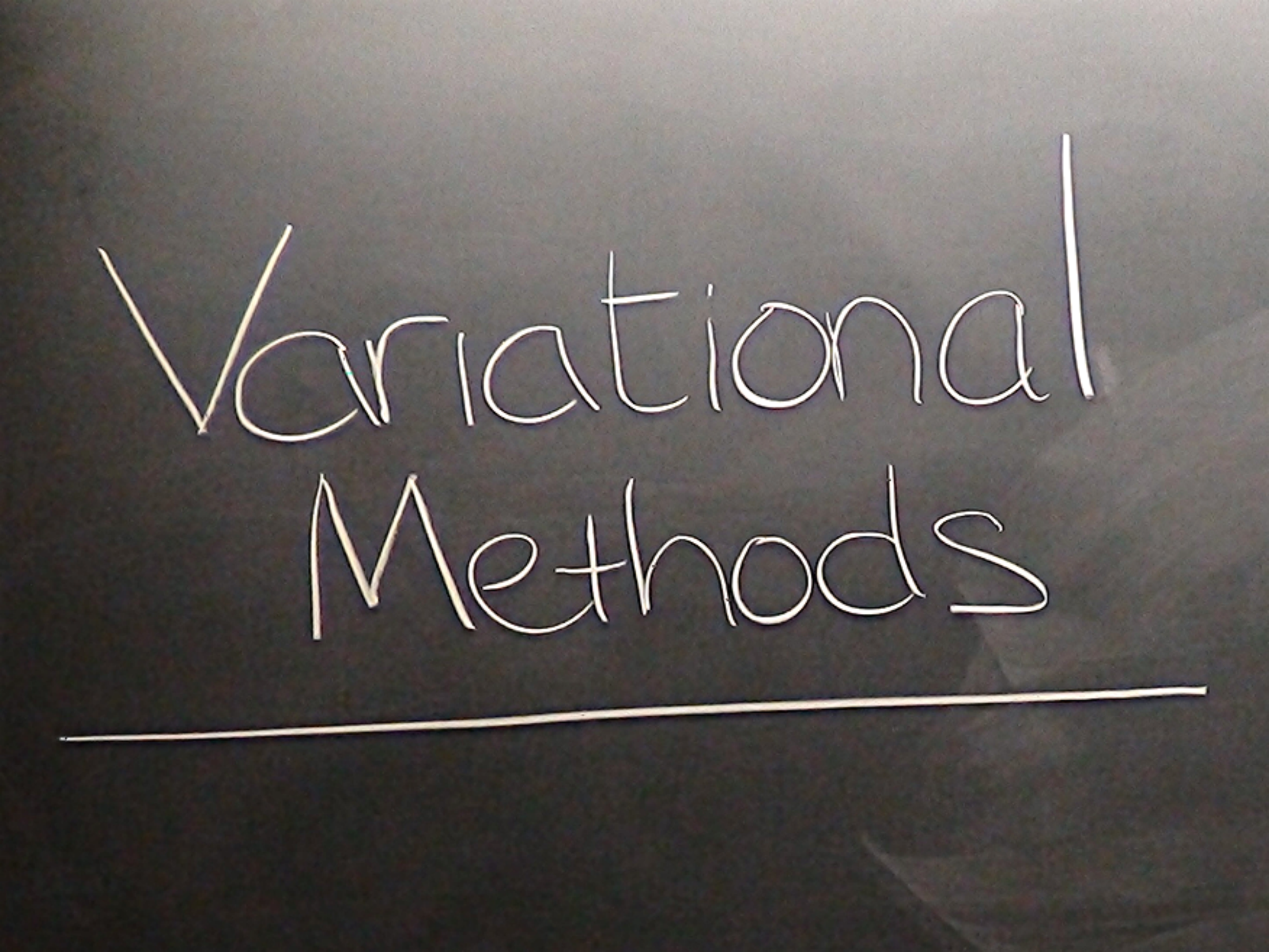
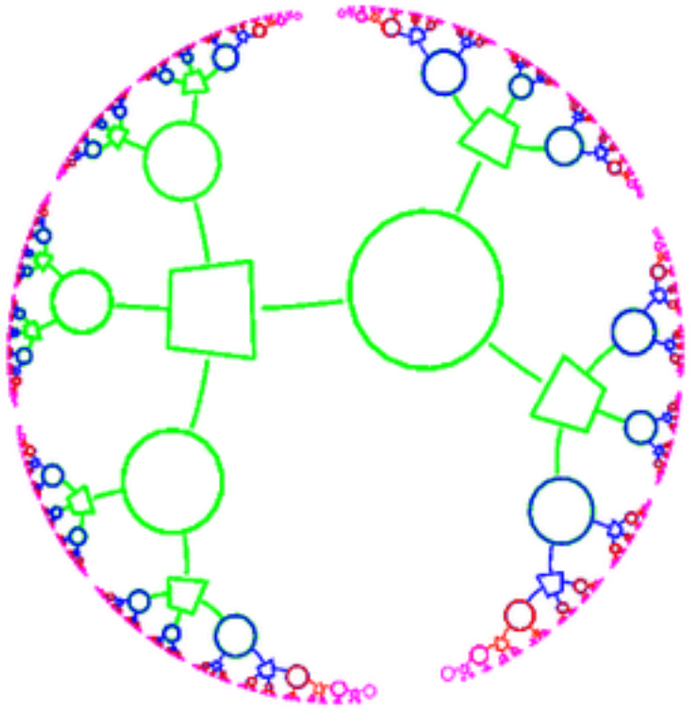

Variational
Methods



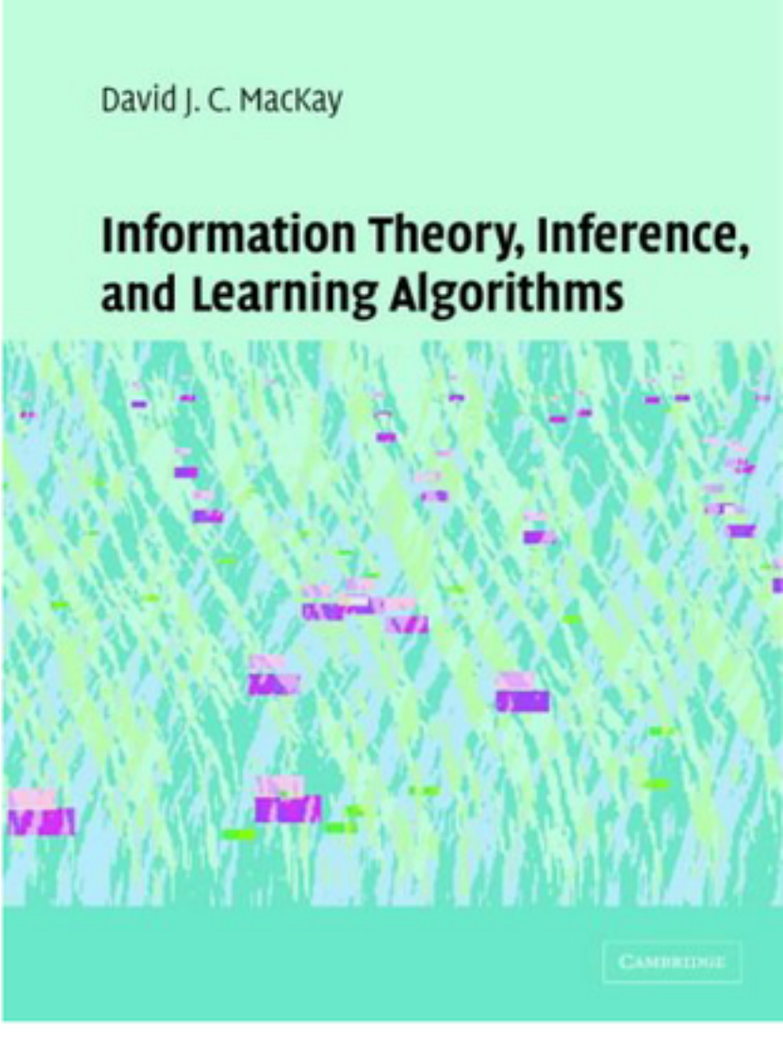
Information theory, pattern recognition, and neural networks

- Source coding (Data compression)
- Noisy-channel coding
- Inference + probabilistic methods
  - 9-10 Inference
  - 11 Clustering
  - 12 Monte Carlo methods
  - 13 Advanced Monte Carlo methods
  - 14 Variational methods
- Neural networks
- State-of-the-art error-correcting codes

www.inference.phy.cam.ac.uk/itprnn/
www.inference.phy.cam.ac.uk/itila/

Overview

- Data compression
- Noisy-channel coding
  - Chs 1-6, 8-10, 14
- Inference, data modelling
  - clustering, pattern recognition
    - Chs 20, 22
- Probability toolbox
  - Monte Carlo methods
    - Ch 29
  - Variational methods
    - Ch 33
- Neural networks
  - Chs 38, 39, (& perhaps 41, 44), 42
- State-of-the-art error-correcting codes

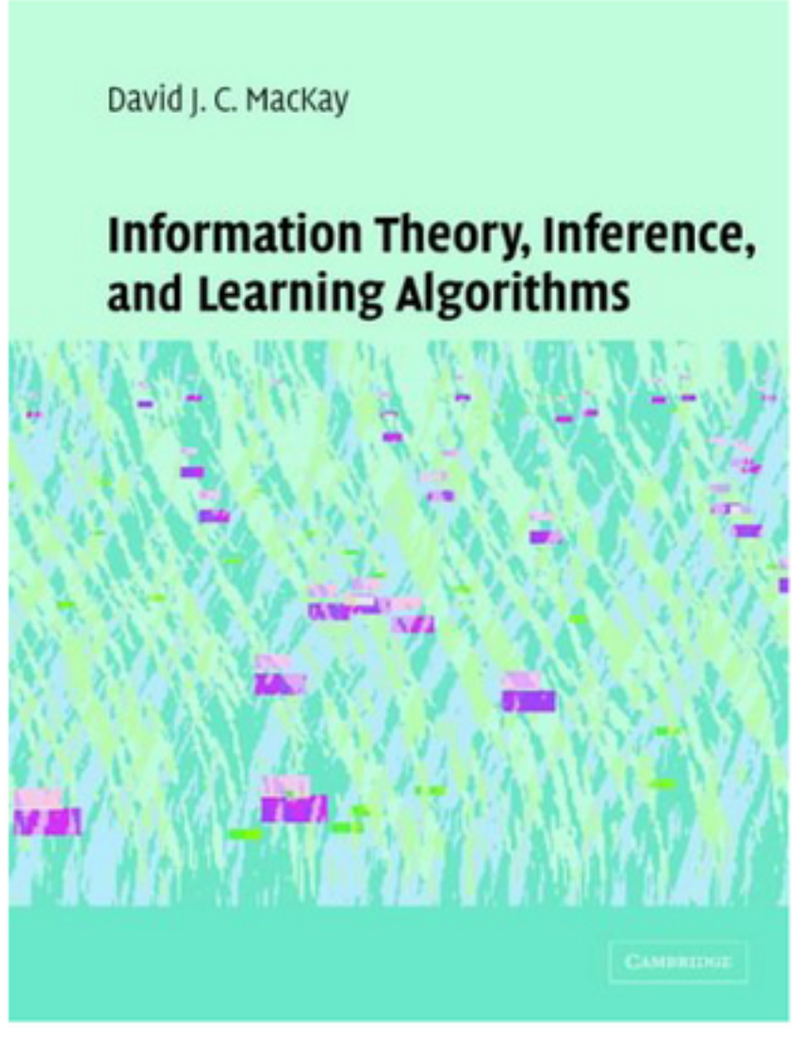


Additional reading

- Laplace's method (Ch 27)
- Ising models (Ch 31)

The course

www.inference.phy.cam.ac.uk/itprnn/



The book

www.inference.phy.cam.ac.uk/itila/

Variational methods

Interested in $P(\mathbf{x}) = \frac{1}{Z} P^*(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})}$.

$E(\mathbf{x})$ is simple, but not simple enough.

Idea Approximate $P(\mathbf{x})$ by a simpler distribution $Q(\mathbf{x}; \theta)$.

Adjust θ to get the 'best' approximation.

Then approximate $\sum_{\mathbf{x}} \phi(\mathbf{x}) P(\mathbf{x})$ by $\sum_{\mathbf{x}} \phi(\mathbf{x}) Q(\mathbf{x}; \theta^*)$

How to measure 'best'? Possible ideas:

$$D_{\mathrm{KL}}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

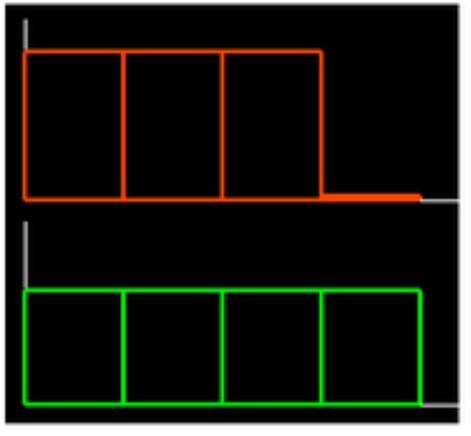
$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Distances between probability distributions

Q:

If $P(x) = \{1/3, 1/3, 1/3, \epsilon\}$

and $Q(x) = \{1/4, 1/4, 1/4, 1/4\}$,



which is bigger:

$$D_{\mathrm{KL}}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

or

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

?

Example $\sum P \log \frac{P}{Q} \approx$

$\sum Q \log \frac{Q}{P} \approx$

$D(Q\|P) \quad > \checkmark \quad D(P\|Q)$

$<$

$$\sum P \log \frac{P}{Q} \simeq \log \frac{4}{3}$$

$$\sum Q \log \frac{Q}{P} \simeq \frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4\varepsilon}$$

Bad news

$$\simeq \log \frac{4}{3}$$

BIG

$$\frac{Q}{P} \simeq \frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4\varepsilon}$$

= Bad news"

$\log \frac{4}{3}$

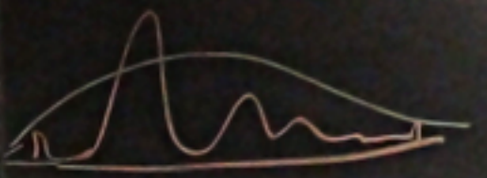
if $\sum Q \log \frac{Q}{P}$

BIG

$\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4\varepsilon}$

$\approx$

= Bad news"

$\sum P \log \frac{P}{Q} \rightarrow$ 

$\log \frac{4}{3}$

if $\sum Q \log \frac{Q}{P}$

BIG

$\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4\varepsilon}$

= Bad news"

$$\sum_X P \log \frac{P}{Q}$$

not on

Variational methods

Interested in $P(\mathbf{x}) = \frac{1}{Z} P^*(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})}$.

$E(\mathbf{x})$ is simple, but not simple enough.

Idea Approximate $P(\mathbf{x})$ by a simpler distribution $Q(\mathbf{x}; \theta)$.

Adjust θ to get the 'best' approximation.

Then approximate $\sum_{\mathbf{x}} \phi(\mathbf{x}) P(\mathbf{x})$ by $\sum_{\mathbf{x}} \phi(\mathbf{x}) Q(\mathbf{x}; \theta^*)$

- Objective function: Variational free energy

$$\tilde{F}(\theta) = \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) E(\mathbf{x}) - \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) \ln \frac{1}{Q(\mathbf{x}; \theta)}$$

$\tilde{F}(\theta)$ is lower-bounded by $-\log Z$

$$\tilde{F}(\theta) = D_{\mathrm{KL}}(Q||P) - \log Z$$

$$-\log Z$$

$$\sum_x P \log \frac{P}{Q} \quad \text{not on} \quad P(x) = \frac{e^{-E(x)}}{Z}$$

$$D_{KL}(Q\|P) = \sum_x Q \log \frac{Q}{P(x)} = \sum_x Q(x) E(x)$$

$$P(x) = \frac{e^{-E(x)}}{Z}$$

$$= \sum_x Q(x) E(x) + \log_e Z - H_Q(x)$$

$\sum_x$ ×

$Q(x) E(x)$ ✓

$\log_e Z$ × doesn't matter

$H_Q(x)$ ✓

$$\tilde{F}(\theta) = \sum_x q_\theta(x) E(x) - \sum_x q_\theta(x) \ln \frac{1}{q_\theta(x)}$$

not on

$$P(x) = \frac{e^{-E(x)}}{Z}$$

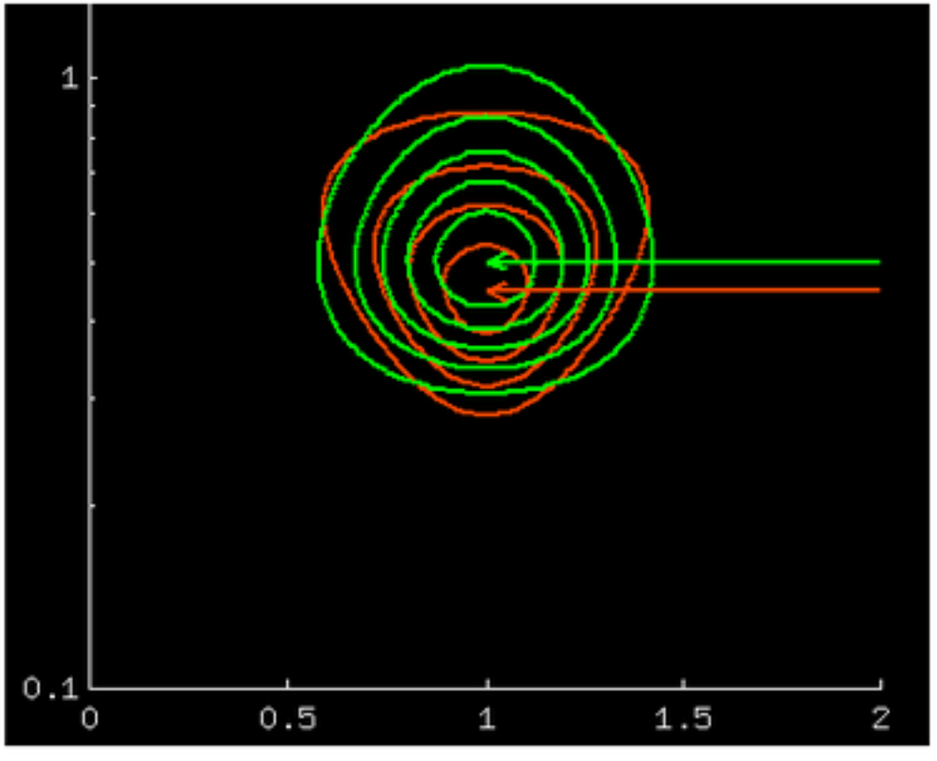
$$\underbrace{\sum_x Q \log \frac{Q}{P(x)}}_{\geq 0} = \left(\sum_x Q(x) E(x) \right) \checkmark$$

$$\tilde{F}(\theta) = \sum_x q_\theta(x) E(x) - \sum_x q_\theta(x) \ln \frac{1}{q_\theta(x)}$$

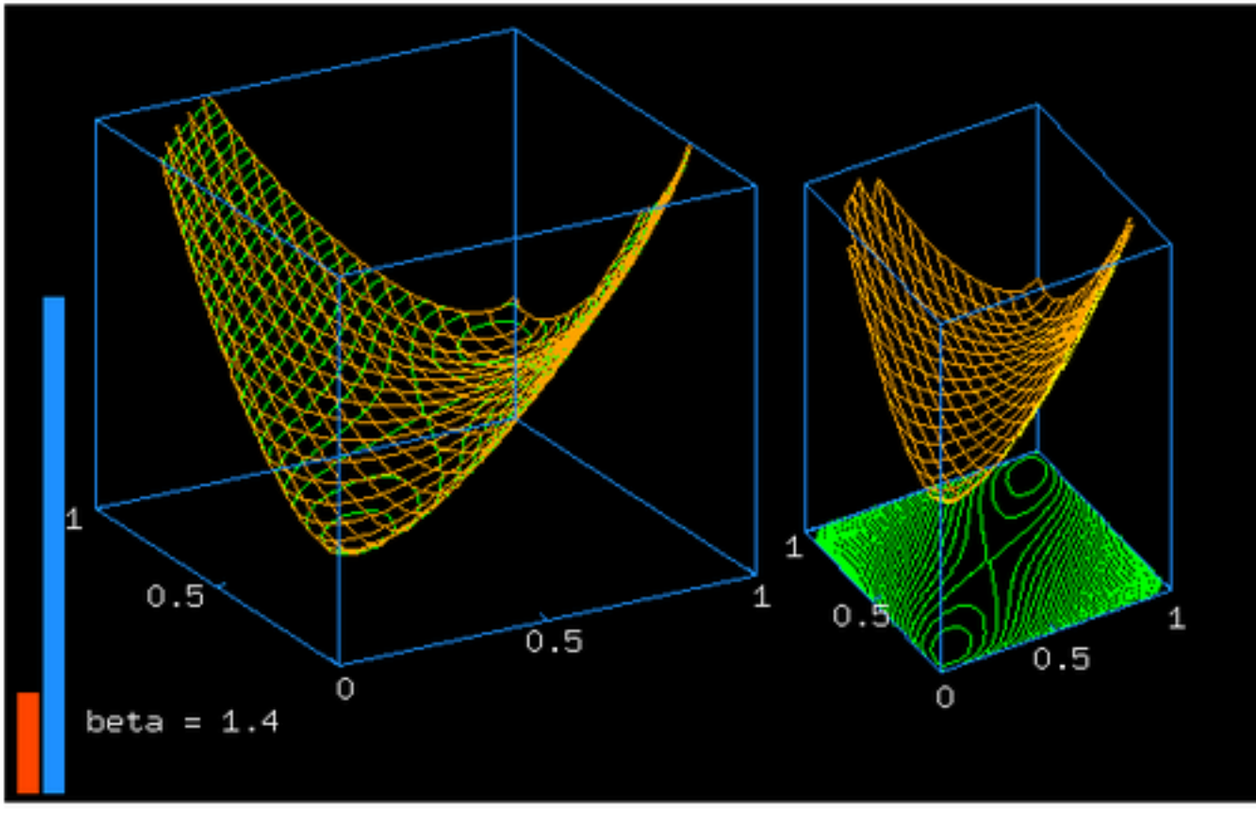
variational free energy

$$\tilde{F}(\theta) \geq -\log Z$$

Examples

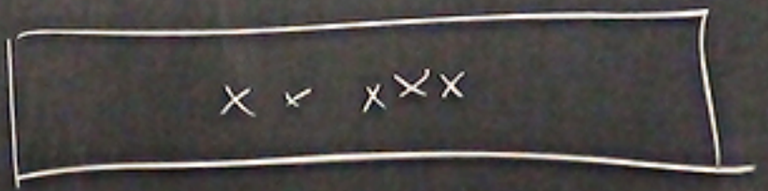


Inferring mu and sigma

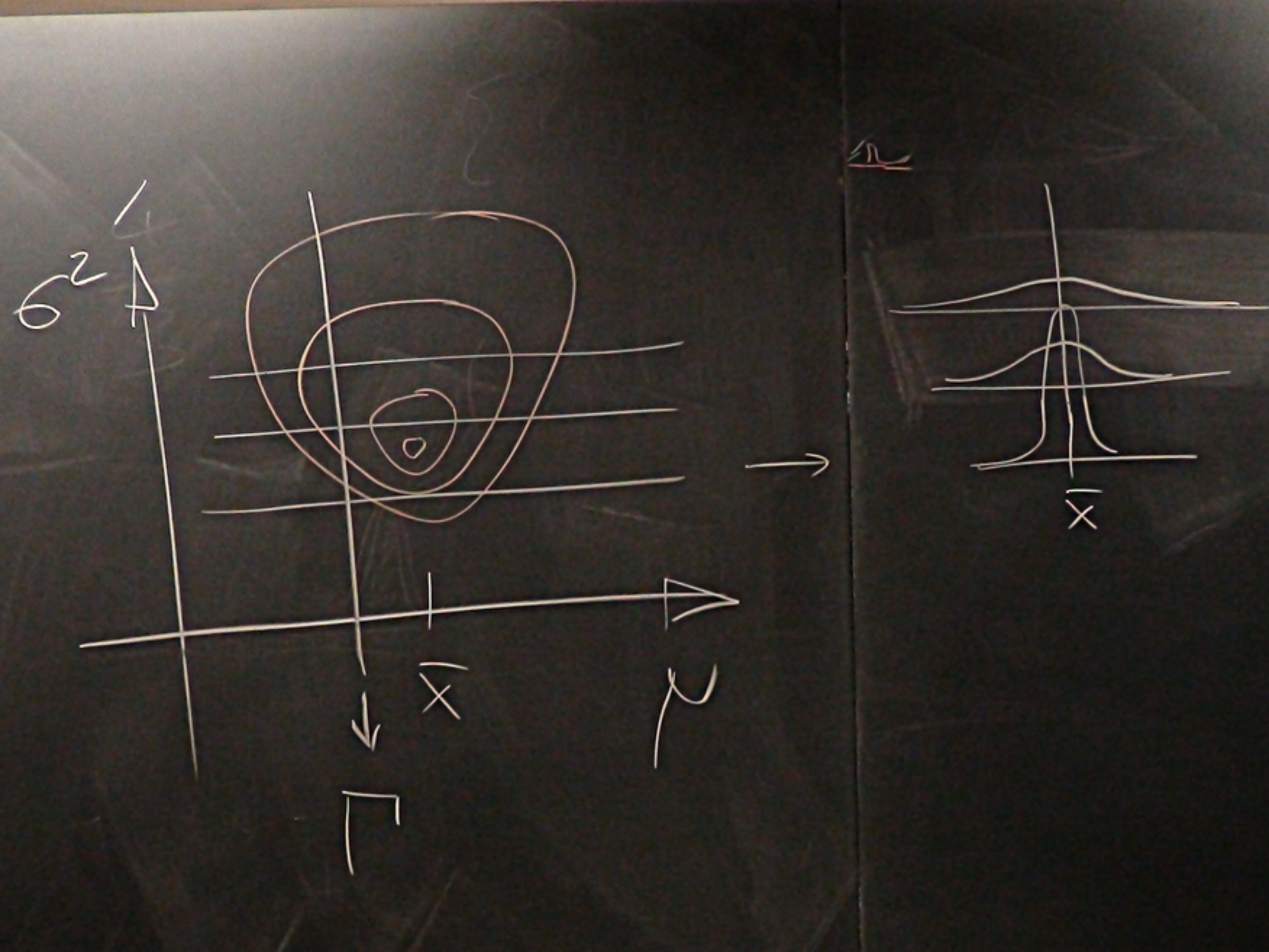


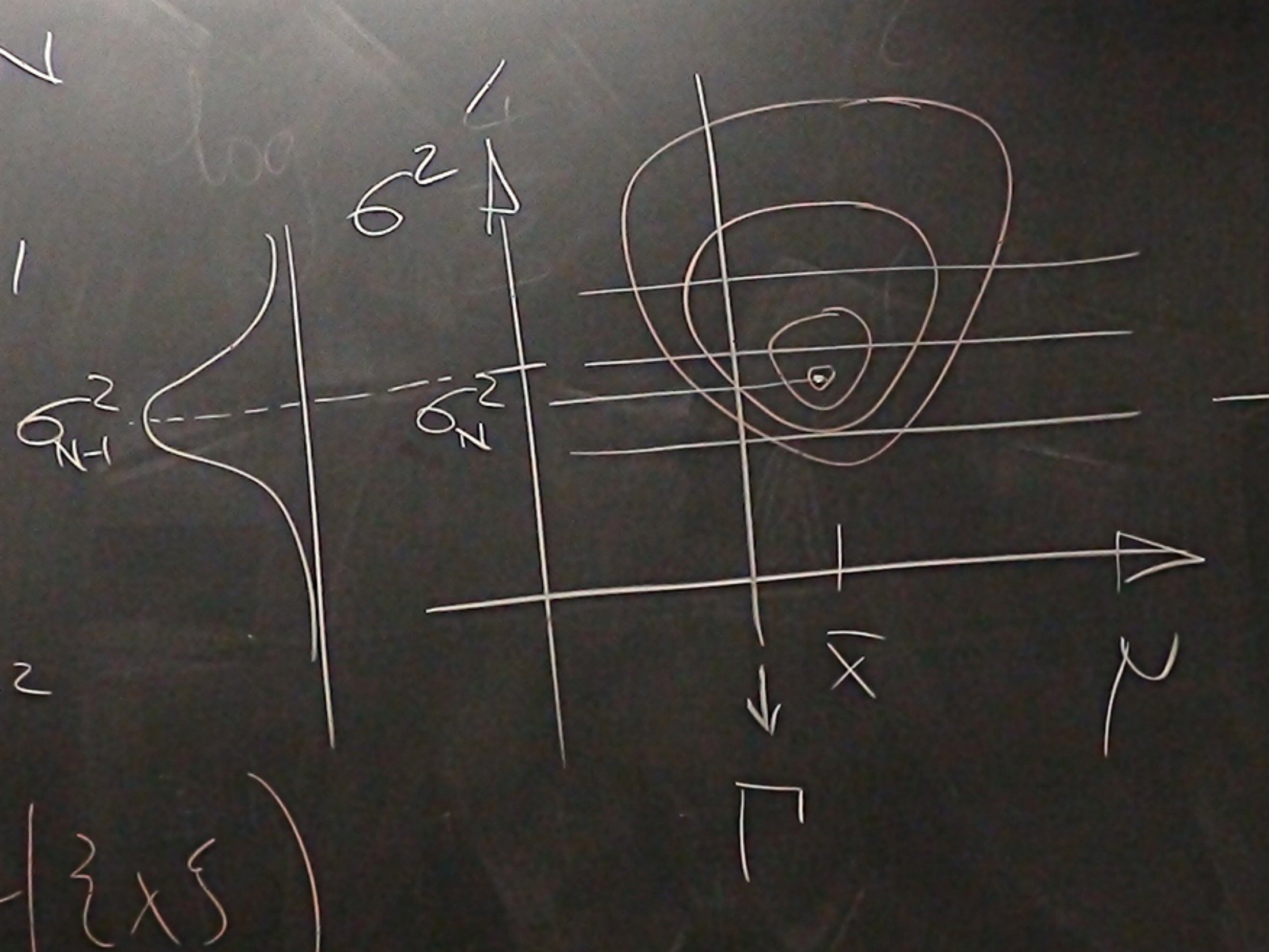
Two coupled spins

$N(\mu, \sigma^2)$ $\longrightarrow$ $\{x_1, \ldots, x_N\}$

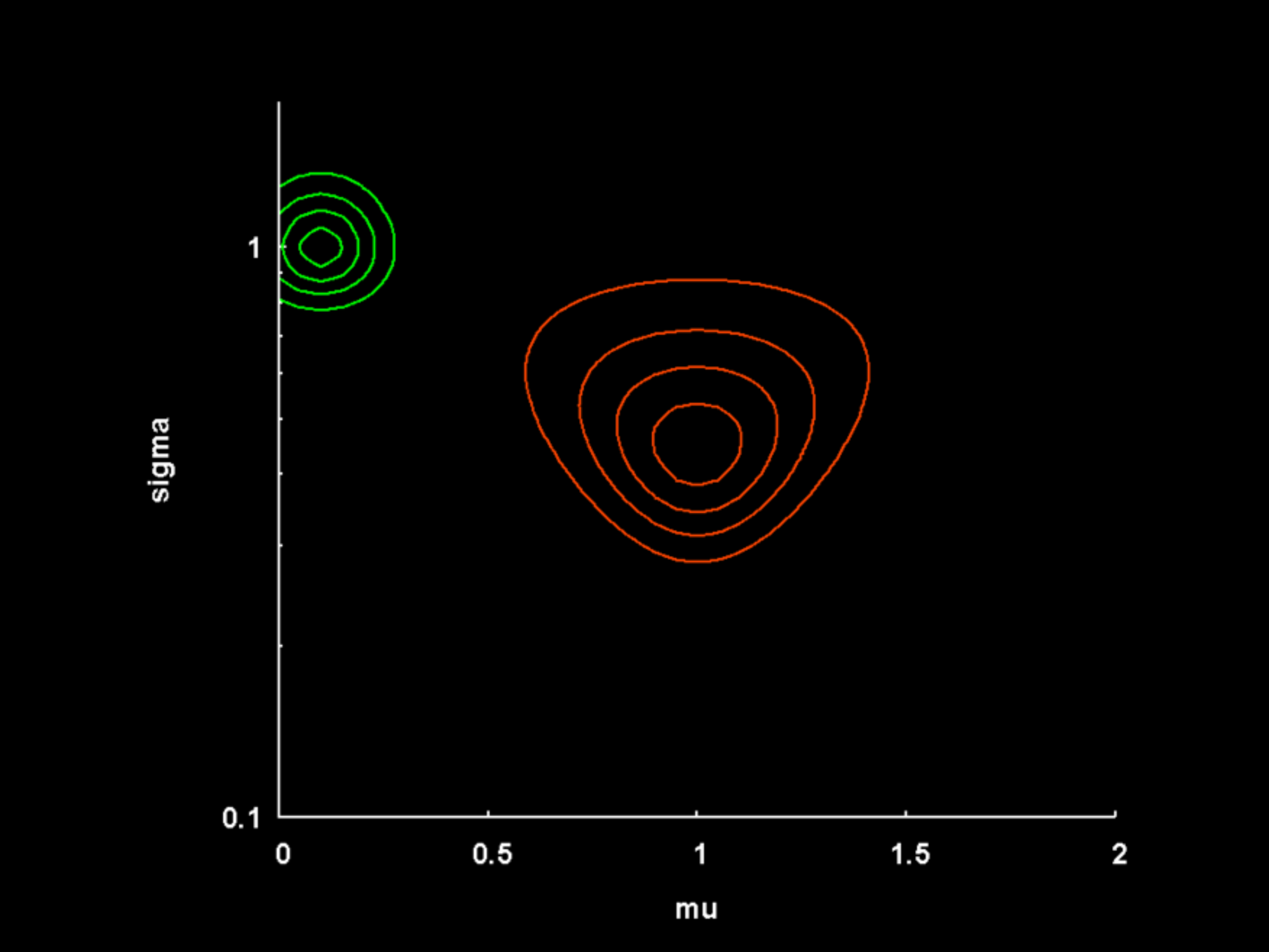


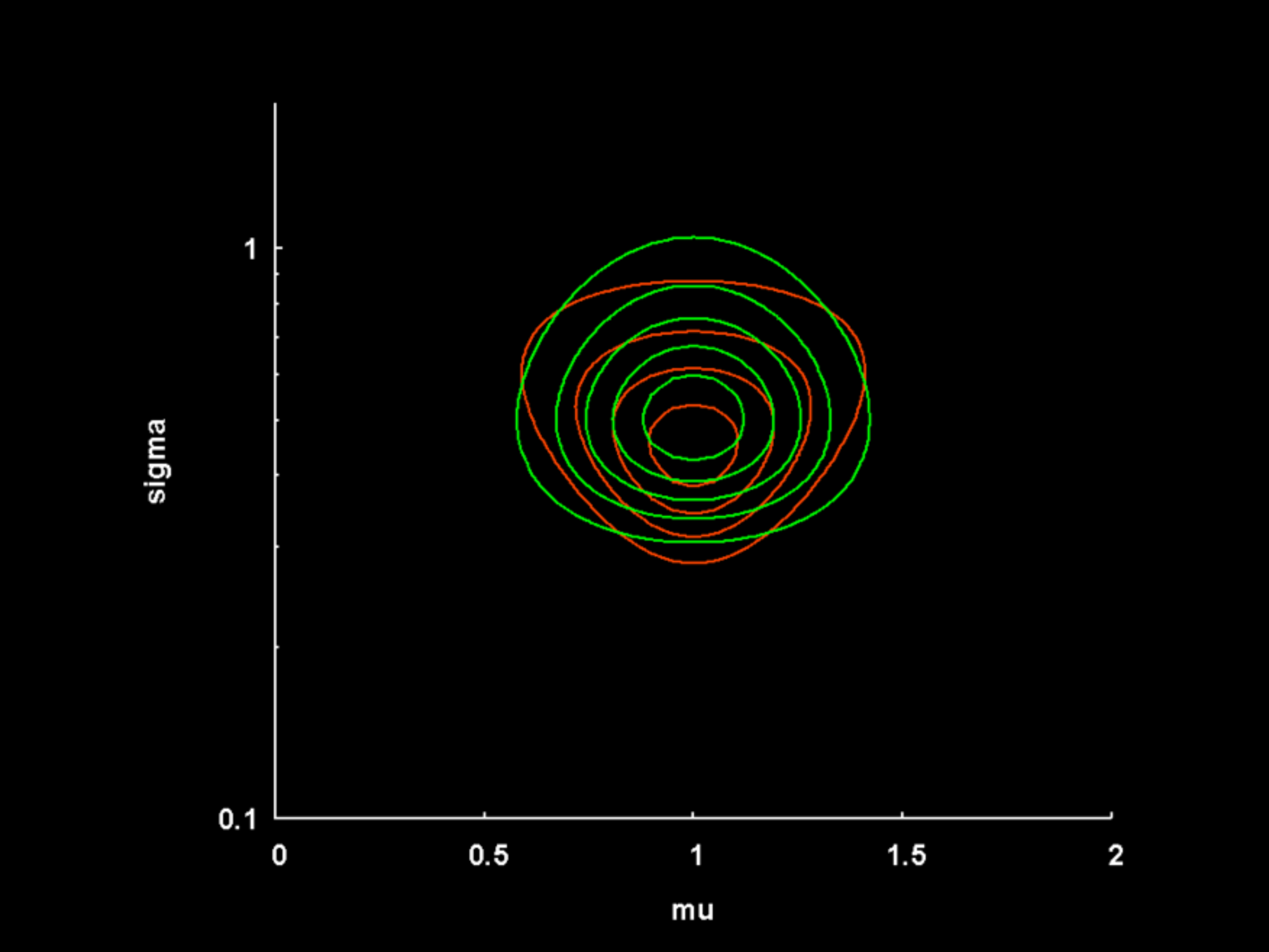
$P \equiv$ Inference of μ, σ^2

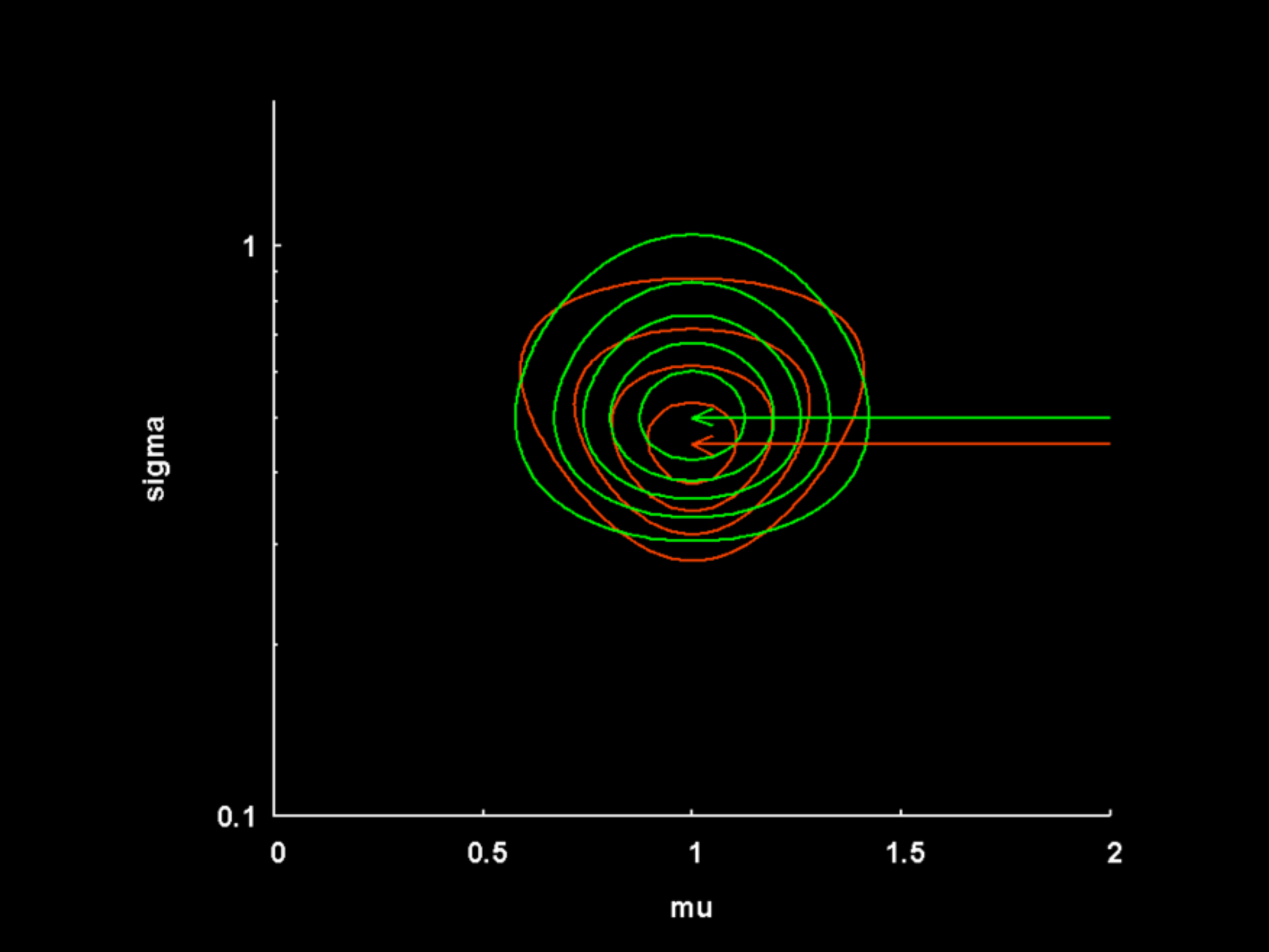

σ^2
$\bar{x}$
μ
Γ
$\bar{x}$


σ^2
σ^2_{N-1}
σ^2_N
μ
Γ

$$Q(\mu, \sigma^2) = Q_\mu(\mu) Q_{\sigma^2}(\sigma^2)$$


sigma
1
0.1
0
0.5
1
1.5
2
mu


sigma
1
0.1
0
0.5
1
1.5
2
mu


sigma
1
0.1
0
0.5
1
1.5
2
mu

Variational methods - example

Approximate the spin system whose energy function is

$$E(\mathbf{x};\mathbf{J}) = -\frac{1}{2}\sum_{m,n} J_{mn}x_m x_n - \sum_n h_n x_n$$

with a separable distribution

$$Q(\mathbf{x};\mathbf{a}) = \frac{1}{Z_Q}\exp\left(\sum_n a_n x_n\right).$$

We optimize Q so as to minimize the variational free energy

$$\beta\tilde{F}(\mathbf{a}) = \beta\sum_{\mathbf{x}} Q(\mathbf{x};\mathbf{a})E(\mathbf{x};\mathbf{J}) - \sum_{\mathbf{x}} Q(\mathbf{x};\mathbf{a})\ln\frac{1}{Q(\mathbf{x};\mathbf{a})},$$

$$= \beta\left(-\frac{1}{2}\sum_{m,n} J_{mn}\bar{x}_m\bar{x}_n - \sum_n h_n\bar{x}_n\right) - \sum_n H_2^{(e)}(q_n).$$

– Minimum (with respect to $\mathbf{a}$) can be found by the iterative equations

$$a_m = \beta\left(\sum_n J_{mn}\bar{x}_n + h_m\right) \quad \text{and} \quad \bar{x}_n = \tanh(a_n)$$

$$P(x) = \frac{e^{-\beta E(x;J)}}{Z(\beta)}$$

$$\tilde{F}(\theta,\beta) = \sum_x Q_\theta(x)\beta E(x) - \sum_x Q_\theta(x) \ln \frac{1}{Q_\theta(x)}$$

β

variational free energy

$$\tilde{F}(\theta) \geq -\log Z(\beta)$$

$$E(\underline{x}) = -\frac{1}{2} \sum_{m\,n} J_{mn} x_m x_n$$

$$\underline{X} \in \{-1, +1\}^N$$

$$\frac{1}{2} \sum_{mn} J_{mn} x_m x_n - \sum_n h_n x_n$$

$$Q(\underline{x};\underline{a}) = \frac{1}{Z_Q} e^{\sum_n a_n x_n}$$

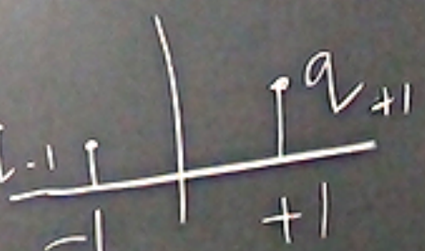
$$= \prod_{n=1}^{N} q_n(x_n; a_n)$$

$$= \frac{1}{Z_Q} \prod_{n=1}^{N} e^{a_n x_n}$$

$h_n x_n$

$\underline{\theta}$

$\sum_n a_n x_n$



$\bar{X}_n = q_{+1} + (-1) q_{-1}$

β

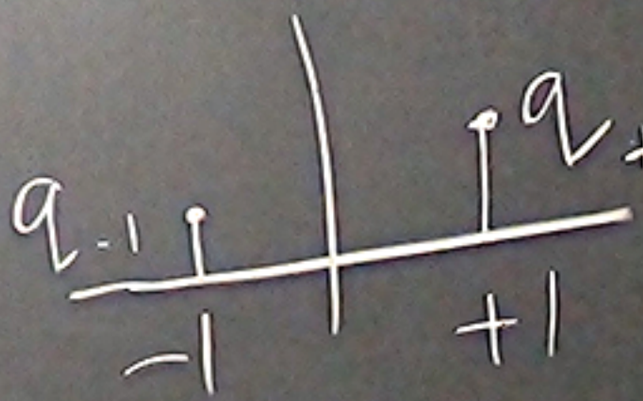
$\tilde{F}(\theta, \beta) = \sum_x Q_\theta(x) \left[\beta E(x) \right]$

$\underline{\theta}$

$\left(x_n ; a_n \right)$

variation

$F(\theta) \geq - \log Z$

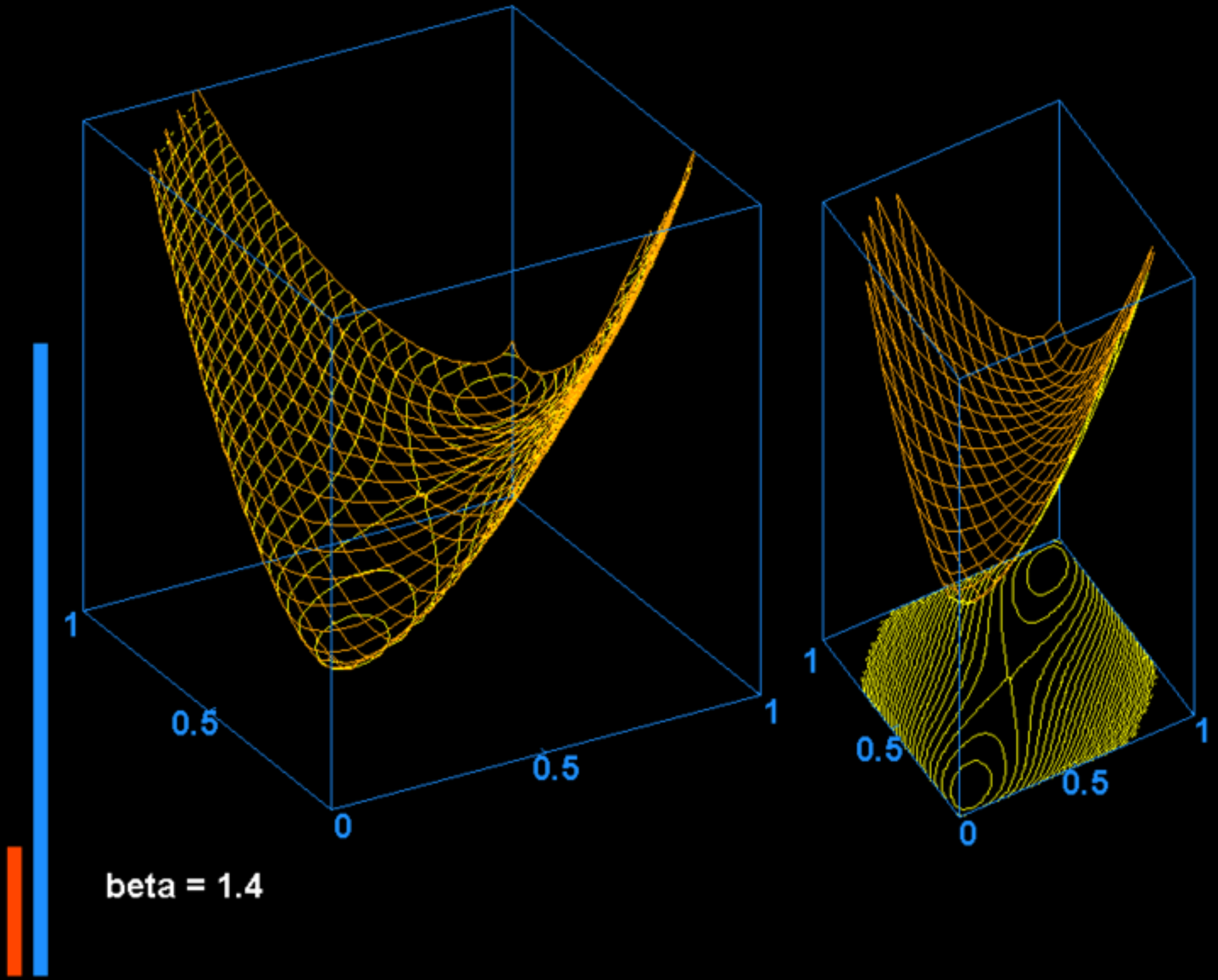


$$\bar{X}_n = q_{+1} + (-1)q_{-1}$$

$$= q - (1-q)$$

$$= \underline{2q-1}$$

$$\tilde{F}(\theta, \beta) = \sum_x Q_\theta(x) \, E($$


1
0.5
0
0.5
1
1
0.5
0
0.5
1
beta = 1.4

$$\{x_1, x_2\} \in \pm 1$$

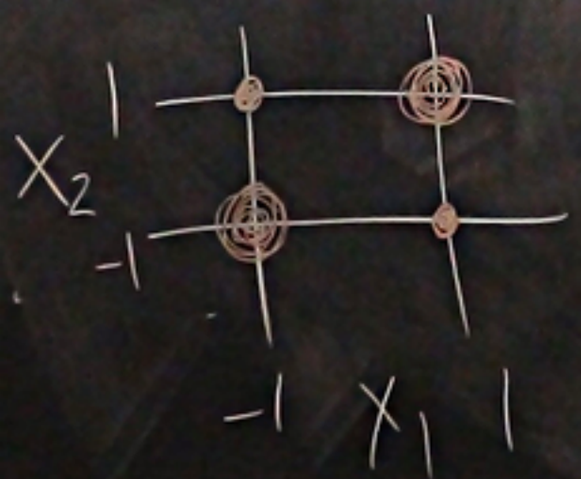
$$E(\underline{x}) = -x_1 x_2$$

$x_1, x_2 \} \in \pm 1$

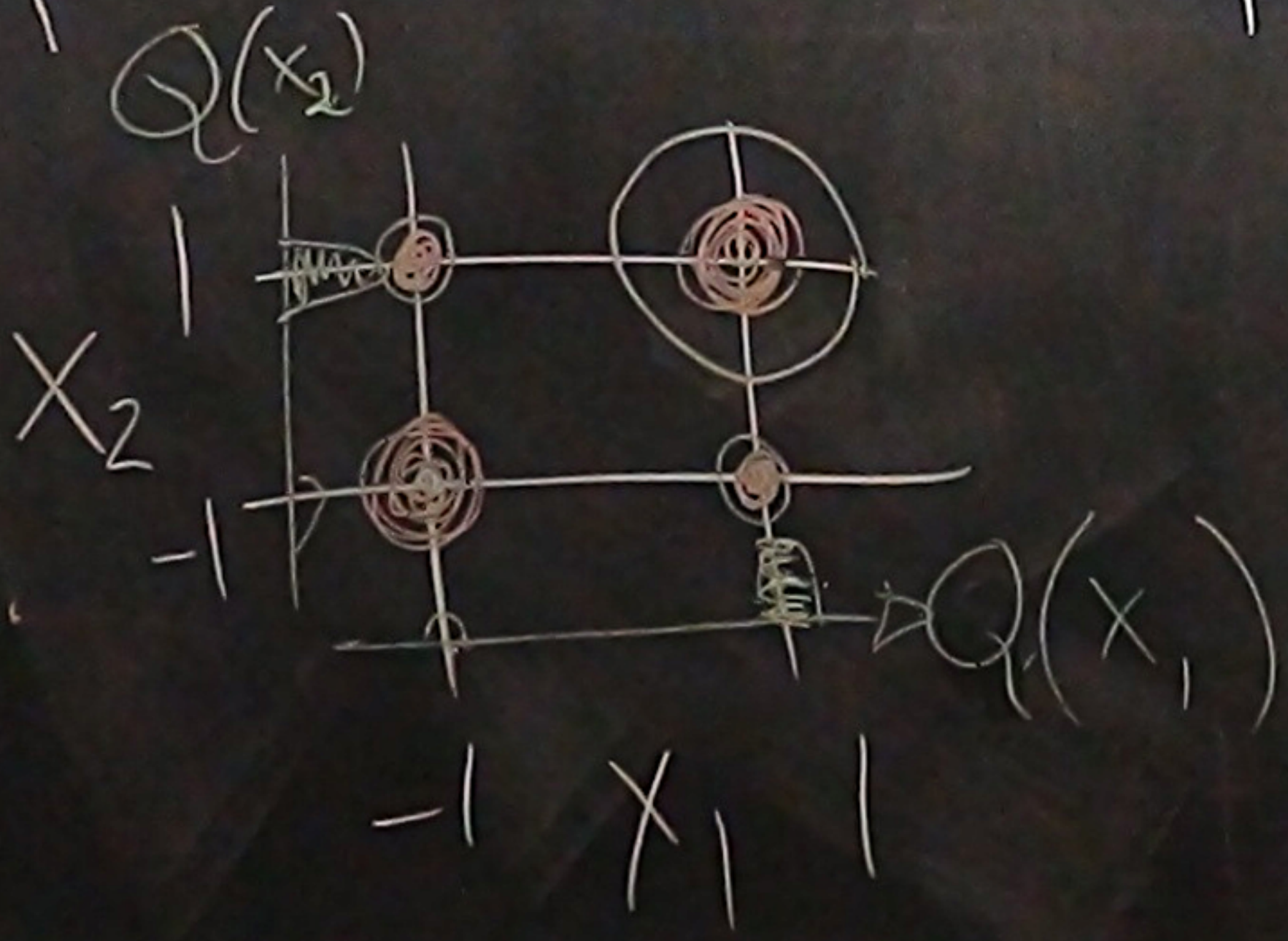
$J = 1$

$\underline{x}) = -x_1 x_2$

$$\beta \tilde{F}(\underline{a}) = \beta \bar{x}_1 \bar{x}_2 - H_2^{(e)}(q_1) - H_2^{(e)}(q_2)$$

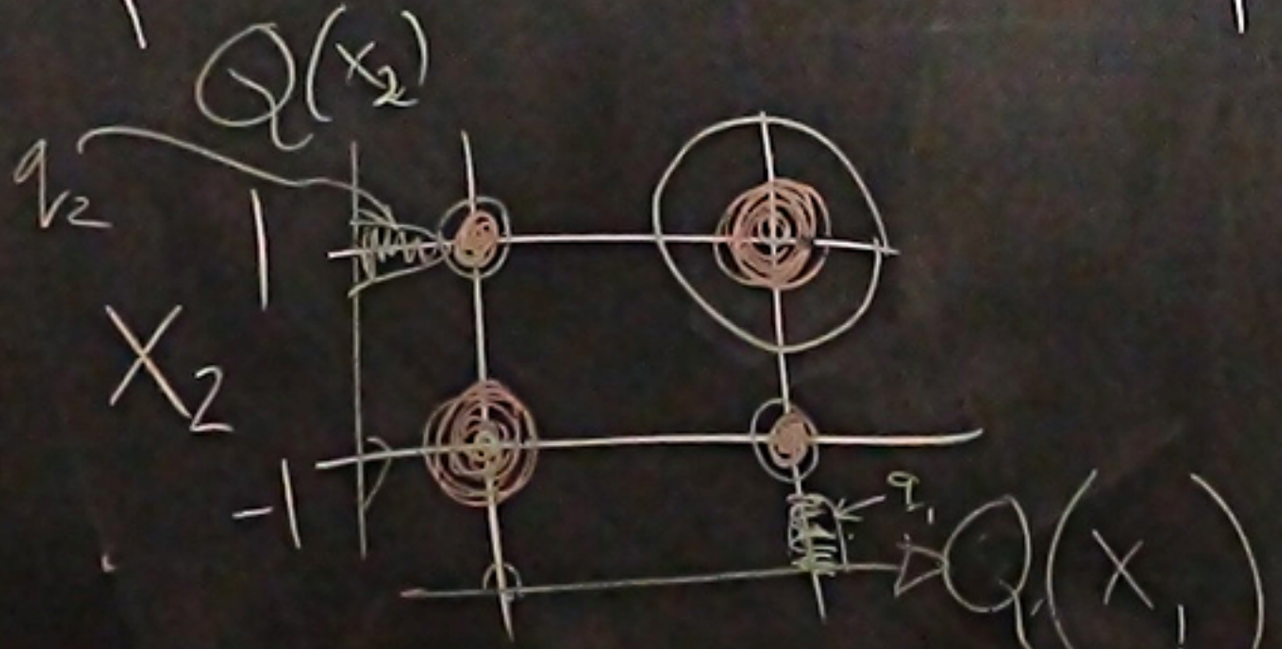


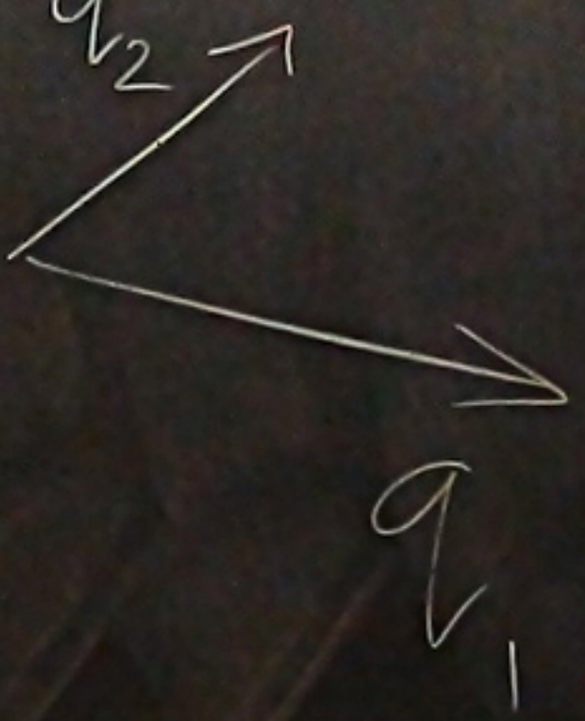
$\beta F(\underline{a}) = \beta$

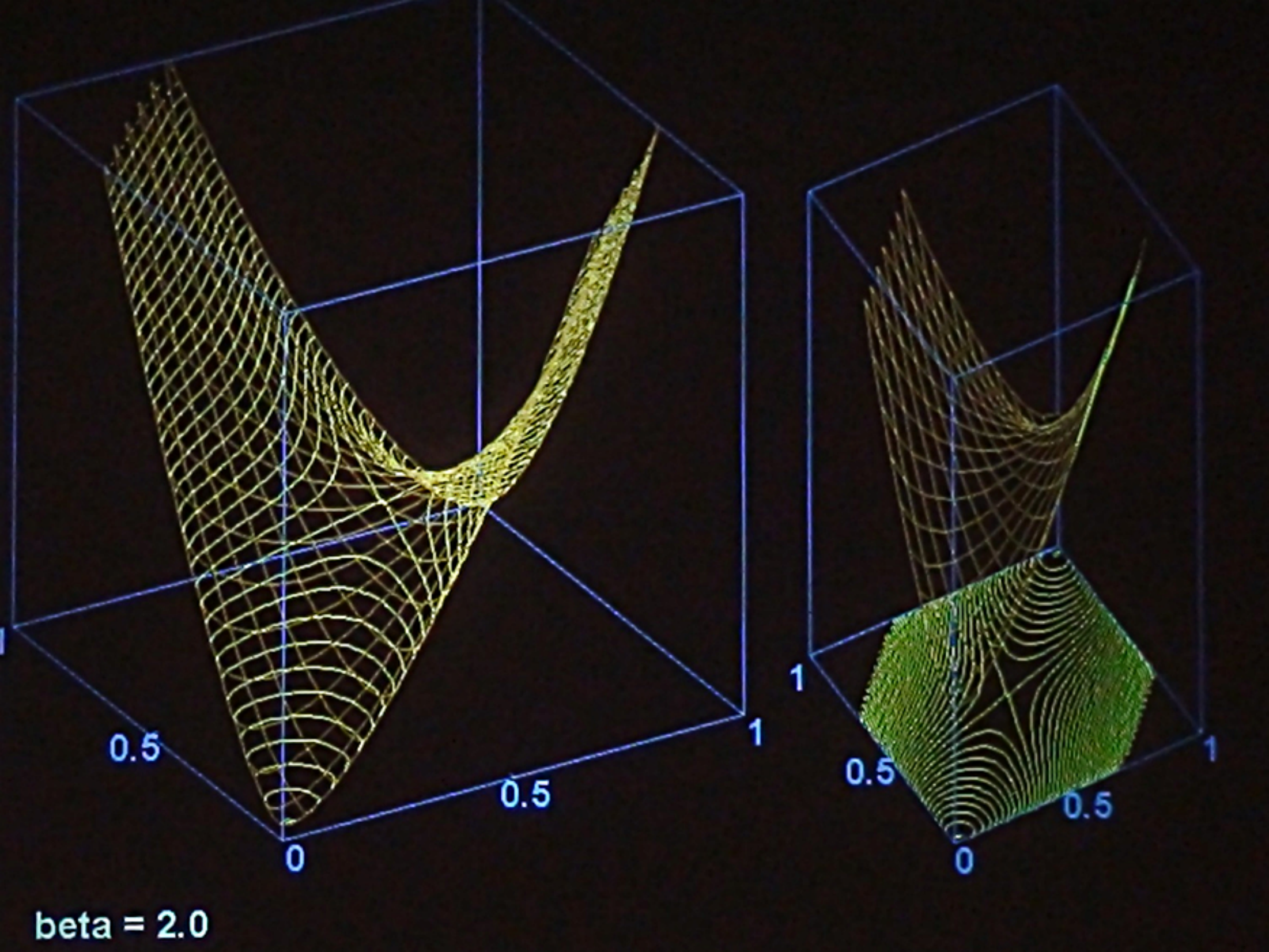


$= x_1 x_2$

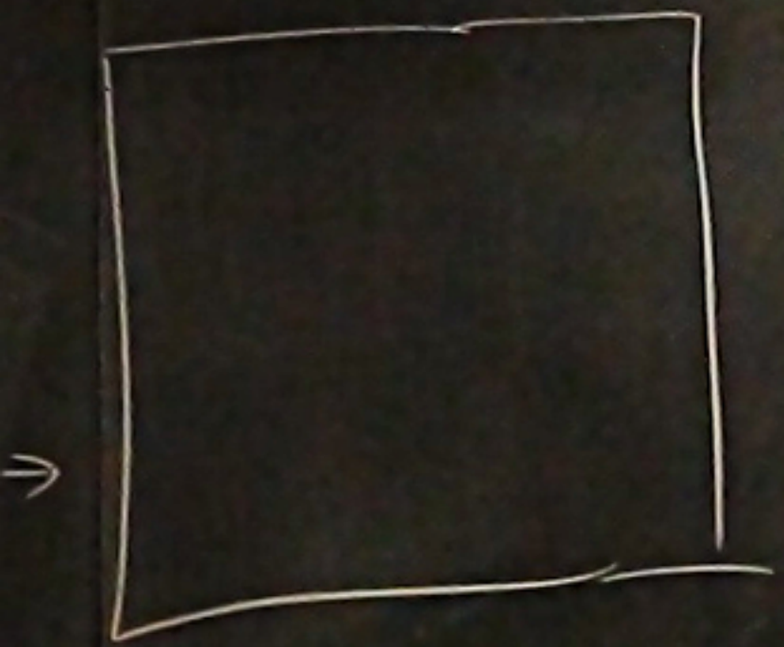
$$\beta \tilde{F}(\underline{a}) = \beta \bar{x}_1 \bar{x}_2 -$$

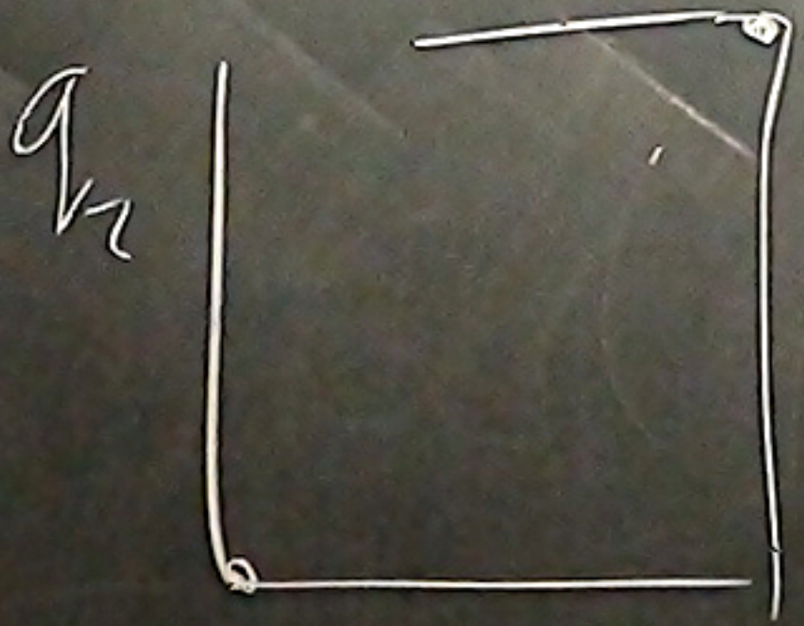






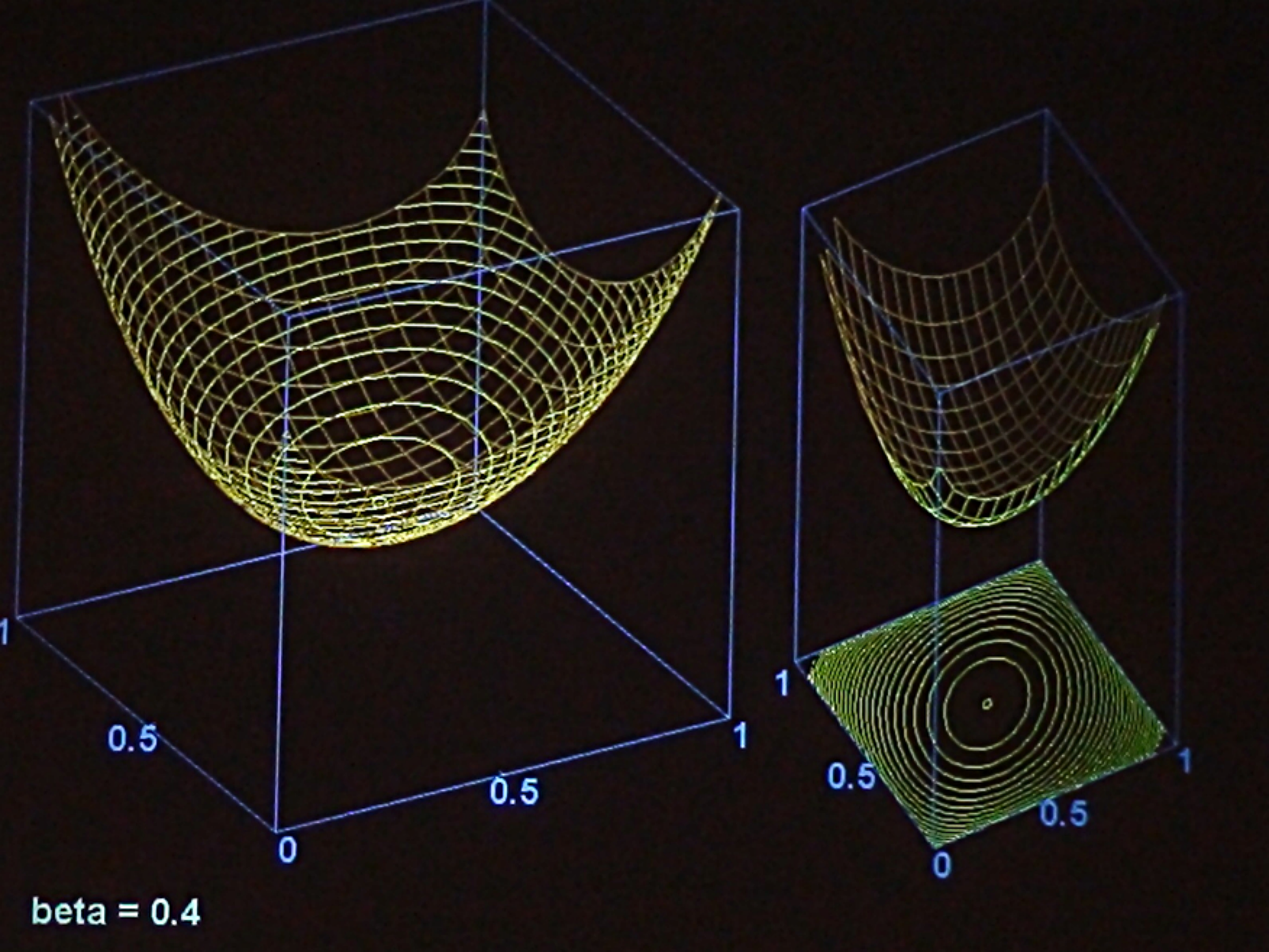
beta = 2.0



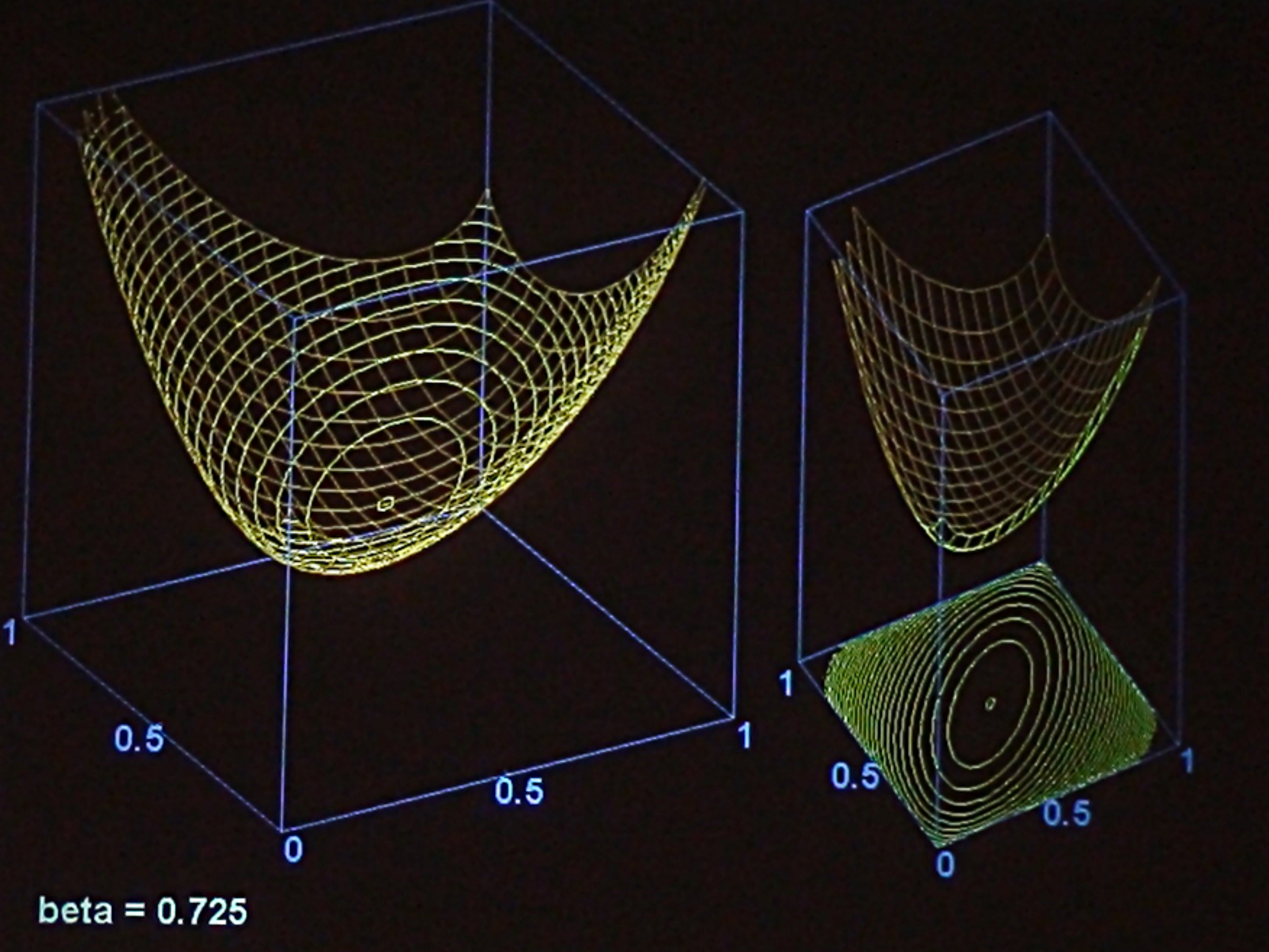


β

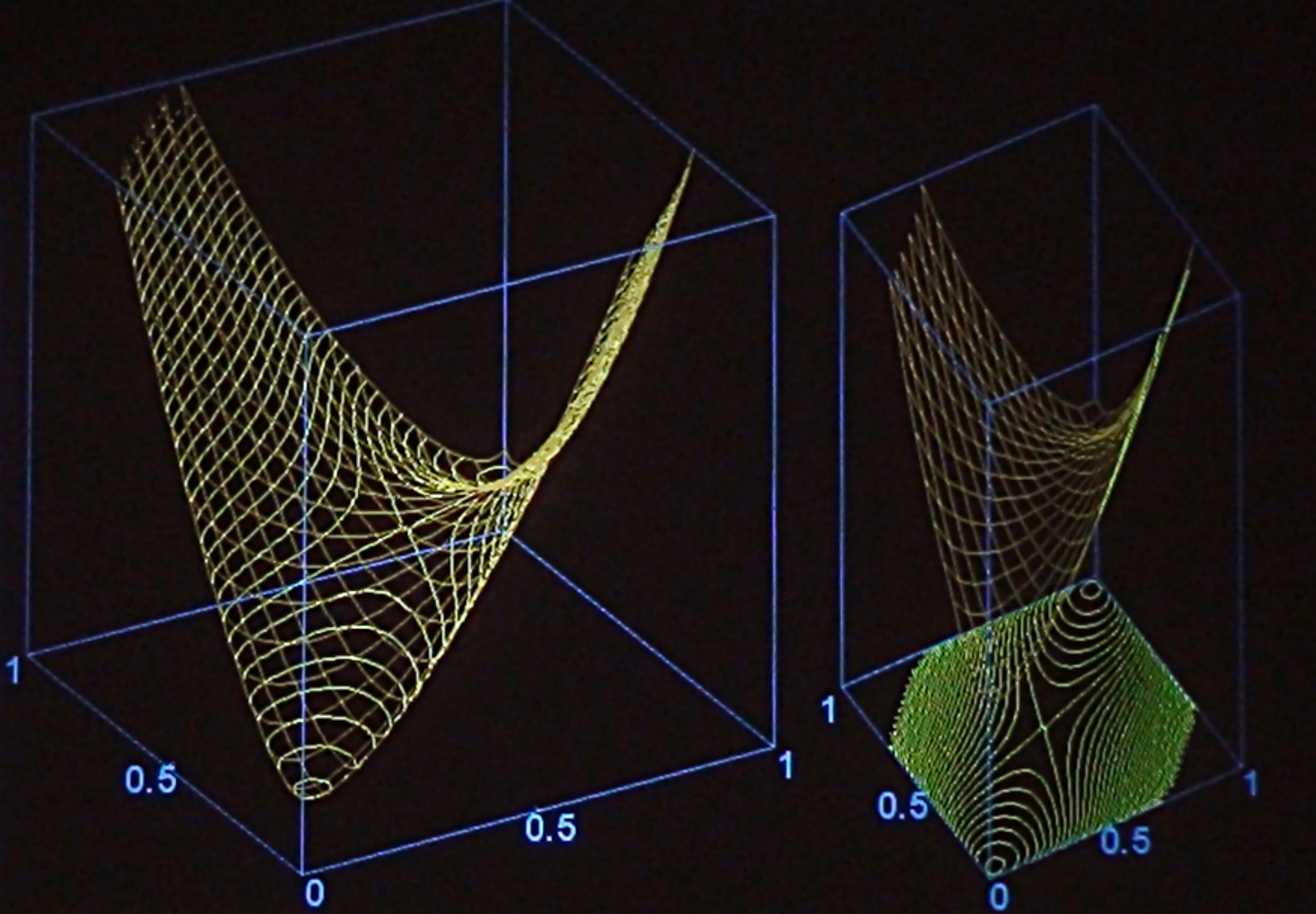
$\beta = 2$



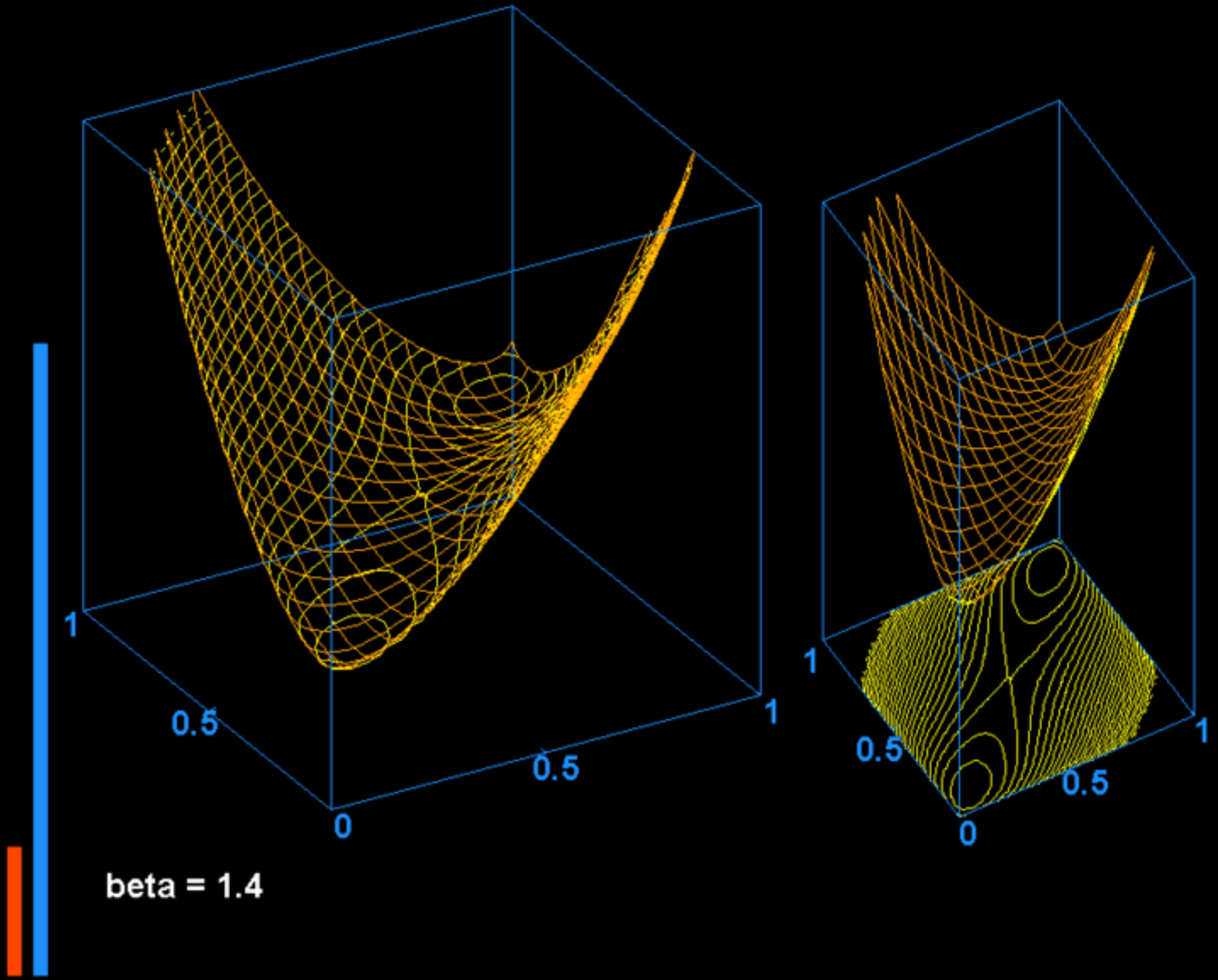
beta = 0.4

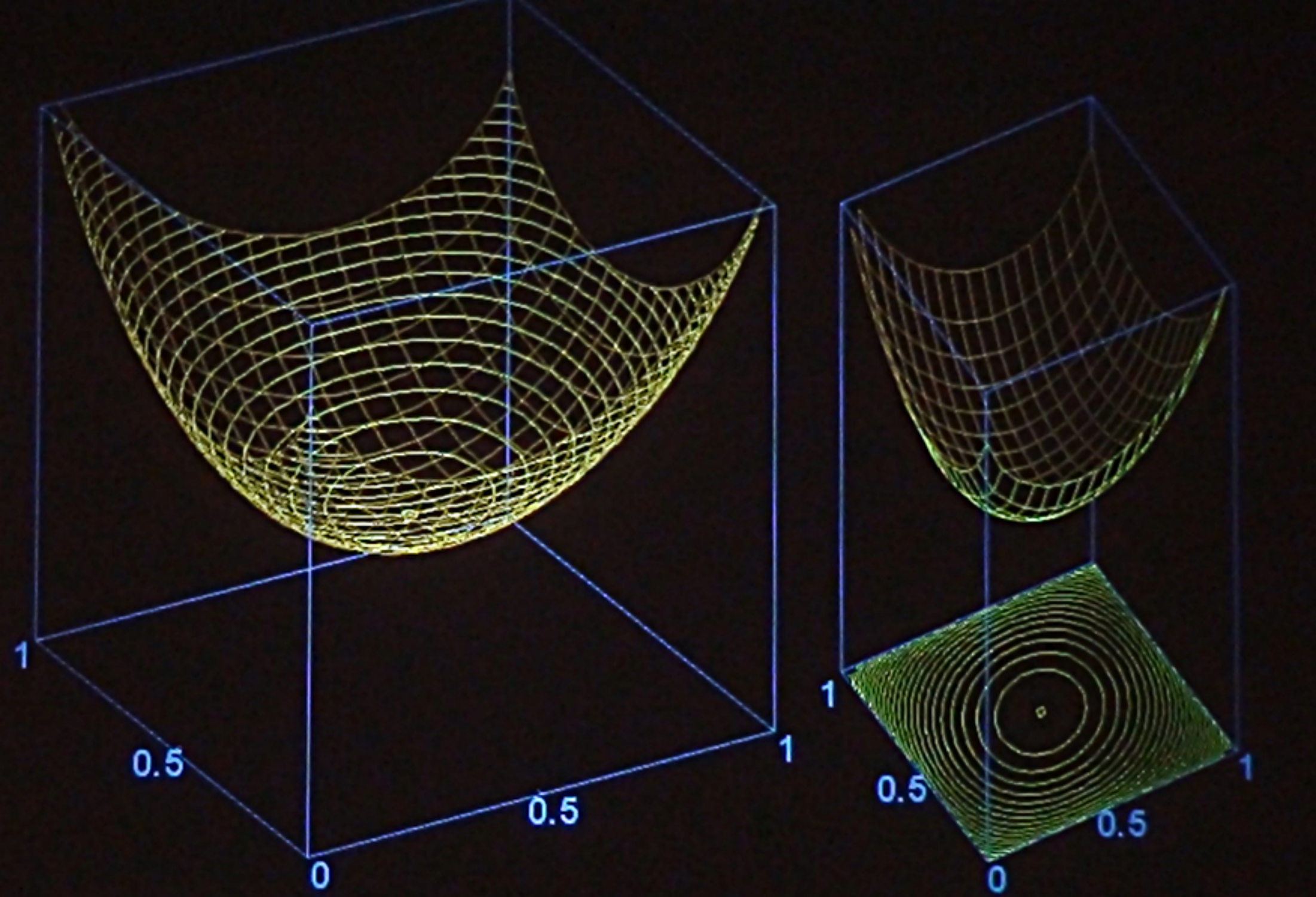


beta = 0.725

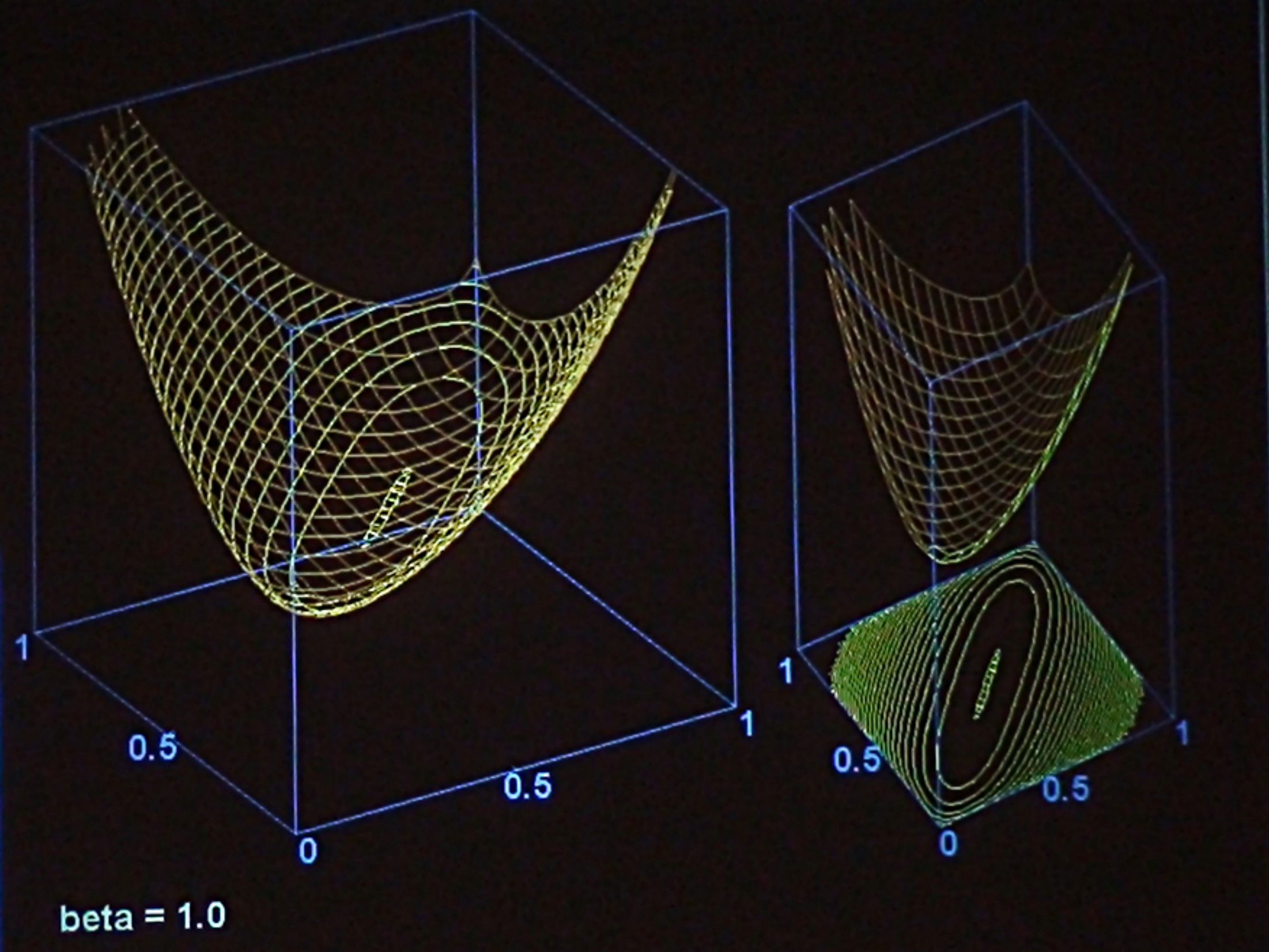


beta = 1.75


1
0.5
0
0.5
1
1
0.5
0
0.5
1
beta = 1.4

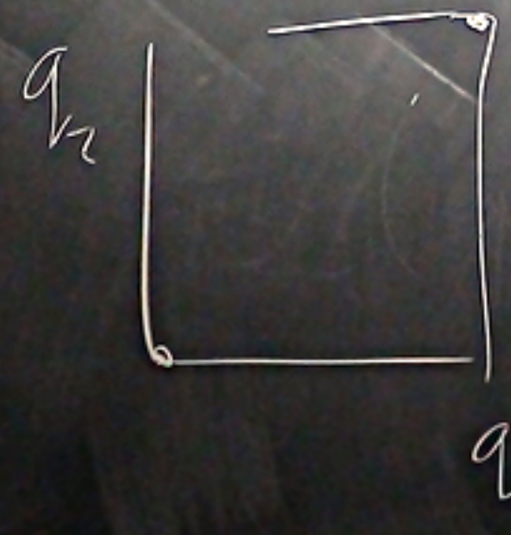


beta = 0.25

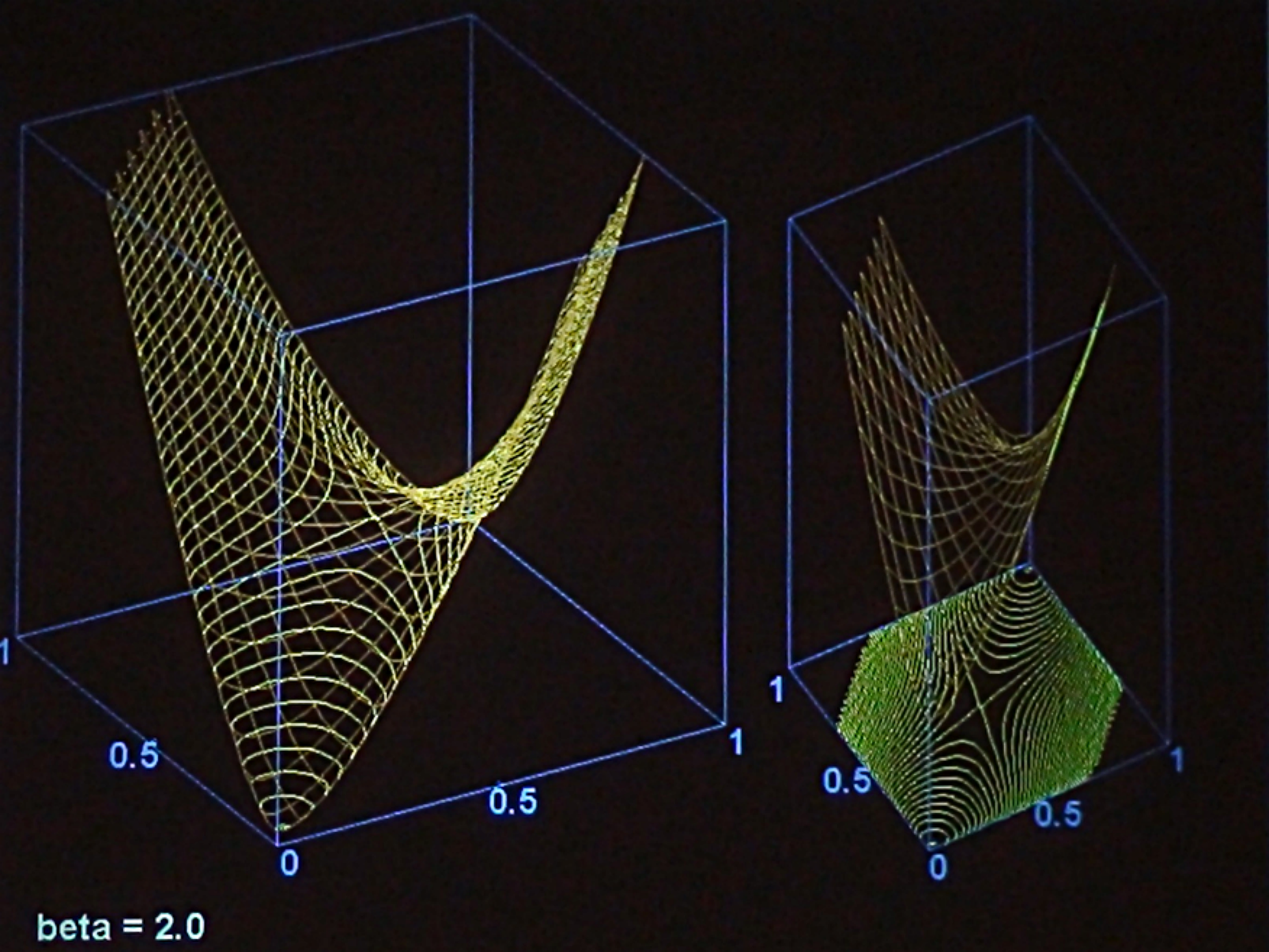


beta = 1.0

$\beta = 0.4$



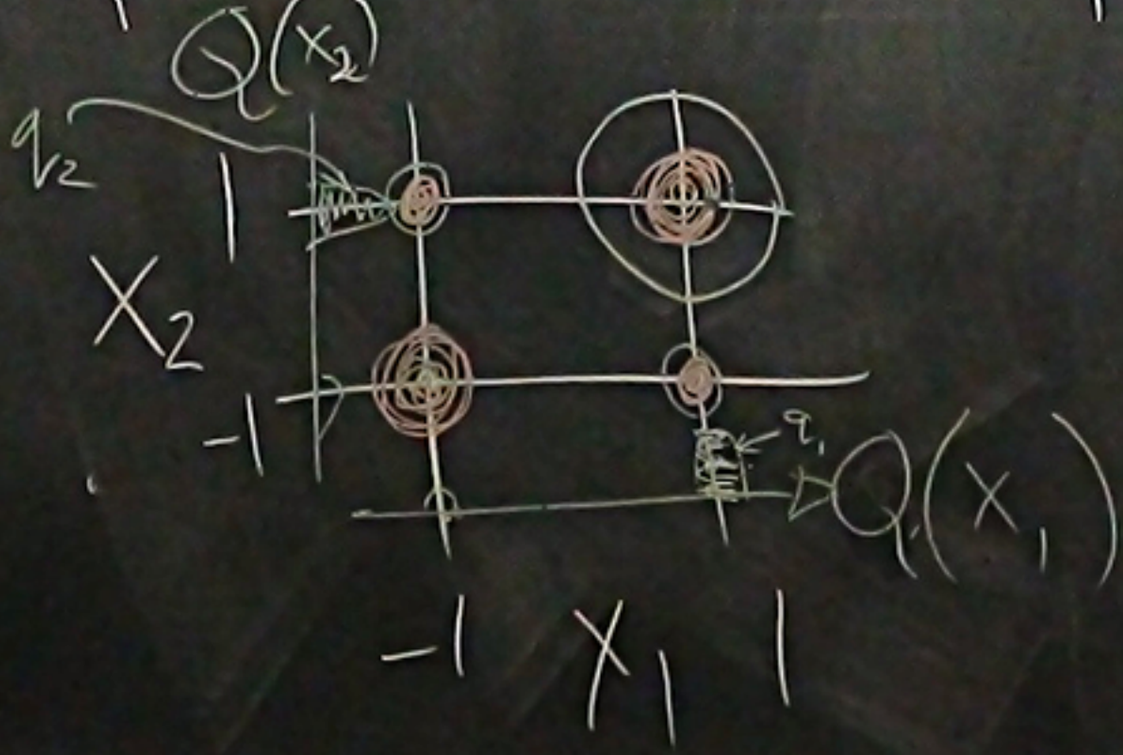
$\beta = 2$

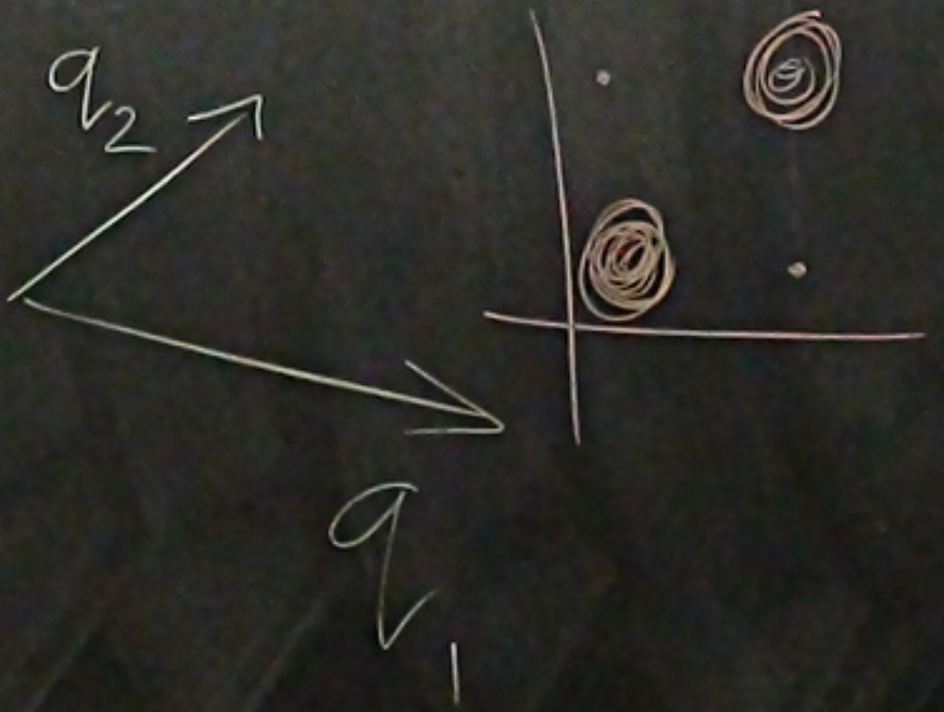


beta = 2.0

$= -\bar{x}_1 \bar{x}_2$

$\beta \tilde{F}(\underline{a}) = \beta \bar{x}_1 \bar{x}_2 - H$



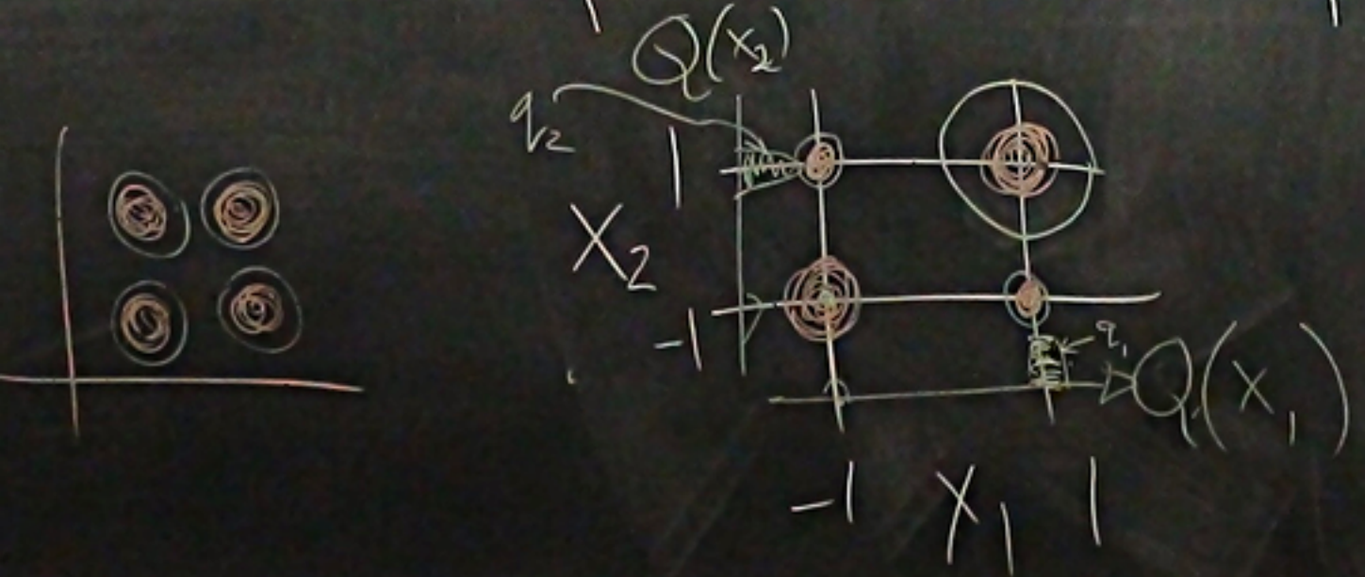


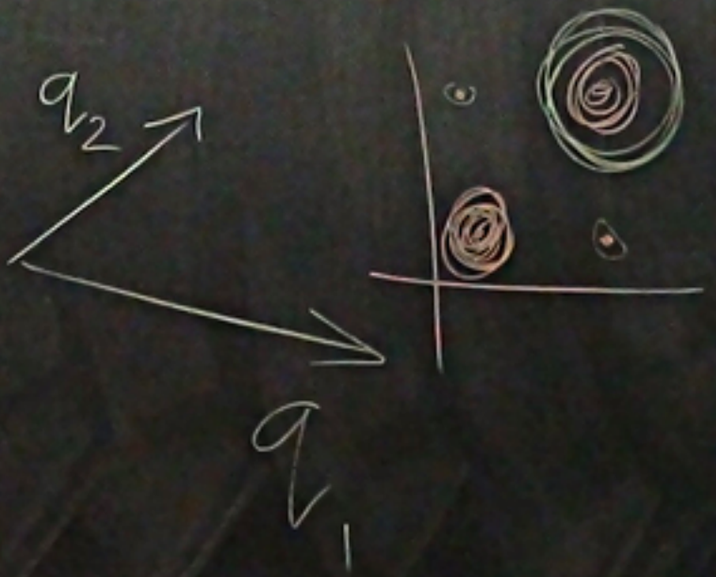
$\{x_1, x_2\} \in \pm 1$

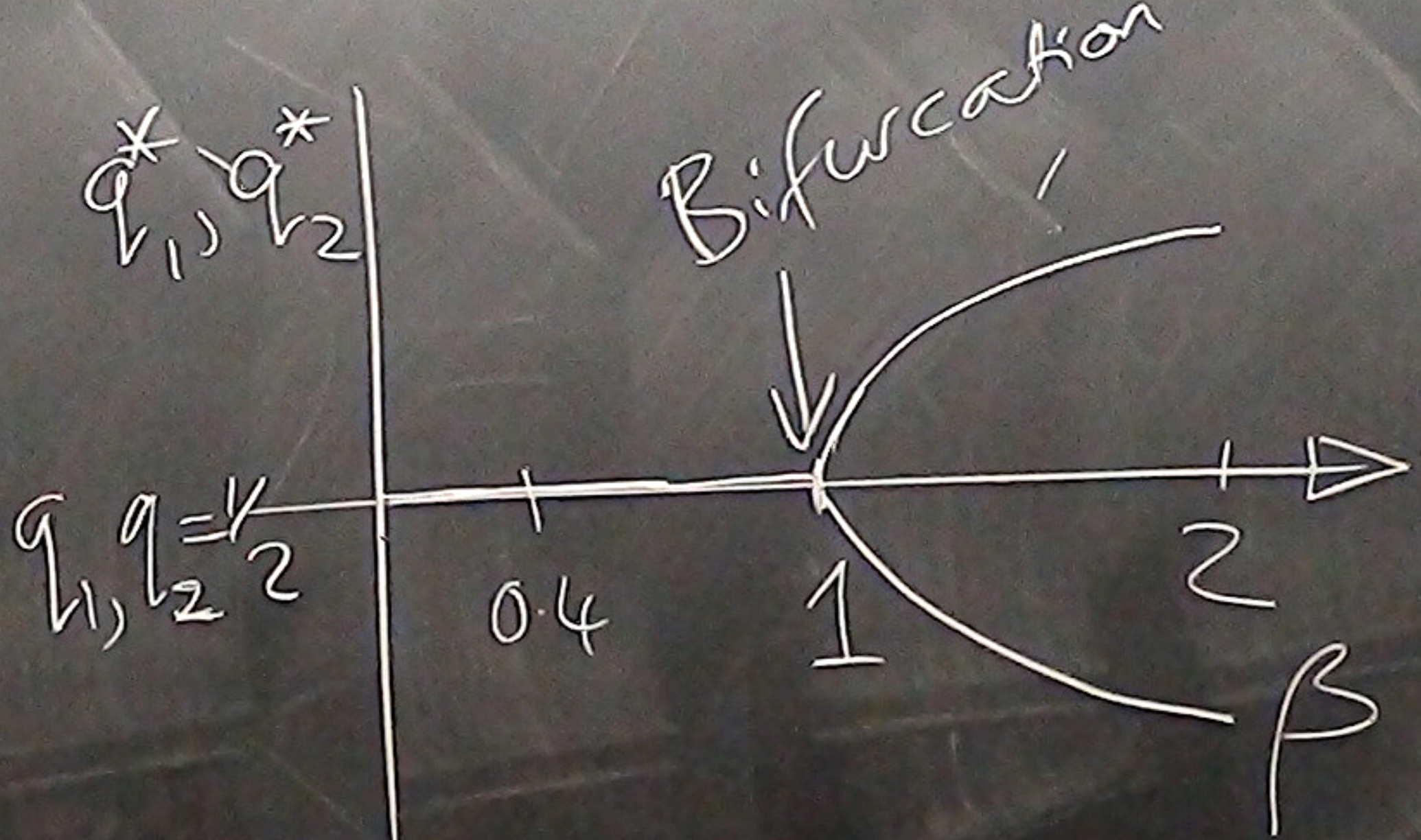
$E(\underline{x}) = -x_1 x_2$

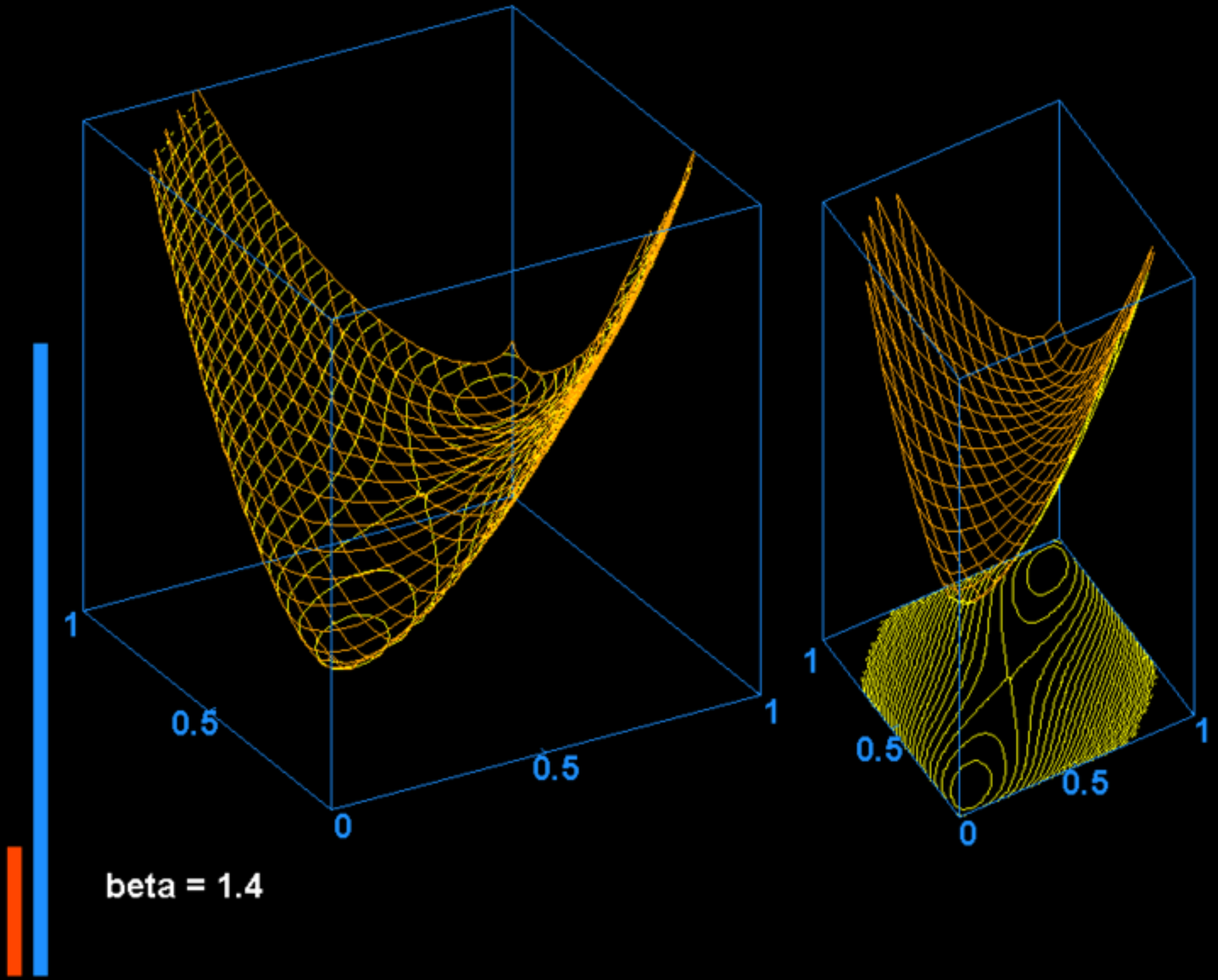
$J=1$

$\beta \tilde{F}(\underline{a}) = \beta \bar{x}_1 \bar{x}_2 - H_2^{(e)}$






q_1^*, q_2^*
Bifurcation
q_1, q_2
$q = 1/2$
0.4
1
2
β


1
0.5
0
0.5
1
1
0.5
0
0.5
1
beta = 1.4

Have you seen this before?

$$a_m = \beta \left(\sum_n J_{mn} \bar{x}_n + h_m \right) \quad \text{and} \quad \bar{x}_n = \tanh(a_n)$$

or equivalently...

$$\bar{x}_m = \tanh \left(\beta h_m + \beta \sum_n J_{mn} \bar{x}_n \right)$$

?

$$\bar{x}_m = \tanh\left(\beta h_m + \beta \sum_n J_{mn}\bar{x}_n\right)$$

'Mean field theory'
is a variational method

Mean Field Theory

Curie-Weiss

Feynman trick

Bogoliubov

view of MFT

1. Clear objective functi

2.

Review of MFT

1. Clear objective function $\Rightarrow$ derivation
2. Could choose more complex $\mathcal{O}$ $\rightarrow$ generalize
3. Shows that MFT gives a bound on Z