathbf{u}$$

$\mathbf{u}$ can have any norm, but we will define it to be unity (i.e., $\mathbf{u}^{\mathsf{T}}\mathbf{u} = 1$).
For a covariance matrix $C = \left\langle \mathbf{x}\mathbf{x}^{\mathsf{T}} \right\rangle$ (which is $D \times D$, symmetric, positive semi-definite):

- In general there are D eigenvector-eigenvalue pairs $(\mathbf{u}_{(i)}, \omega_{(i)})$, except if two or more eigenvectors share the same eigenvalue (in which case the eigenvectors are degenerate — any linear combination is also an eigenvector).
- The D eigenvectors are orthogonal (or orthogonalisable, if $\omega_{(i)} = \omega_{(j)}$). Thus, they form an orthonormal basis. $\sum_i \mathbf{u}_{(i)}\mathbf{u}_{(i)}{}^{\mathsf{T}} = I$.
- Any vector $\mathbf{v}$ can be written as

$$\mathbf{v} = \left(\sum_i \mathbf{u}_{(i)}\mathbf{u}_{(i)}{}^{\mathsf{T}}\right)\mathbf{v} = \sum_i (\mathbf{u}_{(i)}{}^{\mathsf{T}}\mathbf{v})\mathbf{u}_{(i)} = \sum_i v_{(i)}\mathbf{u}_{(i)}$$

- The original matrix C can be written:

$$C = \sum_i \omega_{(i)}\mathbf{u}_{(i)}\mathbf{u}_{(i)}{}^{\mathsf{T}} = USU^{\mathsf{T}}$$

where $U = \left[\mathbf{u}_{(1)}, \mathbf{u}_{(2)}, \ldots, \mathbf{u}_{(D)}\right]$ collects the eigenvectors and $S = \text{diag}\left[(\omega_{(1)}, \omega_{(2)}, \ldots, \omega_{(D)})\right]$.

PCA and Eigenvectors

- The variance in direction $\mathbf{u}_{(i)}$ is

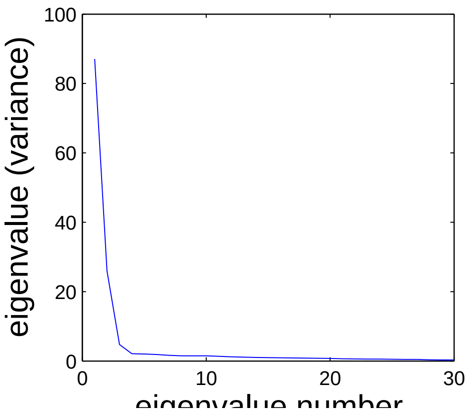
$$\left\langle (\mathbf{x}^\mathsf{T}\mathbf{u}_{(i)})^2 \right\rangle = \left\langle \mathbf{u}_{(i)}{}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{u}_{(i)} \right\rangle = \mathbf{u}_{(i)}{}^\mathsf{T} C \mathbf{u}_{(i)} = \mathbf{u}_{(i)}{}^\mathsf{T}\omega_{(i)}\mathbf{u}_{(i)} = \omega_{(i)}$$

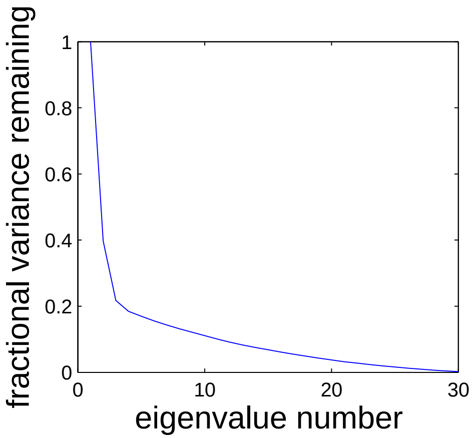
- The variance in an arbitrary direction $\mathbf{v}$ is

$$\left\langle (\mathbf{x}^\mathsf{T}\mathbf{v})^2 \right\rangle = \left\langle \left(\mathbf{x}^\mathsf{T}\left(\sum_i v_{(i)}\mathbf{u}_{(i)}\right)\right)^2 \right\rangle = \sum_{ij} v_{(i)}\mathbf{u}_{(i)}{}^\mathsf{T} C \mathbf{u}_{(j)} v_{(j)}$$

$$= \sum_{ij} v_{(i)}\omega_{(j)}v_{(j)}\mathbf{u}_{(i)}{}^\mathsf{T}\mathbf{u}_{(j)} = \sum_i v_{(i)}^2\omega_{(i)}$$

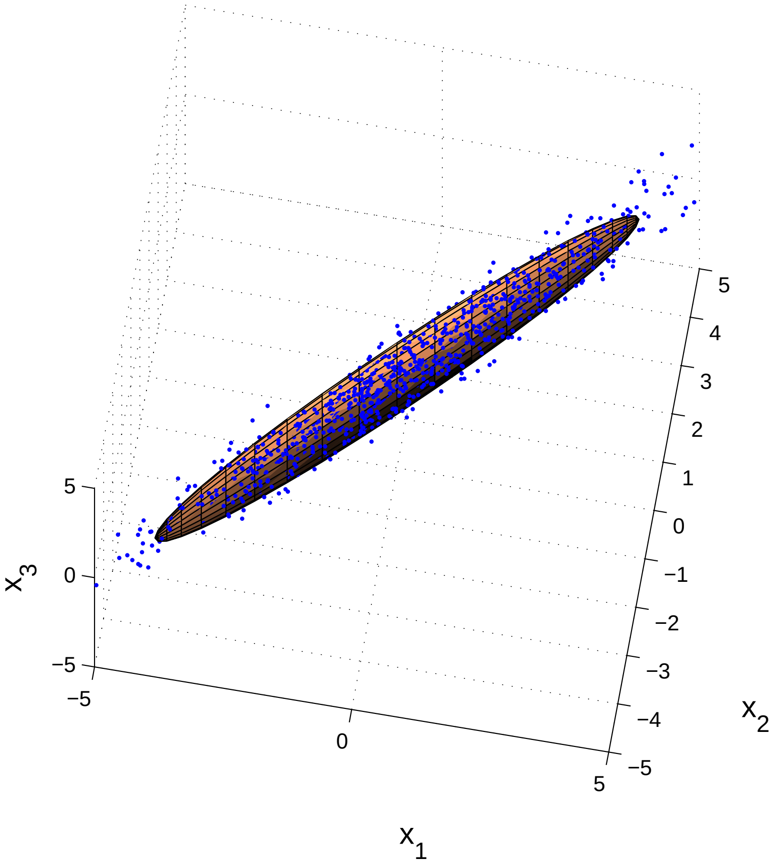
- If $\mathbf{v}^\mathsf{T}\mathbf{v} = 1$, then $\sum_i v_{(i)}^2 = 1$ and so $\operatorname{argmax}_{\|\mathbf{v}\|=1} \left\langle (\mathbf{x}^\mathsf{T}\mathbf{v})^2 \right\rangle = \mathbf{u}_{(\mathrm{max})}$
  The direction of greatest variance is the eigenvector the largest eigenvalue.
- In general, the PCs are exactly the eigenvectors of the empirical covariance matrix, ordered by decreasing eigenvalue.
- The eigenspectrum shows how the variance is distributed across dimensions





PCA subspace

The K principle components define the K-dimensional subspace of greatest variance.

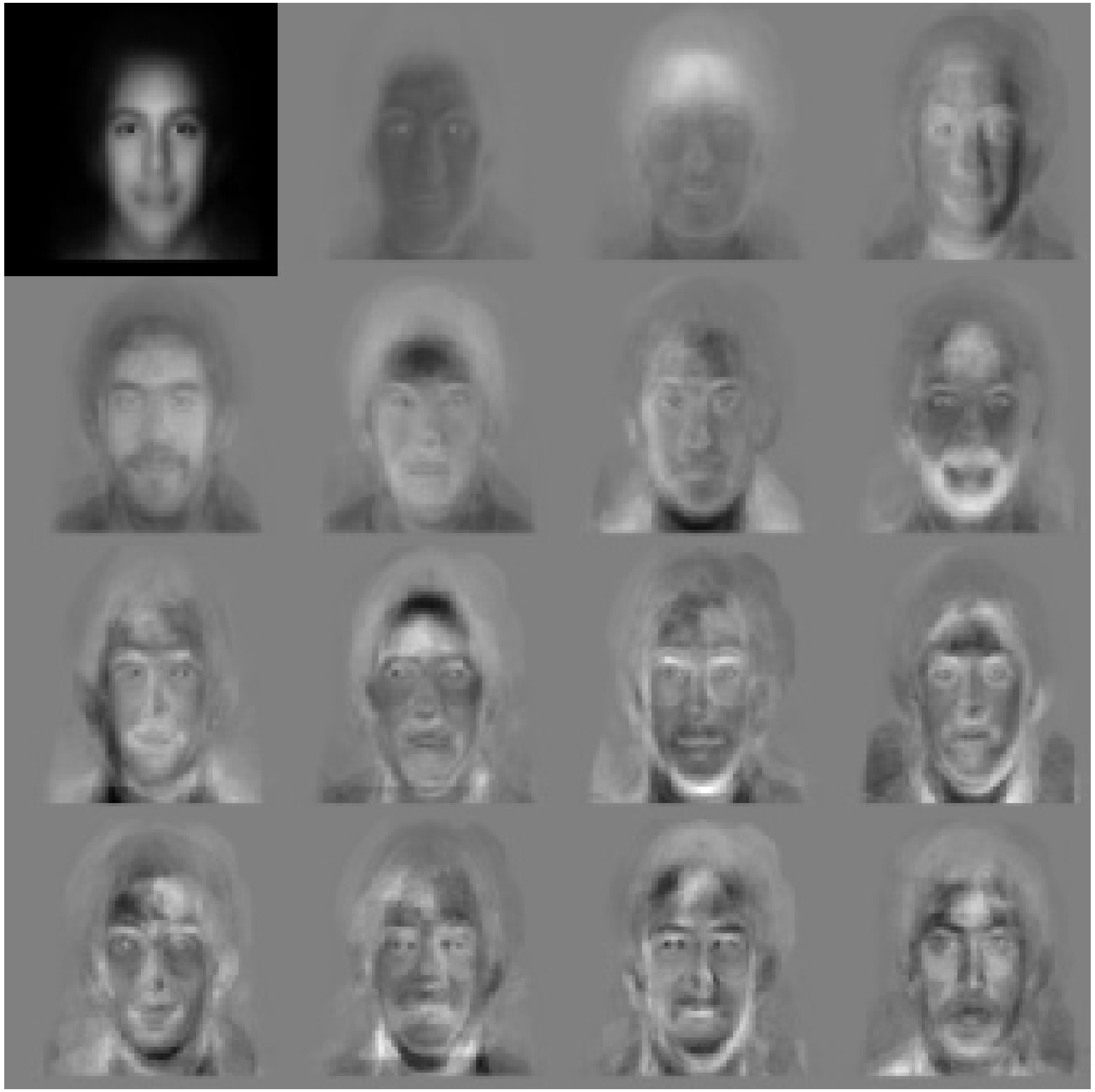


- Each data point $\mathbf{x}_n$ is associated with a projection $\hat{\mathbf{x}}_n$ into the principle subspace.

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} x_{n(k)} \boldsymbol{\lambda}_{(k)}$$

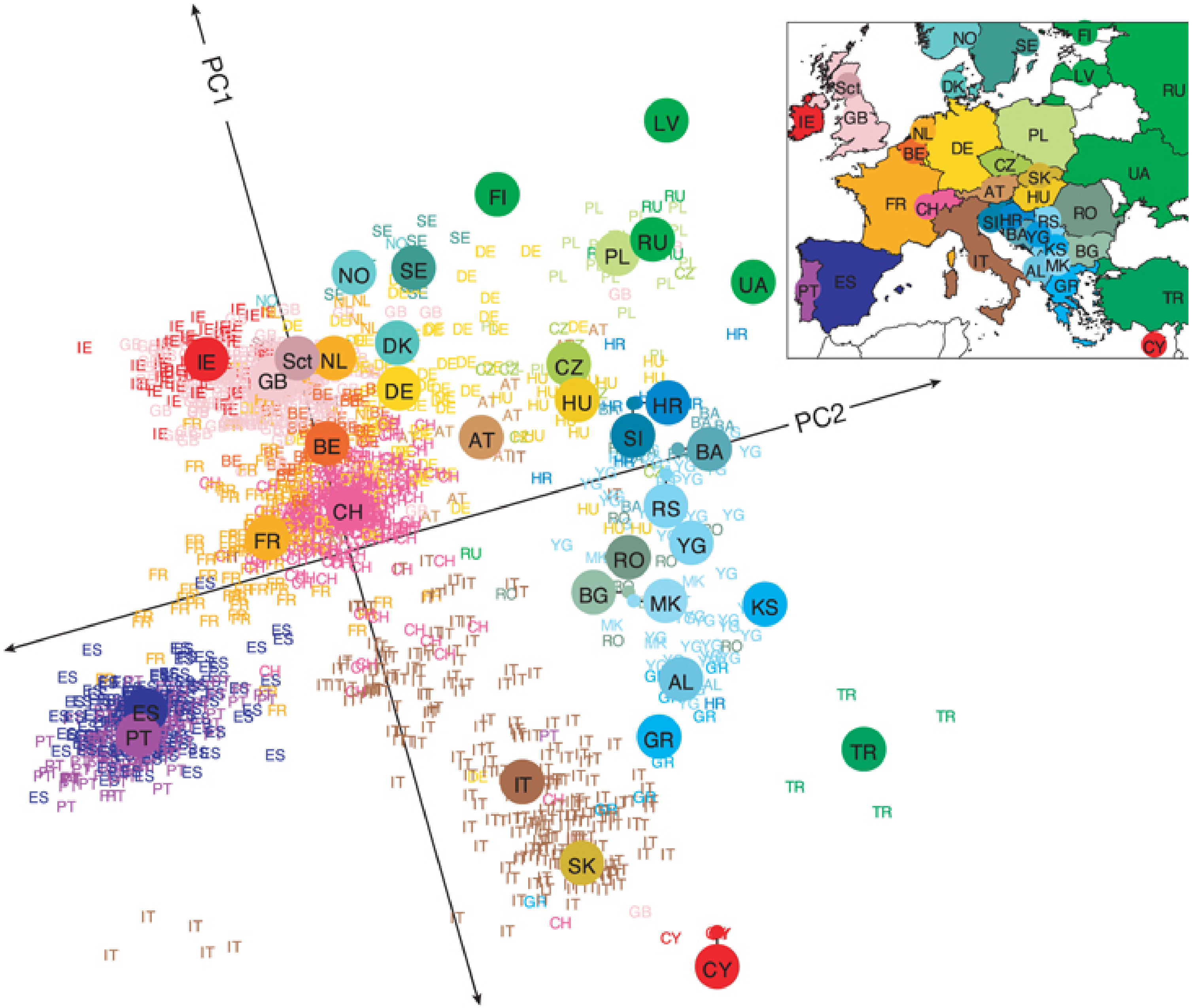
- This can be used for lossy compression, denoising, recognition, . . .

Example of PCA: Eigenfaces



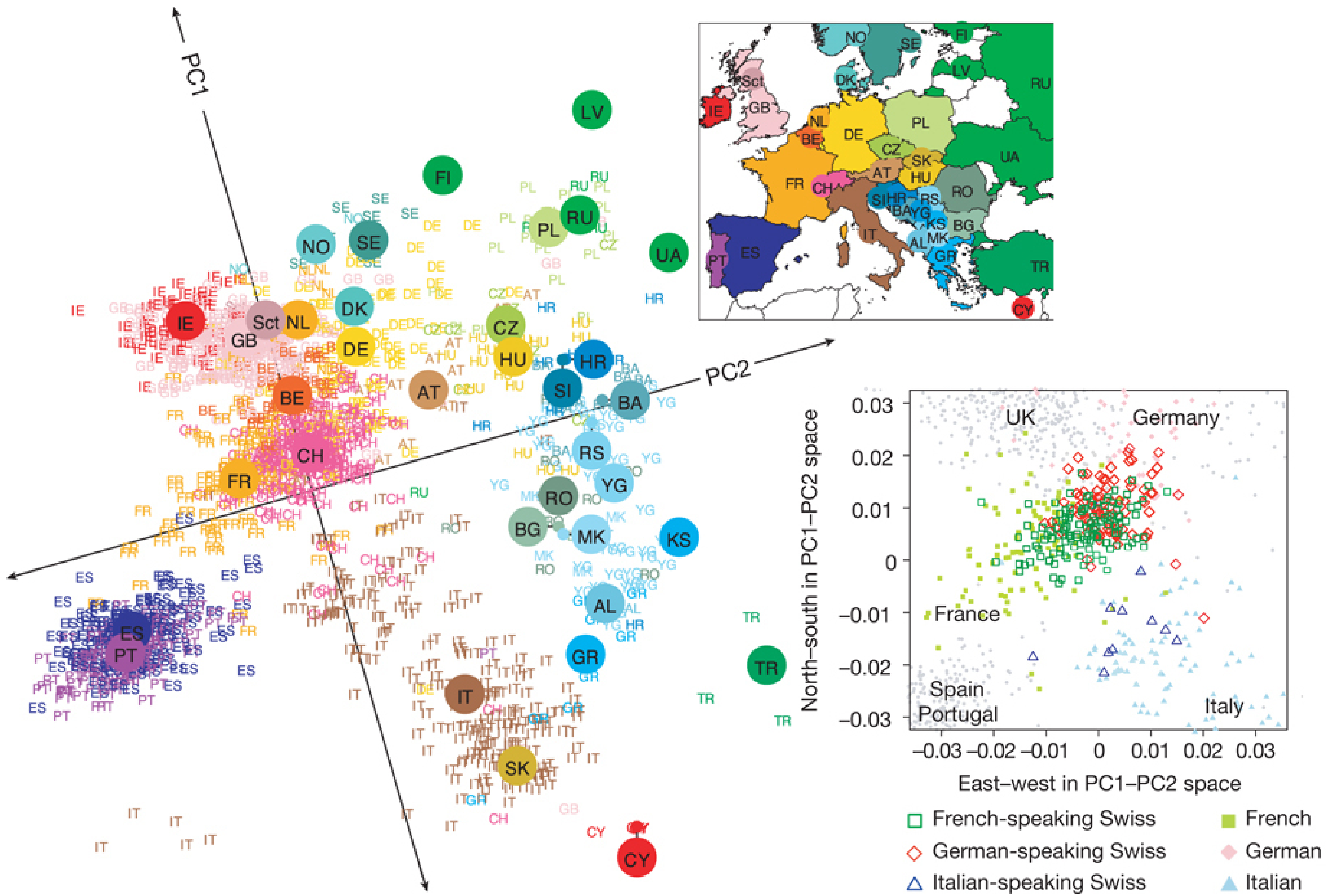
from http://vismod.media.mit.edu/vismod/demos/facerec/basic.html

Example of PCA: Genetic variation within Europe



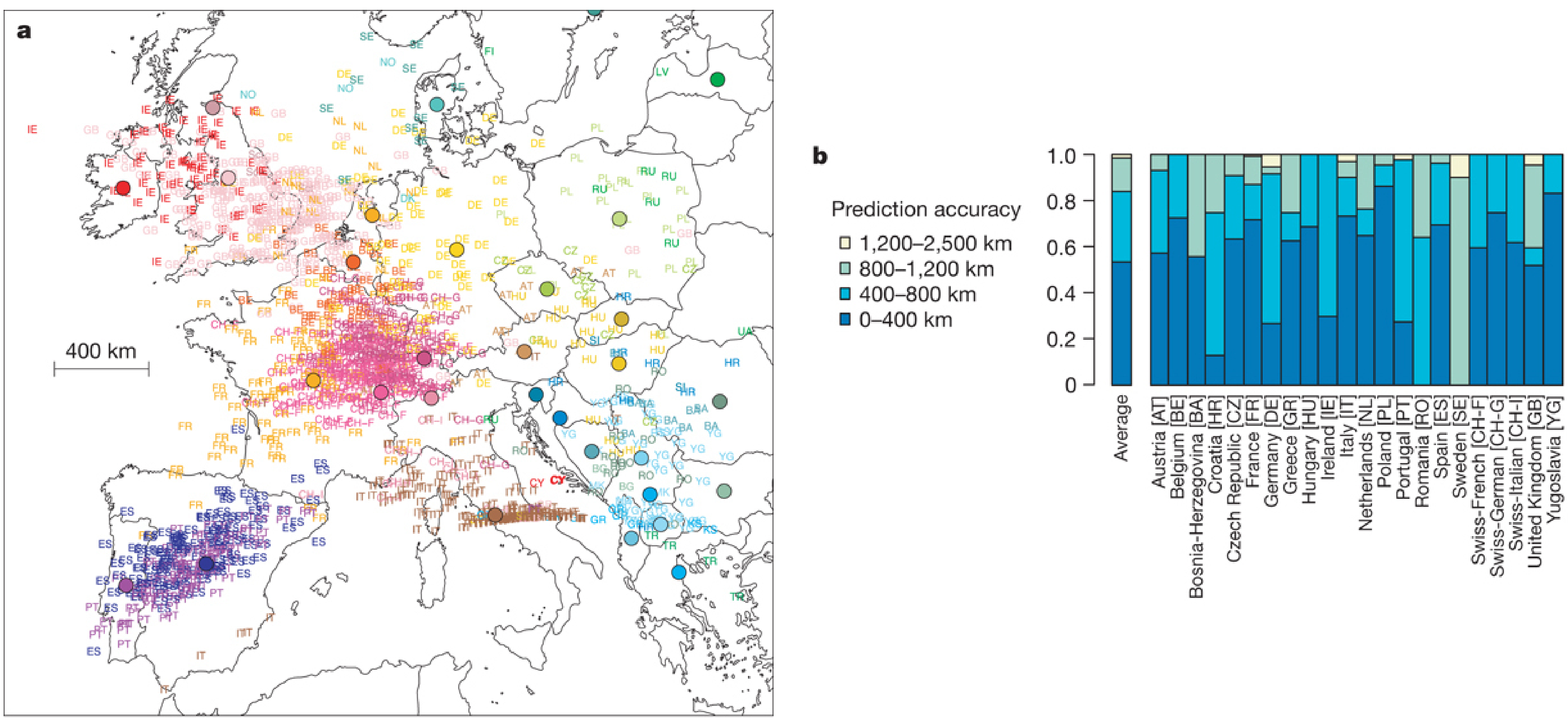
Novembre et al. (2008) Nature 456:98-101

Example of PCA: Genetic variation within Europe



Novembre et al. (2008) Nature 456:98-101

Example of PCA: Genetic variation within Europe



Novembre et al. (2008) Nature 456:98-101

Another view of PCA: Mutual Information

Problem: Given $\mathbf{x}$, find $\mathbf{y} = A\mathbf{x}$ with columns of A unit vectors, s.t. $I(\mathbf{y};\mathbf{x})$ is maximised (assuming that $P(\mathbf{x})$ is Gaussian).

$$I(\mathbf{y};\mathbf{x}) = H(\mathbf{y}) + H(\mathbf{x}) - H(\mathbf{y},\mathbf{x}) = H(\mathbf{y})$$

So we want to maximise the entropy of $\mathbf{y}$. What is the entropy of a Gaussian?

$$H(\mathbf{z}) \;=\; -\int d\mathbf{z}\; p(\mathbf{z}) \ln p(\mathbf{z}) = \frac{1}{2}\ln|\Sigma| + \frac{D}{2}(1 + \ln 2\pi)$$

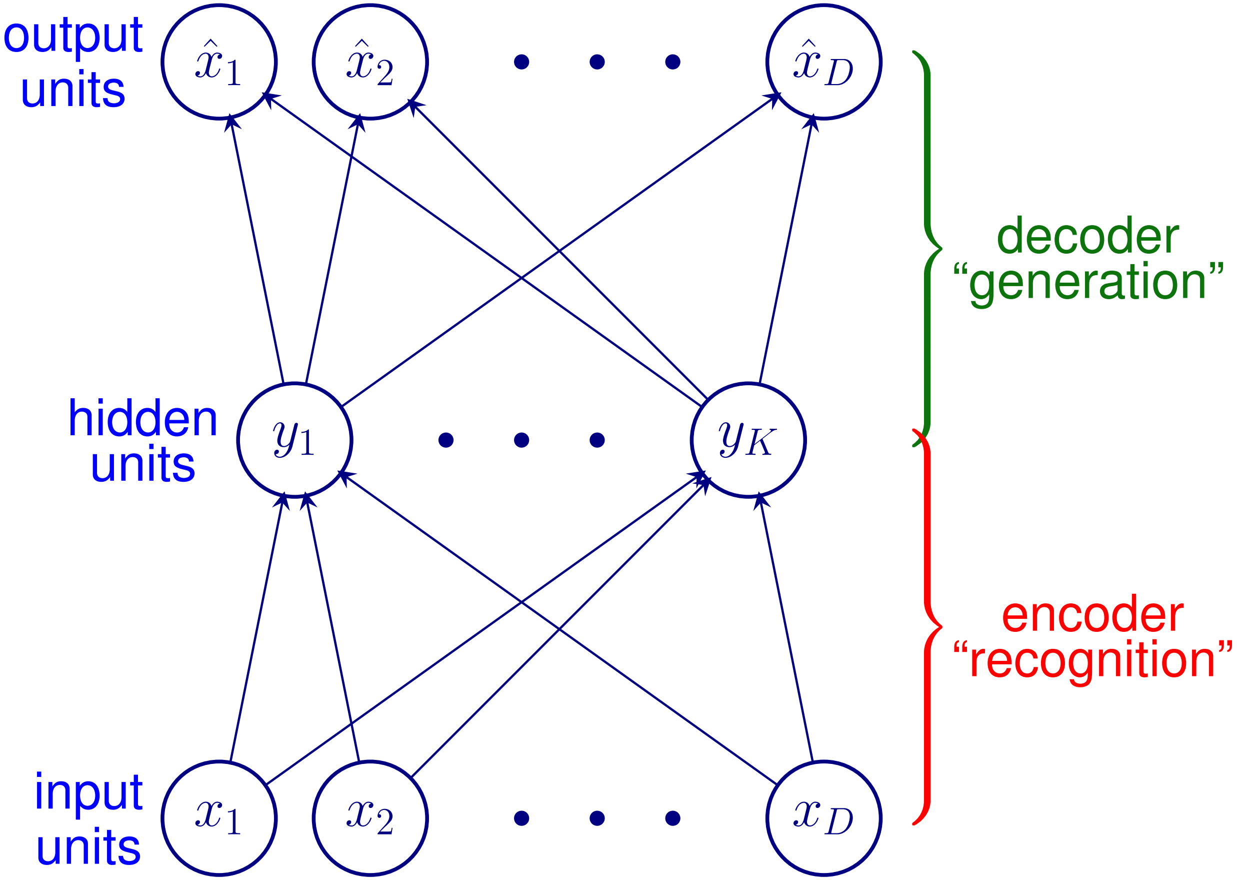
Therefore we want the distribution of $\mathbf{y}$ to have largest volume (i.e. det of covariance matrix).

$$\Sigma_y = A\Sigma_x A^\top = AUS_xU^\top A^\top$$

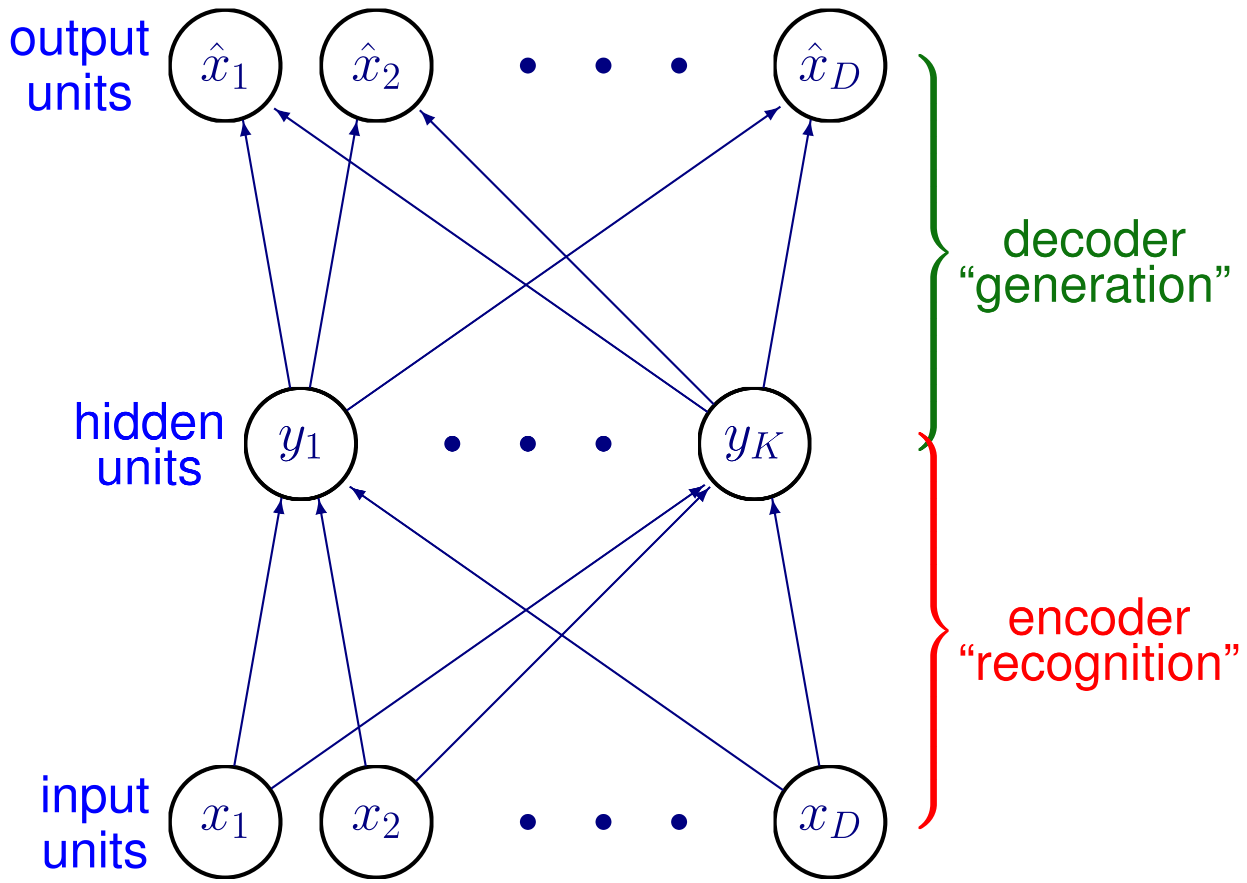
So, A should be aligned with the columns of U which are associated with the largest eigenvalues (variances).

Projection to the principal component subspace preserves the most information about the (Gaussian) data.

Network Interpretations and Encoder-Decoder Duality



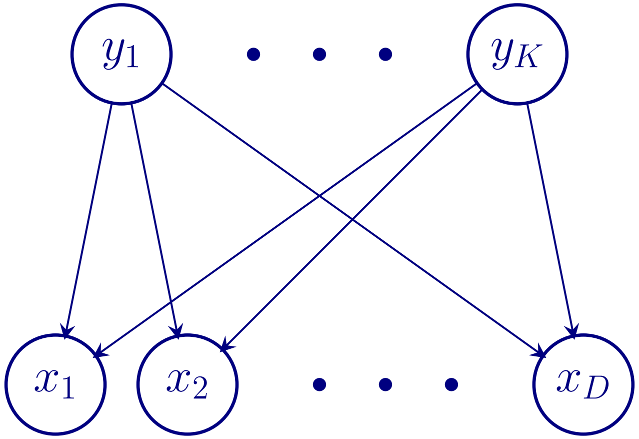
From Supervised Learning to PCA



A linear autoencoder neural network trained to minimise squared error learns to perform PCA (Baldi & Hornik, 1989).

Probabilistic PCA

We could think of the projection $\tilde{\mathbf{x}}_n$ as a latent variable, but it is more usual to construct a standard normal latent $\mathbf{y}$.



Observed vector data $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}; \mathbf{x}_i \in \mathbb{R}^D$
Assumed latent variables $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}; \mathbf{y}_i \in \mathbb{R}^K$

Linear generative model: $x_d = \sum_{k=1}^{K} \Lambda_{dk}\, y_k + \epsilon_d$

- y_k are independent $\mathcal{N}(0,1)$ Gaussian factors
- ϵ_d are independent $\mathcal{N}(0,\psi)$ Gaussian noise
- $K < D$

So, model for observations $\mathbf{x}$ is still Gaussian with:

$$p(\mathbf{y}) = \mathcal{N}(0, I)$$

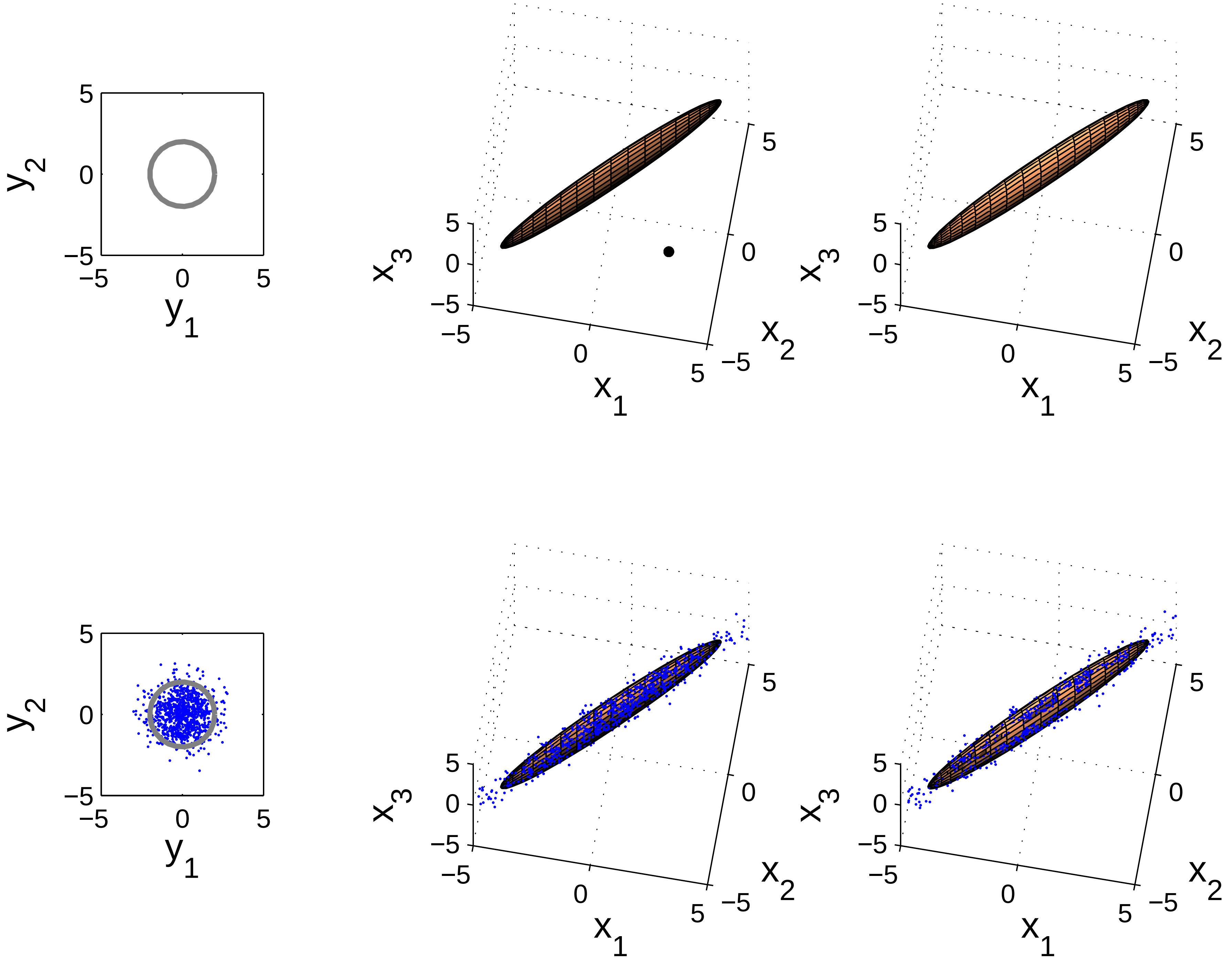
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\Lambda\mathbf{y}, \psi I)$$

$$p(\mathbf{x}) = \int p(\mathbf{y})p(\mathbf{x}|\mathbf{y})d\mathbf{y} = \mathcal{N}\left(0, \Lambda\Lambda^\top + \psi I\right)$$

where Λ is a $D \times K$ matrix.

This is Probabilistic PCA (pPCA). The maximum-likelihood parameter Λ, for fixed ψ, lies in the K-principal subspace.

Probabilistic PCA



PCA and pPCA

- In PCA the “noise” is orthogonal to the subspace, and we can project $\mathbf{x}_n \to \tilde{\mathbf{x}}_n$ trivially.
- In pPCA, the noise is more sensible (equal in all directions). But what is the projection?

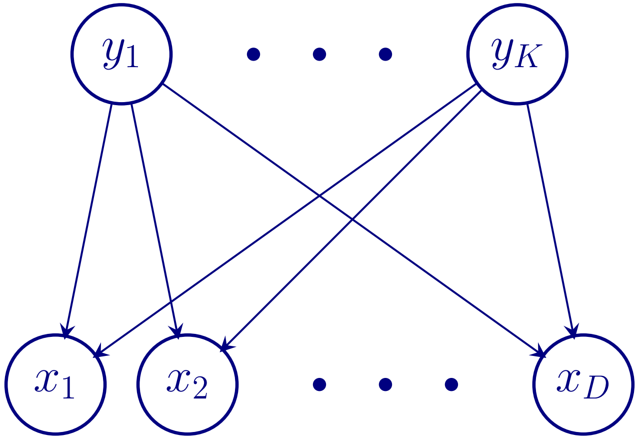
  Find the expected value of $\mathbf{y}_n|\mathbf{x}_n$ and then take $\tilde{\mathbf{x}}_n = \Lambda\bar{\mathbf{y}}_n$.
- **Tactic:** write $p(\mathbf{y}_n, \mathbf{x}_n|\theta)$, consider $\mathbf{x}_n$ to be fixed. What is this as a function of $\mathbf{y}_n$?

$$\begin{aligned}
p(\mathbf{y}_n, \mathbf{x}_n) &= p(\mathbf{y}_n)p(\mathbf{x}_n|\mathbf{y}_n) \\
&= (2\pi)^{-\frac{K}{2}} \exp\{-\frac{1}{2}{\mathbf{y}_n}^\top \mathbf{y}_n\} \; |2\pi\Psi|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}_n - \Lambda\mathbf{y}_n)^\top \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y}_n)\} \\
&= \text{c} \times \exp\{-\frac{1}{2}[{\mathbf{y}_n}^\top \mathbf{y}_n + (\mathbf{x}_n - \Lambda\mathbf{y}_n)^\top \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y}_n)]\} \\
&= \text{c'} \times \exp\{-\frac{1}{2}[{\mathbf{y}_n}^\top (I + \Lambda^\top\Psi^{-1}\Lambda)\mathbf{y}_n - 2{\mathbf{y}_n}^\top \Lambda^\top\Psi^{-1}\mathbf{x}_n]\} \\
&= \text{c''} \times \exp\{-\frac{1}{2}[{\mathbf{y}_n}^\top \Sigma^{-1}\mathbf{y}_n - 2{\mathbf{y}_n}^\top \Sigma^{-1}\mu + \mu^\top\Sigma^{-1}\mu]\}
\end{aligned}$$

  So $\Sigma = (I + \Lambda^\top\Psi^{-1}\Lambda)^{-1} = I - \beta\Lambda$ and $\mu = \Sigma\Lambda^\top\Psi^{-1}\mathbf{x}_n = \beta\mathbf{x}_n$. Where $\beta = \Sigma\Lambda^\top\Psi^{-1}$.
- Thus, $\tilde{\mathbf{x}}_n = \Lambda(I + \Lambda^\top\Psi^{-1}\Lambda)^{-1}\Lambda^\top\Psi^{-1}\mathbf{x}_n = \mathbf{x}_n - \Psi(\Lambda\Lambda^\mathsf{T} + \Psi)^{-1}\mathbf{x}_n$
- This is not the same projection. pPCA takes into account noise in the principal subspace.
- As $\psi \to 0$, pPCA → PCA.

Factor Analysis

If dimensions are not equivalent, equal variance makes no sense.



Observed vector data $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}; \mathbf{x}_i \in \mathbb{R}^D$

Assumed latent variables $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}; \mathbf{y}_i \in \mathbb{R}^K$

Linear generative model: $x_d = \sum_{k=1}^{K} \Lambda_{dk}\, y_k + \epsilon_d$

- y_k are independent $\mathcal{N}(0,1)$ Gaussian factors
- ϵ_d are independent $\mathcal{N}(0,\Psi_{dd})$ Gaussian noise
- $K < D$

So, model for observations $\mathbf{x}$ is still Gaussian with:

$$p(\mathbf{y}) = \mathcal{N}\left(0, I\right)$$

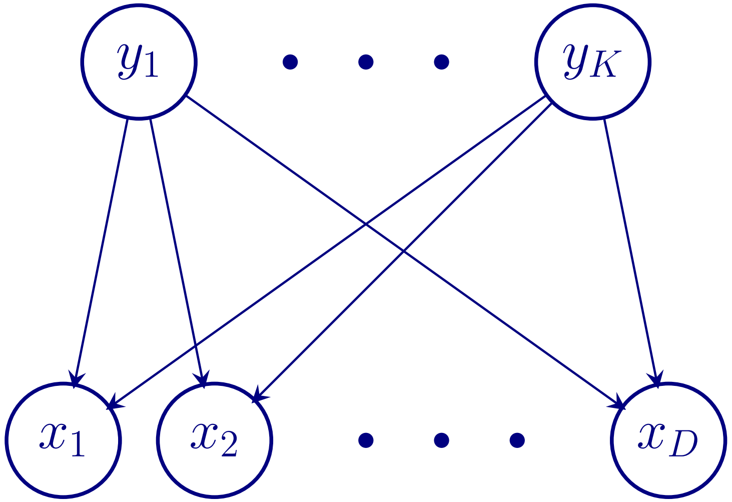
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\Lambda\mathbf{y}, \Psi\right)$$

$$p(\mathbf{x}) = \int p(\mathbf{y})p(\mathbf{x}|\mathbf{y})d\mathbf{y} = \mathcal{N}\left(0, \Lambda\Lambda^\top + \Psi\right)$$

where Λ is a $D \times K$ matrix, and Ψ is $K \times K$ and diagonal.

Dimensionality Reduction: Finds a low-dimensional projection of high dimensional data that captures the correlation structure of the data.

Factor Analysis (cont.)



- ML learning finds Λ and Ψ given data
- parameters (corrected for symmetries): $DK + D - \frac{K(K-1)}{2}$
- If number of parameters $> \frac{D(D+1)}{2}$ model is not identifiable
- no closed form solution for ML params: $\mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$

- [Bayesian treatment would also have priors over Λ and Ψ and would average over them for prediction.]

Factor Analysis Projections

Our analysis for pPCA still applies:

$$\tilde{\mathbf{x}}_n = \Lambda(I + \Lambda^\top \Psi^{-1} \Lambda)^{-1} \Lambda^\top \Psi^{-1} \mathbf{x}_n = \mathbf{x}_n - \Psi(\Lambda\Lambda^\mathsf{T} + \Psi)^{-1} \mathbf{x}_n$$

but now Ψ is diagonal but not spherical.
Note, though, that Λ is generally different from that found by pPCA.

Gradient Methods of Learning FA

Write down negative log likelihood:

$$\frac{1}{2}\log|2\pi(\Lambda\Lambda^\top + \Psi)| + \frac{1}{2}\mathbf{x}^\top(\Lambda\Lambda^\top + \Psi)^{-1}\mathbf{x}$$

Optimise w.r.t. Λ and Ψ (need matrix calculus) subject to constraints

We will soon see an easier way to learn latent variable models...

FA vs PCA

- PCA and pPCA are rotationally invariant; FA is not

$$\text{If } \mathbf{x} \rightarrow U\mathbf{x} \text{ for unitary } U, \quad \text{then } \boldsymbol{\lambda}_{(i)}^{\text{PCA}} \rightarrow U\boldsymbol{\lambda}_{(i)}^{\text{PCA}}$$

- FA is measurement scale invariant; PCA and pPCA are not

$$\text{If } \mathbf{x} \rightarrow S\mathbf{x} \text{ for diagonal } S, \quad \text{then } \boldsymbol{\lambda}_{(i)}^{\text{FA}} \rightarrow S\boldsymbol{\lambda}_{(i)}^{\text{FA}}$$

- FA and pPCA define a probabilistic model; PCA does not

The Expectation Maximisation (EM) algorithm

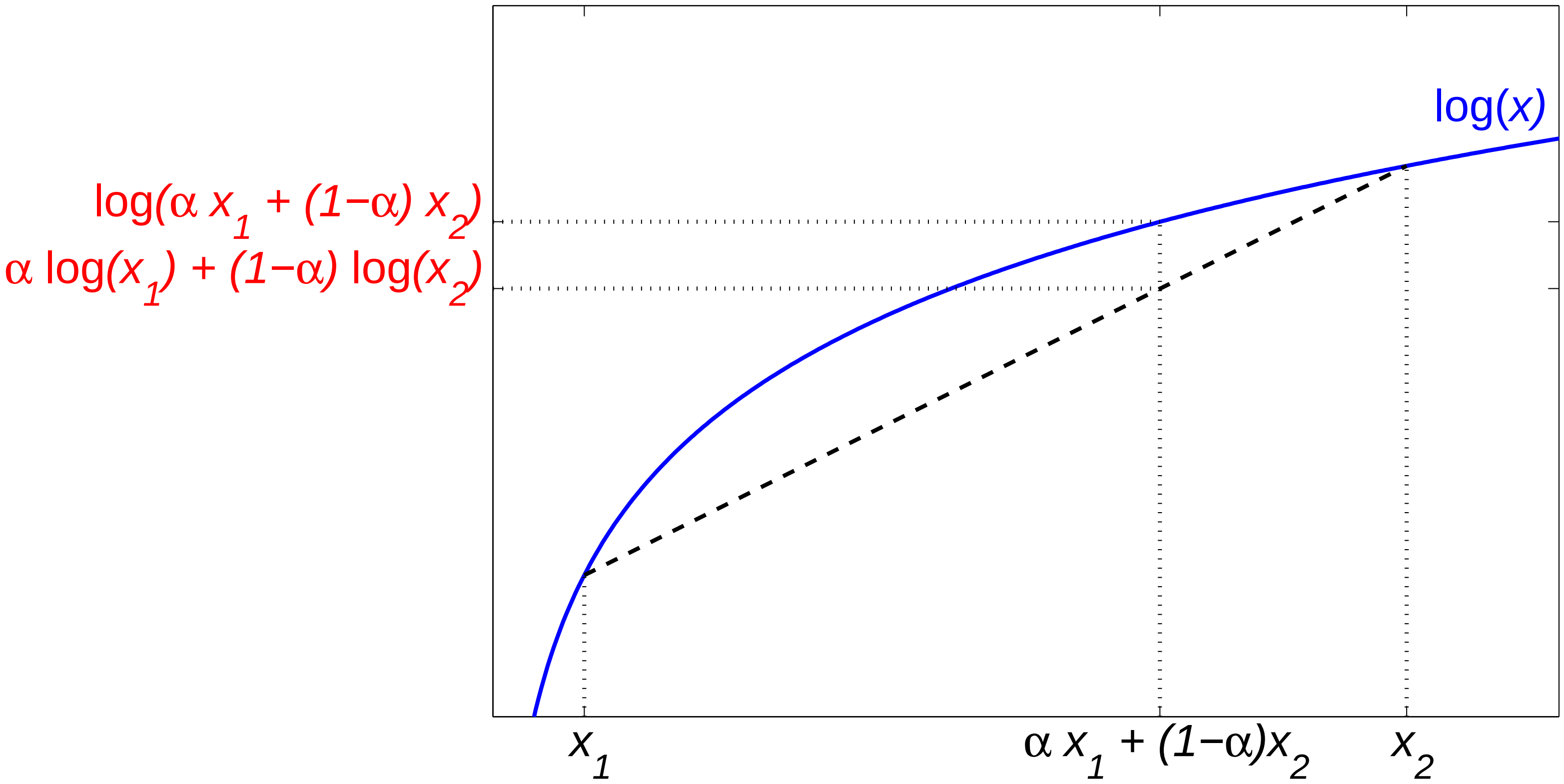
The EM algorithm finds a (local) maximum of a latent variable model likelihood. It starts from arbitrary values of the parameters, and iterates two steps:

E step: Fill in values of latent variables according to posterior given data.

M step: Maximise likelihood as if latent variables were not hidden.

- Useful in models where learning would be easy if hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No learning rate.
- Framework lends itself to principled approximations.

Jensen's Inequality



For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $\{x_i > 0\}$

$$\log\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i \log(x_i)$$

Equality if and only if $\alpha_i = 1$ for some i (and therefore all others are 0).

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \qquad .$$

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned}\int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) \, d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) \, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) \, d\mathcal{Y} + \mathbf{H}[q],\end{aligned}$$

where $\mathbf{H}[q]$ is the entropy of $q(\mathcal{Y})$.

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned}\int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) \, d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) \, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) \, d\mathcal{Y} + \mathbf{H}[q],\end{aligned}$$

where $\mathbf{H}[q]$ is the entropy of $q(\mathcal{Y})$.
So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + \mathbf{H}[q]$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{q(\mathcal{Y})} + \mathbf{H}[q],$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{q(\mathcal{Y})} + \mathbf{H}[q],$$

EM alternates between:

E step: optimize $\mathcal{F}(q,\theta)$ wrt distribution over hidden variables holding parameters fixed:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})}{\operatorname{argmax}} \; \mathcal{F}\big(q(\mathcal{Y}), \theta^{(k-1)}\big).$$

The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{q(\mathcal{Y})} + \mathbf{H}[q],$$

EM alternates between:

E step: optimize $\mathcal{F}(q,\theta)$ wrt distribution over hidden variables holding parameters fixed:

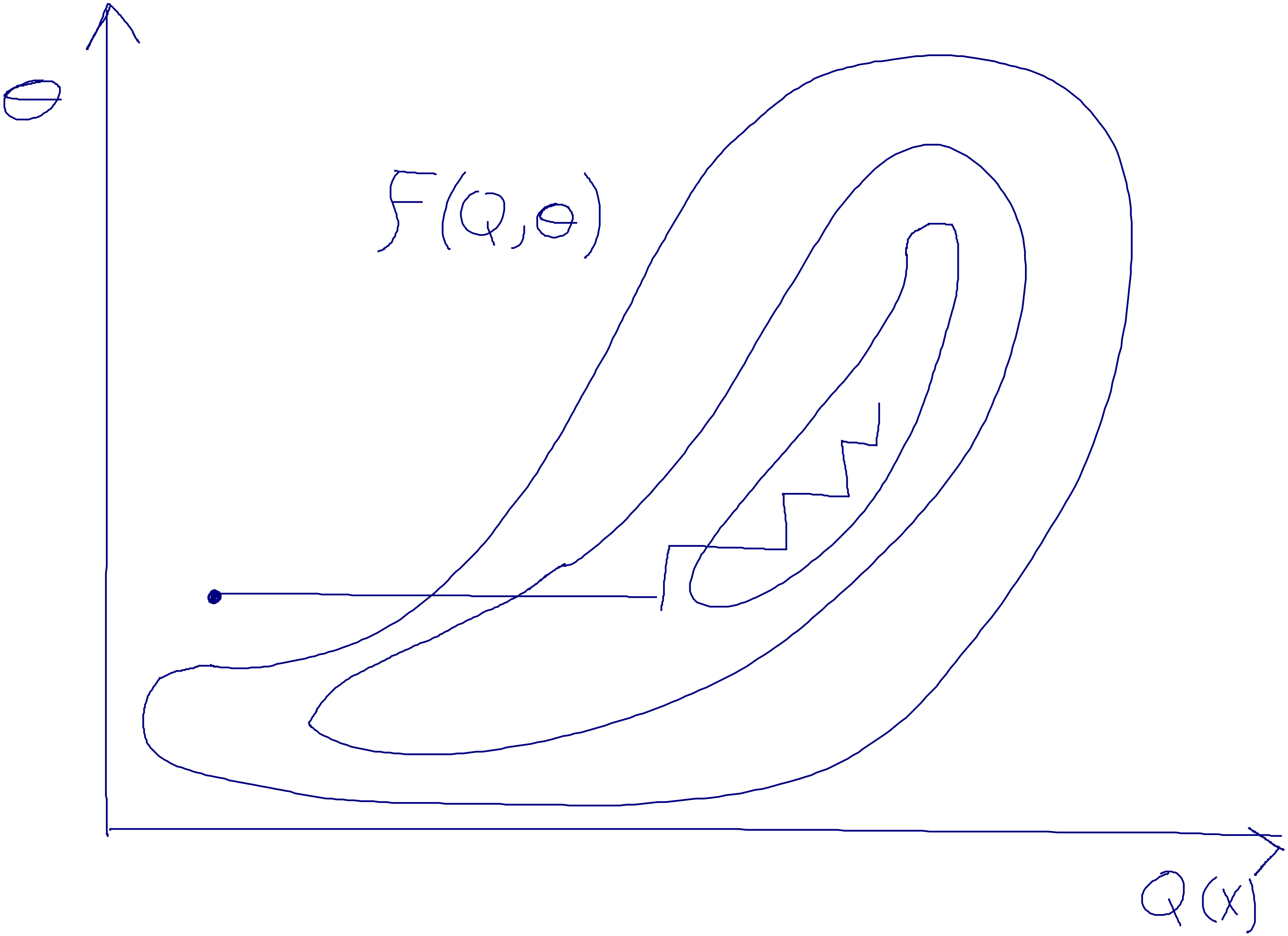
$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})}{\operatorname{argmax}} \; \mathcal{F}\big(q(\mathcal{Y}), \theta^{(k-1)}\big).$$

M step: maximize $\mathcal{F}(q,\theta)$ wrt parameters holding hidden distribution fixed:

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}\big(q^{(k)}(\mathcal{Y}), \theta\big) = \underset{\theta}{\operatorname{argmax}} \; \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{q^{(k)}(\mathcal{Y})}$$

The second equality comes from the fact that the entropy of $q(\mathcal{Y})$ does not depend directly on θ.

EM as Coordinate Ascent in $\mathcal{F}$



The E Step

The free energy can be re-written

$$\mathcal{F}(q,\theta) = \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y}$$

The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)P(\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y}\end{aligned}$$

The E Step

The free energy can be re-written

$$\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)P(\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log P(\mathcal{X}|\theta)\, d\mathcal{Y} + \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)}{q(\mathcal{Y})}\, d\mathcal{Y}
\end{aligned}$$

The E Step

The free energy can be re-written

$$\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)P(\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log P(\mathcal{X}|\theta)\, d\mathcal{Y} + \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \ell(\theta) - \textbf{KL}[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)]
\end{aligned}$$

The second term is the Kullback-Leibler divergence.

The E Step

The free energy can be re-written

$$
\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)P(\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log P(\mathcal{X}|\theta)\, d\mathcal{Y} + \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \ell(\theta) - \textbf{KL}[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)]
\end{aligned}
$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed θ, $\mathcal{F}$ is bounded above by ℓ, and achieves that bound when **KL**$[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)] = 0$.

The E Step

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)P(\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{X}|\theta)\, d\mathcal{Y} + \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\ &= \ell(\theta) - \textbf{KL}[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)]\end{aligned}$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed θ, $\mathcal{F}$ is bounded above by ℓ, and achieves that bound when **KL**$[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)] = 0$.

But **KL**$[q\|p]$ is zero if and only if $q = p$. So, the E step simply sets

$$q^{(k)}(\mathcal{Y}) = P(\mathcal{Y}|\mathcal{X},\theta^{(k-1)})$$

and, after an E step, the free energy equals the likelihood.

The KL$[q(x)\|p(x)]$ is non-negative and zero iff $\forall x : \; p(x) = q(x)$

First let's consider discrete distributions; the Kullback-Liebler divergence is:

$$\textbf{KL}[q\|p] = \sum_i q_i \log \frac{q_i}{p_i}.$$

To find the distribution q which minimizes **KL**$[q\|p]$ we add a Lagrange multiplier to enforce the normalization constraint:

$$E \overset{\text{def}}{=} \textbf{KL}[q\|p] + \lambda\big(1 - \sum_i q_i\big) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda\big(1 - \sum_i q_i\big)$$

We then take partial derivatives and set to zero:

$$\left.\begin{aligned} \frac{\partial E}{\partial q_i} &= \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1) \\ \frac{\partial E}{\partial \lambda} &= 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1 \end{aligned}\right\} \Rightarrow q_i = p_i.$$

The **KL**$[q(x)\|p(x)]$ is non-negative and zero iff $\forall x: \ p(x) = q(x)$

Check that the curvature (Hessian) is positive (definite), corresponding to a minimum:

$$\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0, \qquad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0,$$

showing that $q_i = p_i$ is a genuine minimum.

At the minimum is it easily verified that **KL**$[p\|p] = 0$.

A similar proof holds for **KL**$[\cdot\|\cdot]$ between continuous densities, the derivatives being substituted by functional derivatives.

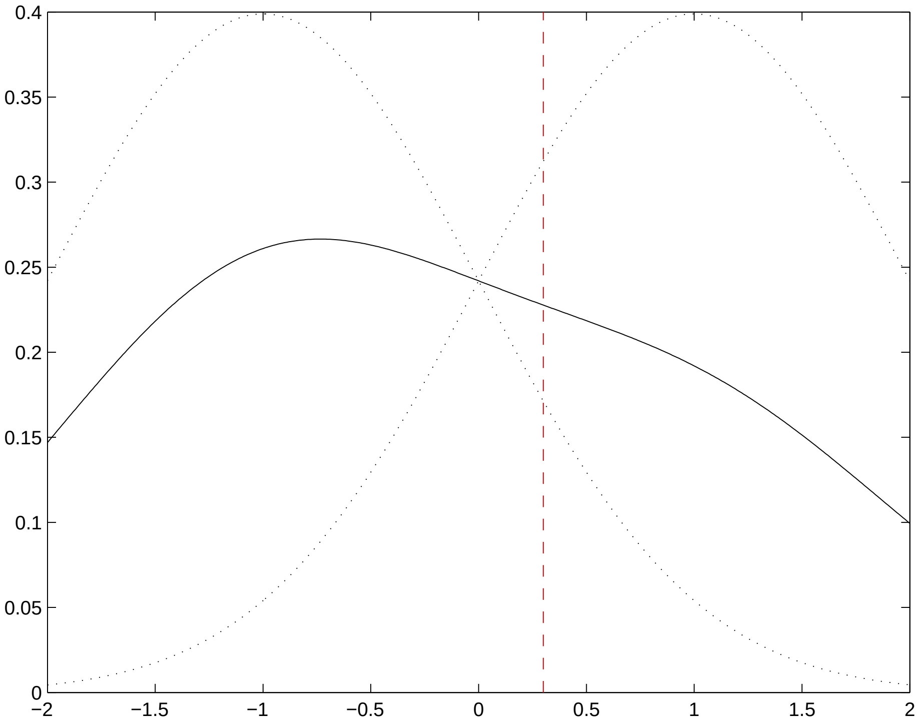
Coordinate Ascent in $\mathcal{F}$ (Demo)

One parameter mixture:

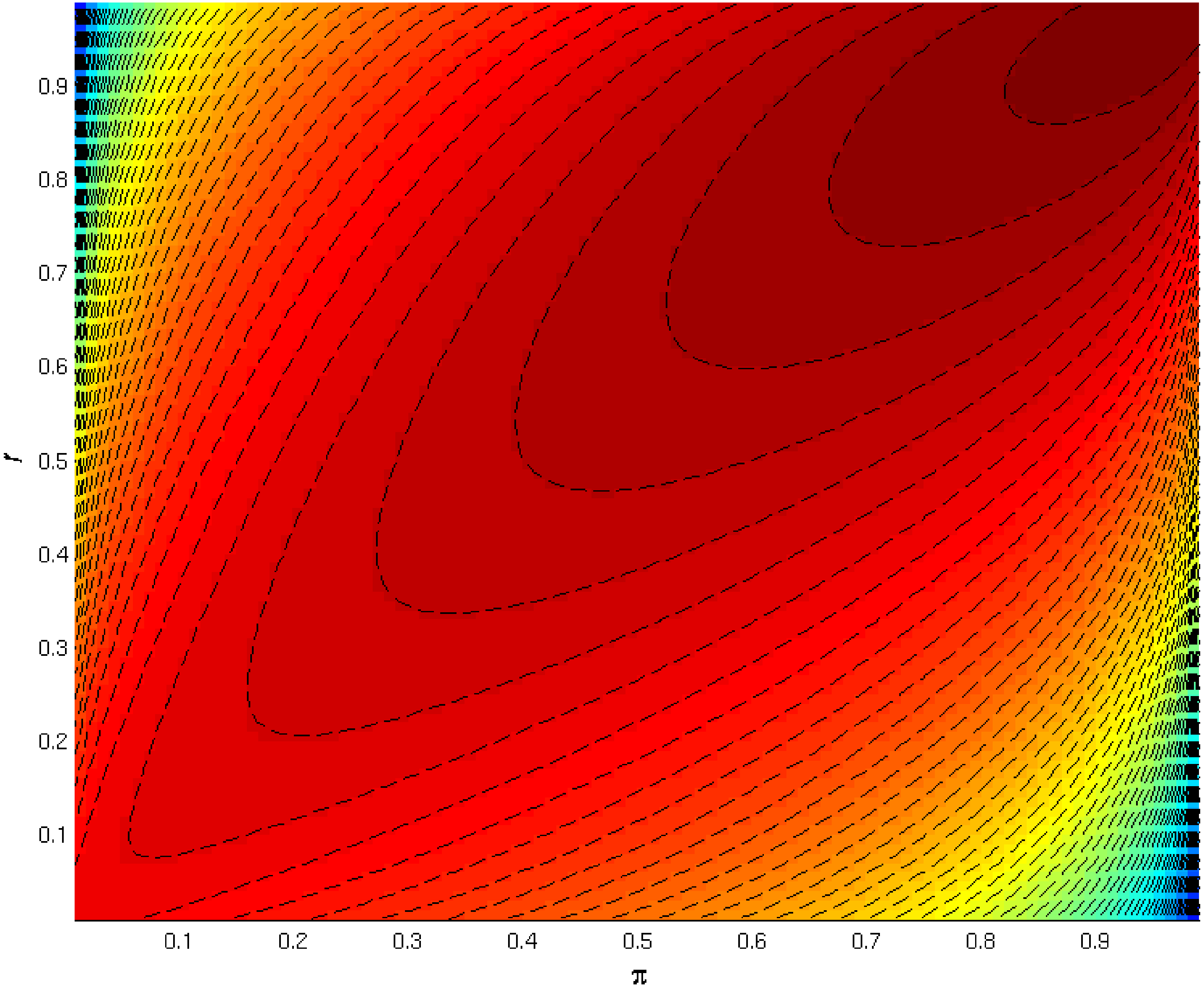
$$s \sim \text{Bernoulli}[\pi]$$

$$x|s = 0 \sim \mathcal{N}[-1, 1] \qquad x|s = 1 \sim \mathcal{N}[1, 1]$$

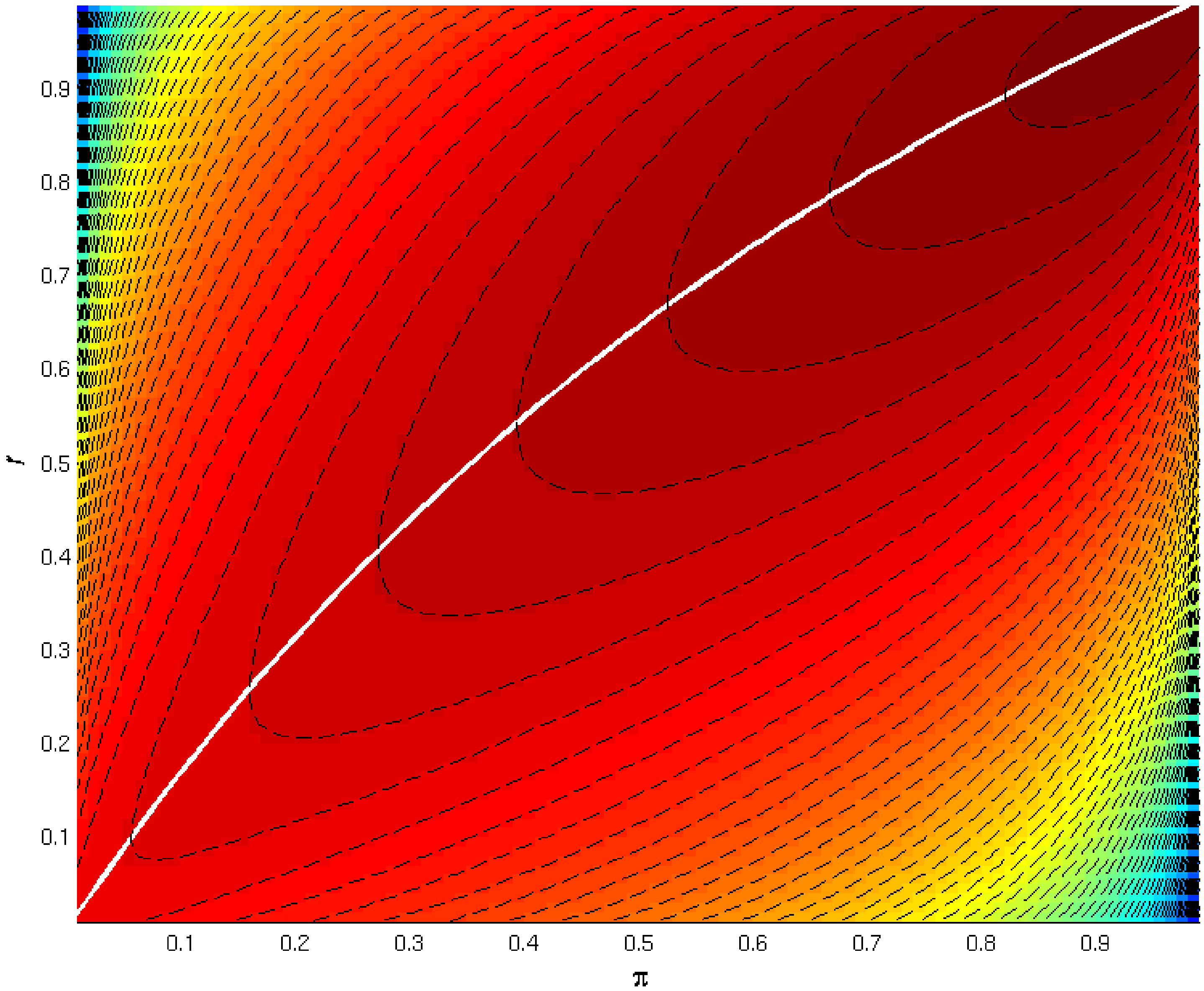
and one data point $x_1 = .3$.
$q(s)$ is a distribution on a single binary latent, and so is represented by $r_1 \in [0, 1]$.



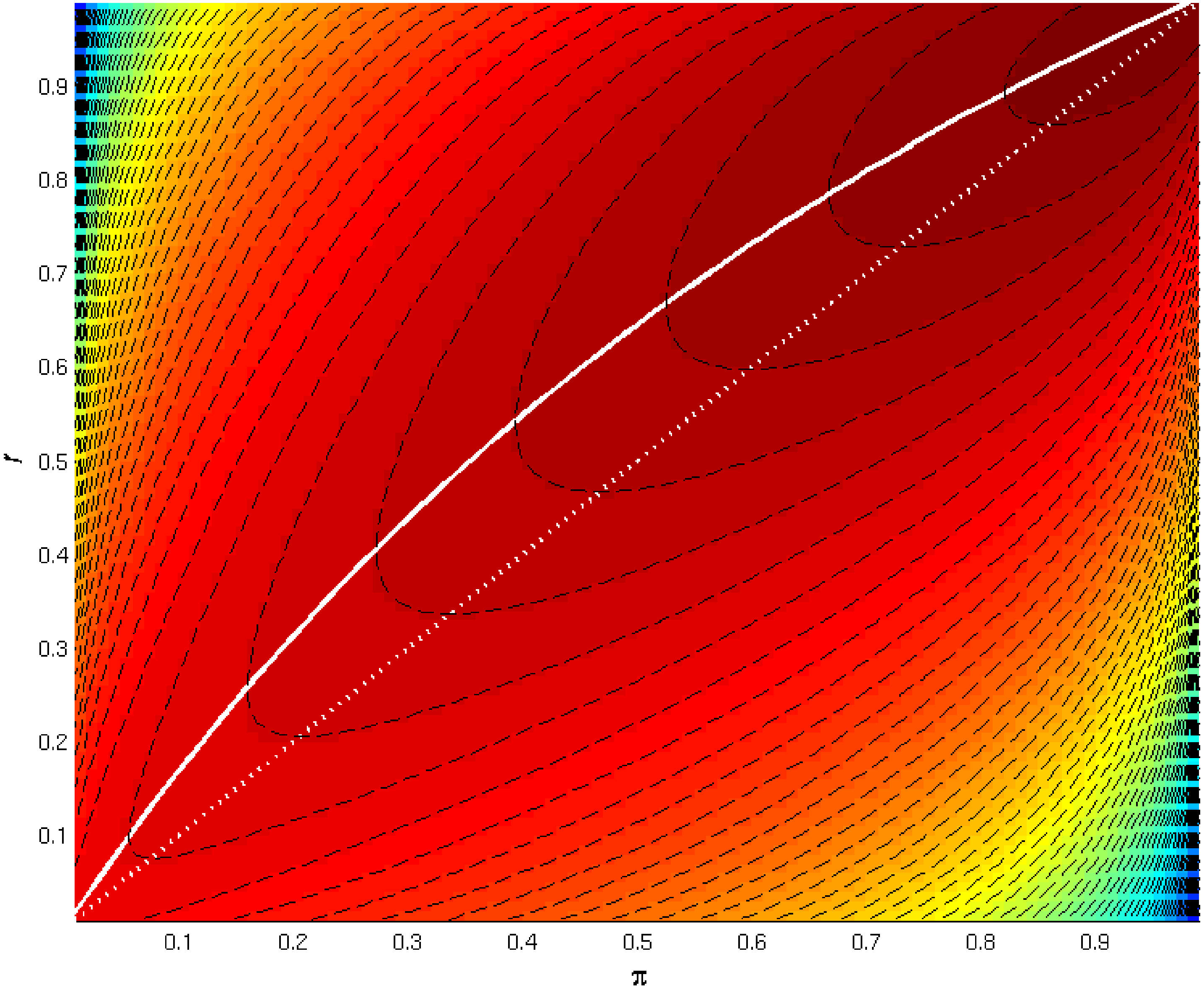
Coordinate Ascent in $\mathcal{F}$ (Demo)



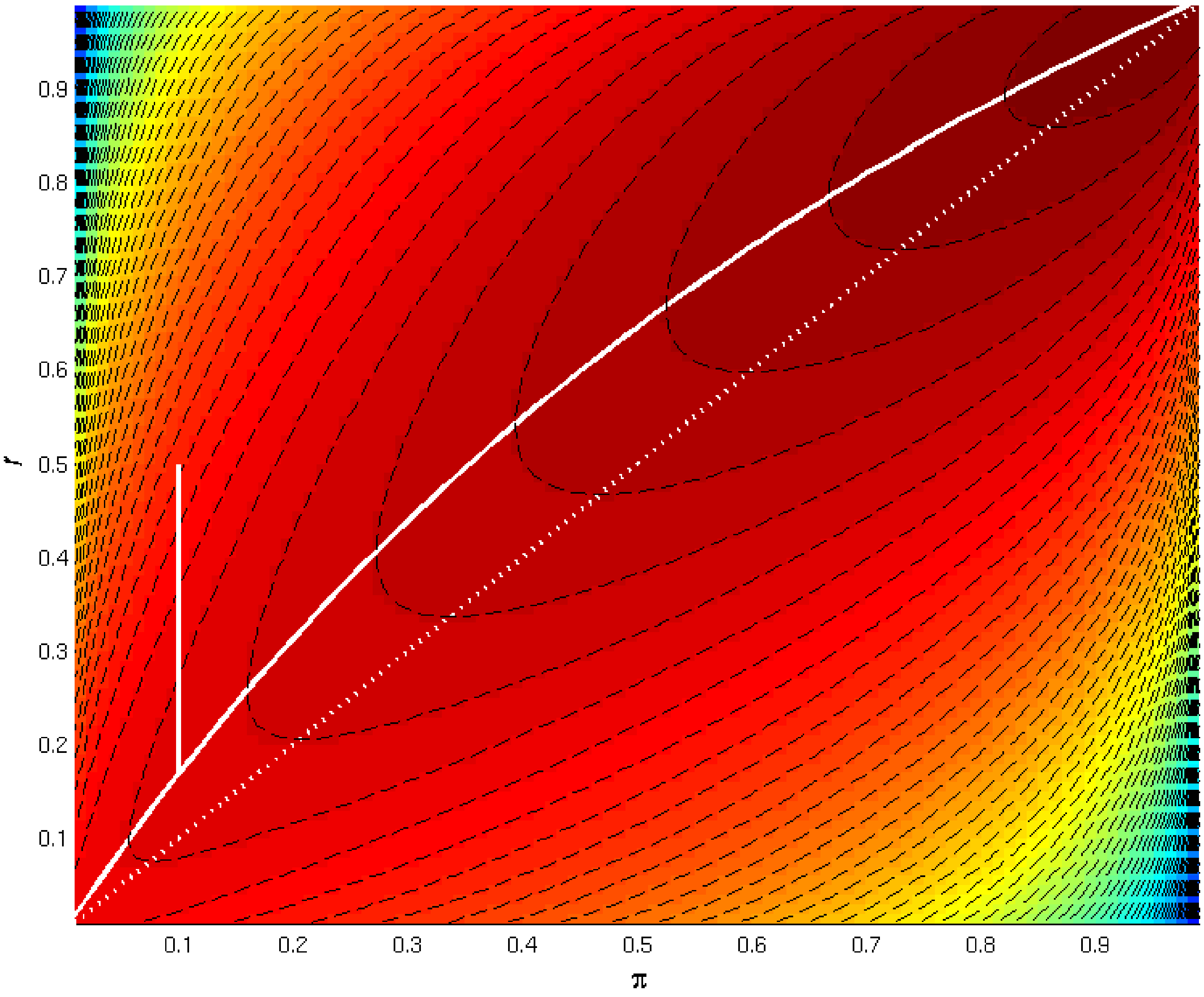
Coordinate Ascent in $\mathcal{F}$ (Demo)



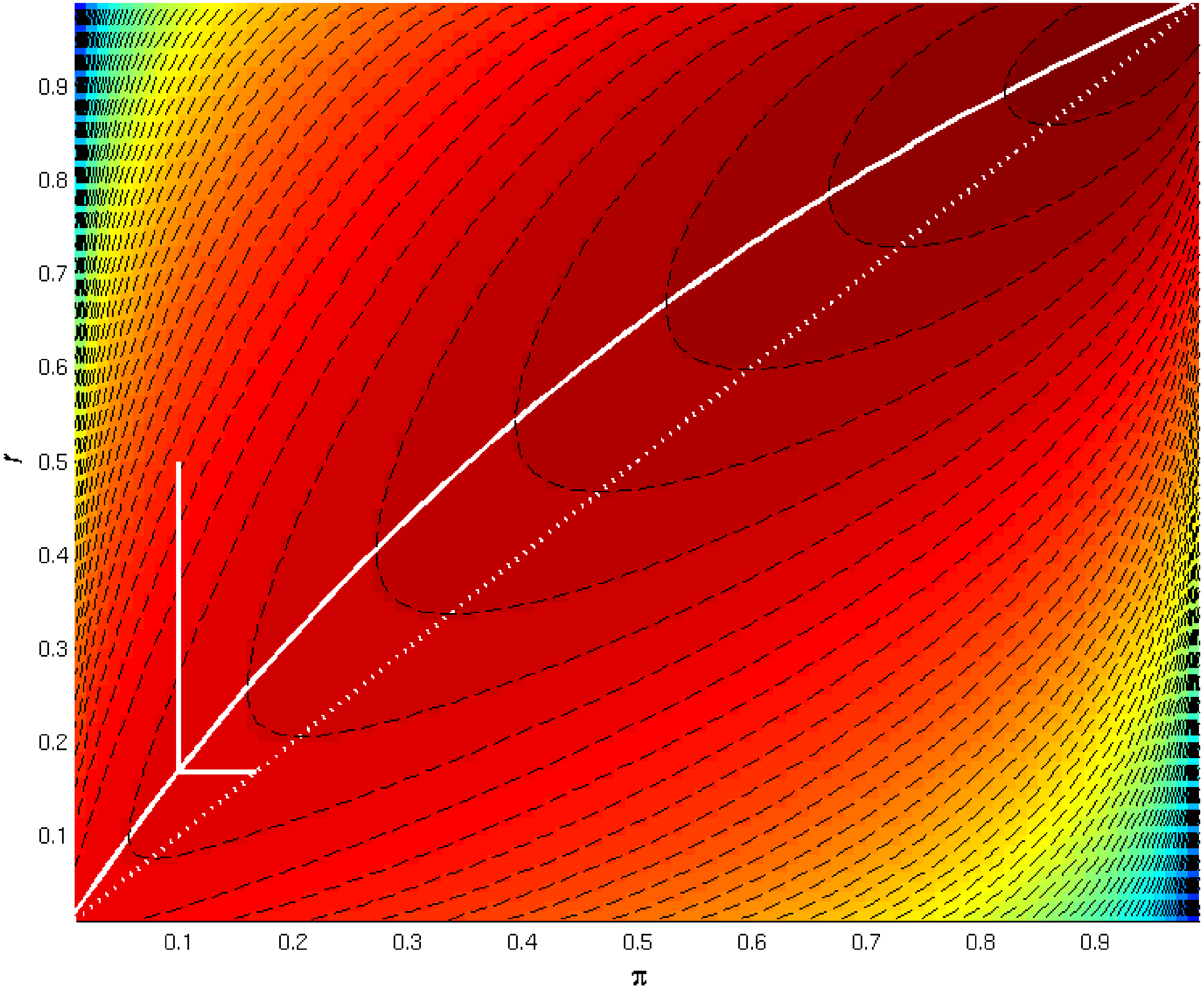
Coordinate Ascent in $\mathcal{F}$ (Demo)



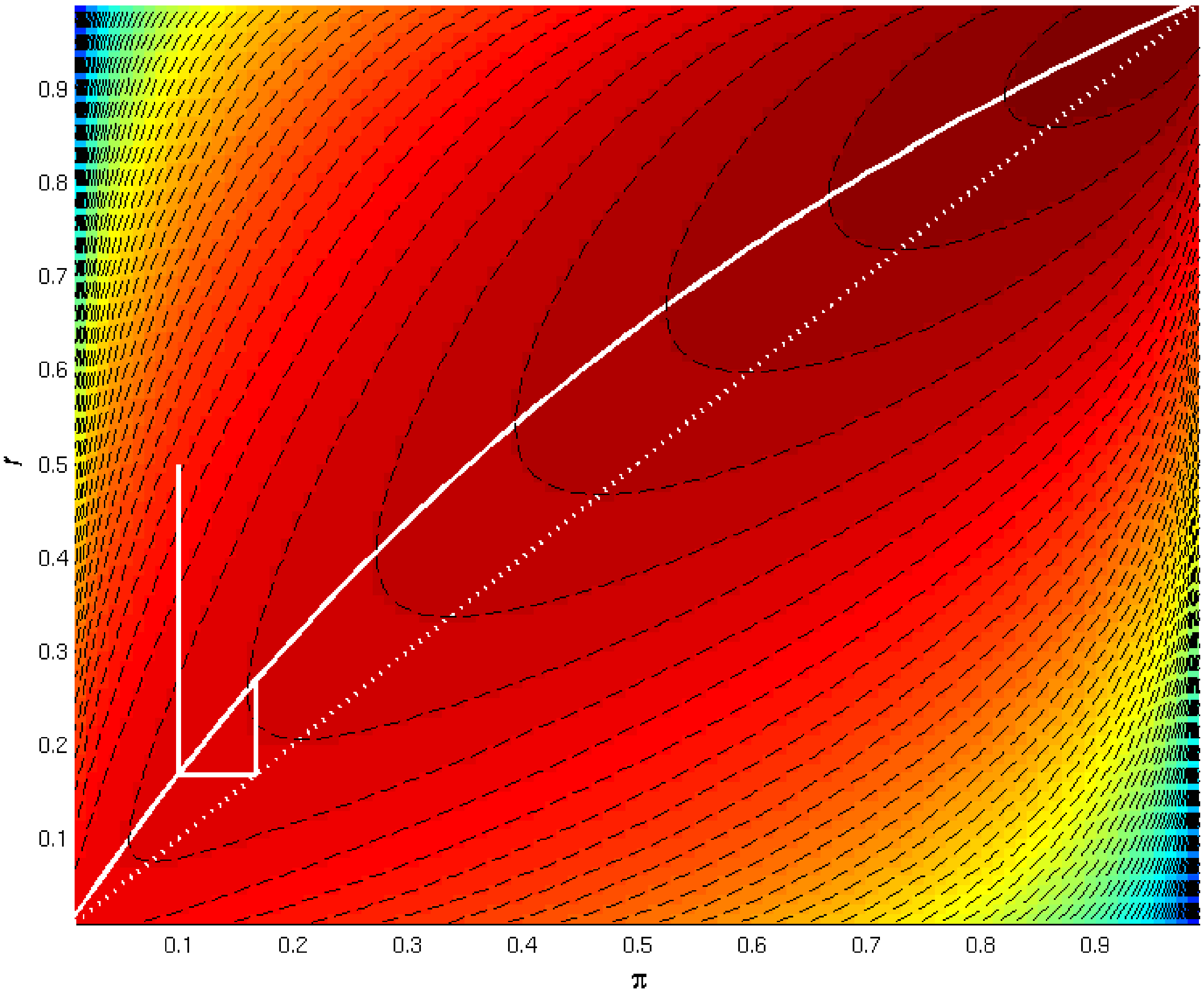
Coordinate Ascent in $\mathcal{F}$ (Demo)



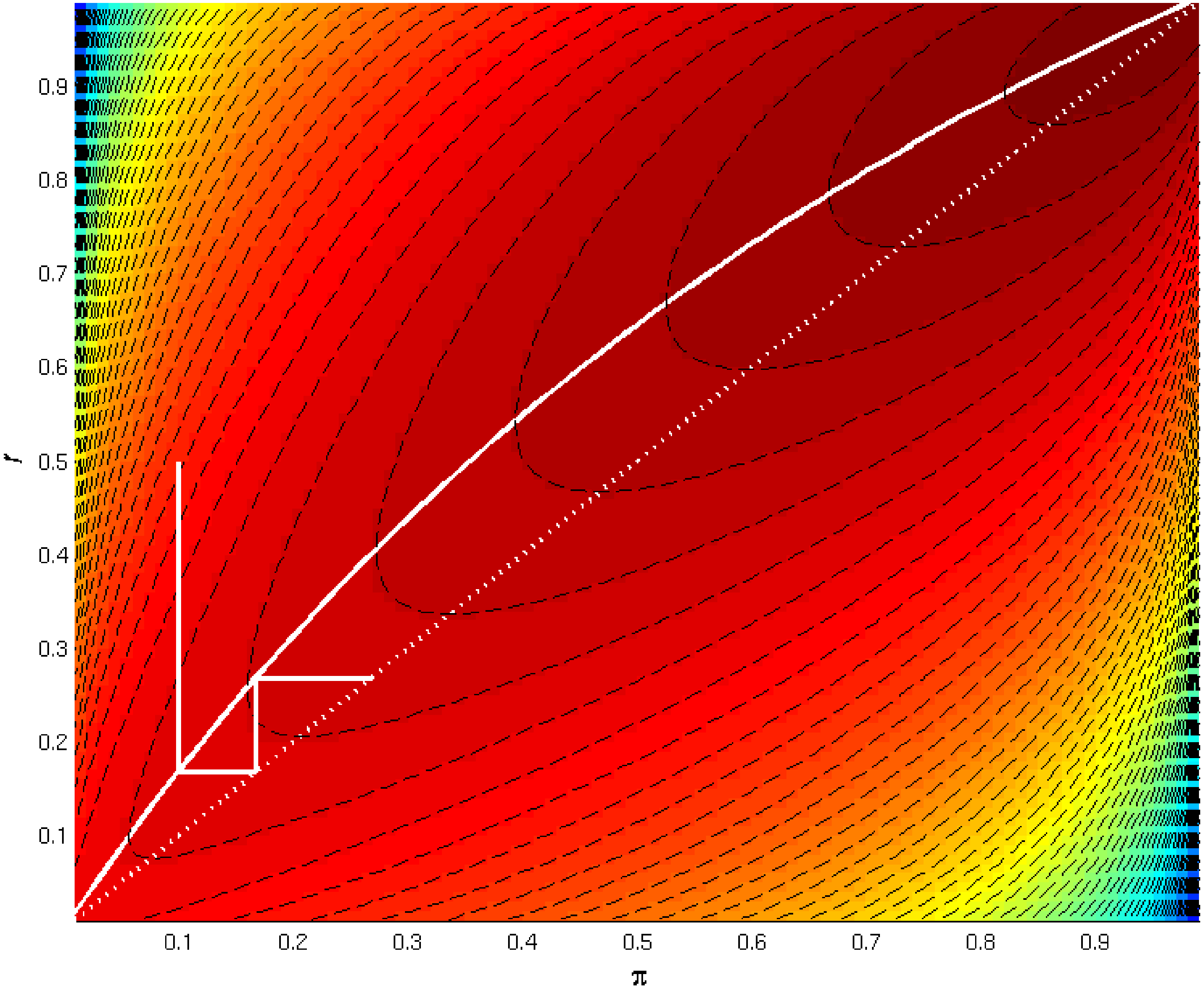
Coordinate Ascent in $\mathcal{F}$ (Demo)



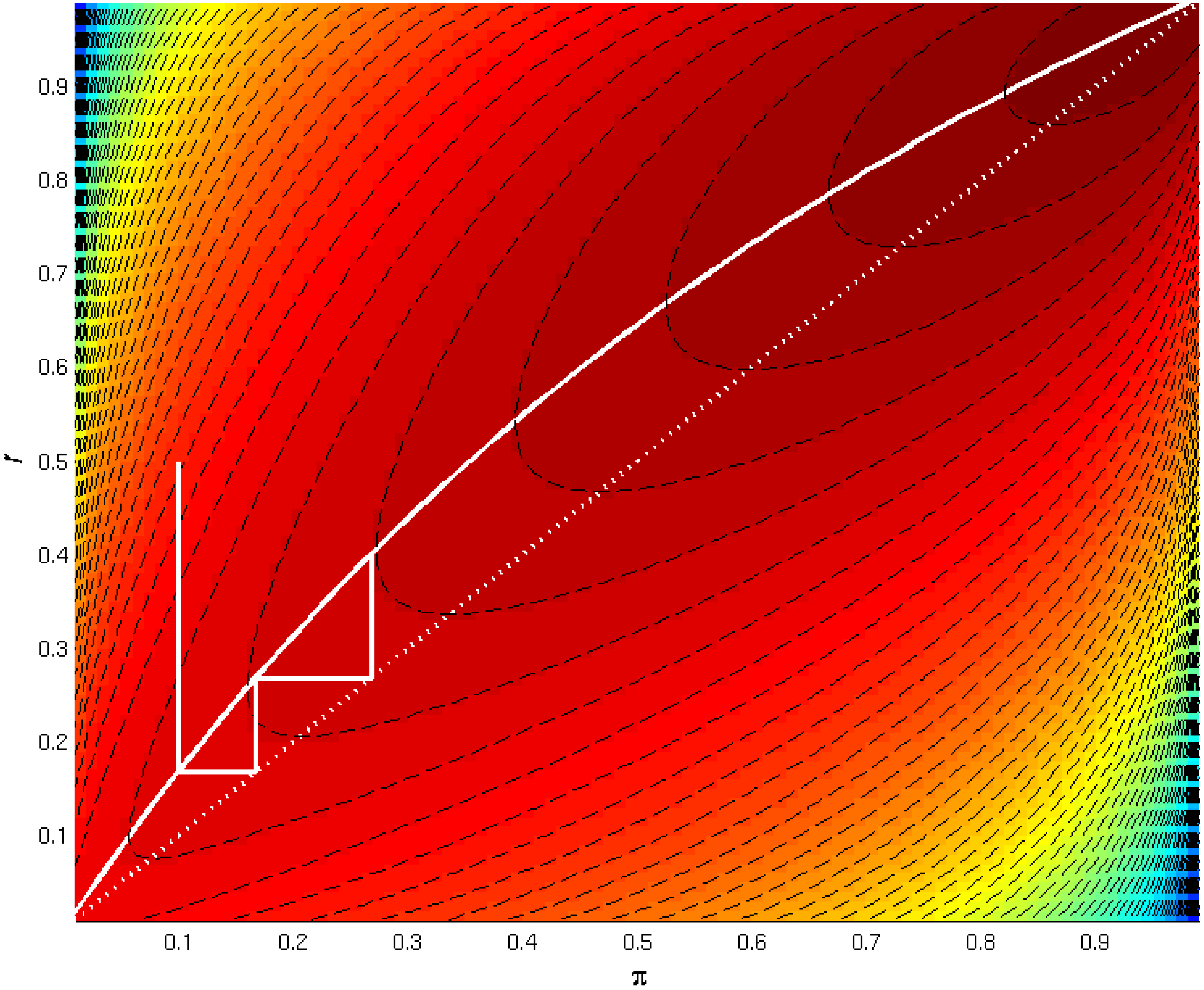
Coordinate Ascent in $\mathcal{F}$ (Demo)



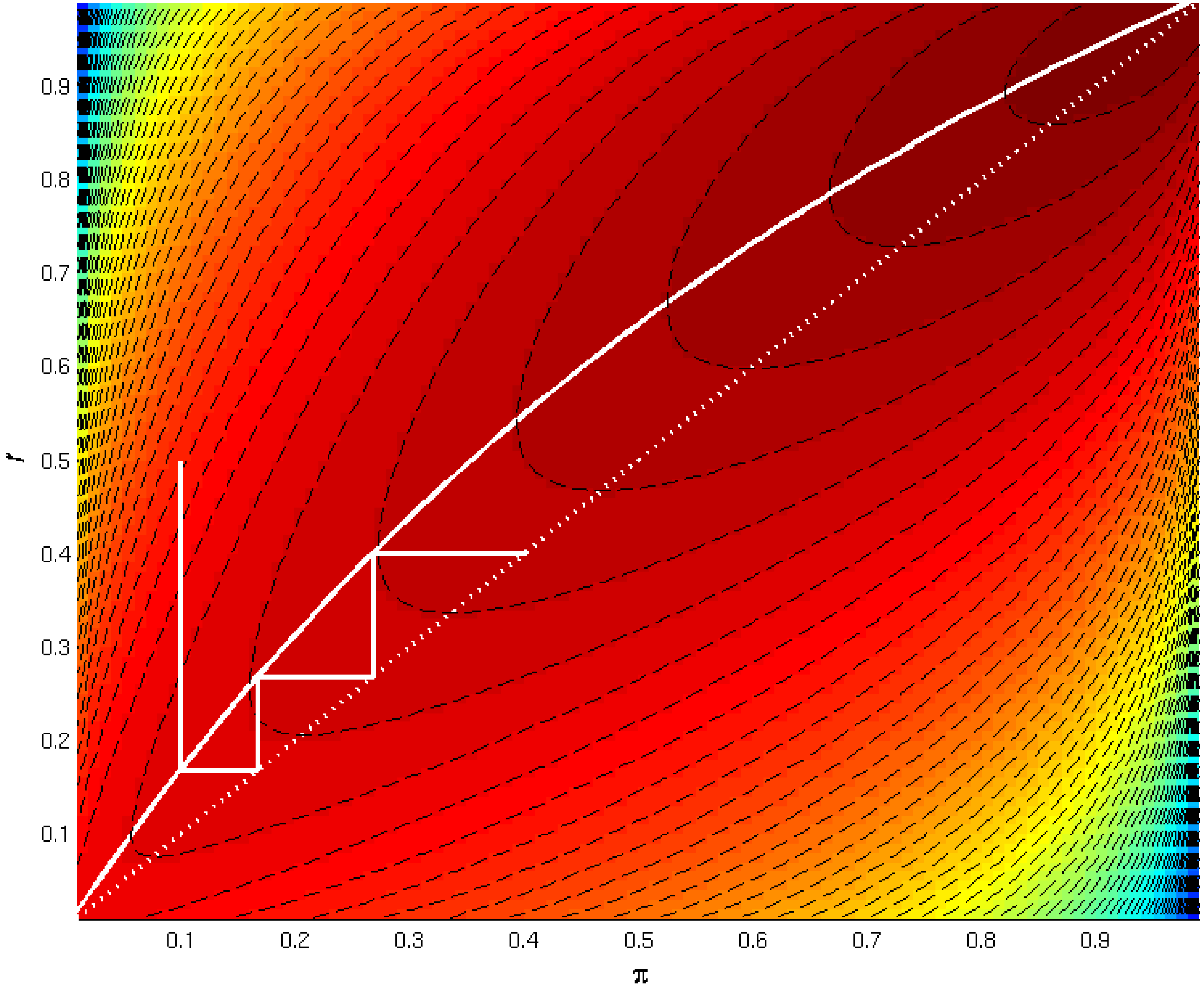
Coordinate Ascent in $\mathcal{F}$ (Demo)



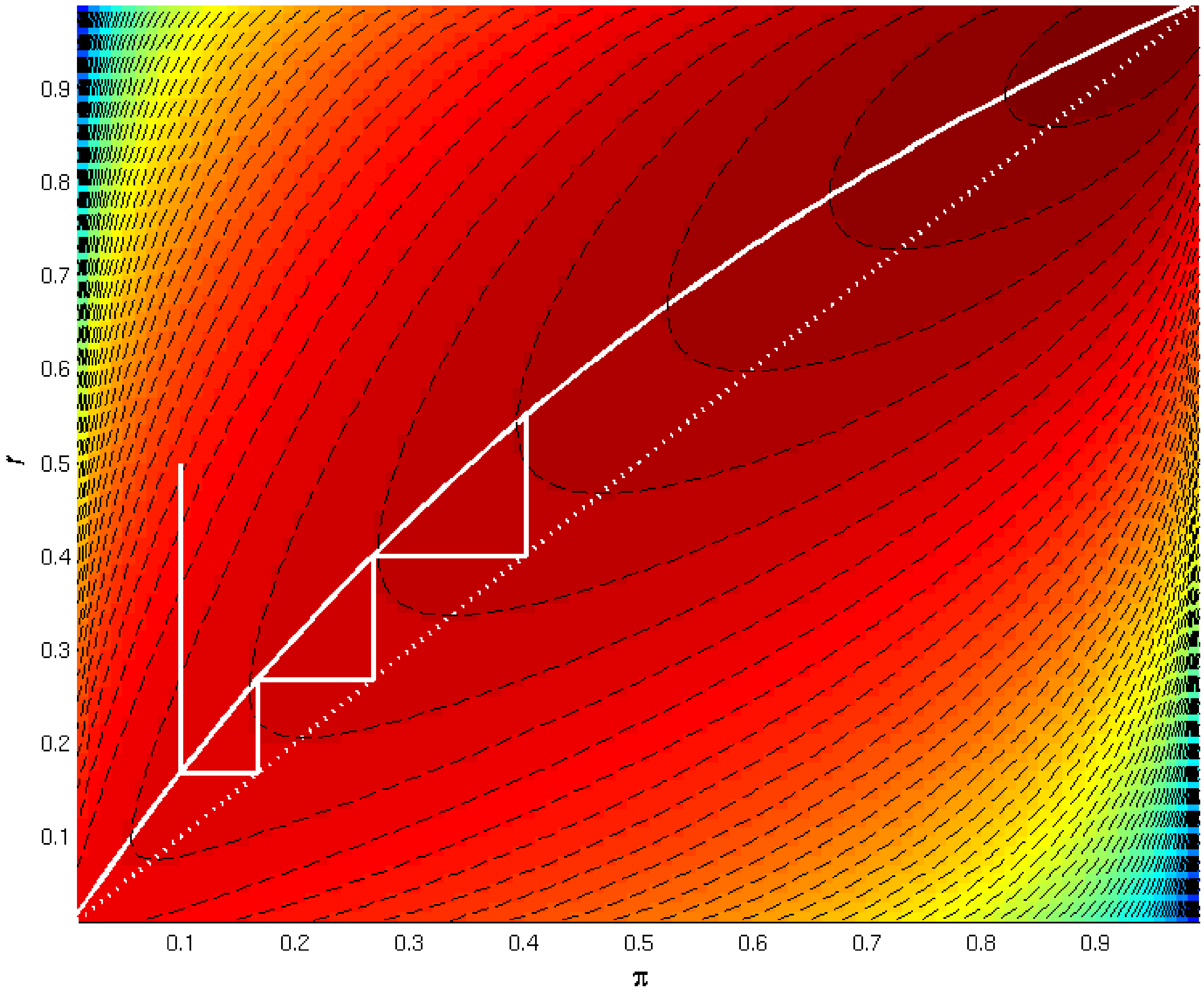
Coordinate Ascent in $\mathcal{F}$ (Demo)



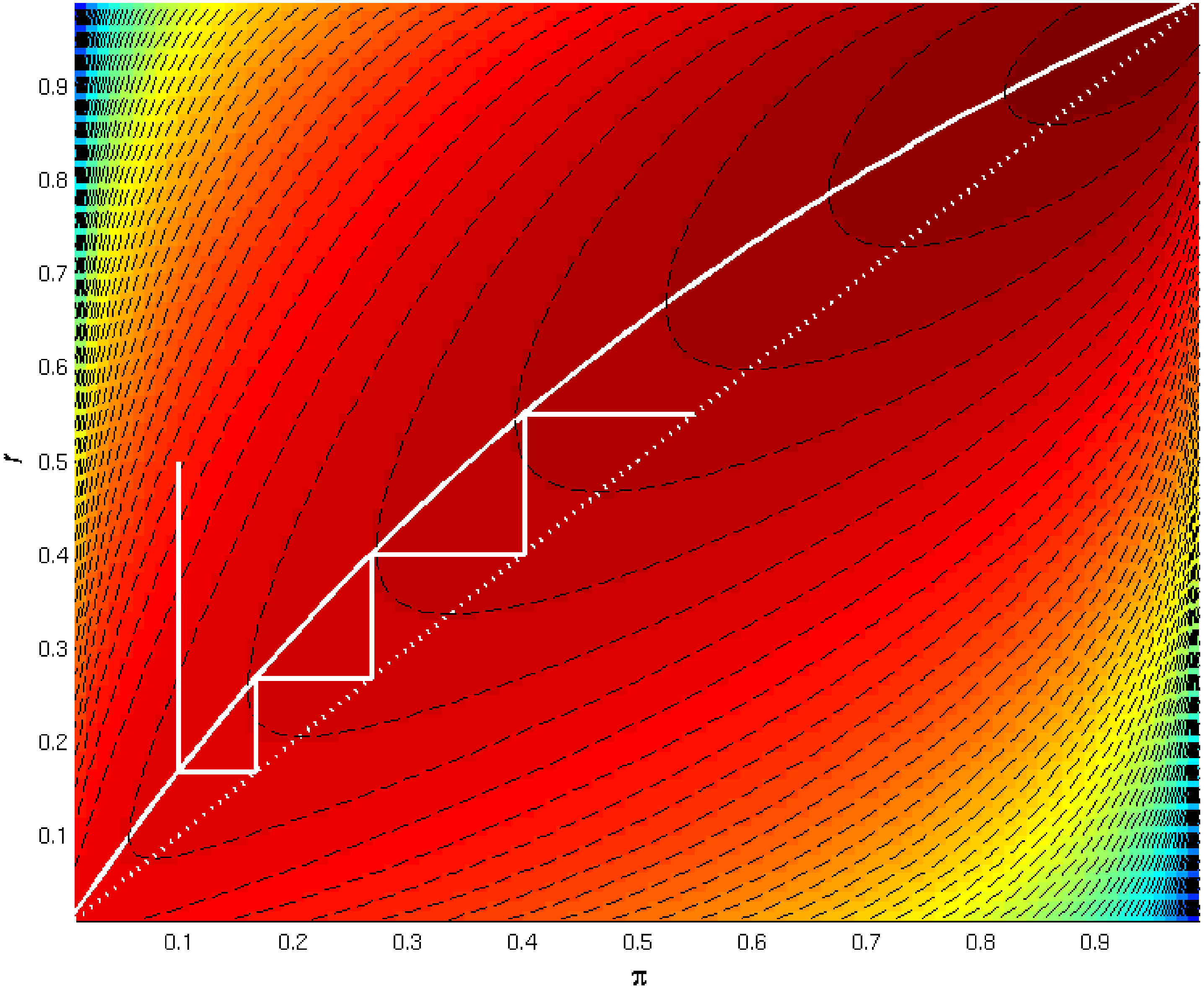
Coordinate Ascent in $\mathcal{F}$ (Demo)



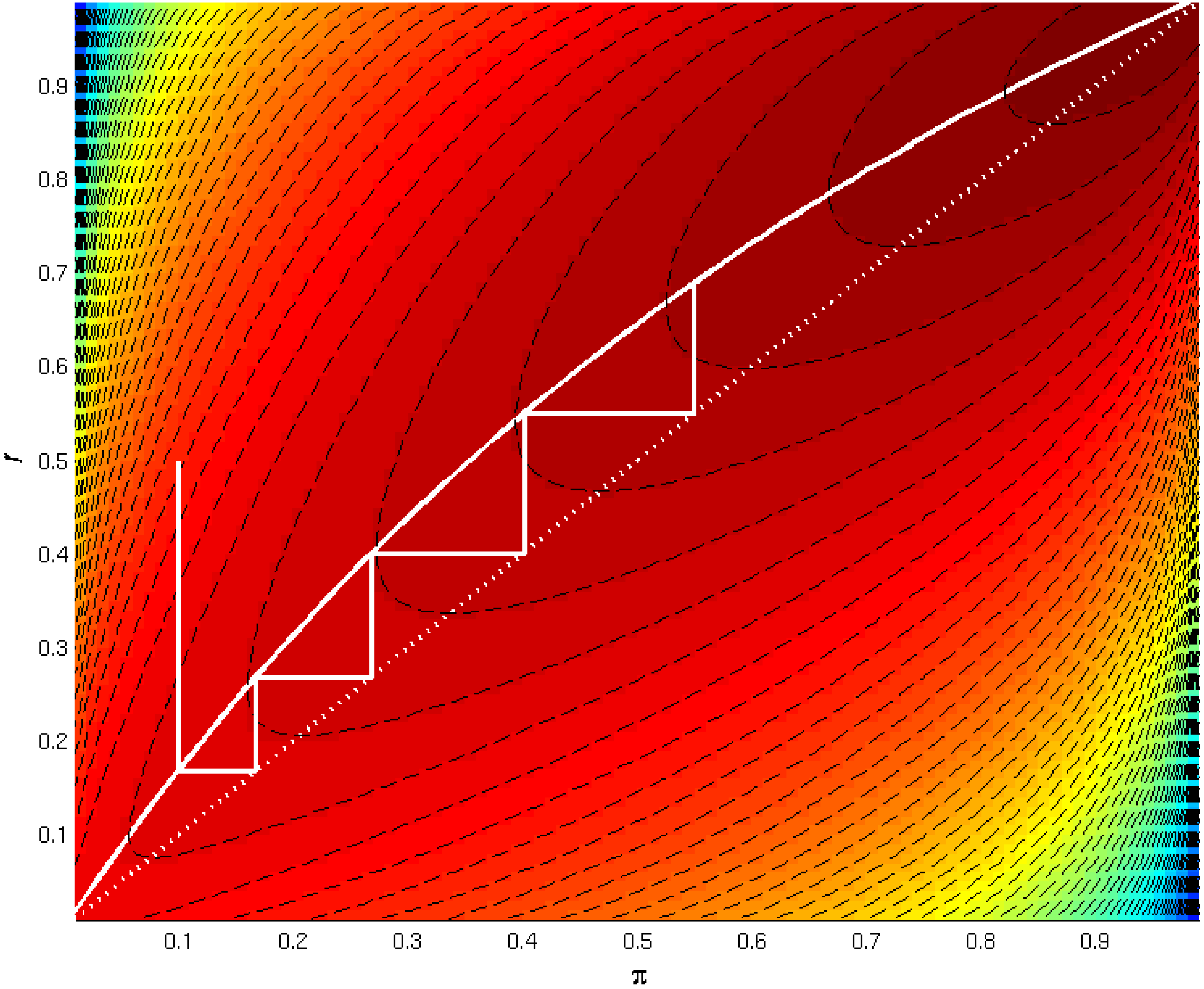
Coordinate Ascent in $\mathcal{F}$ (Demo)



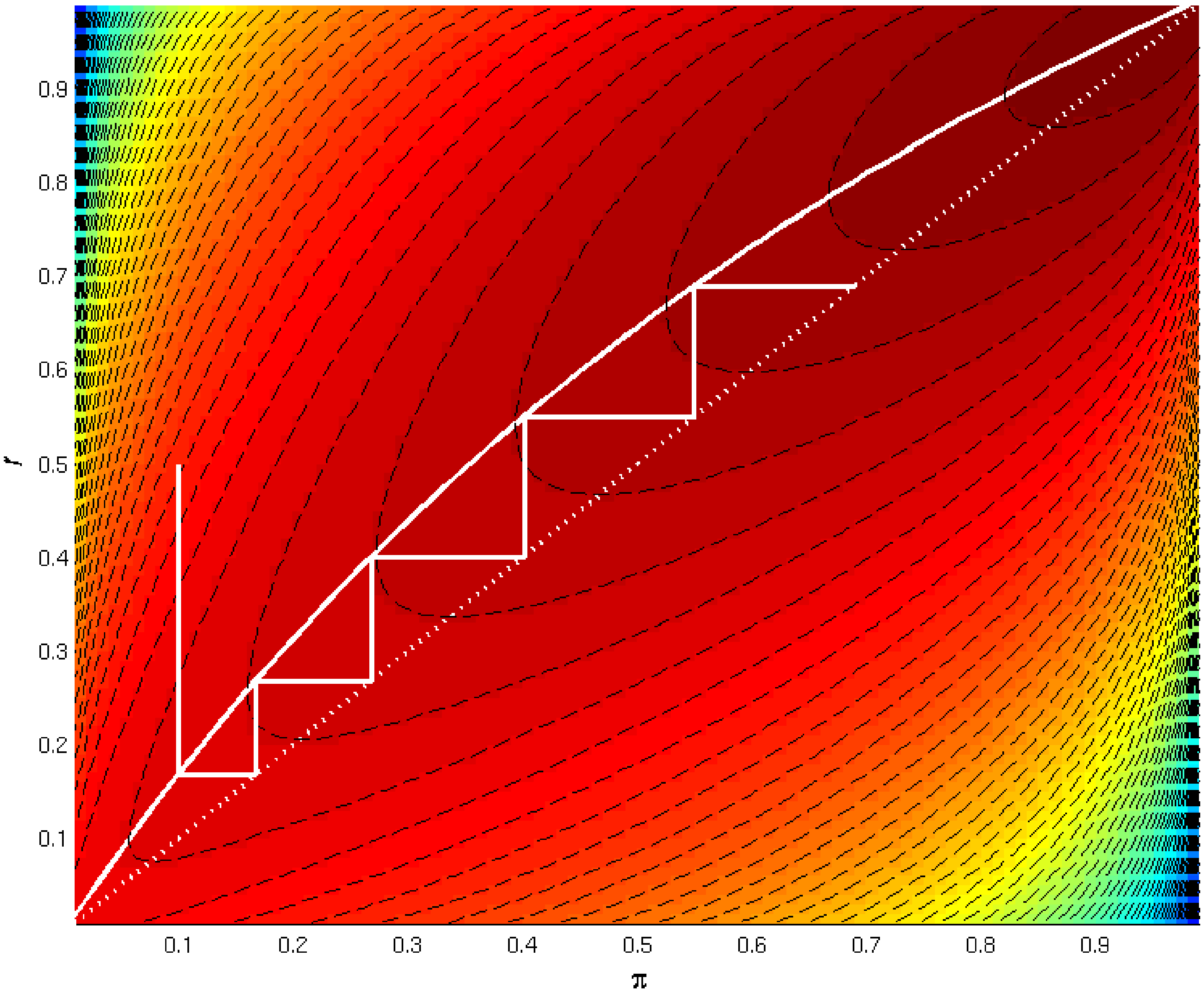
Coordinate Ascent in $\mathcal{F}$ (Demo)



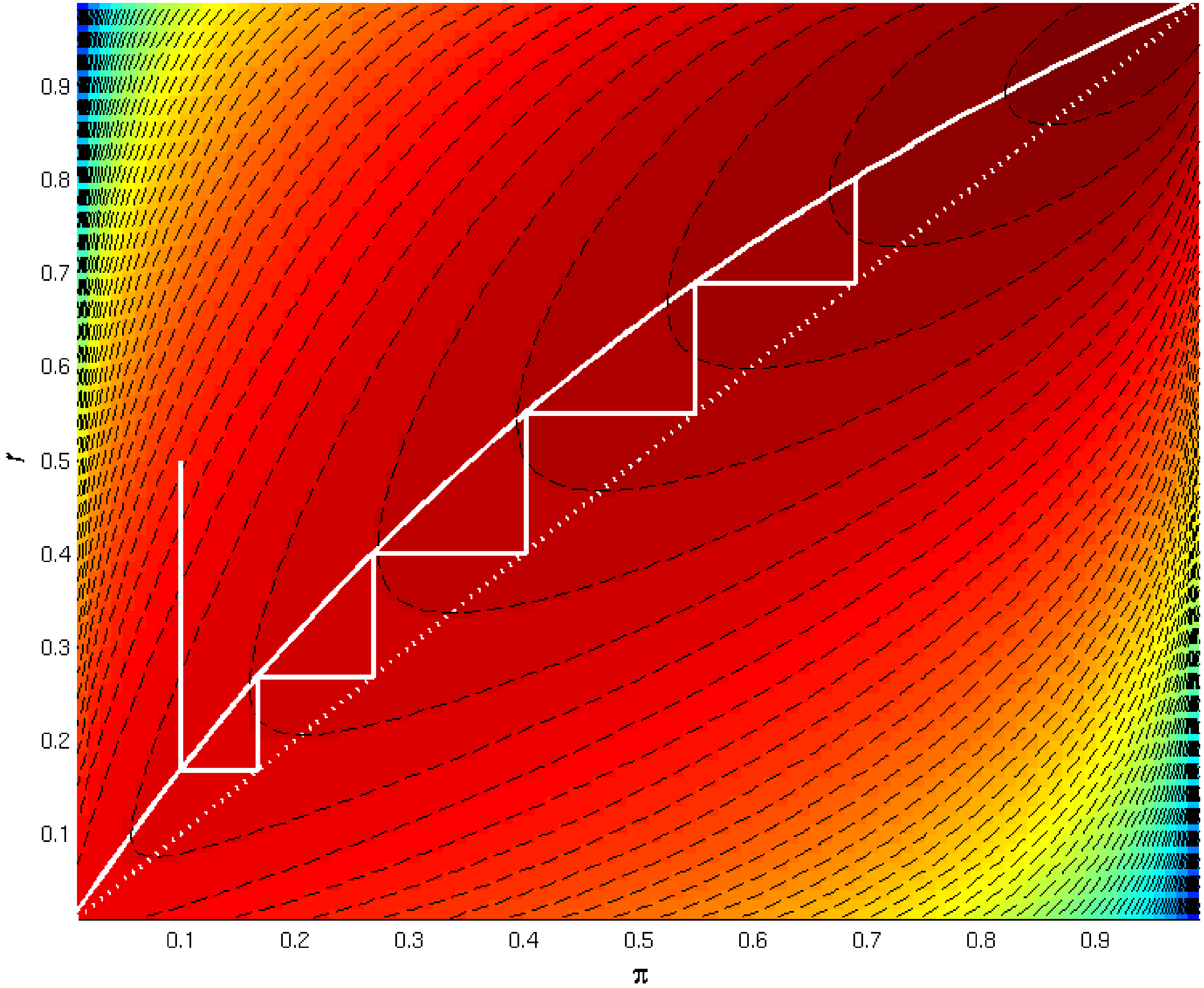
Coordinate Ascent in $\mathcal{F}$ (Demo)



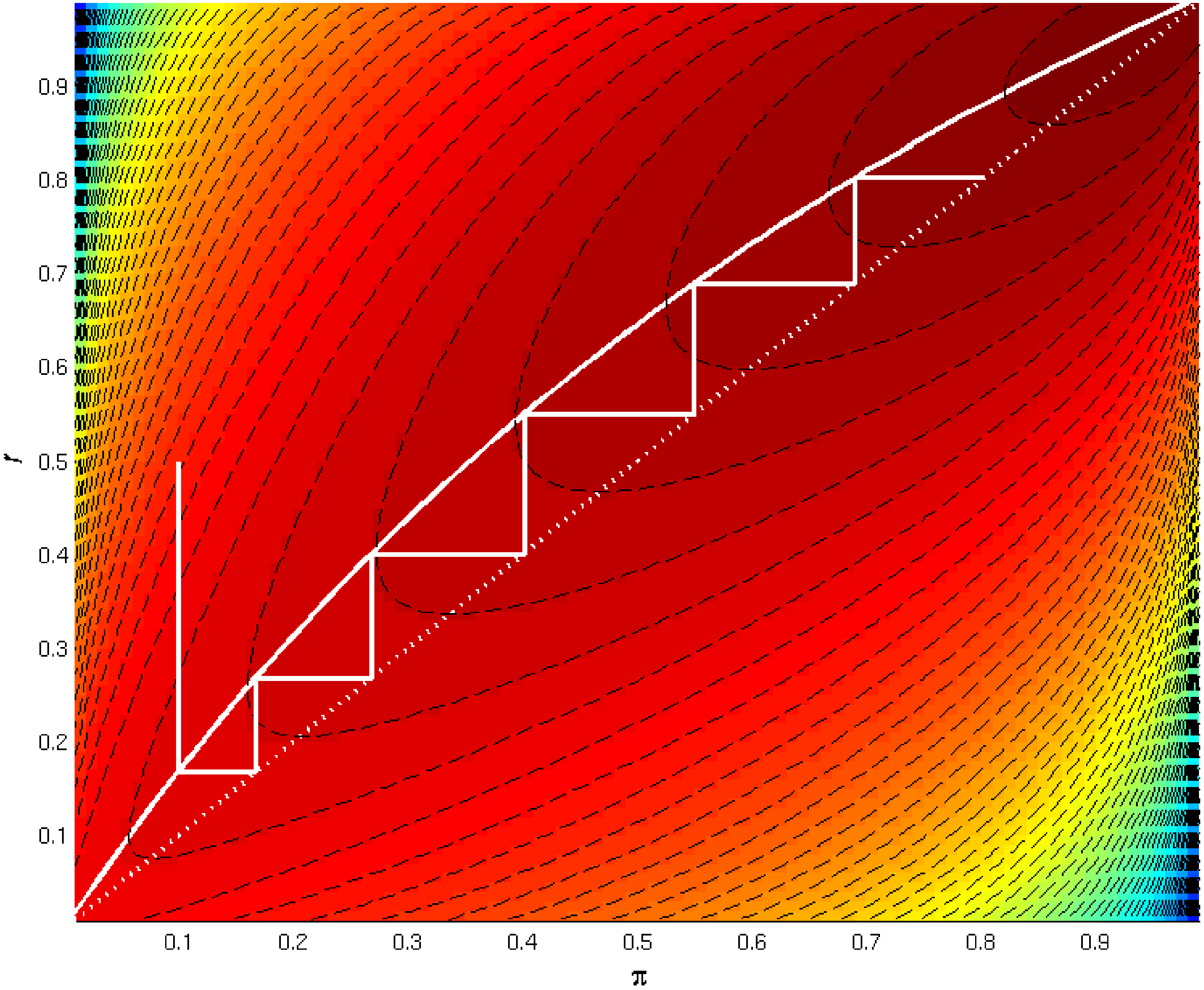
Coordinate Ascent in $\mathcal{F}$ (Demo)



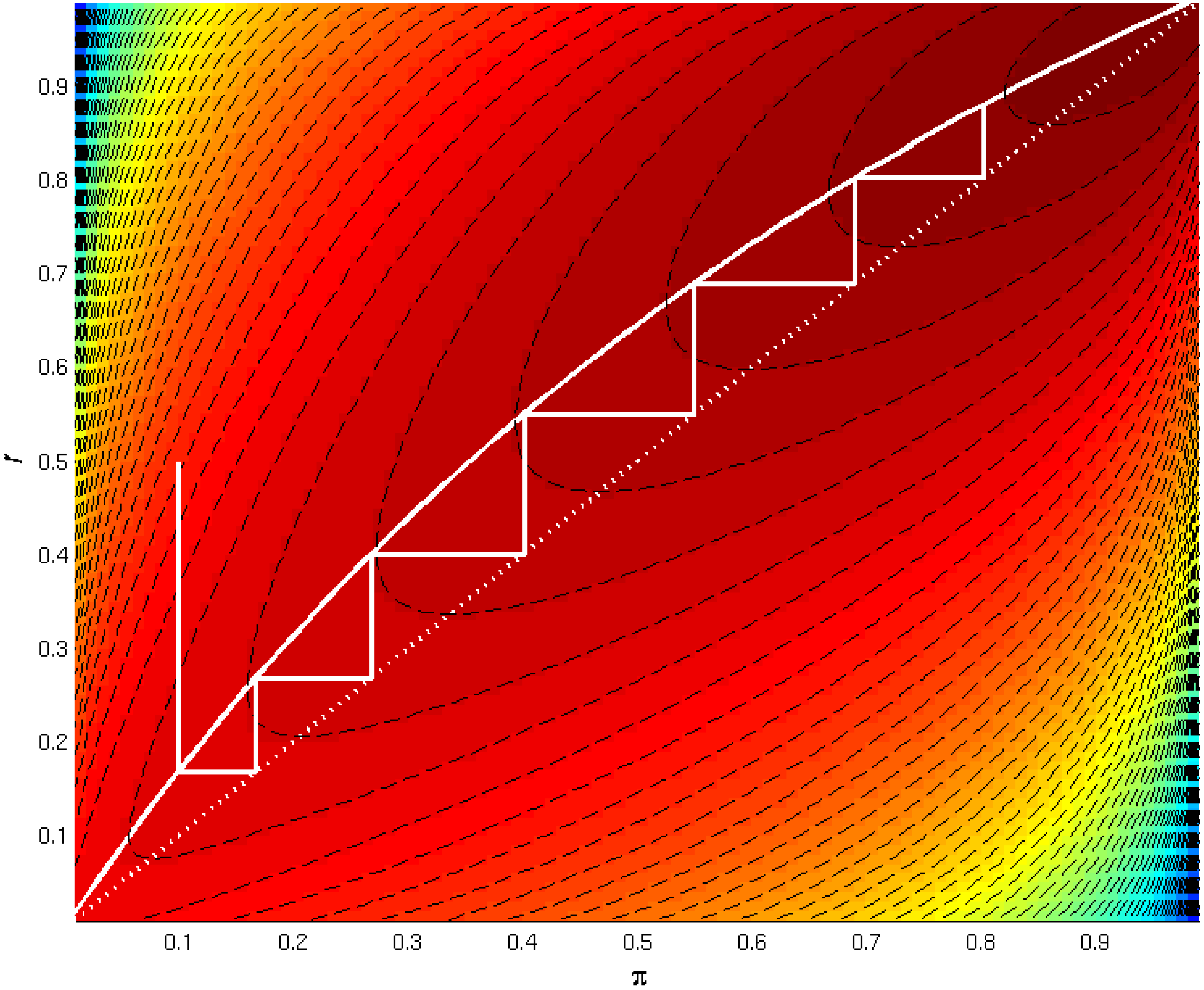
Coordinate Ascent in $\mathcal{F}$ (Demo)



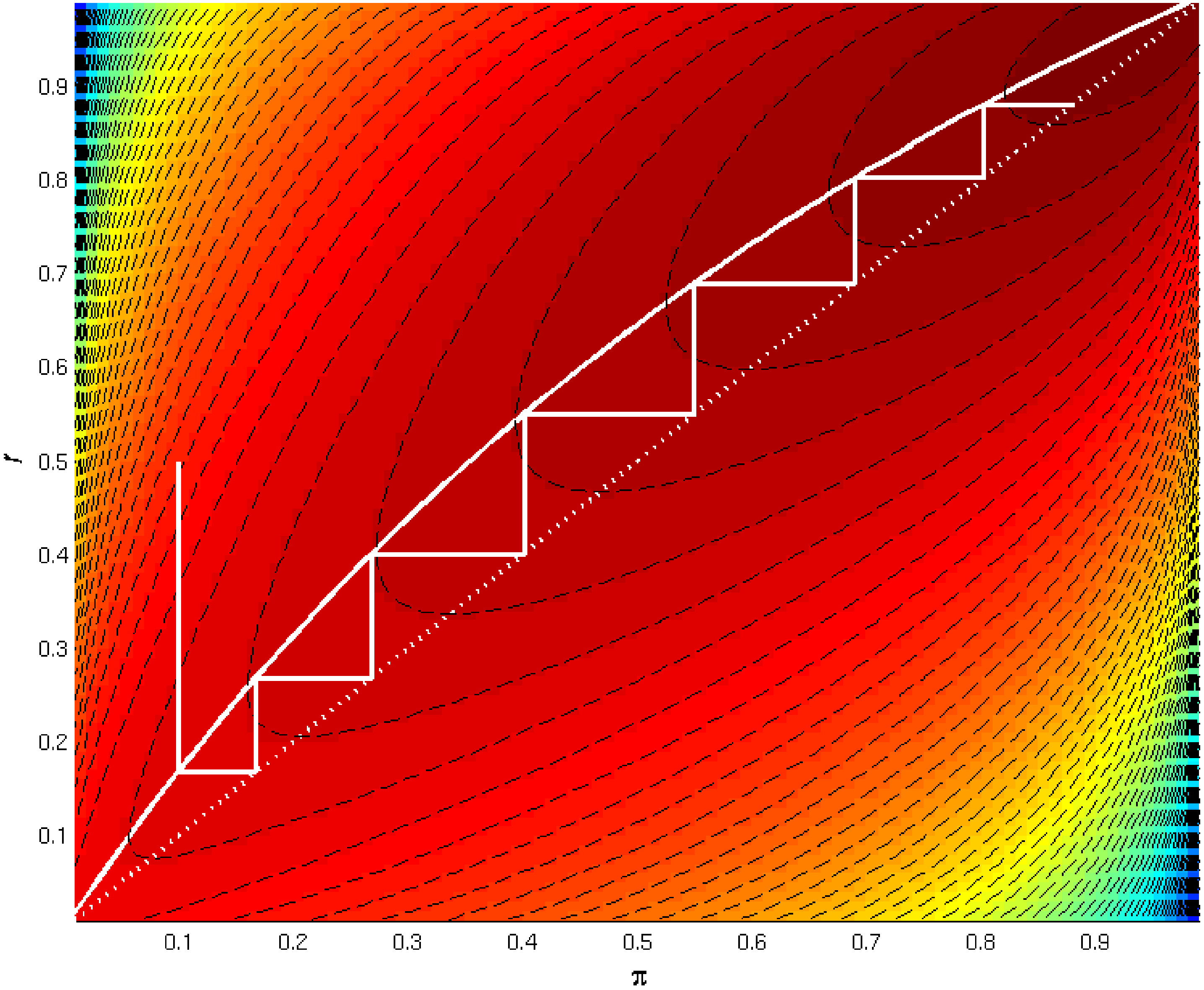
Coordinate Ascent in $\mathcal{F}$ (Demo)



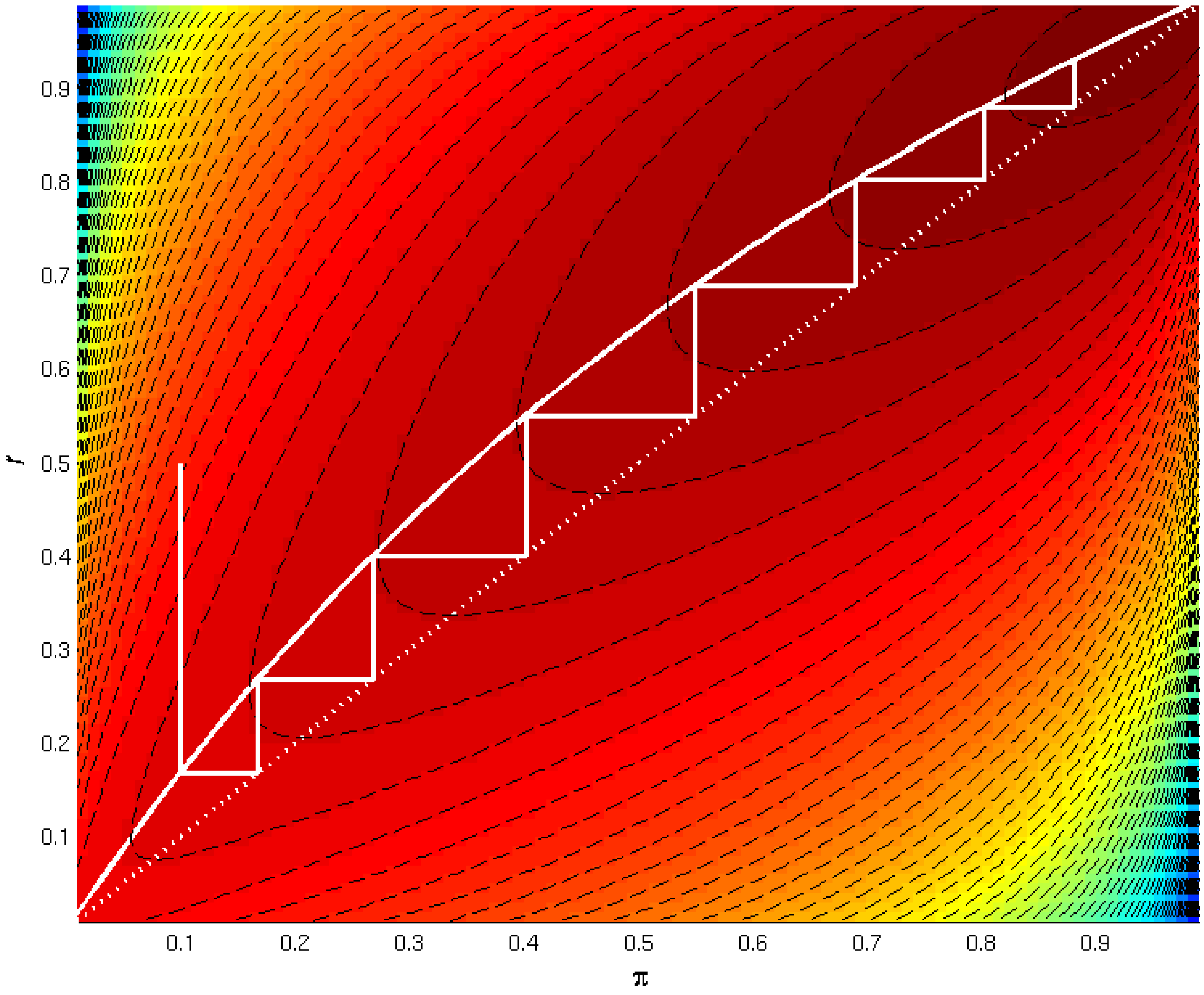
Coordinate Ascent in $\mathcal{F}$ (Demo)



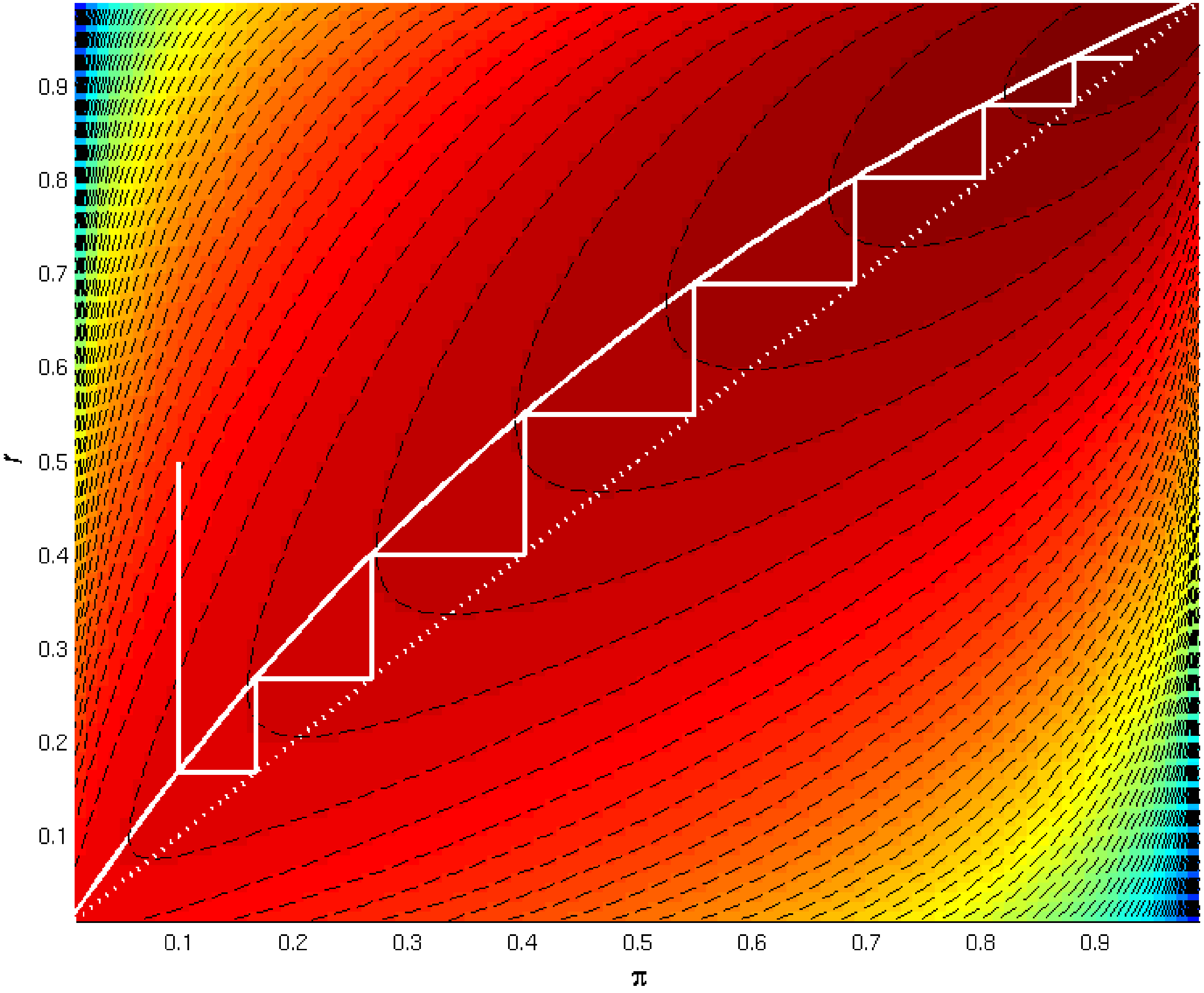
Coordinate Ascent in $\mathcal{F}$ (Demo)



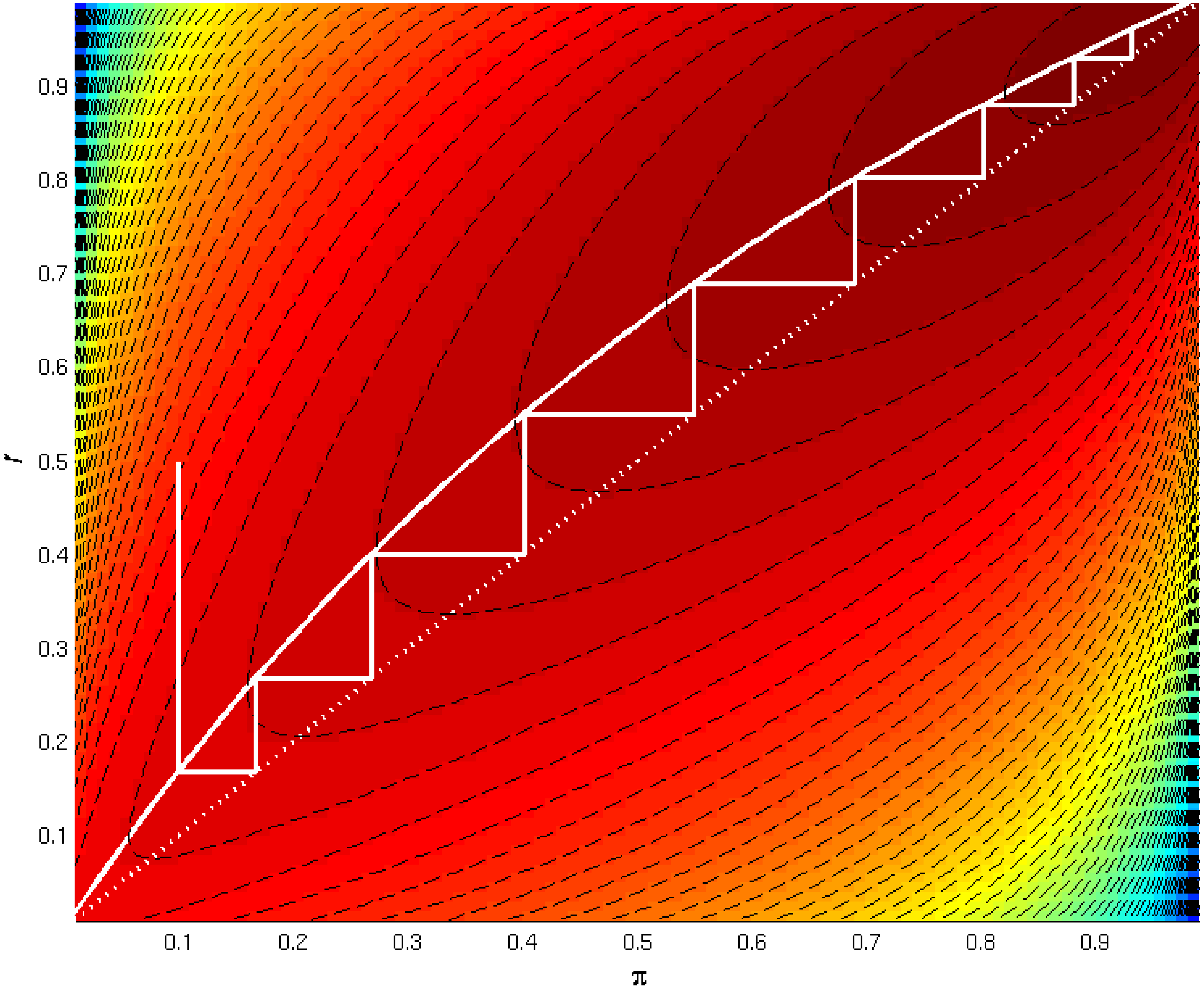
Coordinate Ascent in $\mathcal{F}$ (Demo)



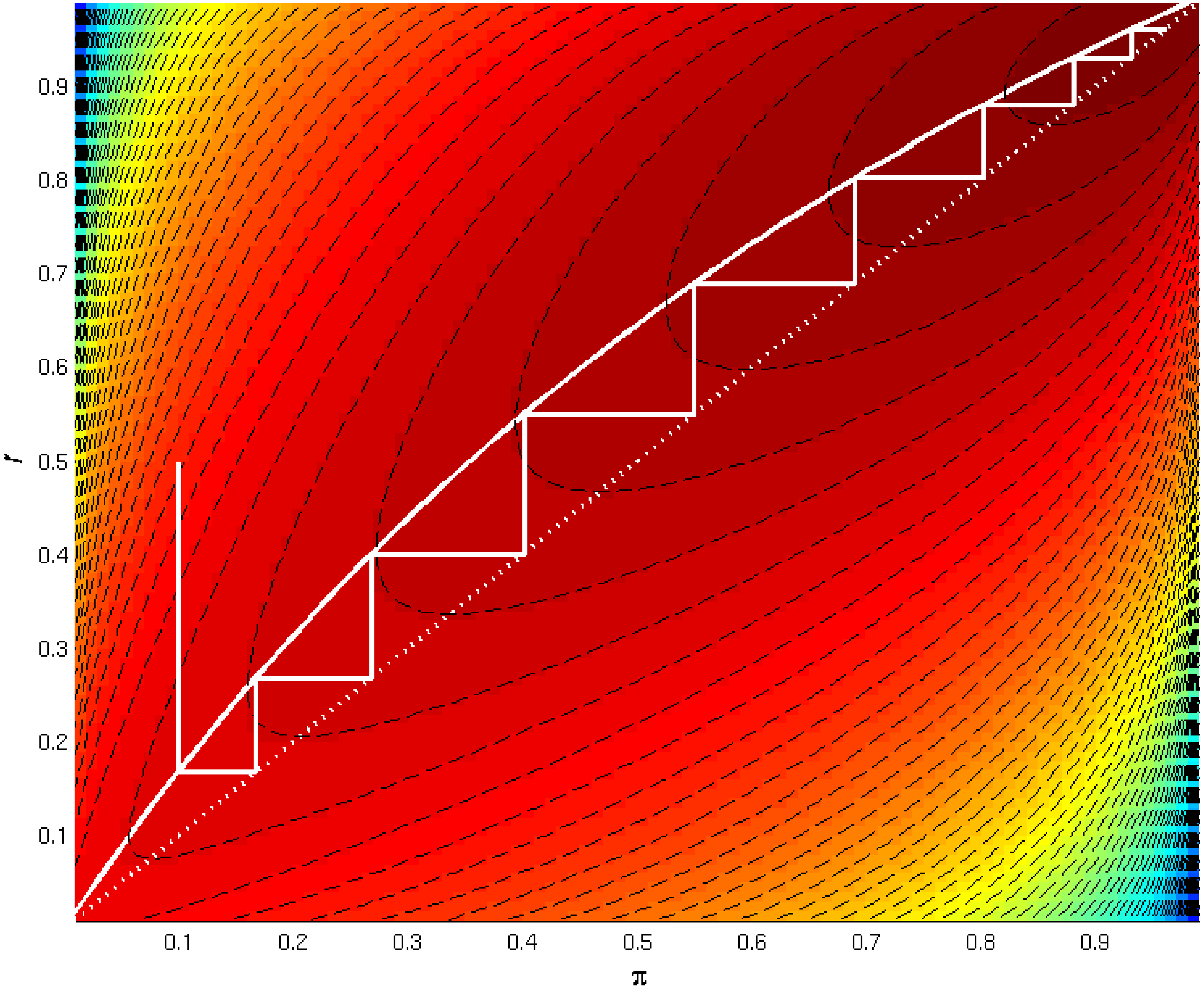
Coordinate Ascent in $\mathcal{F}$ (Demo)



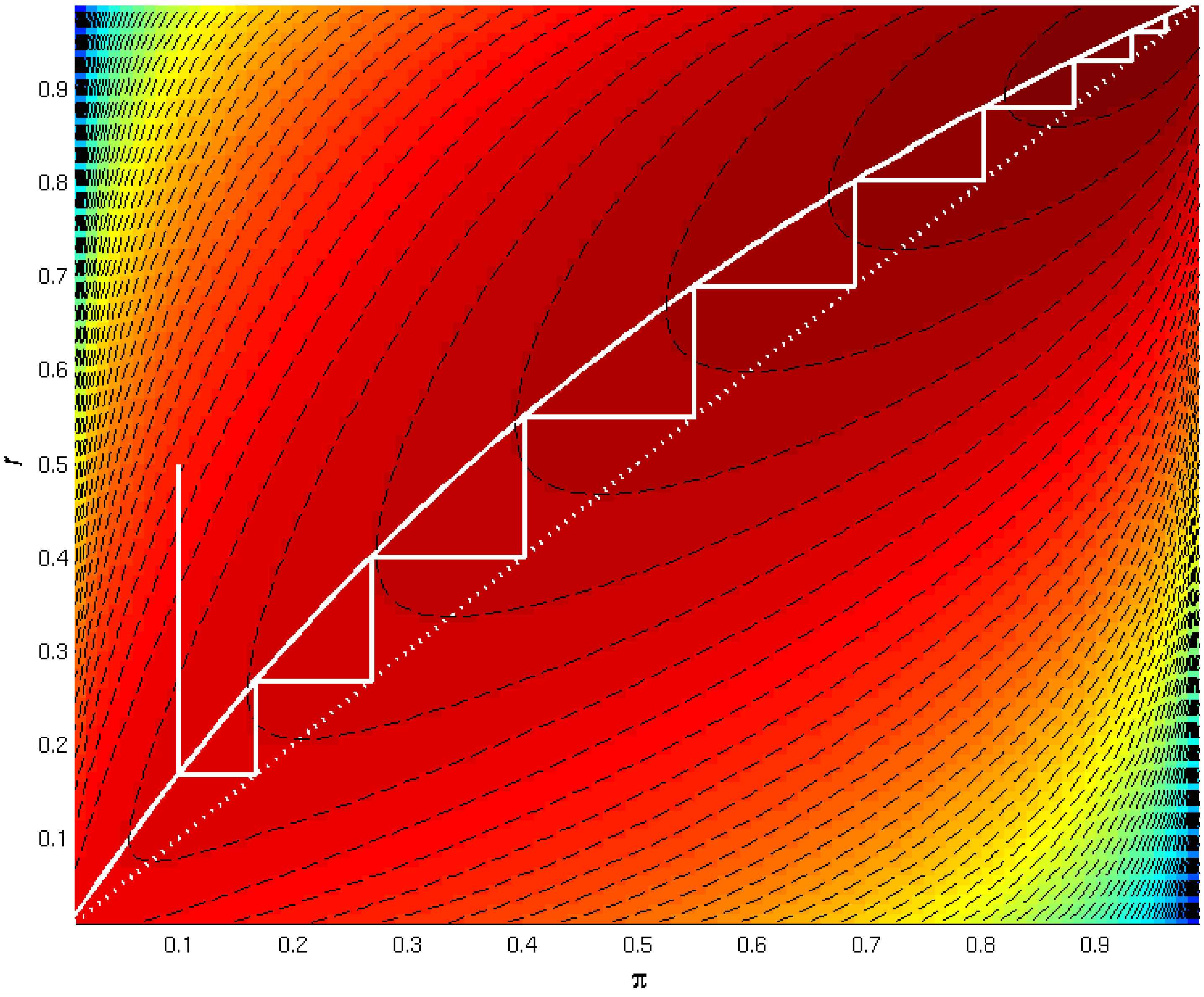
Coordinate Ascent in $\mathcal{F}$ (Demo)



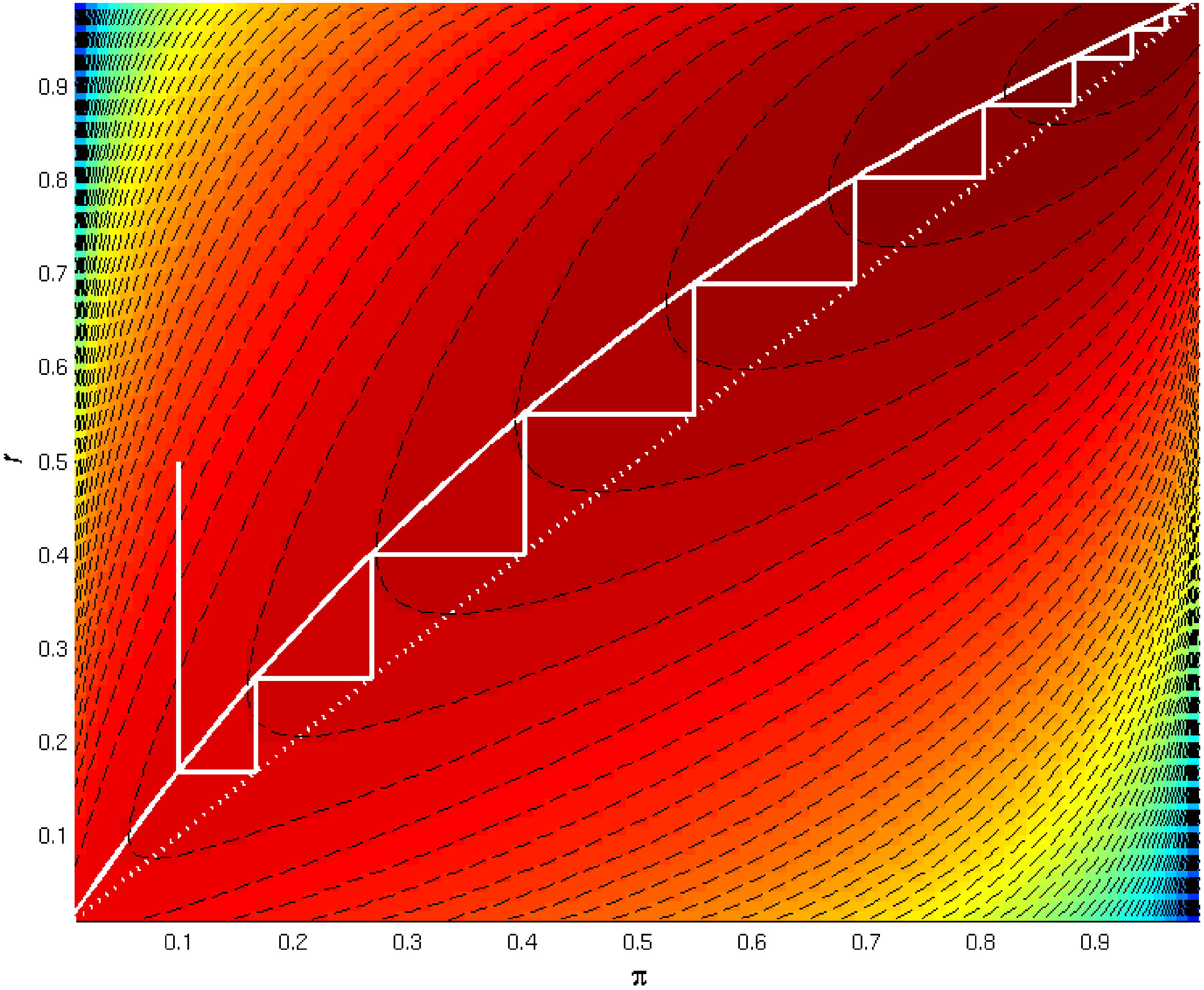
Coordinate Ascent in $\mathcal{F}$ (Demo)



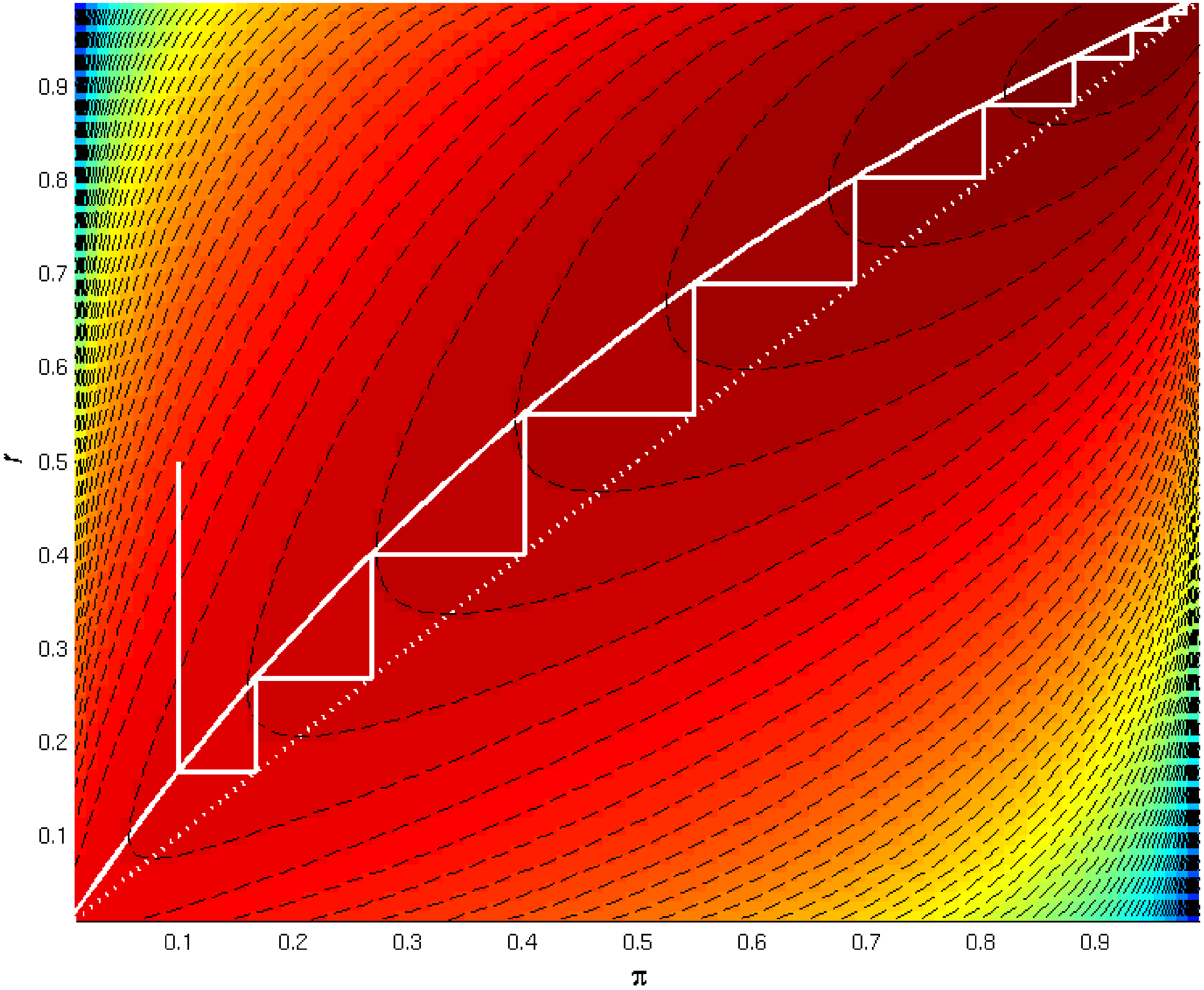
Coordinate Ascent in $\mathcal{F}$ (Demo)



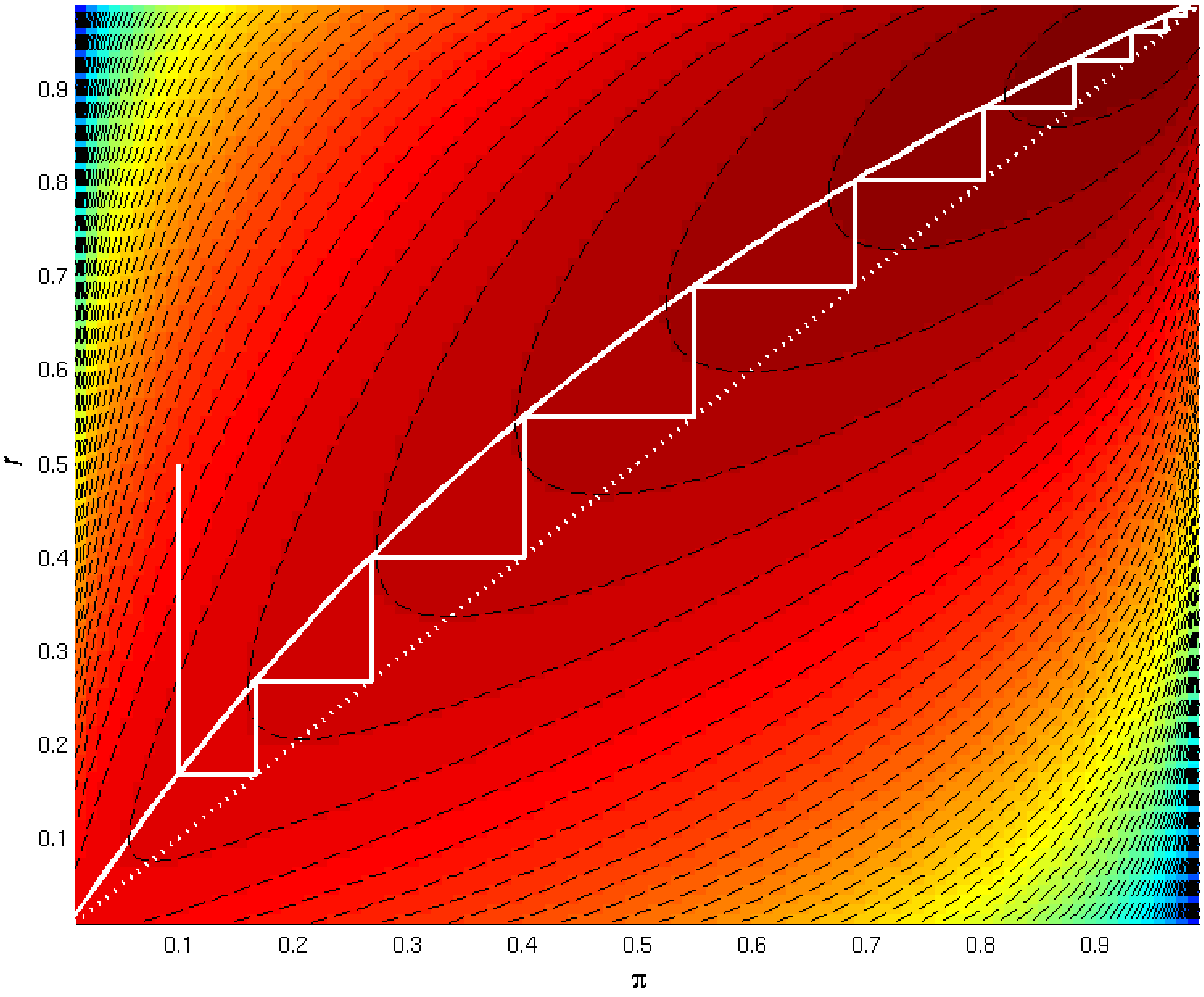
Coordinate Ascent in $\mathcal{F}$ (Demo)



Coordinate Ascent in $\mathcal{F}$ (Demo)



Coordinate Ascent in $\mathcal{F}$ (Demo)



EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\left(\theta^{(k-1)}\right)$$

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big)$$

- The E step brings the free energy to the likelihood.

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)},\theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)},\theta^{(k)}\big)$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)},\theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)},\theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \ell\big(\theta^{(k)}\big),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \ell\big(\theta^{(k)}\big),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\theta^{(k)} \neq \theta^{(k-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of θ iff the likelihood increases.

Fixed Points of EM are Stationary Points in ℓ

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle \log P(\mathcal{Y}, \mathcal{X} \mid \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle\log P(\mathcal{Y},\mathcal{X}\mid\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\Big|_{\theta^*} = 0$$

Now, $\ell(\theta) = \log P(\mathcal{X}|\theta) = \langle\log P(\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle\log P(\mathcal{Y},\mathcal{X}\mid\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\Big|_{\theta^*}=0$$

Now,

$$\begin{aligned}\ell(\theta)&=\log P(\mathcal{X}|\theta)=\langle\log P(\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\\&=\left\langle\log\frac{P(\mathcal{Y},\mathcal{X}|\theta)}{P(\mathcal{Y}|\mathcal{X},\theta)}\right\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\end{aligned}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial \theta} \langle \log P(\mathcal{Y}, \mathcal{X} \mid \theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned}
\ell(\theta) &= \log P(\mathcal{X}|\theta) = \langle \log P(\mathcal{X}|\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\
&= \left\langle \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{P(\mathcal{Y}|\mathcal{X}, \theta)} \right\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\
&= \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \langle \log P(\mathcal{Y}|\mathcal{X}, \theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}
\end{aligned}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta} \langle \log P(\mathcal{Y}, \mathcal{X} \mid \theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned} \ell(\theta) &= \log P(\mathcal{X}|\theta) = \langle \log P(\mathcal{X}|\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \left\langle \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{P(\mathcal{Y}|\mathcal{X},\theta)} \right\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \langle \log P(\mathcal{Y},\mathcal{X}|\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \langle \log P(\mathcal{Y}|\mathcal{X},\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \end{aligned}$$

so,

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta} \langle \log P(\mathcal{Y},\mathcal{X}|\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d}{d\theta} \langle \log P(\mathcal{Y}|\mathcal{X},\theta) \rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle\log P(\mathcal{Y},\mathcal{X} \mid \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned}\ell(\theta) = \log P(\mathcal{X}|\theta) &= \langle\log P(\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \left\langle \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{P(\mathcal{Y}|\mathcal{X},\theta)} \right\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \langle\log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \langle\log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\end{aligned}$$

so,

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\langle\log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d}{d\theta}\langle\log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

The second term is 0 at θ^* if the derivative exists (minimum of **KL**$[\cdot\|\cdot]$), and thus:

$$\frac{d}{d\theta}\ell(\theta)\bigg|_{\theta^*} = \frac{d}{d\theta}\langle\log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial\theta}\langle \log P(\mathcal{Y},\mathcal{X} \mid \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\bigg|_{\theta^*} = 0$$

Now,

$$\begin{aligned}\ell(\theta) = \log P(\mathcal{X}|\theta) &= \langle \log P(\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \left\langle \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{P(\mathcal{Y}|\mathcal{X},\theta)}\right\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \langle \log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\end{aligned}$$

so,

$$\frac{d}{d\theta}\ell(\theta) = \frac{d}{d\theta}\langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d}{d\theta}\langle \log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

The second term is 0 at θ^* if the derivative exists (minimum of **KL**$[\cdot\|\cdot]$), and thus:

$$\frac{d}{d\theta}\ell(\theta)\bigg|_{\theta^*} = \frac{d}{d\theta}\langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\bigg|_{\theta^*} = 0$$

So, EM converges to a stationary point of $\ell(\theta)$.

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle\log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle\log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle\log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle\log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum).

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathcal{Y}, \mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathcal{Y}|\mathcal{X}, \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum). Thus the curvature of the likelihood is negative and

θ^* is a maximum of ℓ.

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

$$\frac{d^2}{d\theta^2}\ell(\theta) = \frac{d^2}{d\theta^2}\langle \log P(\mathcal{Y}, \mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \frac{d^2}{d\theta^2}\langle \log P(\mathcal{Y}|\mathcal{X}, \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum). Thus the curvature of the likelihood is negative and

θ^* is a maximum of ℓ.

[...as long as the derivatives exist. They sometimes don't (zero-noise ICA)].

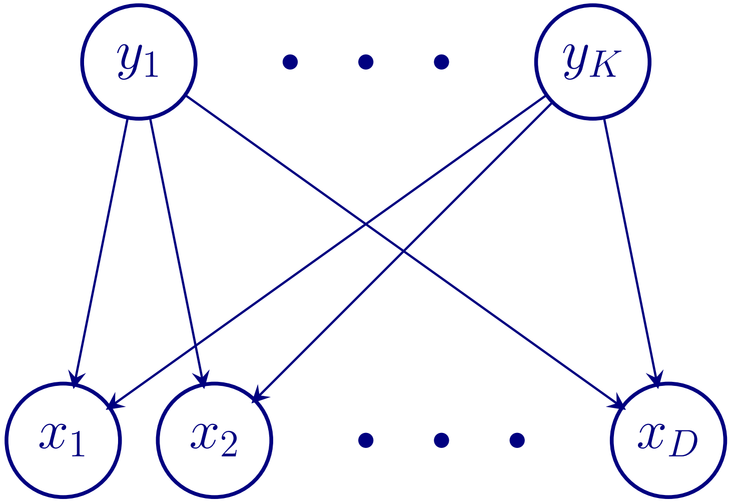
Partial M steps and Partial E steps

Partial M steps: The proof holds even if we just *increase* $\mathcal{F}$ wrt θ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

Partial E steps: We can also just *increase* $\mathcal{F}$ wrt to some of the qs.

For example, sparse or online versions of the EM algorithm would compute the posterior for a subset of the data points or as the data arrives, respectively. You can also update the posterior over a subset of the hidden variables, while holding others fixed...

Factor Analysis



Linear generative model: $x_d = \sum_{k=1}^{K} \Lambda_{dk}\, y_k + \epsilon_d$

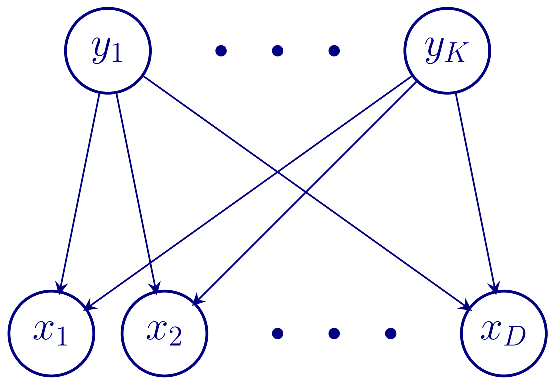
- y_k are independent $\mathcal{N}(0, 1)$ Gaussian factors
- ϵ_d are independent $\mathcal{N}(0, \Psi_{dd})$ Gaussian noise
- $K < D$

So, **x** is Gaussian with: $p(\mathbf{x}) = \int p(\mathbf{y}) p(\mathbf{x}|\mathbf{y}) d\mathbf{y} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$

where Λ is a $D \times K$ matrix, and Ψ is diagonal.

Dimensionality Reduction: Finds a low-dimensional projection of high dimensional data that captures the correlation structure of the data.

EM for Factor Analysis



The model for **x**:

$$p(\mathbf{x}|\theta) = \int p(\mathbf{y}|\theta)p(\mathbf{x}|\mathbf{y},\theta)d\mathbf{y} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$$

Model parameters: $\theta = \{\Lambda, \Psi\}$.

E step: For each data point $\mathbf{x}_n$, compute the posterior distribution of hidden factors given the observed data: $q_n(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}_n, \theta_t)$.

M step: Find the θ_{t+1} that maximises $\mathcal{F}(q, \theta)$:

$$\begin{aligned} \mathcal{F}(q,\theta) &= \sum_n \int q_n(\mathbf{y}) \left[\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y},\theta) - \log q_n(\mathbf{y})\right] d\mathbf{y} \\ &= \sum_n \int q_n(\mathbf{y}) \left[\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y},\theta)\right] d\mathbf{y} + \text{c}. \end{aligned}$$

The E step for Factor Analysis

E step: For each data point $\mathbf{x}_n$, compute the posterior distribution of hidden factors given the observed data: $q_n(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}_n, \theta) = p(\mathbf{y}, \mathbf{x}_n|\theta)/p(\mathbf{x}_n|\theta)$

Tactic: write $p(\mathbf{y}, \mathbf{x}_n|\theta)$, consider $\mathbf{x}_n$ to be fixed. What is this as a function of $\mathbf{y}$?

$$\begin{aligned}
p(\mathbf{y}, \mathbf{x}_n) &= p(\mathbf{y})p(\mathbf{x}_n|\mathbf{y}) \\
&= (2\pi)^{-\frac{K}{2}} \exp\{-\frac{1}{2}\mathbf{y}^\top \mathbf{y}\} \; |2\pi\Psi|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}_n - \Lambda\mathbf{y})^\top \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y})\} \\
&= \text{c} \times \exp\{-\frac{1}{2}[\mathbf{y}^\top \mathbf{y} + (\mathbf{x}_n - \Lambda\mathbf{y})^\top \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y})]\} \\
&= \text{c'} \times \exp\{-\frac{1}{2}[\mathbf{y}^\top (I + \Lambda^\top \Psi^{-1}\Lambda)\mathbf{y} - 2\mathbf{y}^\top \Lambda^\top \Psi^{-1}\mathbf{x}_n]\} \\
&= \text{c''} \times \exp\{-\frac{1}{2}[\mathbf{y}^\top \Sigma^{-1}\mathbf{y} - 2\mathbf{y}^\top \Sigma^{-1}\mu + \mu^\top \Sigma^{-1}\mu]\}
\end{aligned}$$

So $\Sigma = (I + \Lambda^\top \Psi^{-1}\Lambda)^{-1} = I - \beta\Lambda$ and $\mu = \Sigma\Lambda^\top \Psi^{-1}\mathbf{x}_n = \beta\mathbf{x}_n$. Where $\beta = \Sigma\Lambda^\top \Psi^{-1}$.
Note that μ is a linear function of $\mathbf{x}_n$ and Σ does not depend on $\mathbf{x}_n$.

The M step for Factor Analysis

M step: Find θ_{t+1} maximising $\mathcal{F} = \sum_n \int q_n(\mathbf{y}) \left[\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y}, \theta)\right] d\mathbf{y} + \text{c}$

$$\begin{aligned}\log p(\mathbf{y}|\theta) + \log p(\mathbf{x}_n|\mathbf{y}, \theta) &= \text{c} - \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\mathbf{x}_n - \Lambda\mathbf{y})^\top\Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y}) \\ &= \text{c'} - \frac{1}{2}\log|\Psi| - \frac{1}{2}[\mathbf{x}_n{}^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n{}^\top\Psi^{-1}\Lambda\mathbf{y} + \mathbf{y}^\top\Lambda^\top\Psi^{-1}\Lambda\mathbf{y}] \\ &= \text{c'} - \frac{1}{2}\log|\Psi| - \frac{1}{2}[\mathbf{x}_n{}^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n{}^\top\Psi^{-1}\Lambda\mathbf{y} + \text{Tr}\left[\Lambda^\top\Psi^{-1}\Lambda\mathbf{y}\mathbf{y}^\top\right]]\end{aligned}$$

Taking expectations over $q_n(\mathbf{y})\ldots$

$$= \text{c'} - \frac{1}{2}\log|\Psi| - \frac{1}{2}[\mathbf{x}_n{}^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n{}^\top\Psi^{-1}\Lambda\mu_n + \text{Tr}\left[\Lambda^\top\Psi^{-1}\Lambda(\mu_n\mu_n{}^\top + \Sigma)\right]]$$

Note that we don't need to know everything about q, just the expectations of $\mathbf{y}$ and $\mathbf{y}\mathbf{y}^\top$ under q (i.e. the expected sufficient statistics).

The M step for Factor Analysis (cont.)

$$\mathcal{F} = \mathrm{c'} - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \left[\mathbf{x}_n{}^\top \Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n{}^\top\Psi^{-1}\Lambda\mu_n + \mathrm{Tr}\left[\Lambda^\top\Psi^{-1}\Lambda(\mu_n\mu_n{}^\top + \Sigma)\right]\right]$$

Taking derivatives w.r.t. Λ and Ψ^{-1}, using $\frac{\partial \mathrm{Tr}[AB]}{\partial B} = A^\top$ and $\frac{\partial \log|A|}{\partial A} = A^{-\top}$:

$$\frac{\partial \mathcal{F}}{\partial \Lambda} = \Psi^{-1}\sum_n \mathbf{x}_n\mu_n{}^\top - \Psi^{-1}\Lambda\left(N\Sigma + \sum_n \mu_n\mu_n{}^\top\right) = 0$$

$$\hat{\Lambda} = \left(\sum_n \mathbf{x}_n\mu_n{}^\top\right)\left(N\Sigma + \sum_n \mu_n\mu_n{}^\top\right)^{-1}$$

$$\frac{\partial \mathcal{F}}{\partial \Psi^{-1}} = \frac{N}{2}\Psi - \frac{1}{2}\sum_n \left[\mathbf{x}_n\mathbf{x}_n{}^\top - \Lambda\mu_n\mathbf{x}_n{}^\top - \mathbf{x}_n\mu_n{}^\top\Lambda^\top + \Lambda(\mu_n\mu_n{}^\top + \Sigma)\Lambda^\top\right]$$

$$\hat{\Psi} = \frac{1}{N}\sum_n \left[\mathbf{x}_n\mathbf{x}_n{}^\top - \Lambda\mu_n\mathbf{x}_n{}^\top - \mathbf{x}_n\mu_n{}^\top\Lambda^\top + \Lambda(\mu_n\mu_n{}^\top + \Sigma)\Lambda^\top\right]$$

$$\hat{\Psi} = \Lambda\Sigma\Lambda^\top + \frac{1}{N}\sum_n (\mathbf{x}_n - \Lambda\mu_n)(\mathbf{x}_n - \Lambda\mu_n)^\top \qquad \text{(squared residuals)}$$

Note: we should actually only take derivarives w.r.t. Ψ_{dd} since Ψ is diagonal.
When $\Sigma \to 0$ these become the equations for linear regression!

EM for exponential families

Defn: p is in the exponential family for $\mathbf{z} = (\mathbf{y}, \mathbf{x})$ if it can be written:

$$p(\mathbf{z}|\theta) = b(\mathbf{z}) \exp\{\theta^\top s(\mathbf{z})\}/\alpha(\theta)$$

where $\alpha(\theta) = \int b(\mathbf{z}) \exp\{\theta^\top s(\mathbf{z})\} d\mathbf{z}$

E step: $q(\mathbf{y}) = p(\mathbf{y}|\mathbf{x}, \theta)$

M step: $\theta^{(k)} := \underset{\theta}{\operatorname{argmax}}\ \mathcal{F}(q, \theta)$

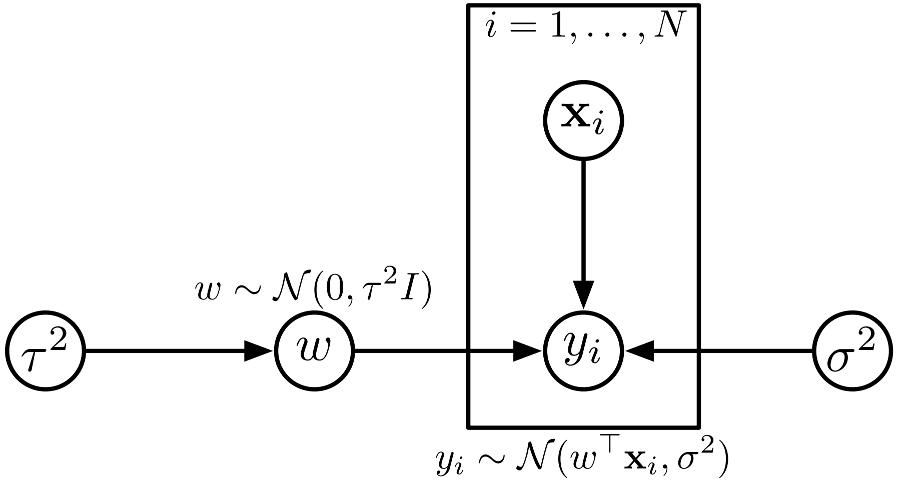
$$\begin{aligned} \mathcal{F}(q,\theta) &= \int q(\mathbf{y}) \log p(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{y} - \mathcal{H}(q) \\ &= \int q(\mathbf{y}) [\theta^\top s(\mathbf{z}) - \log \alpha(\theta)] d\mathbf{y} + \text{const} \end{aligned}$$

It is easy to verify that: $\dfrac{\partial \log \alpha(\theta)}{\partial \theta} = E[s(\mathbf{z})|\theta]$

Therefore, M step solves: $\dfrac{\partial \mathcal{F}}{\partial \theta} = E_{q(\mathbf{y})}[s(\mathbf{z})] - E[s(\mathbf{z})|\theta] = 0$

Gaussian Processes

Bayesian Linear Regression


$i = 1, \ldots, N$
$\mathbf{x}_i$
$w \sim \mathcal{N}(0, \tau^2 I)$
τ^2
w
y_i
σ^2
$y_i \sim \mathcal{N}(w^\top \mathbf{x}_i, \sigma^2)$

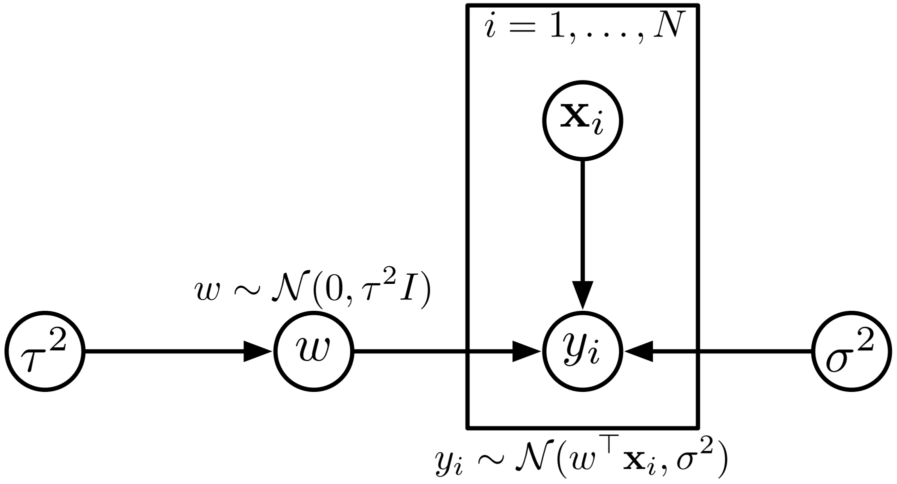
Given observed data $\mathcal{D} = \{X = [\mathbf{x}_1 \ldots \mathbf{x}_N], Y = [y_1 \ldots y_N]\}$, the posterior on $\mathbf{w}$ is:

$$\mathbf{w}|\mathcal{D} \sim \mathcal{N}\left(\underbrace{\frac{1}{\sigma^2}\Sigma_{\mathbf{w}} X Y^\top}_{\mu_{\mathbf{w}}}, \underbrace{\left(\frac{1}{\sigma^2} X X^\top + \frac{1}{\tau^2} I\right)^{-1}}_{\Sigma_{\mathbf{w}}} \right)$$

The Bayesian predictive distribution for $y'|\mathbf{x}'$ is obtained by integrating out $\mathbf{w}$:

$$\begin{aligned} p(y'|\mathbf{x}', \mathcal{D}) &= \int d\mathbf{w}\, p(y'|\mathbf{w}, \mathbf{x}') p(\mathbf{w}|\mathcal{D}) \\ &= \int d\mathbf{w}\, \mathcal{N}\left(y'|\mathbf{w}^\top \mathbf{x}', \sigma^2\right) \mathcal{N}\left(\mathbf{w}|\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}\right) \\ &= \mathcal{N}(\mu_{\mathbf{w}}^\top \mathbf{x}', \mathbf{x}'^\top \Sigma_{\mathbf{w}} \mathbf{x}' + \sigma^2). \end{aligned}$$

Alternative View of Linear Regression


$i = 1, \ldots, N$
$\mathbf{x}_i$
$w \sim \mathcal{N}(0, \tau^2 I)$
τ^2
w
y_i
σ^2
$y_i \sim \mathcal{N}(w^\top \mathbf{x}_i, \sigma^2)$

Integrating out **w**, the joint distribution of $y_1, \ldots, y_N$ given $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is Gaussian. The means and covariances are:

$$\begin{aligned}
E[y_i] &= E[\mathbf{w}^\mathsf{T}\mathbf{x}_i] = 0^\mathsf{T}\mathbf{x}_i = 0 \\
E[(y_i - 0)^2] &= E[(\mathbf{x}_i^\mathsf{T}\mathbf{w})(\mathbf{w}^\mathsf{T}\mathbf{x}_i)] + \sigma^2 = \tau^2\mathbf{x}_i^\mathsf{T}\mathbf{x}_i + \sigma^2 \\
E[(y_i - 0)(y_j - 0)] &= E[(\mathbf{x}_i^\mathsf{T}\mathbf{w})(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] = \tau^2\mathbf{x}_i^\mathsf{T}\mathbf{x}_j
\end{aligned}$$

$$\left.\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}\right| \mathbf{x}_1, \ldots, \mathbf{x}_N \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2\mathbf{x}_1^\mathsf{T}\mathbf{x}_1 + \sigma^2 & \tau^2\mathbf{x}_1^\mathsf{T}\mathbf{x}_2 & \cdots & \tau^2\mathbf{x}_1^\mathsf{T}\mathbf{x}_N \\ \tau^2\mathbf{x}_2^\mathsf{T}\mathbf{x}_1 & \tau^2\mathbf{x}_2^\mathsf{T}\mathbf{x}_2 + \sigma^2 & & \tau^2\mathbf{x}_2^\mathsf{T}\mathbf{x}_N \\ \vdots & & \ddots & \vdots \\ \tau^2\mathbf{x}_N^\mathsf{T}\mathbf{x}_1 & \tau^2\mathbf{x}_N^\mathsf{T}\mathbf{x}_2 & \cdots & \tau^2\mathbf{x}_N^\mathsf{T}\mathbf{x}_N + \sigma^2 \end{bmatrix} \right)$$

$$Y^\mathsf{T} | X \sim \mathcal{N}(0_N, \tau^2 X^\mathsf{T} X + \sigma^2 I_N)$$

Alternative View of Linear Regression

Now, include the test input vector $\mathbf{x}'$ and test output y':

$$\begin{bmatrix} Y^\top \\ y' \end{bmatrix} \bigg| X, \mathbf{x}' \sim \mathcal{N}\left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^\top X + \sigma^2 I & \tau^2 X^\top \mathbf{x}' \\ \tau^2 \mathbf{x}'^\top X & \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 \end{bmatrix} \right)$$

Alternative View of Linear Regression

Now, include the test input vector $\mathbf{x}'$ and test output y':

$$\begin{bmatrix} Y^\top \\ y' \end{bmatrix} \Bigg| X, \mathbf{x}' \sim \mathcal{N}\left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^\top X + \sigma^2 I & \tau^2 X^\top \mathbf{x}' \\ \tau^2 \mathbf{x}'^\top X & \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 \end{bmatrix} \right)$$

We can find $y'|Y$ by the standard multivariate Gaussian result:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \right) \quad \Rightarrow \quad \mathbf{b}|\mathbf{a} \sim \mathcal{N}\left(C^\top A^{-1} \mathbf{a}, B - C^\top A^{-1} C \right)$$

Alternative View of Linear Regression

Now, include the test input vector $\mathbf{x}'$ and test output y':

$$\begin{bmatrix} Y^\top \\ y' \end{bmatrix} \Bigg| X, \mathbf{x}' \sim \mathcal{N}\left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^\top X + \sigma^2 I & \tau^2 X^\top \mathbf{x}' \\ \tau^2 \mathbf{x}'^\top X & \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 \end{bmatrix} \right)$$

We can find $y'|Y$ by the standard multivariate Gaussian result:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \right) \quad \Rightarrow \quad \mathbf{b}|\mathbf{a} \sim \mathcal{N}\left(C^\top A^{-1} \mathbf{a}, B - C^\top A^{-1} C \right)$$

So

$$y'|Y, X, \mathbf{x}' \sim \mathcal{N}\left(\tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} Y^\top, \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 - \tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} \tau^2 X^\top \mathbf{x}' \right)$$

Alternative View of Linear Regression

Now, include the test input vector $\mathbf{x}'$ and test output y':

$$\left.\begin{bmatrix} Y^\top \\ y' \end{bmatrix}\right| X, \mathbf{x}' \sim \mathcal{N}\left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^\top X + \sigma^2 I & \tau^2 X^\top \mathbf{x}' \\ \tau^2 \mathbf{x}'^\top X & \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 \end{bmatrix}\right)$$

We can find $y'|Y$ by the standard multivariate Gaussian result:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}\right) \quad \Rightarrow \quad \mathbf{b}|\mathbf{a} \sim \mathcal{N}\left(C^\top A^{-1} \mathbf{a}, B - C^\top A^{-1} C\right)$$

So

$$\begin{aligned} y'|Y, X, \mathbf{x}' &\sim \mathcal{N}\Big(\tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} Y^\top, \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 - \tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} \tau^2 X^\top \mathbf{x}' \\ &\sim \mathcal{N}\left(\tfrac{1}{\sigma^2} \mathbf{x}'^\top \Sigma X Y^\top, \mathbf{x}'^\top \Sigma \mathbf{x}' + \sigma^2\right) \qquad \Sigma = \left(\tfrac{1}{\sigma^2} X X^\top + \tfrac{1}{\tau^2} I\right)^{-1} \end{aligned}$$

Same answer as when we integrated posterior over $\mathbf{w}$ to obtain predictive distribution over y'.

Alternative View of Linear Regression

Now, include the test input vector $\mathbf{x}'$ and test output y':

$$\begin{bmatrix} Y^\top \\ y' \end{bmatrix} \Bigg| X, \mathbf{x}' \sim \mathcal{N}\left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^\top X + \sigma^2 I & \tau^2 X^\top \mathbf{x}' \\ \tau^2 \mathbf{x}'^\top X & \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 \end{bmatrix} \right)$$

We can find $y'|Y$ by the standard multivariate Gaussian result:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \right) \quad \Rightarrow \quad \mathbf{b}|\mathbf{a} \sim \mathcal{N}\left(C^\top A^{-1} \mathbf{a}, B - C^\top A^{-1} C \right)$$

So

$$\begin{aligned} y'|Y, X, \mathbf{x}' &\sim \mathcal{N}\left(\tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} Y^\top, \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 - \tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} \tau^2 X^\top \mathbf{x}' \right) \\ &\sim \mathcal{N}\left(\tfrac{1}{\sigma^2} \mathbf{x}'^\top \Sigma X Y^\top, \mathbf{x}'^\top \Sigma \mathbf{x}' + \sigma^2 \right) \qquad \Sigma = \left(\tfrac{1}{\sigma^2} X X^\top + \tfrac{1}{\tau^2} I \right)^{-1} \end{aligned}$$

Same answer as when we integrated posterior over $\mathbf{w}$ to obtain predictive distribution over y'.

Similarly, evidence $P(Y|X)$ is just probability under Gaussian, and reduces to previous expression.

Alternative View of Linear Regression

Now, include the test input vector $\mathbf{x}'$ and test output y':

$$\begin{bmatrix} Y^{\top} \\ y' \end{bmatrix} \Bigg| X, \mathbf{x}' \sim \mathcal{N}\left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^{\top} X + \sigma^2 I & \tau^2 X^{\top} \mathbf{x}' \\ \tau^2 \mathbf{x}'^{\top} X & \tau^2 \mathbf{x}'^{\top} \mathbf{x}' + \sigma^2 \end{bmatrix}\right)$$

We can find $y'|Y$ by the standard multivariate Gaussian result:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} A & C \\ C^{\top} & B \end{bmatrix}\right) \quad \Rightarrow \quad \mathbf{b}|\mathbf{a} \sim \mathcal{N}\left(C^{\top} A^{-1} \mathbf{a}, B - C^{\top} A^{-1} C\right)$$

So

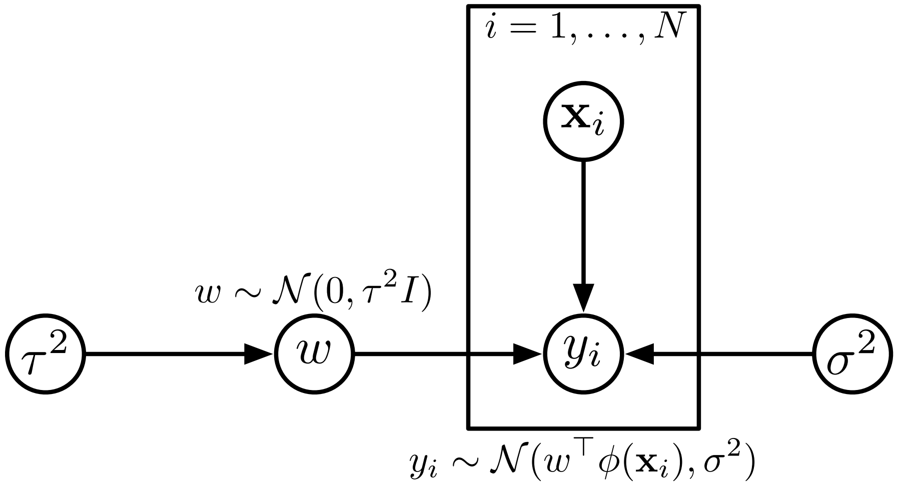
$$\begin{aligned} y'|Y, X, \mathbf{x}' &\sim \mathcal{N}\left(\tau^2 \mathbf{x}'^{\top} X (\tau^2 X^{\top} X + \sigma^2 I)^{-1} Y^{\top}, \tau^2 \mathbf{x}'^{\top} \mathbf{x}' + \sigma^2 - \tau^2 \mathbf{x}'^{\top} X (\tau^2 X^{\top} X + \sigma^2 I)^{-1} \tau^2 X^{\top} \mathbf{x}'\right. \\ &\sim \mathcal{N}\left(\tfrac{1}{\sigma^2} \mathbf{x}'^{\top} \Sigma X Y^{\top}, \mathbf{x}'^{\top} \Sigma \mathbf{x}' + \sigma^2\right) \qquad \Sigma = \left(\tfrac{1}{\sigma^2} X X^{\top} + \tfrac{1}{\tau^2} I\right)^{-1} \end{aligned}$$

Same answer as when we integrated posterior over $\mathbf{w}$ to obtain predictive distribution over y'.

Similarly, evidence $P(Y|X)$ is just probability under Gaussian, and reduces to previous expression.

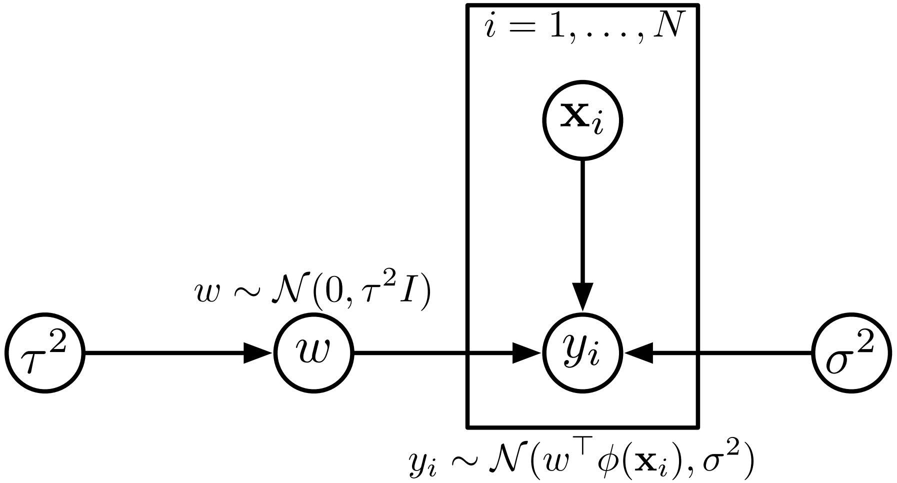
The point: Bayesian regression can be derived from a joint, parameter-free distribution on the outputs conditioned on the inputs.

Nonlinear Regression


$i = 1, \ldots, N$
$\mathbf{x}_i$
$w \sim \mathcal{N}(0, \tau^2 I)$
τ^2
w
y_i
σ^2
$y_i \sim \mathcal{N}(w^\top \phi(\mathbf{x}_i), \sigma^2)$

What if we introduce a nonlinear mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$? Each element of $\phi(\mathbf{x})$ is a (nonlinear) feature extracted from **x**. May be many more features than elements on **x**.

Nonlinear Regression


$i = 1, \ldots, N$
$\mathbf{x}_i$
$w \sim \mathcal{N}(0, \tau^2 I)$
τ^2
w
y_i
σ^2
$y_i \sim \mathcal{N}(w^\top \phi(\mathbf{x}_i), \sigma^2)$

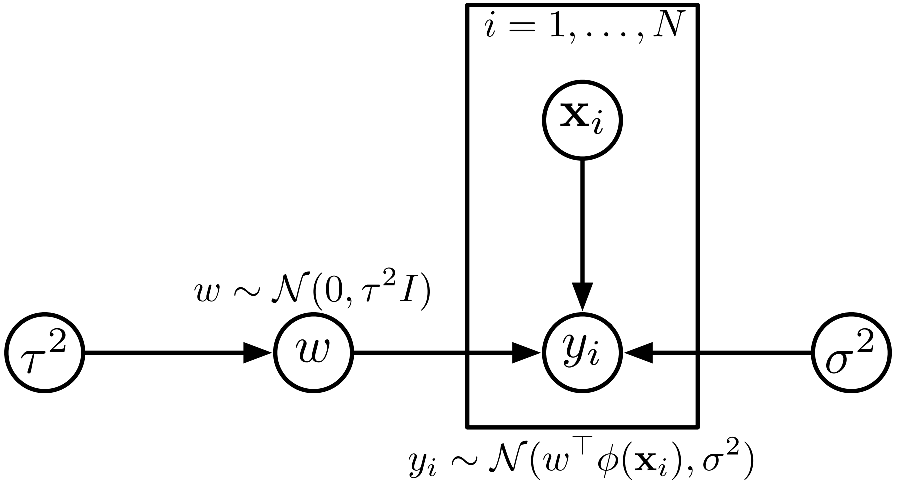
What if we introduce a nonlinear mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$? Each element of $\phi(\mathbf{x})$ is a (nonlinear) feature extracted from $\mathbf{x}$. May be many more features than elements on $\mathbf{x}$.

The regression function $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\phi(\mathbf{x})$ is nonlinear, but outputs Y still jointly Gaussian!

$$Y^\mathsf{T}|X \sim \mathcal{N}(0_N, \tau^2\Phi^\mathsf{T}\Phi + \sigma^2 I_N)$$

where the i^{th} column of matrix Φ is $\phi(\mathbf{x}_i)$.

Nonlinear Regression



$i = 1, \ldots, N$
$\mathbf{x}_i$
$w \sim \mathcal{N}(0, \tau^2 I)$
τ^2
w
y_i
σ^2
$y_i \sim \mathcal{N}(w^\top \phi(\mathbf{x}_i), \sigma^2)$

What if we introduce a nonlinear mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$? Each element of $\phi(\mathbf{x})$ is a (nonlinear) feature extracted from $\mathbf{x}$. May be many more features than elements on $\mathbf{x}$.

The regression function $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\phi(\mathbf{x})$ is nonlinear, but outputs Y still jointly Gaussian!

$$Y^\mathsf{T}|X \sim \mathcal{N}(0_N, \tau^2\Phi^\mathsf{T}\Phi + \sigma^2 I_N)$$

where the i^{th} column of matrix Φ is $\phi(\mathbf{x}_i)$.

Proceeding as before, the predictive distribution over y' on a test input $\mathbf{x}'$ is:

$$y'|Y, X, \mathbf{x}' \sim \mathcal{N}\left(\tau^2\phi(\mathbf{x}')^\mathsf{T}\Phi K^{-1}Y^\mathsf{T}, \tau^2\phi(\mathbf{x}')^\mathsf{T}\phi(\mathbf{x}') + \sigma^2 - \tau^4\phi(\mathbf{x})^\mathsf{T}\Phi K^{-1}\Phi^\mathsf{T}\phi(\mathbf{x}')\right)$$

$$K = \tau^2\Phi^\mathsf{T}\Phi + \sigma^2 I$$

The Covariance Kernel

$$Y^{\mathsf{T}}|X \sim \mathcal{N}\left(\mathbf{0}_N, \tau^2\Phi^{\mathsf{T}}\Phi + \sigma^2 I_N\right)$$

The covariance of the output vector Y plays a central role in the development of the theory of Gaussian processes.

The Covariance Kernel

$$Y^{\mathsf{T}} | X \sim \mathcal{N}\left(\mathbf{0}_N, \tau^2 \Phi^{\mathsf{T}} \Phi + \sigma^2 I_N\right)$$

The covariance of the output vector Y plays a central role in the development of the theory of Gaussian processes.

Define the covariance kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that if $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$ are two input vectors with corresponding outputs y, y', then

$$K(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[y, y'] = E[yy'] - E[y]E[y']$$

In the nonlinear regression example we have $K(\mathbf{x}, \mathbf{x}') = \tau^2 \phi(\mathbf{x})^{\mathsf{T}} \phi(\mathbf{x}') + \sigma^2 \delta_{\mathbf{x}=\mathbf{x}'}$.

The Covariance Kernel

$$Y^{\mathsf{T}} | X \sim \mathcal{N}\left(\mathbf{0}_N, \tau^2 \Phi^{\mathsf{T}} \Phi + \sigma^2 I_N\right)$$

The covariance of the output vector Y plays a central role in the development of the theory of Gaussian processes.

Define the covariance kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that if $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$ are two input vectors with corresponding outputs y, y', then

$$K(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[y, y'] = E[yy'] - E[y]E[y']$$

In the nonlinear regression example we have $K(\mathbf{x}, \mathbf{x}') = \tau^2 \phi(\mathbf{x})^{\mathsf{T}} \phi(\mathbf{x}') + \sigma^2 \delta_{\mathbf{x}=\mathbf{x}'}$.

The covariance kernel has two properties:

- Symmetric: $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}'$.
- Positive semidefinite: the matrix $[K(\mathbf{x}_i, \mathbf{x}_j)]$ formed by any finite set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is positive semidefinite.

The Covariance Kernel

$$Y^{\mathsf{T}}|X \sim \mathcal{N}\left(\mathbf{0}_N, \tau^2\Phi^{\mathsf{T}}\Phi + \sigma^2 I_N\right)$$

The covariance of the output vector Y plays a central role in the development of the theory of Gaussian processes.

Define the covariance kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that if $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$ are two input vectors with corresponding outputs y, y', then

$$K(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[y, y'] = E[yy'] - E[y]E[y']$$

In the nonlinear regression example we have $K(\mathbf{x}, \mathbf{x}') = \tau^2\phi(\mathbf{x})^{\mathsf{T}}\phi(\mathbf{x}') + \sigma^2\delta_{\mathbf{x}=\mathbf{x}'}$.

The covariance kernel has two properties:

- Symmetric: $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}'$.
- Positive semidefinite: the matrix $[K(\mathbf{x}_i, \mathbf{x}_j)]$ formed by any finite set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is positive semidefinite.

Theorem: A covariance kernel $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is symmetric and positive semidefinite if and only if there is a feature map $\phi : \mathbb{X} \to \mathbb{H}$ such that

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathsf{T}}\phi(\mathbf{x}')$$

The feature space $\mathbb{H}$ can potentially be infinite dimensional.

Gaussian Process Regression

For non-linear regression, all operations depended on $K(\mathbf{x}, \mathbf{x}')$ rather than explicitly on $\phi(\mathbf{x})$.

So we can define the joint in terms of K *implicitly* using a (potentially infinite-dimensional) feature map $\phi(\mathbf{x})$.

$$Y|X, K \sim \mathcal{N}(0_N, K(X, X))$$

where the i, j entry in the covariance matrix $K(X, X)$ is $K(\mathbf{x}_i, \mathbf{x}_j)$.

This is called the kernel trick.

Prediction: compute the predictive distribution of y' conditioned on Y:

$$y'|\mathbf{x}', X, Y, K \sim \mathcal{N}(\underbrace{K(\mathbf{x}', X)K(X, X)^{-1}Y^{\mathsf{T}}}_{\text{mean}}, \underbrace{K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}', X)K(X, X)^{-1}K(X, \mathbf{x}')}_{\text{variance}})$$

Evidence: this is just the Gaussian likelihood:

$$P(Y|X, K) = |2\pi K(X, X)|^{-\frac{1}{2}} e^{-\frac{1}{2} Y K(X,X)^{-1} Y^{\mathsf{T}}}$$

Evidence optimisation: the covariance kernel K often has parameters, and these can be optimized by gradient ascent in $\log P(Y|X, K)$.

The Gaussian Process

A **Gaussian process** (GP) is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.

In our regression setting, corresponding to each input vector $\mathbf{x}$ we have an output $f(\mathbf{x})$. Given $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, the joint distribution of the outputs $F = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]$ is:

$$F|X, K \sim \mathcal{N}(0, K(X, X))$$

Thus the random function $f(\mathbf{x})$ (as a collection of random variables, one $f(\mathbf{x})$ for each $\mathbf{x}$) is a Gaussian process.

In general, a Gaussian process is parametrized by a mean function $m(\mathbf{x})$ and covariance kernel $K(\mathbf{x}, \mathbf{x}')$, and we write

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Posterior Gaussian process: on observing X and F, the conditional joint distribution of $F' = [f(\mathbf{x}'_1), \ldots, f(\mathbf{x}'_M)]$ on another set of input vectors $\mathbf{x}'_1, \ldots, \mathbf{x}'_M$ is still Gaussian:

$$F'|X', X, F, K \sim \mathcal{N}(K(X', X)K(X, X)^{-1}F^{\mathsf{T}}, K(X', X') - K(X', X)K(X, X)^{-1}K(X, X'))$$

thus the posterior over functions $f(\cdot)|X, F$ is still a Gaussian process!

Regression with Gaussian Processes

We wish to model the joint distribution of outputs $y_1, \ldots, y_N$ given inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$.
Use a GP prior over functions:

$$f(\cdot) \sim \mathcal{GP}(0, K(\cdot, \cdot))$$

Usually, instead of treating y_i as direct observation of the function value $f(\mathbf{x}_i)$, we add Gaussian observation noise:

$$y_i | \mathbf{x}_i, f(\cdot) \sim \mathcal{N}(f(\mathbf{x}_i), \sigma^2)$$

Evidence: again this is just a multivariate Gaussian likelihood,

$$P(Y|X) = |2\pi(K(X,X) + \sigma^2 I)|^{-\frac{1}{2}} e^{-\frac{1}{2} Y (K(X,X)+\sigma^2 I)^{-1} Y^\mathsf{T}}$$

Posterior: the posterior function is still a GP,

$$f(\cdot)|X, Y \sim \mathcal{GP}(K(\cdot, X)(K(X,X) + \sigma^2 I)^{-1} Y^\mathsf{T}, K(\cdot, \cdot) - K(\cdot, X)(K(X,X) + \sigma^2 I)^{-1} K(X, \cdot))$$

Prediction: the predictive distribution is just posterior plus observation noise:

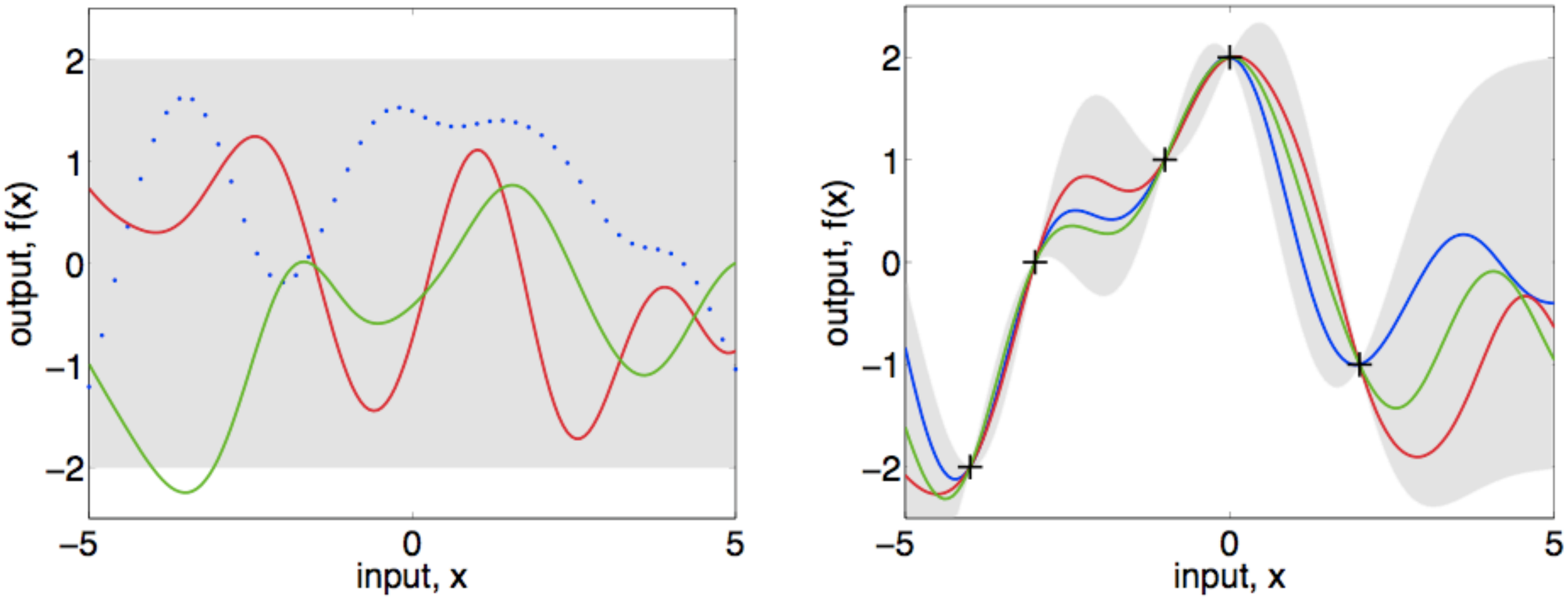
$$y' | X, Y, \mathbf{x}' \sim \mathcal{N}(E[f(\mathbf{x}')|X, Y], \mathrm{Var}[f(\mathbf{x}')|X, Y] + \sigma^2)$$

Evidence Optimisation: we can do this by gradient ascent in $\log P(Y|X)$.

Samples from a Gaussian Process

We can draw sample functions from a GP by fixing a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and drawing a sample $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)$ from the corresponding multivariate Gaussian. This can then be plotted.

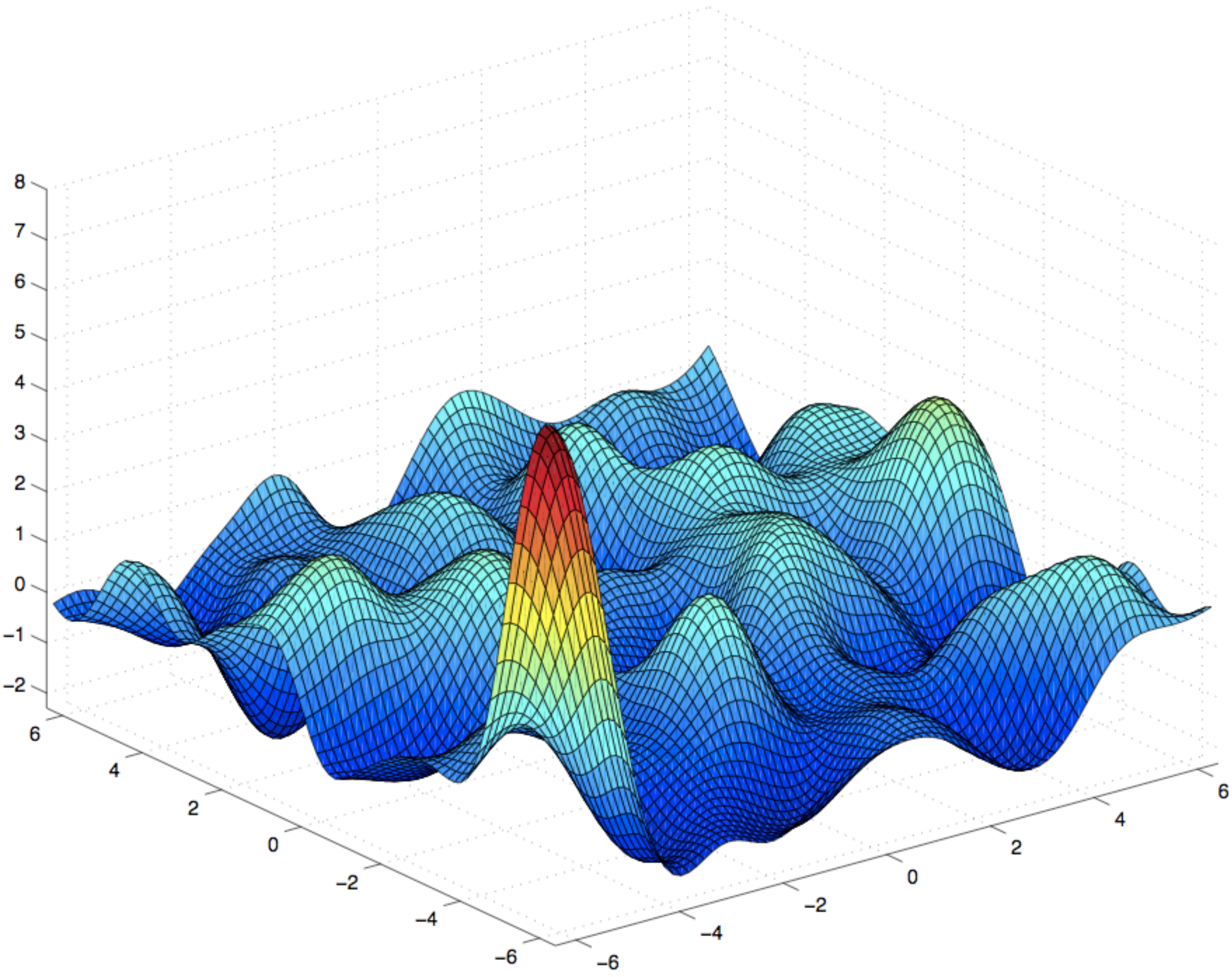
Below we plot samples from an example prior and corresponding posterior GP.



Another approach is to

- sample $f(\mathbf{x}_1)$ first,
- then $f(\mathbf{x}_2)|f(\mathbf{x}_1)$,
- and generally $f(\mathbf{x}_n)|f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{n-1})$ for $n = 1, 2, \ldots$.

Sample from a 2D Gaussian Process



Covariance Kernels

Examples of covariance kernels:

- Polynomial:

$$K(\mathbf{x},\mathbf{x}') = (1+\mathbf{x}^{\mathsf{T}}\mathbf{x}')^m \qquad\qquad m = 1, 2, \ldots$$

- Squared-exponential:

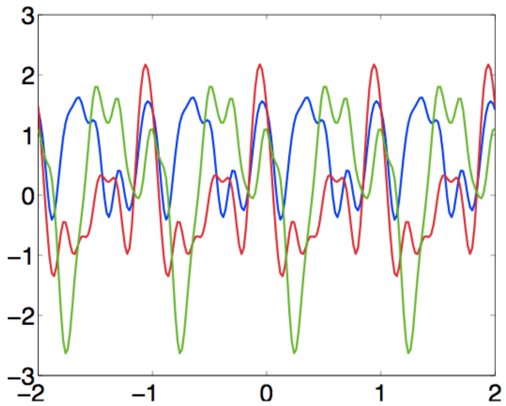
$$K(\mathbf{x},\mathbf{x}') = \theta^2 e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\eta^2}}$$

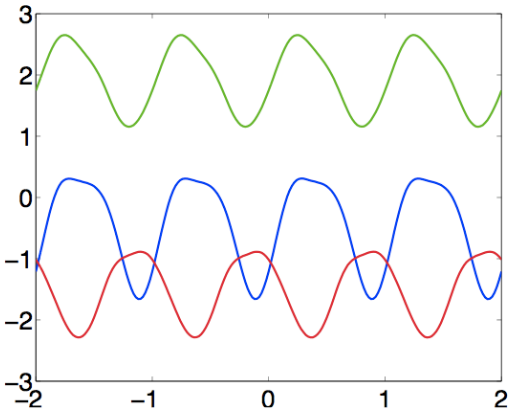
- Periodic:

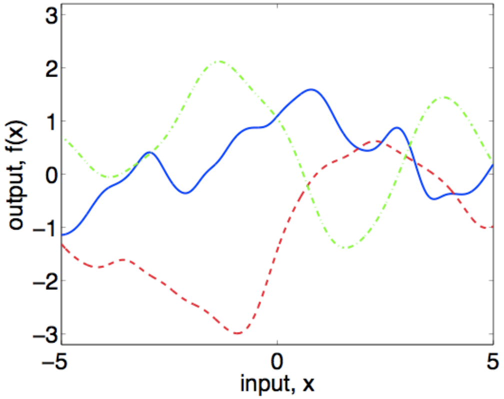
$$K(x,x') = \theta^2 e^{-\frac{2\sin^2(\pi(x-x')/\tau)}{\eta^2}}$$

- Rational Quadratic:

$$K(\mathbf{x},\mathbf{x}') = \left(1+\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\alpha\eta^2}\right)^{-\alpha} \qquad\qquad \alpha > 0$$







Covariance Kernels

If K_1 and K_2 are covariance kernels, then so are:

- Rescaling: αK_1 for $\alpha > 0$.
- Addition: $K_1 + K_2$
- Elementwise product: $K_1 K_2$
- Mapping: $K_1(\phi(\mathbf{x}), \phi(\mathbf{x}'))$ for some function ϕ.

We say a covariance kernel is translation-invariant if

$$K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$$

A GP with a translation-invariant covariance kernel is stationary: if $f(\cdot) \sim \mathcal{GP}(0, K)$, then so is $f(\cdot - \mathbf{x}) \sim \mathcal{GP}(0, K)$ for each $\mathbf{x}$.

We say a covariance kernel is radial if

$$K(\mathbf{x}, \mathbf{x}') = h(\|\mathbf{x} - \mathbf{x}'\|)$$

A GP with a radial covariance kernel is stationary with respect to translations, rotations, and reflections of the input space.

Nonparametric Bayesian Models and Occam's Razor

Overparameterised models can overfit.

But the Bayesian treatment integrates parameters out, so they cannot be adjusted to overfit the data! In the GP, the parameter is the function $f(\mathbf{x})$ which can be infinite-dimensional.

The Gaussian process is an example of a larger class of **nonparametric Bayesian models**.

- Infinite number of parameters.
- Often constructed as the infinite limit of a nested family of finite models (sometimes equivalent to infinite model averaging).
- Parameters integrated out, so effective number of parameters to overfit is zero or small (hyperparameters).
- No need for model selection. Bayesian posterior on parameters will concentrate on "submodel" with largest integral automatically.
- No explicit need for Occam's razor, validation or added regularisation penalty.

End Notes

Automatic relevance determination appeared in MacKay (1993) Bayesian Methods for Backpropagation Networks and Neal (1993) Bayesian Learning for Neural Networks. Gaussian processes can also be used in classification and latent variable models. We will consider classification in the second half of course.

Many of the figures have been copied from a Gaussian process tutorial by Carl Rasmussen (MLSS 2007) at http://agbs.kyb.tuebingen.mpg.de/wikis/mlss07/CarlERasmussen

An excellent text book on Gaussian processes is Gaussian processes for Machine Learning by Rasmussen and Williams, available online at http://www.gaussianprocess.org/gpml/

End Notes