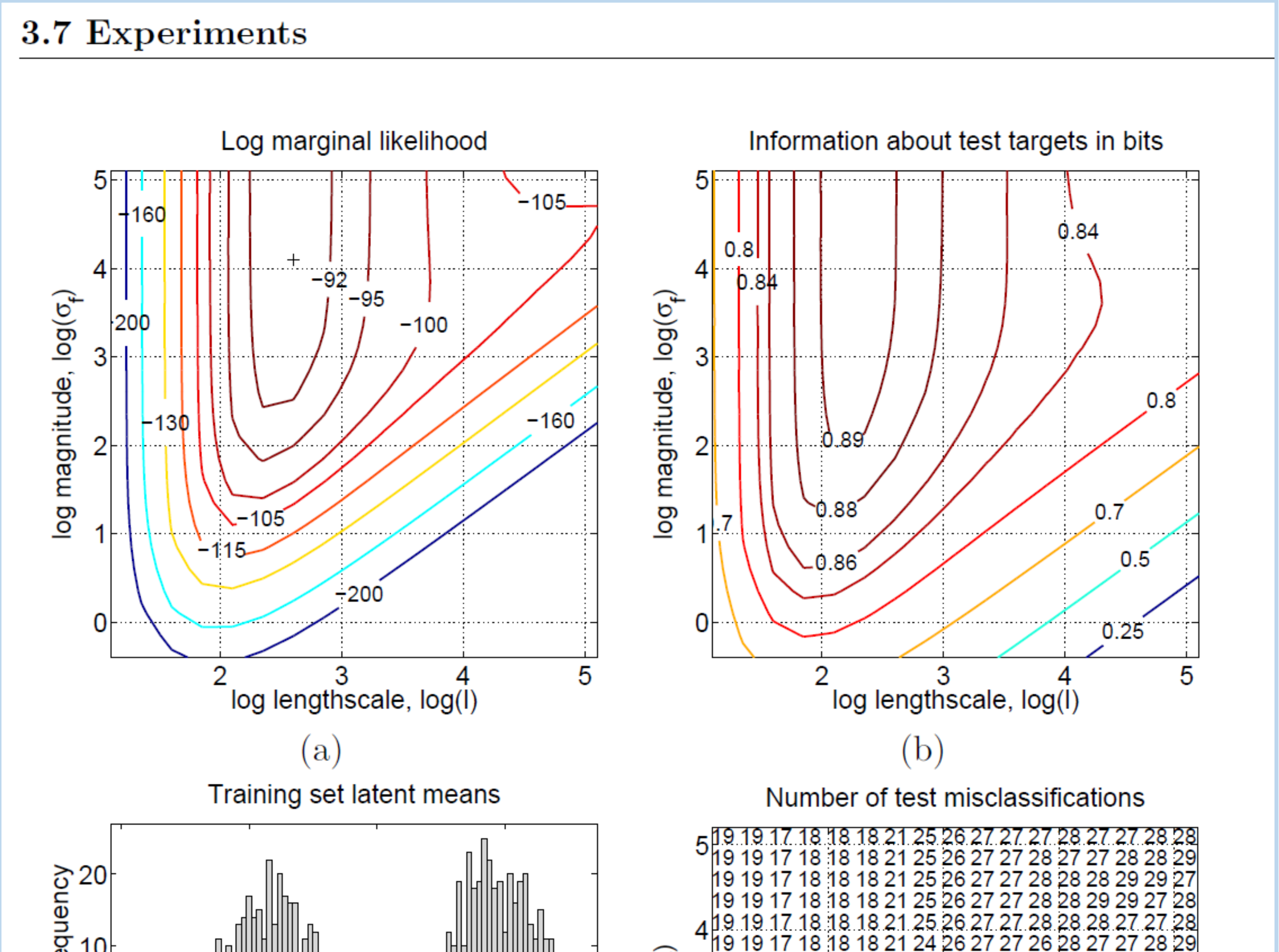
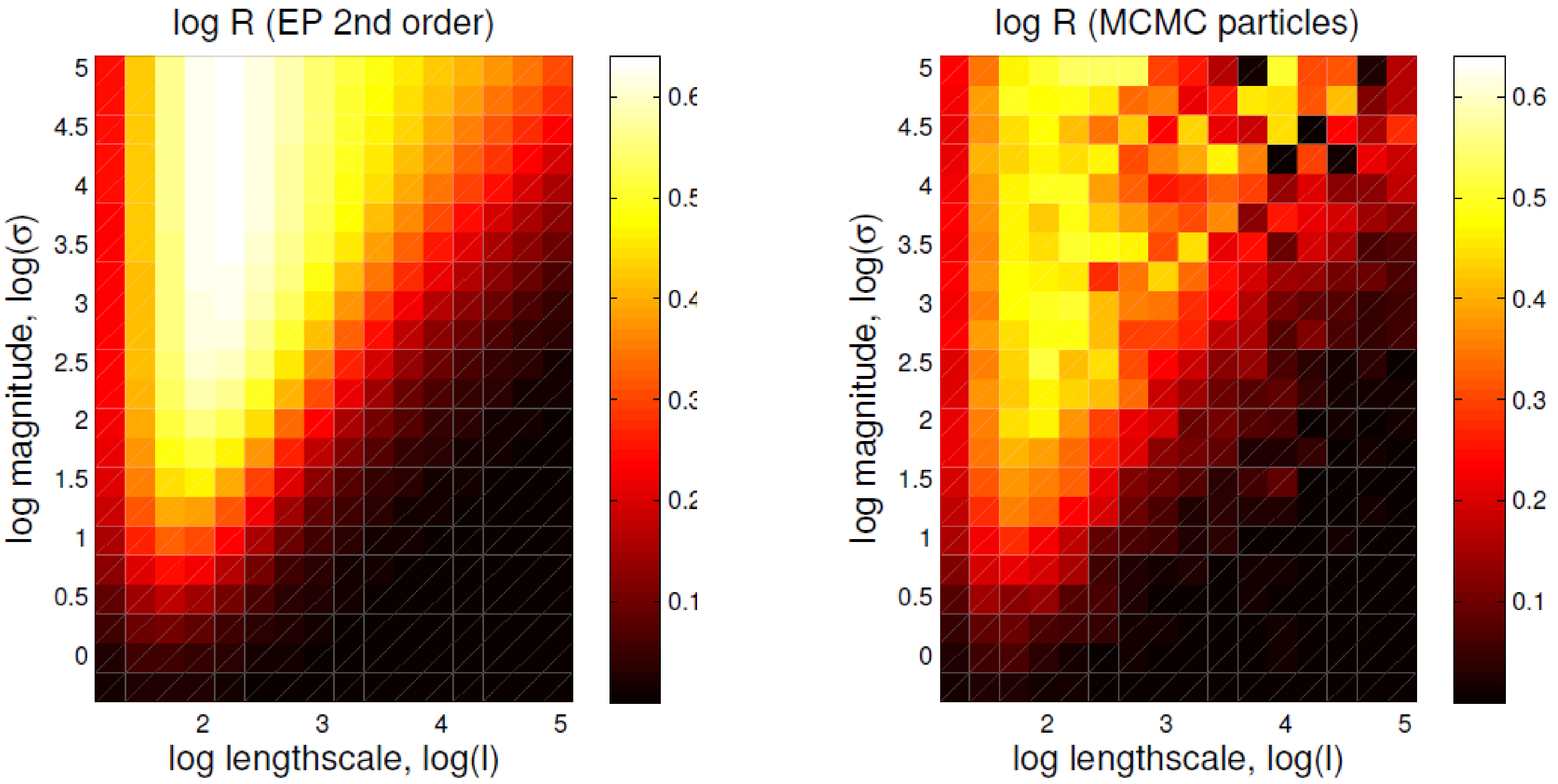
Towards (?) marginal likelihood lower bounds with EP

Ulrich Paquet, Adrian Weller, Ole Winther, Nick Ruozzi

Kuss & Rasmussen’s grid



$\log Z - \log Z_{EP}$ (GP classification)



Belief Propagation: done

- Loop Series and Bethe Variational Bounds in Attractive Graphical Models
  - Sudderth et al, 2007 [Loop series]
- The Bethe Partition Function of Log-supermodular Graphical Models
  - Ruozzi, 2012 [Graph covers]
- Clamping Variables and Approximate Inference
  - Weller & Jebara, 2014 [Clamp and integrate]

  [First principles]

Notation / whiteboard

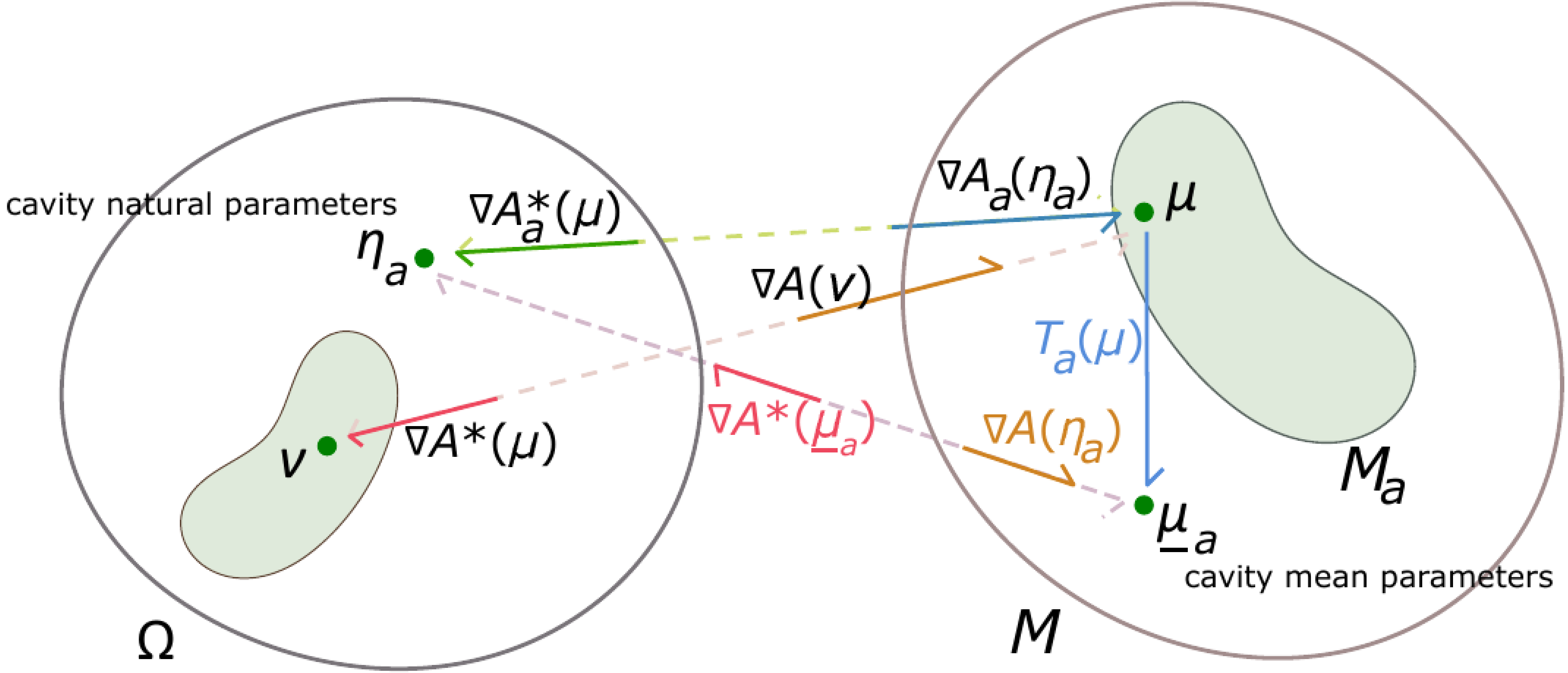
EP…

(Tom Minka's "The EP Free Energy and Minimization Schemes")

$$F_{\mathrm{EP}}(\nu, \{\eta_a\}) = (n-1)\log Z_q(\nu) - \sum_a \log Z_a(\eta_a)$$

$$-\log Z_{\mathrm{EP}} = \min_\nu \max_{\{\eta_a\}} F_{\mathrm{EP}}(\nu, \{\eta_a\}) \quad \text{subject to } \sum_a \eta_a = (n-1)\nu.$$

Notation



cavity natural parameters
$\nabla A_a^*(\mu)$
η_a
$\nabla A_a(\eta_a)$
μ
$\nabla A(v)$
$T_a(\mu)$
$\nabla A^*(\underline{\mu}_a)$
$\nabla A(\eta_a)$
v
$\nabla A^*(\mu)$
M_a
$\underline{\mu}_a$
cavity mean parameters
Ω
M

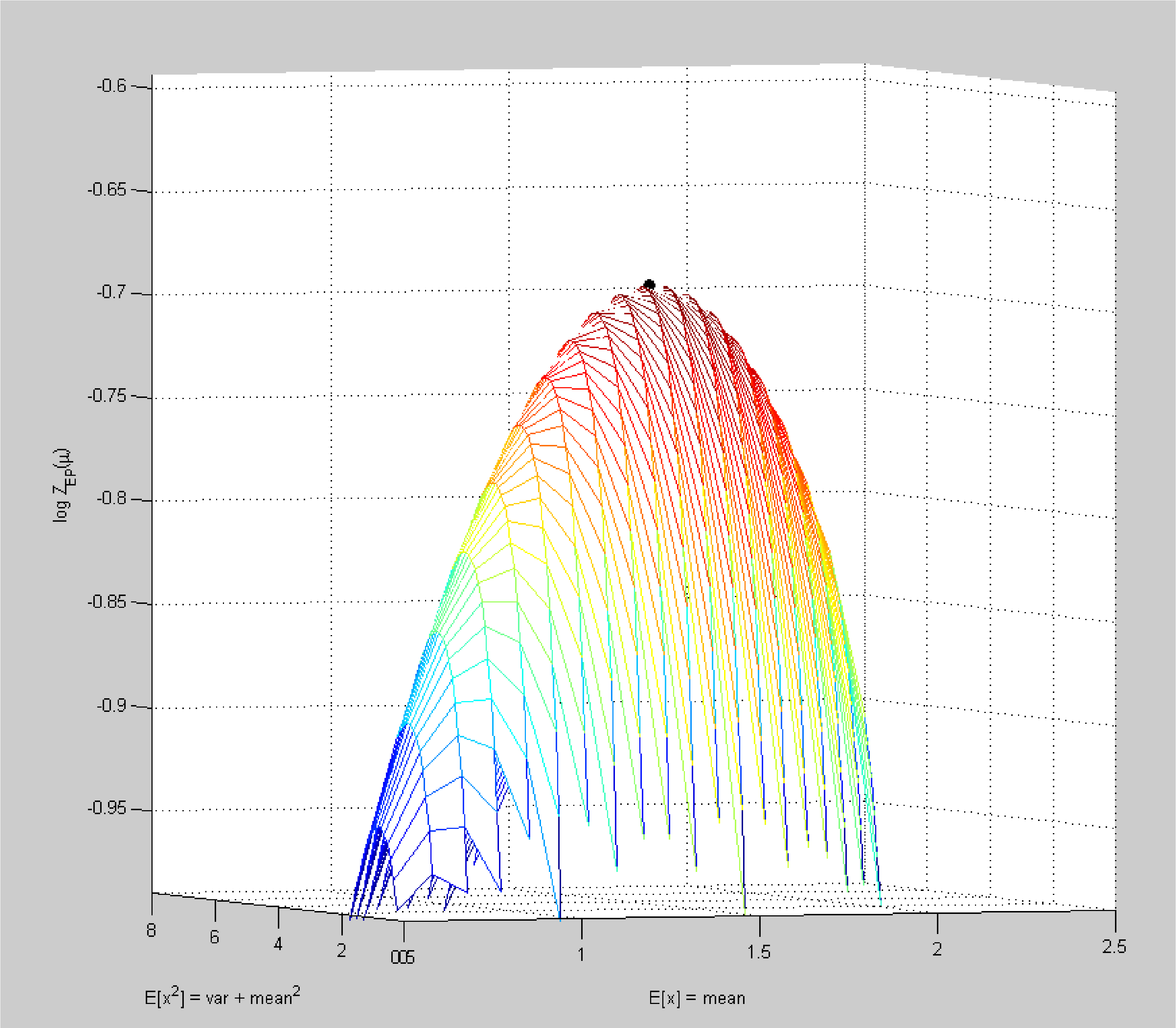
log $Z_{EP}(\mu)$: maximizing over μ only!

- EP approximation: $q(x|\mu) = N(x; m, V)$

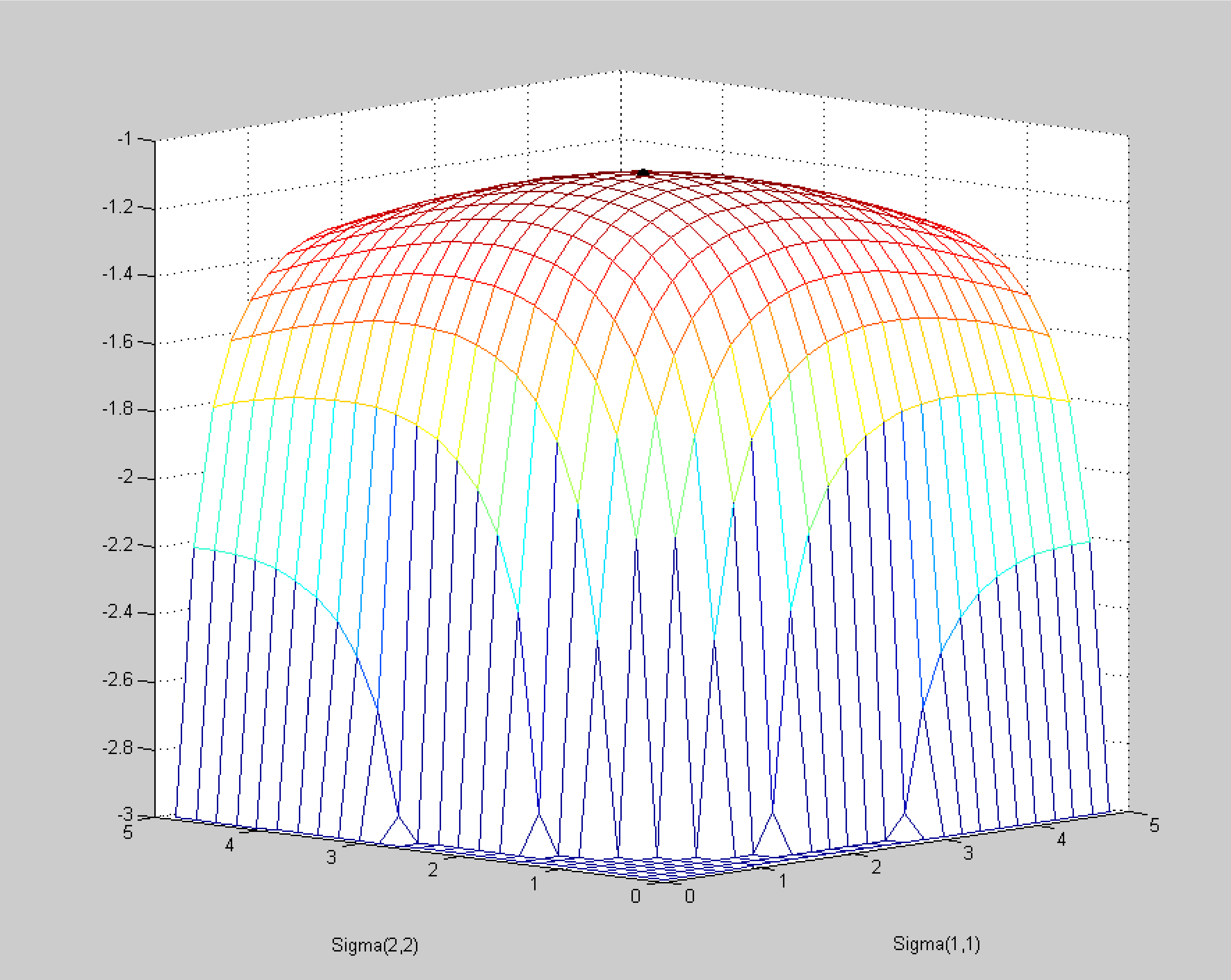
$$\log Z_{\mathrm{EP}} = \max_{\mu} \left[\sum_a A_a(\nabla A_a^*(\mu)) - (n-1)A(\nabla A^*(\mu)) + \mu^T \left((n-1)\nabla A^*(\mu) - \sum_a \nabla A_a^*(\mu) \right) \right]$$
$$= \max_{\mu} \log Z_{\mathrm{EP}}(\mu) \ . \qquad (5)$$

$\mu \in \bigcap_a \mathcal{M}_a$.

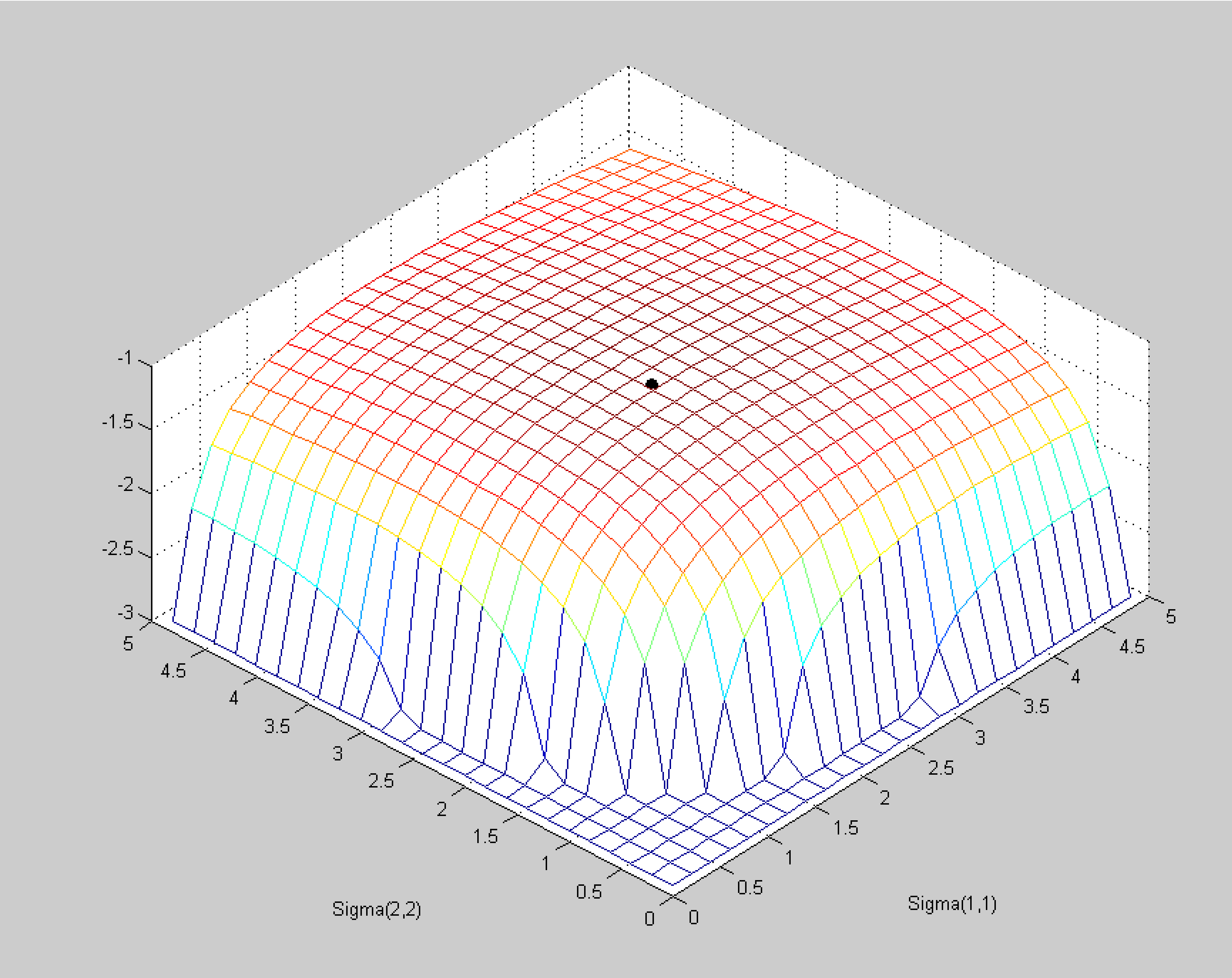
$$\mathcal{M}_a = \{\mu = \nabla A_a(\eta_a) : \eta_a \in \Omega\}$$



-0.6
-0.65
-0.7
-0.75
-0.8
-0.85
-0.9
-0.95
log Z_EP(μ)
8
6
4
2
0.5
1
1.5
2
2.5
E[x^2] = var + mean^2
E[x] = mean



-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
-2.4
-2.6
-2.8
-3
5
4
3
2
1
0
0
1
2
3
4
5
Sigma(2,2)
Sigma(1,1)



-1
-1.5
-2
-2.5
-3
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Sigma(2,2)
Sigma(1,1)

Clamping a variable

$$Z = \int \mathcal{N}(x; 0, K) \prod_j \Phi(y_j x_j)\,\mathrm{d}x$$

$$= \int \mathcal{N}(x_i^*; 0, k_{ii})\Phi(y_i x_i^*) \left[\int \mathcal{N}\left(x_{\setminus i}\,;\, k_{i\setminus i}\frac{x_i^*}{k_{ii}},\, K_{\setminus i} - \frac{k_{i\setminus i}k_{i\setminus i}^T}{k_{ii}} \right) \prod_{j \neq i} \Phi(y_j x_j)\,\mathrm{d}x_{\setminus i} \right] \mathrm{d}x_i^*$$

$$= \int \mathcal{N}(x_i^*; 0, k_{ii})\Phi(y_i x_i^*)\, Z_{|x_i^*}\,\mathrm{d}x_i^*$$

Clamping a variable

$$Z = \int \mathcal{N}(x; 0, K) \prod_j \Phi(y_j x_j) \, \mathrm{d}x$$

$$= \int \mathcal{N}(x_i^*; 0, k_{ii}) \Phi(y_i x_i^*) \left[\int \mathcal{N}\left(x_{\setminus i} \, ; \, k_{i \setminus i} \frac{x_i^*}{k_{ii}}, \, K_{\setminus i} - \frac{k_{i \setminus i} k_{i \setminus i}^T}{k_{ii}} \right) \prod_{j \neq i} \Phi(y_j x_j) \, \mathrm{d}x_{\setminus i} \right] \mathrm{d}x_i^*$$

$$= \int \mathcal{N}(x_i^*; 0, k_{ii}) \Phi(y_i x_i^*) \, Z_{|x_i^*} \mathrm{d}x_i^*$$

$$Z_{\mathrm{EP}}^{\sim x_i^*} \approx \int \mathcal{N}\left(x_{\setminus i} \, ; \, k_{i \setminus i} \frac{x_i^*}{k_{ii}}, \, K_{\setminus i} - \frac{k_{i \setminus i} k_{i \setminus i}^T}{k_{ii}} \right) \prod_{j \neq i} \Phi(y_j x_j) \, \mathrm{d}x_{\setminus i}$$

When can we show that...?

$$
\begin{aligned}
Z_{\mathrm{EP}} &\leq \int \mathcal{N}(x_i^*; 0, k_{ii})\, \Phi(y_i x_i^*)\, Z_{\mathrm{EP}}^{\sim x_i^*}\, \mathrm{d}x_i^* \\
&\leq \int \mathcal{N}(x_i^*, x_j^*; 0, K_{[ij]})\, \Phi(y_i x_i^*)\, \Phi(y_j x_j^*)\, Z_{\mathrm{EP}}^{\sim x_i^*, x_i^*}\, \mathrm{d}x_i^*\, \mathrm{d}x_j^* \\
&\leq \cdots \\
&\leq \int \mathcal{N}(x^*; 0, K) \prod_j \Phi(y_j x_j^*)\, \mathrm{d}x^* \\
&= Z
\end{aligned}
$$

When can we show that...?

$$Z_{\mathrm{EP}} \leq \int \mathcal{N}(x_i^*; 0, k_{ii})\, \Phi(y_i x_i^*)\, Z_{\mathrm{EP}}^{\sim x_i^*}\, \mathrm{d}x_i^*$$

$$\leq \int \mathcal{N}(x_i^*, x_j^*; 0, K_{[ij]})\, \Phi(y_i x_i^*)\, \Phi(y_j x_j^*)\, Z_{\mathrm{EP}}^{\sim x_i^*, x_i^*}\, \mathrm{d}x_i^*\, \mathrm{d}x_j^*$$

$$\leq \cdots$$

$$\leq \int \mathcal{N}(x^*; 0, K) \prod_j \Phi(y_j x_j^*)\, \mathrm{d}x^*$$

$$= Z$$

When can we show that…?

$$Z_{\mathrm{EP}} \leq \int \mathcal{N}(x_i^*; 0, k_{ii})\, \Phi(y_i x_i^*)\, Z_{\mathrm{EP}}^{\sim x_i^*}\, \mathrm{d}x_i^*$$

$$\mathrm{LHS}(\mu) = \sum_j A_j(\nabla A_j^*(\mu)) - N A(\nabla A^*(\mu)) + \left[N \nabla A^*(\mu) - \sum_j \nabla A_j^*(\mu)\right]^T \mu$$

$$\leq \int \mathcal{N}(x^*; 0, K) \prod_j \Phi(y_j x_j^*)\, \mathrm{d}x^*$$

$$= Z$$

When can we show that…?

$$Z_{\mathrm{EP}} \leq \int \mathcal{N}(x_i^*; 0, k_{ii})\, \Phi(y_i x_i^*)\, Z_{\mathrm{EP}}^{\sim x_i^*}\, \mathrm{d}x_i^*$$

$$\begin{aligned}\mathrm{RHS}(\mu') = \log \int_{\mathcal{X}_i} \exp \Bigg\{ & \log f_i(x_i^*) + \sum_{j \neq i} A_j(\nabla A_j^*(\mu')) - (N-1) A(\nabla A^*(\mu')) + \log \mathcal{N}(x_i^*; 0, k_{ii}) \\ & + \Big[(N-1)\nabla A^*(\mu') - \sum_{j \neq i} \nabla A_j^*(\mu') \Big]^T \mu' \Bigg\} \mathrm{d}x_i^* \ . \qquad (9)\end{aligned}$$

$$= Z$$

max(RHS) ≥ max(LHS)...?

$$\mathrm{RHS}(\mu') = \log \int_{\mathcal{X}_i} \exp\Bigg\{ \log f_i(x_i^*) + \sum_{j \neq i} A_j(\nabla A_j^*(\mu')) - (N-1) A(\nabla A^*(\mu')) + \log \mathcal{N}(x_i^*; 0, k_{ii}) + \Big[(N-1)\nabla A^*(\mu') - \sum_{j \neq i} \nabla A_j^*(\mu') \Big]^T \mu' \Bigg\} \mathrm{d}x_i^* \ . \quad (9)$$

$$\mathrm{LHS}(\mu) = \sum_j A_j(\nabla A_j^*(\mu)) - N A(\nabla A^*(\mu)) + \Big[N \nabla A^*(\mu) - \sum_j \nabla A_j^*(\mu) \Big]^T \mu$$

Getting rid of the integral...

Shortcut 1

Let's write the distribution $\pi(x_i^*) = \frac{1}{Z_i} f_i(x_i^*) \mathcal{N}(x_i^*; 0, k_{ii})$, so that

$$\mathrm{RHS}(\mu') = \log Z_i + \log \int_{\mathcal{X}_i} \pi(x_i^*) \exp\left\{ \sum_{j \neq i} A_j(\nabla A_j^*(\mu')) - (N-1) A(\nabla A^*(\mu')) + \left[(N-1)\nabla A^*(\mu') - \sum_{j \neq i} \nabla A_j^*(\mu') \right]^T \mu' \right\} \mathrm{d}x_i^*$$

Getting rid of the integral...

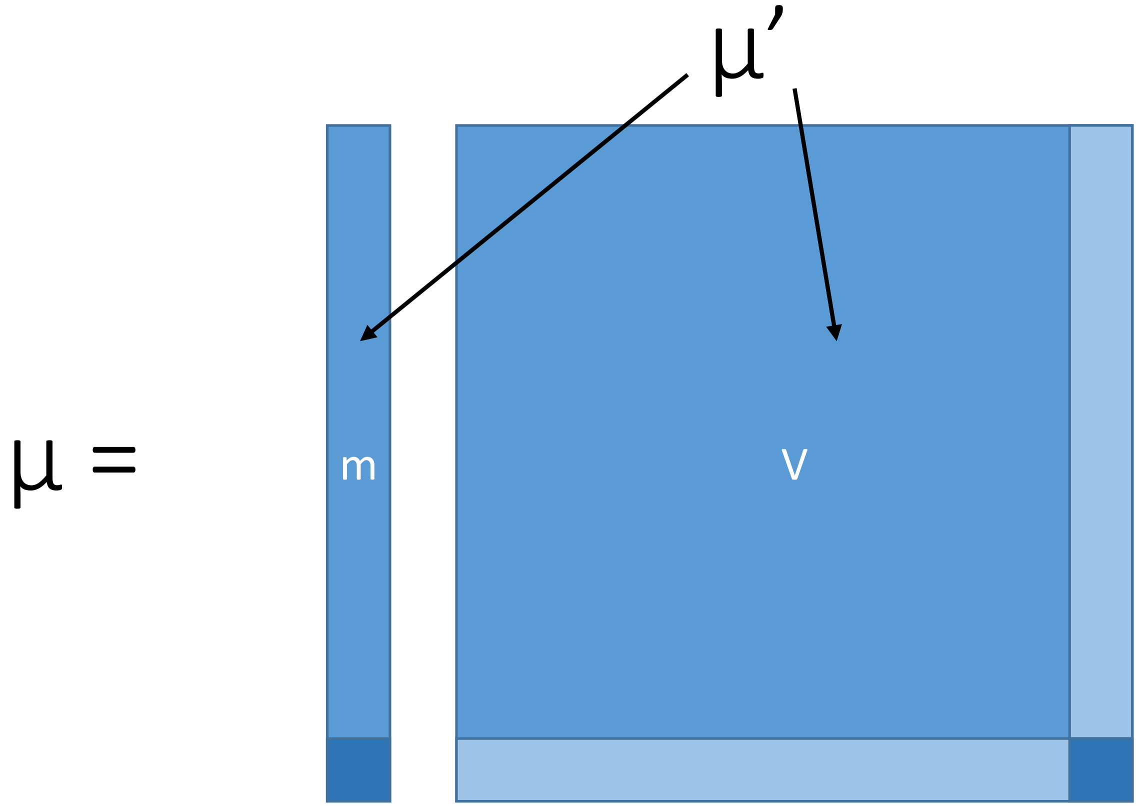
Shortcut 1

Let's write the distribution $\pi(x_i^*) = \frac{1}{Z_i} f_i(x_i^*) \mathcal{N}(x_i^*; 0, k_{ii})$, so that

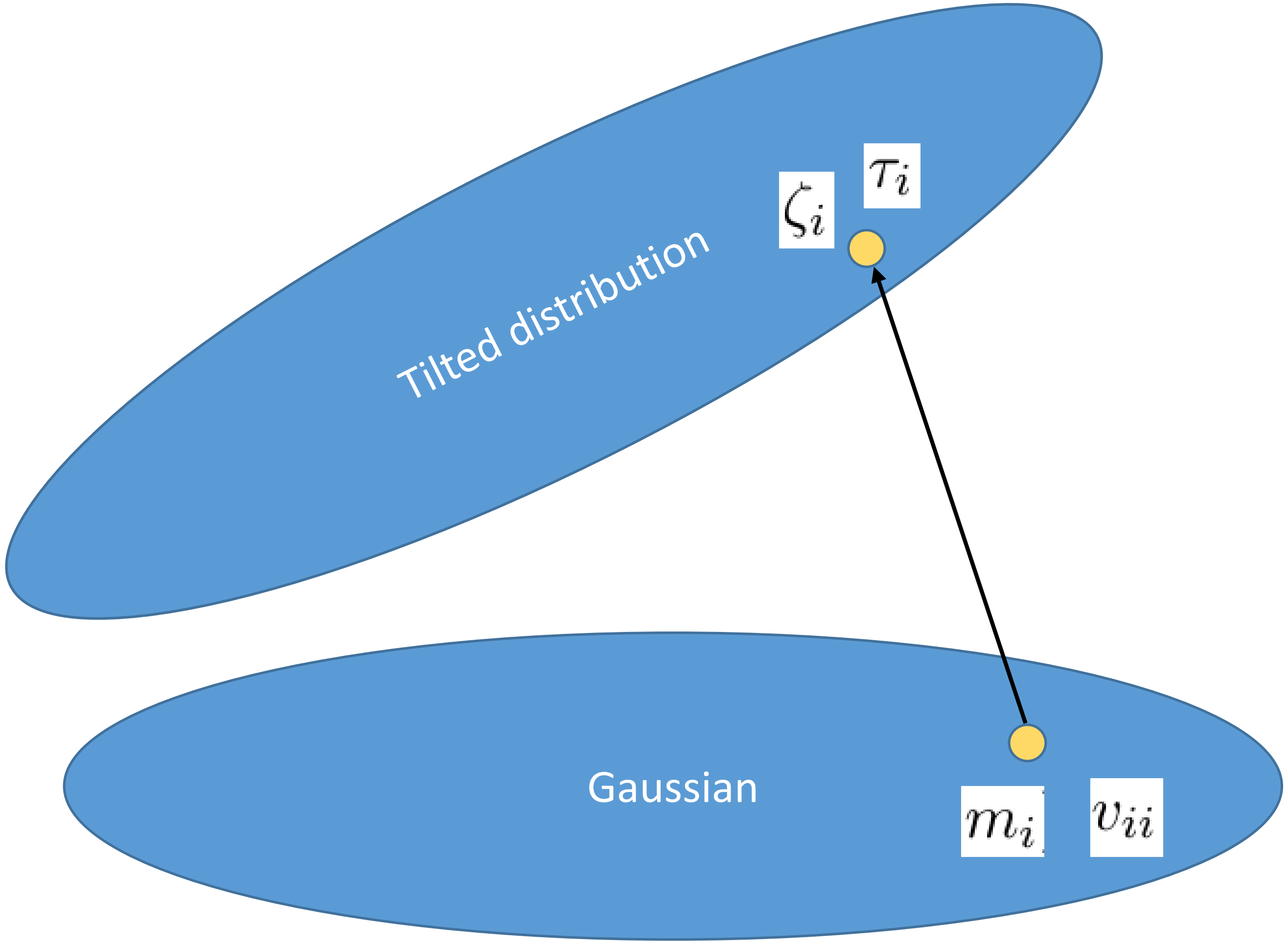
$$\begin{aligned}
\mathrm{RHS}(\mu') &= \log Z_i + \log \int_{\mathcal{X}_i} \pi(x_i^*) \exp\Bigg\{ \sum_{j \neq i} A_j(\nabla A_j^*(\mu')) - (N-1) A(\nabla A^*(\mu')) \\
&\qquad + \Big[(N-1) \nabla A^*(\mu') - \sum_{j \neq i} \nabla A_j^*(\mu') \Big]^T \mu' \Bigg\} \mathrm{d}x_i^* \\
&\geq \log Z_i + \mathbb{E}_{\pi(x_i^*)} \Bigg[\sum_{j \neq i} A_j(\nabla A_j^*(\mu')) - (N-1) A(\nabla A^*(\mu')) \\
&\qquad + \Big[(N-1) \nabla A^*(\mu') - \sum_{j \neq i} \nabla A_j^*(\mu') \Big]^T \mu' \Bigg] \\
&= \log Z_i + \mathbb{E}_{\pi(x_i^*)} \Big[G(\mu'; x_i^*) \Big] \\
&= \mathrm{RHS}^{\dagger}(\mu')
\end{aligned}$$

Filling in the blank...?

$$\log Z_{\text{EP}} = \text{LHS}(\mu) \;\Box\; \text{RHS}^{\dagger}(\mu') \leq \text{RHS}(\mu') \leq \text{RHS}$$



(Reverse) KL projections



$\log Z_{EP}(\mu)$ for GPC

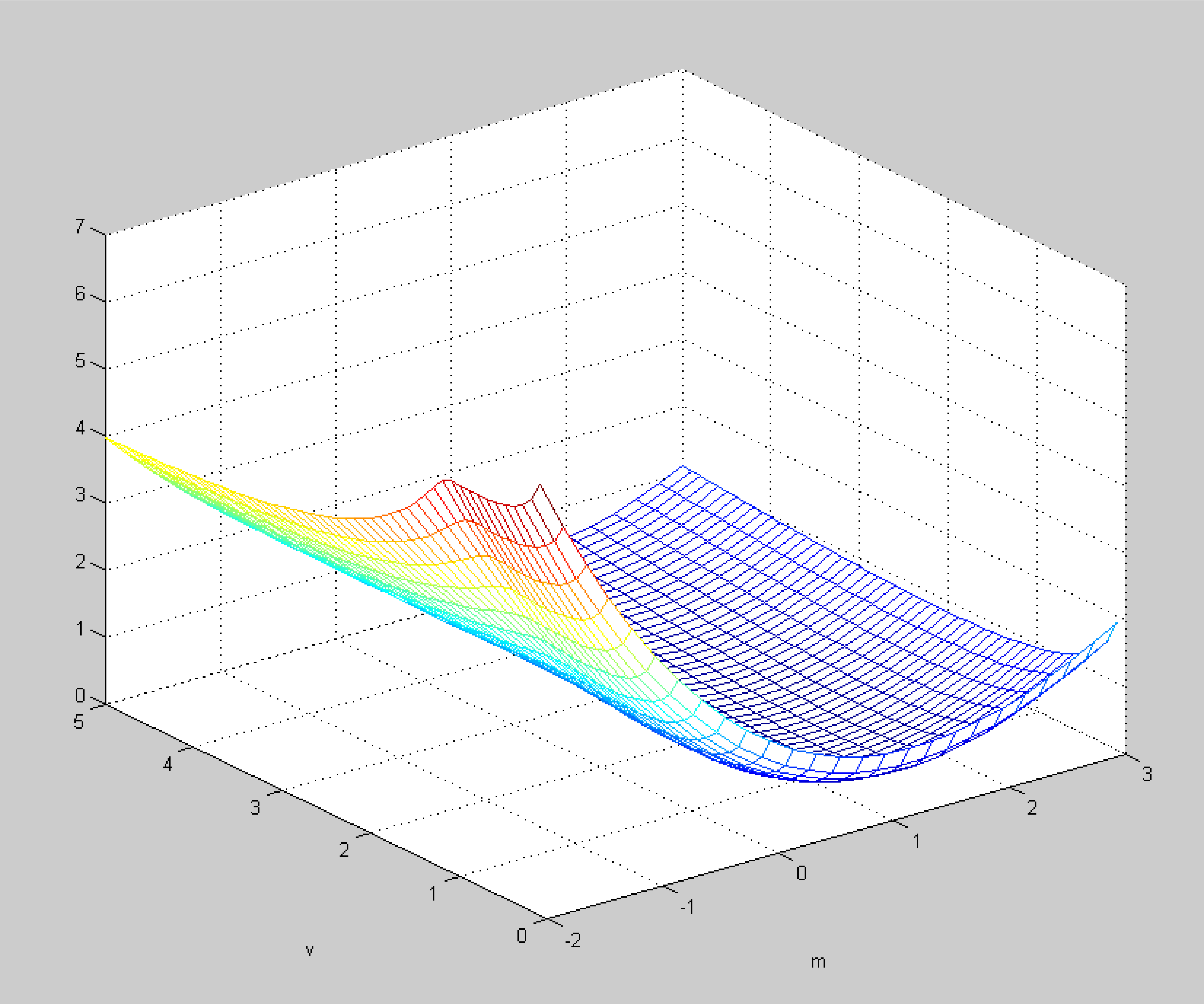
- Fantastic simplification!

$$\mathrm{LHS}(\mu) = \sum_{j \neq 0} \left[\log \Phi\left(\frac{y_j \zeta_j}{\sqrt{1+\tau_j}}\right) + \frac{1}{2}\log \tau_j - \frac{1}{2}\log v_{jj} + \frac{1}{2}\left(\frac{(\zeta_j - m_j)^2 + v_{jj}}{\tau_j}\right)\right] - \frac{1}{2}\log|K| + \frac{1}{2}\log|V| - \frac{1}{2}\mathrm{tr}[K^{-1}(V + mm^T)] \,.$$

RHS – LHS, as function of and

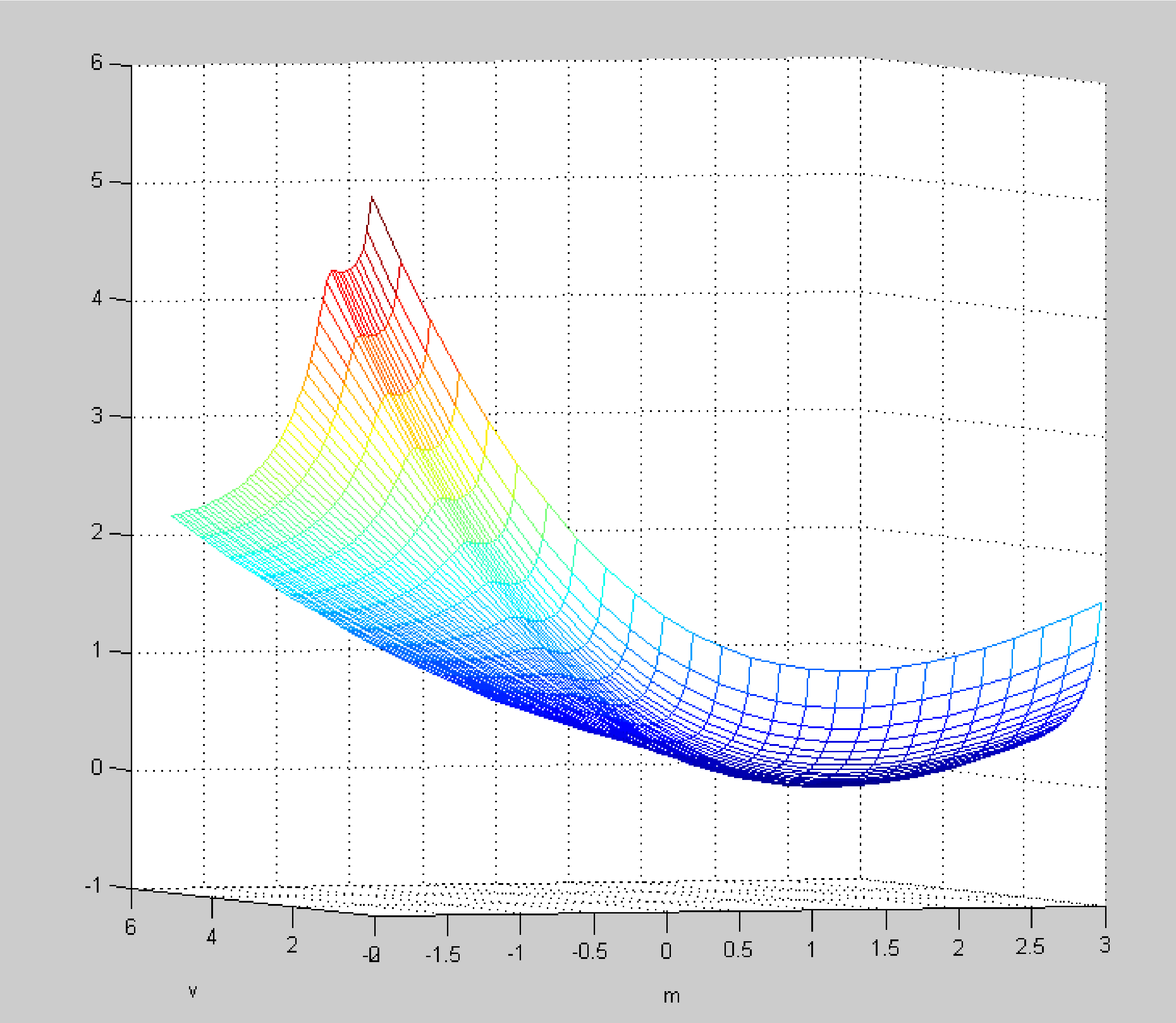
$$
\begin{aligned}
\mathsf{diff}(\mu') &= \mathrm{RHS}^{\dagger}(\mu') - \mathrm{LHS}(\mu') \\
&= \min_{\mu_i} \Bigg\{ \log Z_i - \log \Phi\left(\frac{y_i \zeta_i}{\sqrt{1+\tau_i}}\right) - \frac{1}{2}\log \tau_i + \frac{1}{2}\log v_{ii} - \frac{1}{2}\left(\frac{(\zeta_i - m_i)^2 + v_{ii}}{\tau_i}\right) \\
&\quad + \frac{1}{2}\log k_{ii} - \frac{1}{2}\log\left(v_{ii} - v_{i\backslash i}^T {\color{blue}V_{\backslash i}}^{-1} v_{i\backslash i}\right) \\
&\quad - \frac{1}{2}\mathbb{E}_\pi[{\color{red}x_i^{*}}^2] \left(\frac{k_{i\backslash i}}{k_{ii}}\right)^T \left(K_{\backslash i} - \frac{k_{i\backslash i} k_{i\backslash i}^T}{k_{ii}}\right)^{-1} \left(\frac{k_{i\backslash i}}{k_{ii}}\right) \\
&\quad + \left(\left(\mathbb{E}_\pi[{\color{red}x_i^*}] - m_i\right) {\color{blue}m_{\backslash i}} - v_{i\backslash i}\right)^T \left(K_{\backslash i} - \frac{k_{i\backslash i} k_{i\backslash i}^T}{k_{ii}}\right)^{-1} \left(\frac{k_{i\backslash i}}{k_{ii}}\right) \\
&\quad + \frac{1}{2}(m_i^2 + v_{ii}) \left(k_{ii} - k_{i\backslash i}^T K_{\backslash i}^{-1} k_{i\backslash i}\right)^{-1} \Bigg\} .
\end{aligned}
$$

Shortcut 2



7
6
5
4
3
2
1
0
5
4
3
2
1
0
v
-2
-1
0
1
2
3
m

Counter-example, but only to proof plan ☹



Why?

- No bound at all?
  - Redundancy: John Cunningham’s counter-example

- Too many short-cuts!
  - Jensen’s inequality
  - Keeping μ’ the same on LHS and RHS
    - Counter-example was only to proof plan, and not a real counter-example. No redundancy; one likelihood factor per dimension.

Directions and conclusion

- Still an area where we can’t make definite statements
- Can we characterize log Z_{EP}?

- Whereto now?
  - Use the proof conditions to narrow down the claim