Predictive Coarse-Graining

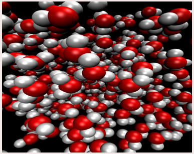
M. Schöberl (TUM), N.Zabaras (Warwick/TUM-IAS),
P.S. Koutsourelakis (TUM)

Predictive Multiscale Materials Modelling
Turing Gateway to Mathematics, Isaac Newton Institute for Mathematical Sciences
Decemeber 1 2015

Problem Definition - Equilibrium Statistical Mechanics

Fine scale

$$p_f(\boldsymbol{x}) \propto e^{-\beta V_f(\boldsymbol{x})}$$



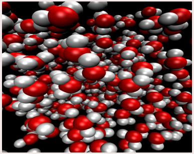
- $\boldsymbol{x}$: *fine*-scale dofs
- $V_f(\boldsymbol{x})$: atomistic potential

Observables: $\mathbb{E}_{p_f}[a] = \int a(\boldsymbol{x})\, p_f(\boldsymbol{x})\, d\boldsymbol{x}$

Problem Definition - Equilibrium Statistical Mechanics

Fine scale

$$p_f(\boldsymbol{x}) \propto e^{-\beta V_f(\boldsymbol{x})}$$

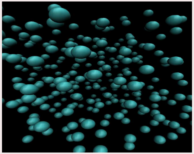


- $\boldsymbol{x}$: *fine*-scale dofs
- $V_f(\boldsymbol{x})$: atomistic potential

Observables: $\mathbb{E}_{p_f}[a] = \int a(\boldsymbol{x})\, p_f(\boldsymbol{x})\, d\boldsymbol{x}$

Coarse scale

$$\boldsymbol{X} = \boldsymbol{R}(\boldsymbol{x}), \quad dim(\boldsymbol{X}) << dim(\boldsymbol{x})$$

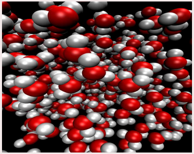


- $\boldsymbol{X}$: *coarse*-scale dofs
- $\boldsymbol{R}$: *restriction* operator (*fine* $\rightarrow$ *coarse*)

Problem Definition - Equilibrium Statistical Mechanics

Fine scale

$$p_f(\boldsymbol{x}) \propto e^{-\beta V_f(\boldsymbol{x})}$$

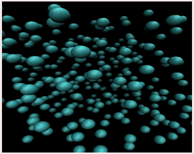


- $\boldsymbol{x}$: *fine*-scale dofs
- $V_f(\boldsymbol{x})$: atomistic potential

Observables: $\mathbb{E}_{p_f}[a] = \int a(\boldsymbol{x})\, p_f(\boldsymbol{x})\, d\boldsymbol{x}$

Coarse scale

$$\boldsymbol{X} = \boldsymbol{R}(\boldsymbol{x}), \quad dim(\boldsymbol{X}) << dim(\boldsymbol{x})$$



- $\boldsymbol{X}$: *coarse*-scale dofs
- $\boldsymbol{R}$: *restriction* operator ($fine \rightarrow coarse$)

Goal:

How can one simulate $\boldsymbol{X}$ and still *predict* $\mathbb{E}_{p_f}[a]$?

Problem Definition - Equilibrium Statistical Mechanics

- Suppose the observable of interest $a(\boldsymbol{x})$ depends on $\boldsymbol{X}$ i.e.:

$$a(\boldsymbol{x}) = A(\boldsymbol{X}) = A(\boldsymbol{R}(\boldsymbol{x}))$$

- Then:

$$\begin{aligned}
\mathbb{E}_{p_f}[a] &= \int a(\boldsymbol{x})\, p_f(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int A(\boldsymbol{R}(\boldsymbol{x}))\, p_f(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int \left(\int A(\boldsymbol{X})\delta(\boldsymbol{X}-\boldsymbol{R}(\boldsymbol{x}))\, d\boldsymbol{X}\right)\ p_f(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int A(\boldsymbol{X}) \left(\int \delta(\boldsymbol{X}-\boldsymbol{R}(\boldsymbol{x}))\, p_f(\boldsymbol{x})\, d\boldsymbol{x}\right)\ d\boldsymbol{X} \\
&= \int A(\boldsymbol{X})\, p_c(\boldsymbol{X})\, d\boldsymbol{X}
\end{aligned}$$

where p_c is the (marginal) PDF of the CG variables $\boldsymbol{X}$:

$$p_c(\boldsymbol{X}) = \int \delta(\boldsymbol{X}-\boldsymbol{R}(\boldsymbol{x}))\, p_f(\boldsymbol{x})\, d\boldsymbol{x} \propto e^{-\beta V_c(\boldsymbol{X})}$$

and the CG potential $V_c(\boldsymbol{X})$ is the potential of mean force (PMF) w.r.t. $\boldsymbol{X}$:

$$V_c(\boldsymbol{X}) = -\beta^{-1} \log \int \delta(\boldsymbol{X}-\boldsymbol{R}(\boldsymbol{x}))\, p_f(\boldsymbol{x})\, d\boldsymbol{x}$$

Existing Methods

- Free-energy methods (for low-dimensional $\boldsymbol{X}$) [Lelièvre et al 2010]
- Lattice systems [Katsoulakis 2003], Soft matter]Peter & Kremer 2010]
- Inversion-based methods: Iterative Boltzmann Inversion [Reith et al. (2003)], Inverse Monte Carlo [Lyubartsev & Laaksonen (1995), Soper (1996)], Molecular RG-CG [Savelyev & Papoian 2009]
- Variational methods: Multiscale CG [Izvekov &et al. (2005), Noid et al. (2007)] , Relative Entropy [Shell (2008)], Ultra-Coarse-Graining [Dama et al. 2013]

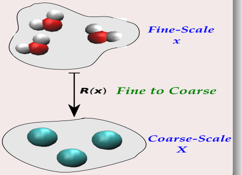
Motivation

- What are *good* coarse-grained variables $\boldsymbol{X}$ (how many, what is fine-to-coarse mapping $\boldsymbol{R}$)
- What is the right CG potential (or CG model)?
- How much information is lost during coarse-graining and how does this affect predictive uncertainty?
- Given finite simulation data at the fine-scale, how (un)certain can we be in our predictions?
- Can one use the same CG variables $\boldsymbol{X}$ to make predictions about observables $a(\boldsymbol{x}) \neq A(\boldsymbol{X})$?

Motivation

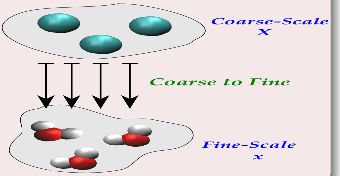
Existing methods

$$\underbrace{p_f(\boldsymbol{x})}_{fine} \xrightarrow{\boldsymbol{R}(\boldsymbol{x})=\boldsymbol{X}} \underbrace{\bar{p}_c(\boldsymbol{X})}_{coarse}$$



Proposed (Generative model)

$$\underbrace{p_c(\boldsymbol{X})}_{coarse} \xrightarrow{p_{cf}(\boldsymbol{x}|\boldsymbol{X})} \underbrace{\bar{p}_f(\boldsymbol{x}) = \int p_{cf}(\boldsymbol{x}|\boldsymbol{X})\, p_c(\boldsymbol{X})\, d\boldsymbol{X}}_{fine}$$



Notes

- No restriction operator (fine-to-coarse $\boldsymbol{R}(\boldsymbol{x}) = \boldsymbol{X}$).
- A probabilistic *coarse-to-fine* map $p_{cf}(\boldsymbol{x}|\boldsymbol{X})$ is prescribed
- The coarse model $p_c(\boldsymbol{X})$ is not the marginal of $\boldsymbol{X}$ (given $\boldsymbol{R}(\boldsymbol{x}) = \boldsymbol{X}$)

Motivation

Relative Entropy CG [Shell, 2008]

- A fine$\rightarrow$coarse map $\boldsymbol{R}$ and an (approximate) CG density $\bar{p}_c(\boldsymbol{X})$ imply:

$$\bar{p}_f(\boldsymbol{x}) = \frac{\delta(\boldsymbol{R}(\boldsymbol{x}) - \boldsymbol{X})}{\Omega(\boldsymbol{R}(\boldsymbol{x}))} \bar{p}_c(\boldsymbol{R}(\boldsymbol{x}))$$

  where: $\Omega(\boldsymbol{R}(\boldsymbol{x})) = \int \delta(\boldsymbol{R}(\boldsymbol{x}) - \boldsymbol{X}) \, d\boldsymbol{x}$

- Find $\bar{p}_c(\boldsymbol{X})$ that minimizes KL-divergence between $p_f(\boldsymbol{x})$ (exact) and $\bar{p}_f(\boldsymbol{x})$ (approximate):

$$KL(p_f(\boldsymbol{x}) || \bar{p}_f(\boldsymbol{x})) = \underbrace{KL(p_c(\boldsymbol{X}) || \bar{p}_c(\boldsymbol{X}))}_{\text{inf. loss due to error in PMF}} + \underbrace{S_{map}(\boldsymbol{R})}_{\text{inf. loss due to map } \boldsymbol{R}}$$

  where $S_{map} = \int p_f(\boldsymbol{x}) \log \Omega(\boldsymbol{R}(\boldsymbol{x})) \, d\boldsymbol{x}$

Motivation

Proposed Probabilistic Generative model

- A probabilistic *coarse*$\rightarrow$*fine* map $p_{cf}(\boldsymbol{x}|\boldsymbol{X})$ and a CG model $p_c(\boldsymbol{X})$, imply:

$$\bar{p}_f(\boldsymbol{x}) = \int p_{cf}(\boldsymbol{x}|\boldsymbol{X})\ p_c(\boldsymbol{X})\ d\boldsymbol{X}$$

- Find $p_{cf}(\boldsymbol{x}|\boldsymbol{X})$ and $p_c(\boldsymbol{X})$ that minimize:

$$KL(p_f(\boldsymbol{x})||\bar{p}_f(\boldsymbol{x})) = -\int p_f(\boldsymbol{x}) \log \frac{\int p_{cf}(\boldsymbol{x}|\boldsymbol{X})\ p_c(\boldsymbol{X})\ d\boldsymbol{X}}{p_f(\boldsymbol{x})}\ d\boldsymbol{x}$$

Learning

Proposed Probabilistic Generative model

- Parametrize: $\underbrace{p_c(\boldsymbol{X}|\theta_c)}_{\text{coarse model}}, \quad \underbrace{p_{cf}(\boldsymbol{x}|\boldsymbol{X},\ \theta_{cf})}_{\text{coarse}\rightarrow\text{fine map}}$
- Optimize:

$$
\begin{aligned}
& \min_{\theta_c,\theta_{cf}} KL(p_f(\boldsymbol{x}) \,||\, \bar{p}_f(\boldsymbol{x}|\theta_c,\theta_{cf})) \\
\leftrightarrow & \min_{\theta_c,\theta_{cf}} - \int p_f(\boldsymbol{x}) \log \frac{\int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\theta_{cf})\ p_c(\boldsymbol{X}|\theta_c)\ d\boldsymbol{X}}{p_f(\boldsymbol{x})}\ d\boldsymbol{x} \\
\leftrightarrow & \max_{\theta_c,\theta_{cf}} \int p_f(\boldsymbol{x}) \left(\log \int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\theta_{cf})\ p_c(\boldsymbol{X}|\theta_c)\ d\boldsymbol{X}\right)\ d\boldsymbol{x} \\
\leftrightarrow & \max_{\theta_c,\theta_{cf}} \sum_{i=1}^{N} \log \int p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X},\theta_{cf})\ p_c(\boldsymbol{X}|\theta_c)\ d\boldsymbol{X} \\
\leftrightarrow & \max_{\theta_c,\theta_{cf}} \mathcal{L}(\theta_c,\theta_{cf}), \quad \text{(MLE)}
\end{aligned}
$$

- MAP estimate: $\max_{\theta_c,\theta_{cf}} \mathcal{L}(\theta_c,\theta_{cf}) + \underbrace{\log p(\theta_c,\theta_{cf})}_{log-prior}$
- Fully Bayesian i.e. posterior: $p(\theta_c,\theta_{cf}|\boldsymbol{x}^{(1:N)}) \propto exp\{\mathcal{L}(\theta_c,\theta_{cf})\ p(\theta_c,\theta_{cf})\}$

Learning

Proposed Probabilistic Generative model

- Parametrize: $\underbrace{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}_{\text{coarse model}}, \quad \underbrace{p_{cf}(\boldsymbol{x}|\boldsymbol{X},\ \boldsymbol{\theta}_{cf})}_{\text{coarse}\rightarrow\text{fine map}}$

- Optimize:

$$
\begin{aligned}
& \min_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} KL(p_f(\boldsymbol{x}) \mid\mid \bar{p}_f(\boldsymbol{x}|\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})) \\
\leftrightarrow & \min_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} -\int p_f(\boldsymbol{x}) \log \frac{\int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}}{p_f(\boldsymbol{x})}\ d\boldsymbol{x} \\
\leftrightarrow & \max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \int p_f(\boldsymbol{x}) \left(\log \int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}\right)\ d\boldsymbol{x} \\
\leftrightarrow & \max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \sum_{i=1}^{N} \log \int p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X},\boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X} \\
\leftrightarrow & \max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \ \mathcal{L}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}), \quad \text{(MLE)}
\end{aligned}
$$

- MAP estimate: $\max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \ \mathcal{L}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}) + \underbrace{\log p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})}_{log-prior}$

- Fully Bayesian i.e. posterior: $p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)}) \propto exp\{\mathcal{L}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})\ p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})\}$

Learning

Proposed Probabilistic Generative model

- Parametrize: $\underbrace{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}_{\text{coarse model}}, \quad \underbrace{p_{cf}(\boldsymbol{x}|\boldsymbol{X},\ \boldsymbol{\theta}_{cf})}_{\text{coarse}\rightarrow\text{fine map}}$

- Optimize:

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} KL(p_f(\boldsymbol{x})\ ||\ \bar{p}_f(\boldsymbol{x}|\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}))\\
\leftrightarrow &\min_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} -\int p_f(\boldsymbol{x}) \log \frac{\int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}}{p_f(\boldsymbol{x})}\ d\boldsymbol{x}\\
\leftrightarrow &\max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \int p_f(\boldsymbol{x}) \left(\log \int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}\right)\ d\boldsymbol{x}\\
\leftrightarrow &\max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \sum_{i=1}^{N} \log \int p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X},\boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}\\
\leftrightarrow &\max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}}\ \mathcal{L}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}), \quad \text{(MLE)}
\end{aligned}
$$

- MAP estimate: $\max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}}\ \mathcal{L}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}) + \underbrace{\log p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})}_{log-prior}$

- Fully Bayesian i.e. posterior: $p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)}) \propto exp\{\mathcal{L}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})\ p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})\}$

Prediction

Probabilistic Prediction

- For an observable $a(\boldsymbol{x})$ (reconstruction, [Katsoulakis et al. 2006, Trashorras et al. 2010]):

$$\begin{aligned} \mathbb{E}_{p_f}[a] & \approx \mathbb{E}_{\bar{p}_f}[a|\overbrace{\boldsymbol{x}^{(1:N)}}^{data}] \\ & = \int a(\boldsymbol{x})\, \bar{p}_f(\boldsymbol{x}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{x} \end{aligned}$$

- For each θ_c, θ_{cf} from the posterior, one gets an estimate of the observable
- Not just point-estimates anymore, but whole distributions!

Prediction

Probabilistic Prediction

- For an observable $a(\boldsymbol{x})$ (reconstruction, [Katsoulakis et al. 2006, Trashorras et al. 2010]):

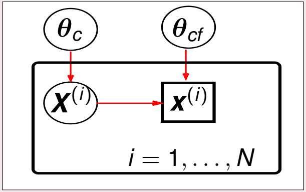
$$
\begin{aligned}
\mathbb{E}_{p_f}[a] &\approx \mathbb{E}_{\bar{p}_f}[a|\overbrace{\boldsymbol{x}^{(1:N)}}^{data}] \\
&= \int a(\boldsymbol{x})\ \bar{p}_f(\boldsymbol{x}|\boldsymbol{x}^{(1:N)})\ d\boldsymbol{x} \\
&= \int a(\boldsymbol{x})\ \bar{p}_f(\boldsymbol{x},\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)})\ d\boldsymbol{\theta}_c\ d\boldsymbol{\theta}_{cf}\ d\boldsymbol{x} \\
&= \int a(\boldsymbol{x})\ \bar{p}_f(\boldsymbol{x}|\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)})\ d\boldsymbol{\theta}_c\ d\boldsymbol{\theta}_{cf}\ d\boldsymbol{x} \\
&= \int a(\boldsymbol{x})\ \left(\int p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf})p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}\right)p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)})\ d\boldsymbol{\theta}_c\ d\boldsymbol{\theta}_{cf}\ d\boldsymbol{x} \\
&= \int \left(\int a(\boldsymbol{x})p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf})p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}d\boldsymbol{x}\right)p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)})\ d\boldsymbol{\theta}_c\ d\boldsymbol{\theta}_{cf} \\
&= \int \hat{a}(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf})\ \underbrace{p(\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)})}_{posterior}\ d\boldsymbol{\theta}_c\ d\boldsymbol{\theta}_{cf}
\end{aligned}
$$

- For each $\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf}$ from the posterior, one gets an estimate of the observable
- Not just point-estimates anymore, but whole distributions!

Learning/Inference

MCMC-SA (Expectation-Maximization) [Gu & Kong 1998]

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf}) &= \sum_{i=1}^N \log \int p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\, p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)\, d\boldsymbol{X}^{(i)} \\
&= \sum_{i=1}^N \log \int q(\boldsymbol{X}^{(i)}) \frac{p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\, p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)}{q(\boldsymbol{X}^{(i)})}\, d\boldsymbol{X}^{(i)} \\
&\geq \sum_{i=1}^N \int q(\boldsymbol{X}^{(i)}) \log \frac{p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\, p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)}{q(\boldsymbol{X}^{(i)})}\, d\boldsymbol{X}^{(i)} \\
&= \sum_{i=1}^N \mathcal{F}(q(\boldsymbol{X}^{(i)}), \boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})
\end{aligned}
$$



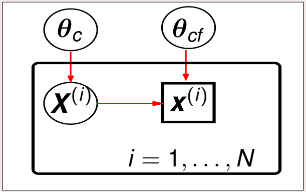
- E-step (given $(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})$) : Sample each $\boldsymbol{X}^{(i)}$ from $q^{opt}(\boldsymbol{X}^{(i)})$:

$$q^{opt}(\boldsymbol{X}^{(i)}) \propto p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\, p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)$$

- M-step: Compute gradients $\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_c} \mathcal{F}$, $\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_{cf}} \mathcal{F}$, (and Hessian) and update $(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})$

MCMC-SA (Expectation-Maximization) [Gu & Kong 1998]

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf}) &= \sum_{i=1}^N \log \int p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)\ d\boldsymbol{X}^{(i)} \\ &= \sum_{i=1}^N \log \int q(\boldsymbol{X}^{(i)}) \frac{p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)}{q(\boldsymbol{X}^{(i)})}\ d\boldsymbol{X}^{(i)} \\ &\geq \sum_{i=1}^N \int q(\boldsymbol{X}^{(i)}) \log \frac{p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)}{q(\boldsymbol{X}^{(i)})}\ d\boldsymbol{X}^{(i)} \\ &= \sum_{i=1}^N \mathcal{F}(q(\boldsymbol{X}^{(i)}),\ \boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})\end{aligned}$$



- E-step (given $(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})$) : Sample each $\boldsymbol{X}^{(i)}$ from $q^{opt}(\boldsymbol{X}^{(i)})$:

$$q^{opt}(\boldsymbol{X}^{(i)}) \propto p_{cf}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf})\ p_c(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_c)$$

- M-step: Compute gradients $\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_c} \mathcal{F}$, $\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_{cf}} \mathcal{F}$, (and Hessian) and update $(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})$

$$\text{Robbins-Monro:} \quad \begin{aligned}&\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \alpha_t \nabla_{\boldsymbol{\theta}} \mathcal{F} \\ &\sum \alpha_t = \infty, \sum \alpha_t^2 < \infty\end{aligned}$$

Learning/Inference

- For exponential-family distributions:

$$
\begin{aligned}
p_c(\boldsymbol{X}|\boldsymbol{\theta}_c) &= \exp\{\boldsymbol{\theta}_c^T \phi(\boldsymbol{X}) - A(\boldsymbol{\theta}_c)\} & \left(e^{A(\boldsymbol{\theta}_c)} = \int e^{\boldsymbol{\theta}_c^T \phi(\boldsymbol{X})}\, d\boldsymbol{X}\right) \\
p_{cf}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{cf}) &= \exp\{\boldsymbol{\theta}_{cf}^T \psi(\boldsymbol{x}, \boldsymbol{X}) - B(\boldsymbol{X}, \boldsymbol{\theta}_{cf})\} & \left(e^{B(\boldsymbol{X}, \boldsymbol{\theta}_{cf})} = \int e^{\boldsymbol{\theta}_{cf}^T \psi(\boldsymbol{x}, \boldsymbol{X})}\, d\boldsymbol{x}\right)
\end{aligned}
$$

- Gradients:

$$\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_c} \mathcal{F} = \sum_{i=1}^N < \phi(\boldsymbol{X}^{(i)}) >_{q(\boldsymbol{X}^{(i)})} - N < \phi(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}$$

$$\left(\nabla_{\boldsymbol{\theta}_c} KL = \sum_{i=1}^N \phi(\boldsymbol{R}(\boldsymbol{x}^{(i)})) - N < \phi(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)} \text{ Relative Entropy [Shell 2008]} \right)$$

$$\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_{cf}} \mathcal{F} = \sum_{i=1}^N (< \psi(\boldsymbol{x}^{(i)}, \boldsymbol{X}^{(i)}) >_{q(\boldsymbol{X}^{(i)})} - < \psi(\boldsymbol{x}, \boldsymbol{X}^{(i)}) >_{p_{cf}(\boldsymbol{x}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf}) q(\boldsymbol{X}^{(i)})}$$

- Hessian:

$$
\left.
\begin{aligned}
\sum_{i=1}^N \nabla^2_{\boldsymbol{\theta}_c} \mathcal{F} &= -N\, Cov_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}[\phi(\boldsymbol{X})] \\
\sum_{i=1}^N \nabla^2_{\boldsymbol{\theta}_{cf}} \mathcal{F} &= -\sum_{i=1}^N Cov_{p_{cf}(\boldsymbol{x}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{cf}) q(\boldsymbol{X}^{(i)})}[\psi(\boldsymbol{x}, \boldsymbol{X}^{(i)})]
\end{aligned}
\right\} \longrightarrow \text{Concave}
$$

Learning/Inference

- For exponential-family distributions:

$$p_c(\boldsymbol{X}|\boldsymbol{\theta}_c) = \exp\{\boldsymbol{\theta}_c^T \phi(\boldsymbol{X}) - A(\boldsymbol{\theta}_c)\} \qquad (e^{A(\boldsymbol{\theta}_c)} = \int e^{\boldsymbol{\theta}_c^T \phi(\boldsymbol{X})}\, d\boldsymbol{X})$$
$$p_{cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{cf}) = \exp\{\boldsymbol{\theta}_{cf}^T \psi(\boldsymbol{x},\boldsymbol{X}) - B(\boldsymbol{X},\boldsymbol{\theta}_{cf})\} \quad (e^{B(\boldsymbol{X},\boldsymbol{\theta}_{cf})} = \int e^{\boldsymbol{\theta}_{cf}^T \psi(\boldsymbol{x},\boldsymbol{X})}\, d\boldsymbol{x})$$

- Gradients:

$$\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_c}\mathcal{F} = \sum_{i=1}^N < \phi(\boldsymbol{X}^{(i)}) >_{q(\boldsymbol{X}^{(i)})} - N < \phi(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}$$
$$\left(\nabla_{\boldsymbol{\theta}_c} KL = \sum_{i=1}^N \phi(\boldsymbol{R}(\boldsymbol{x}^{(i)})) - N < \phi(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)} \text{ Relative Entropy [Shell 2008]} \right)$$
$$\sum_{i=1}^N \nabla_{\boldsymbol{\theta}_{cf}}\mathcal{F} = \sum_{i=1}^N (< \psi(\boldsymbol{x}^{(i)},\boldsymbol{X}^{(i)}) >_{q(\boldsymbol{X}^{(i)})} - < \psi(\boldsymbol{x},\boldsymbol{X}^{(i)}) >_{p_{cf}(\boldsymbol{x}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{cf})q(\boldsymbol{X}^{(i)})}$$

- Hessian:

$$\left.\begin{aligned} \sum_{i=1}^N \nabla^2_{\boldsymbol{\theta}_c}\mathcal{F} &= -N\, Cov_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}[\phi(\boldsymbol{X})] \\ \sum_{i=1}^N \nabla^2_{\boldsymbol{\theta}_{cf}}\mathcal{F} &= -\sum_{i=1}^N Cov_{p_{cf}(\boldsymbol{x}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{cf})q(\boldsymbol{X}^{(i)})}[\psi(\boldsymbol{x},\boldsymbol{X}^{(i)})] \end{aligned}\right\} \longrightarrow \text{Concave}$$

Learning/Inference

- MAP-estimates:

$$\max_{\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf}} \mathcal{L}(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf}) + \underbrace{\log p(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})}_{log-prior}$$

- *Approximate* Bayesian posterior using *Laplace approximation*

$$p(\boldsymbol{\theta}_{cf} | \boldsymbol{x}^{(1:N)}) \approx \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$$

Figure : Laplace approximation

where:

- $\boldsymbol{\mu} = \boldsymbol{\theta}_{cf,MAP}$
- $\boldsymbol{S}^{-1} = -\sum_{i=1}^{N} \nabla^2_{\boldsymbol{\theta}_{cf}} \mathcal{F}$

Learning/Inference

- MAP-estimates:

$$\max_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_{cf}} \mathcal{L}(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf}) + \underbrace{\log p(\boldsymbol{\theta}_c, \boldsymbol{\theta}_{cf})}_{log-prior}$$

- *Approximate* Bayesian posterior using *Laplace approximation*

$$p(\boldsymbol{\theta}_{cf}|\boldsymbol{x}^{(1:N)}) \approx \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$$

where:

- $\boldsymbol{\mu} = \boldsymbol{\theta}_{cf,MAP}$
- $\boldsymbol{S}^{-1} = -\sum_{i=1}^{N} \nabla^2_{\boldsymbol{\theta}_{cf}} \mathcal{F}$

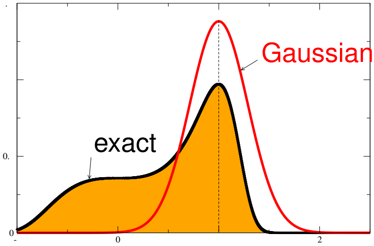


Figure : Laplace approximation

Ising Model - Fine-Scale

Fine-scale variables $x_i \in \{-1, 1\}$ following $p_f(\boldsymbol{x}) \propto e^{-\beta V_f(\boldsymbol{x})}$:

Fine-scale potential

$$V_f(\boldsymbol{x}) = -\frac{1}{2} \sum_{k=1}^{L_f} J_k \sum_{|i-j|=k} x_i x_j - \mu \sum_{i=1}^{n_f} x_i$$

with $i, j \in \{1, \dots, n_f\}$ having n_f lattice sites.

- Maximal interactions of L_f sites apart are regarded in the potential.
- $|i - j| = k$ neighbors over k-sites apart
- J_k, strength of the k-th interaction.

with J_k following a power law for a given overall strength J_0 and exponent a, $J_k = \frac{K}{Lk^a}$ with,

$$K = J_0 L^{1-a} \sum_{k=1}^{L} k^{-a}$$

in order to normalize the interaction strength [Katsoulakis et al 2007].

Ising Model - *Coarse* $\rightarrow$ *Fine* map

Coarse-to-fine mapping $p_{cf}(\boldsymbol{x}|\boldsymbol{X}, \theta_{cf})$

$$p_{cf}(\boldsymbol{x}|\boldsymbol{X}, \theta_{cf}) \prod_{parent\ r} \prod_{child\ s} \theta_{cf}^{\frac{1+x_{r,s}X_r}{2}} (1-\theta_{cf})^{\frac{1+x_{r,s}X_r}{2}}$$

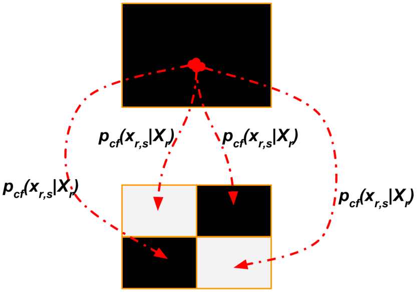
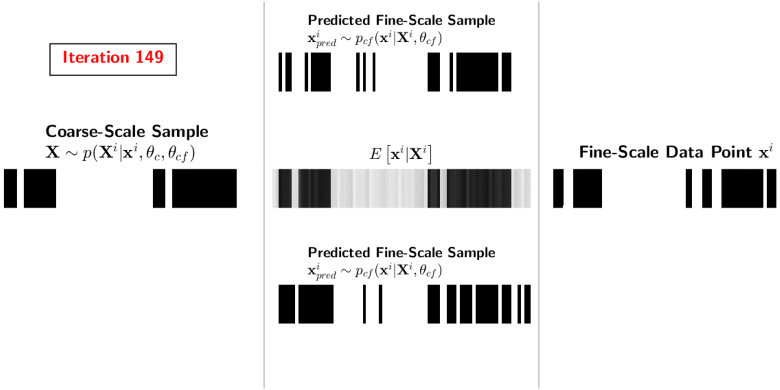


Figure : Probabilistic *coarse* $\rightarrow$ *fine* map

Illustration



Overview of results

Overview:

- Comparison with Relative Entropy at various μ - magnetization
- Probabilistic predictions:
  - with various amounts of data
  - various levels of coarse-graining $dim(\boldsymbol{x})/dim(\boldsymbol{X})$
- Model Selection

Comparison

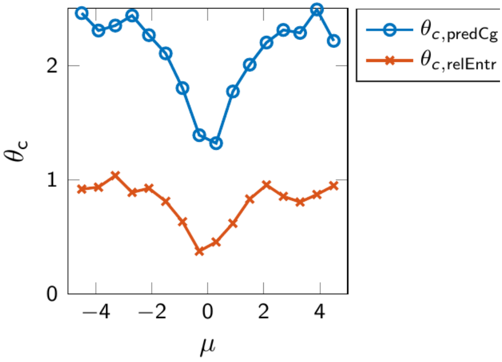


Figure : $\boldsymbol{\theta}_c$

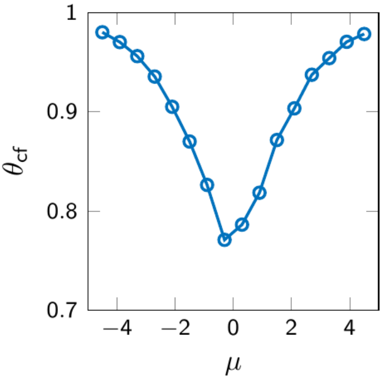


Figure : $\boldsymbol{\theta}_{cf}$

$$V_f(\boldsymbol{x}) = -\frac{1}{2}\sum_{k=1}^{L_f} J_k \sum_{|i-j|=k} x_i x_j - \mu \sum_{i=1}^{n_f} x_i$$

Comparison of predicted magnetization

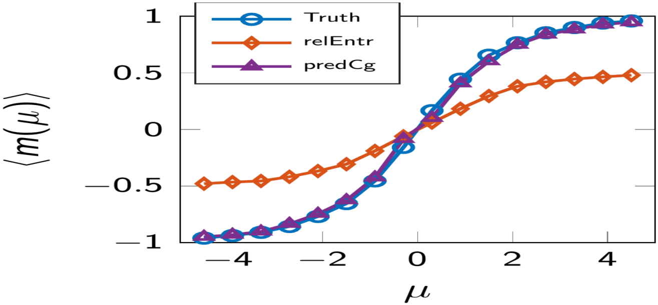


Figure : Predicted magnetization $< m(\mu) >$ with Rel. Entropy (relEntr) and proposed method (predCg)

Fine-to-coarse map in Rel. Entropy

$$X_r = \begin{cases} +1, & \frac{1}{S}\sum_s^S x_{r,s} > 0 \\ -1, & \frac{1}{S}\sum_s^S x_{r,s} < 0 \\ U(-1,+1), & \text{otherwise} \end{cases}$$

Probabilistic Predictitons

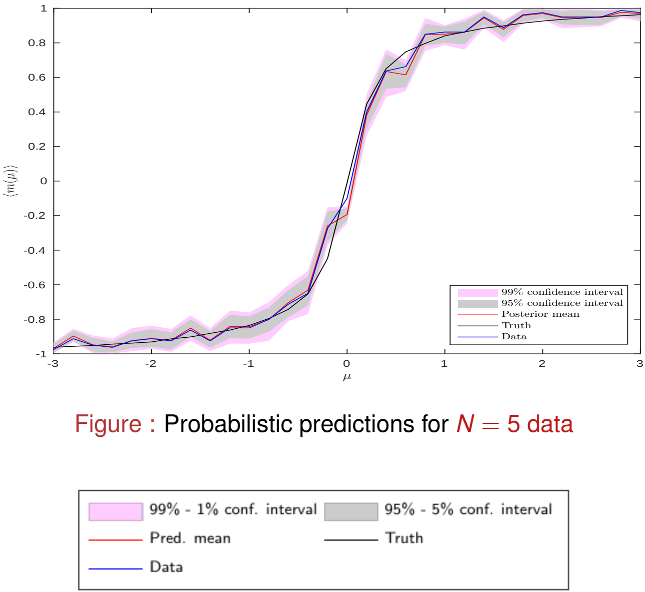


Figure : Probabilistic predictions for $N = 5$ data

Probabilistic Predictitons

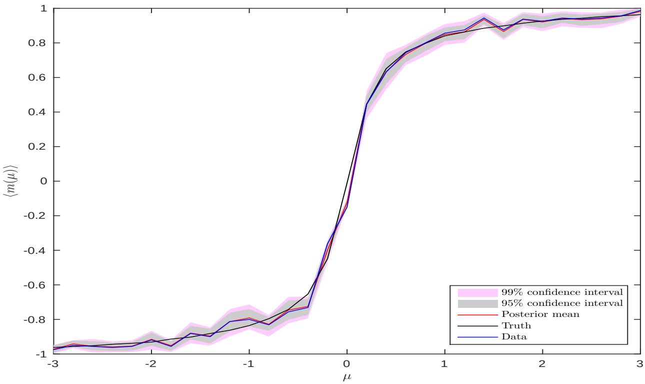
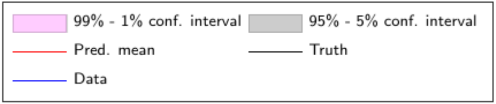


Figure : Probabilistic predictions for $N = 10$ data



Probabilistic Predictitons

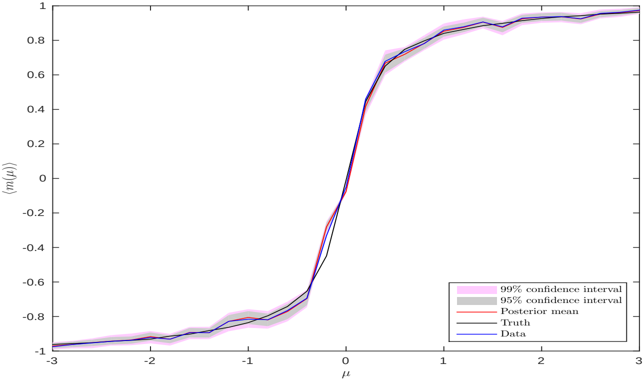
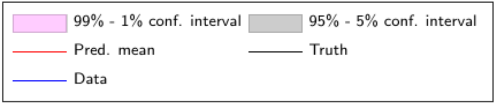


Figure : Probabilistic predictions for $N = 20$ data



Probabilistic Predictitons

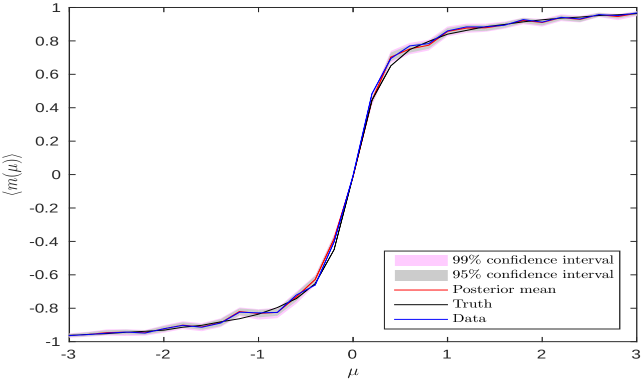
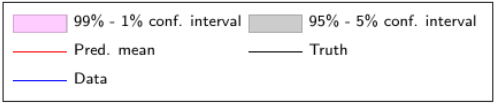


Figure : Probabilistic predictions for $N = 50$ data



Probabilistic Predictitons

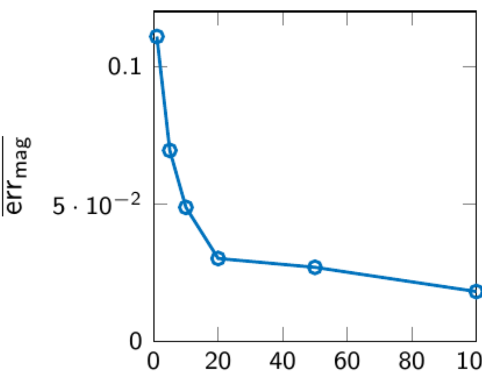


Figure : Error in magnetization as a function of training data N

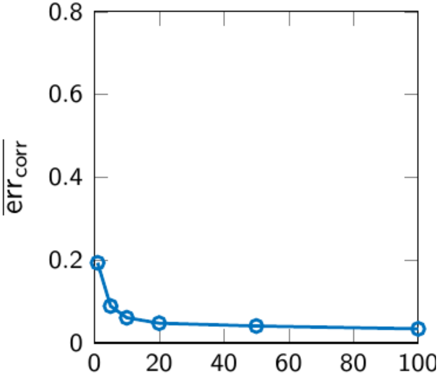


Figure : Error in correlation as a function of training data N

Effect of Coarse-Graining level

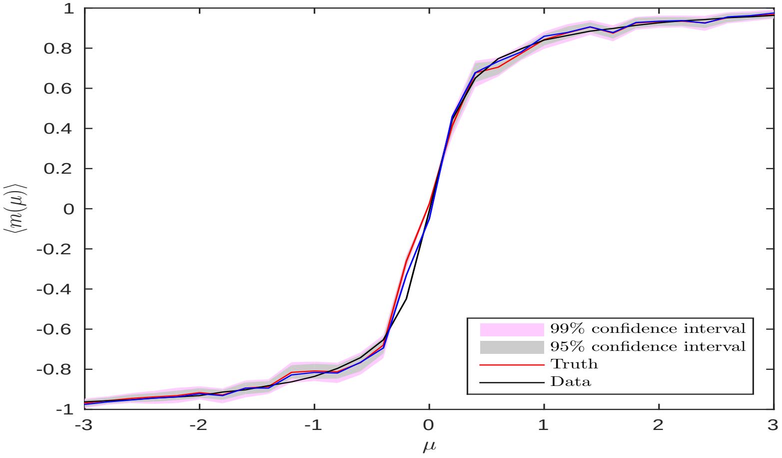


Figure : Probabilistic predictions for $dim(\boldsymbol{x})/dim(\boldsymbol{X}) = 2$

Effect of Coarse-Graining level

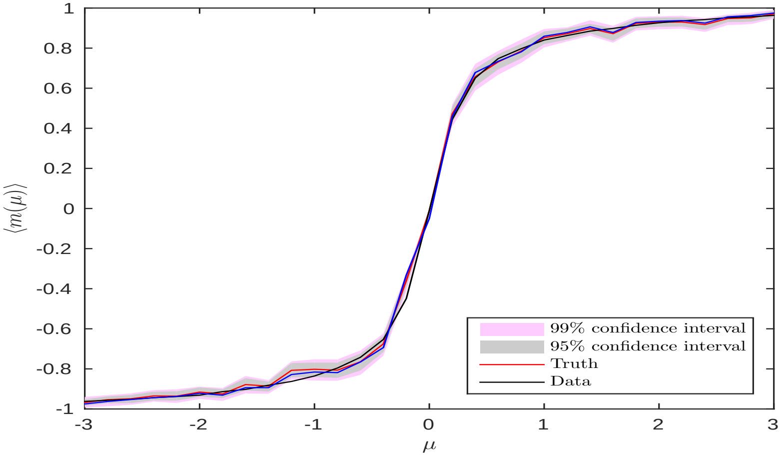


Figure : Probabilistic predictions for $dim(\boldsymbol{x})/dim(\boldsymbol{X}) = 4$

Effect of Coarse-Graining level

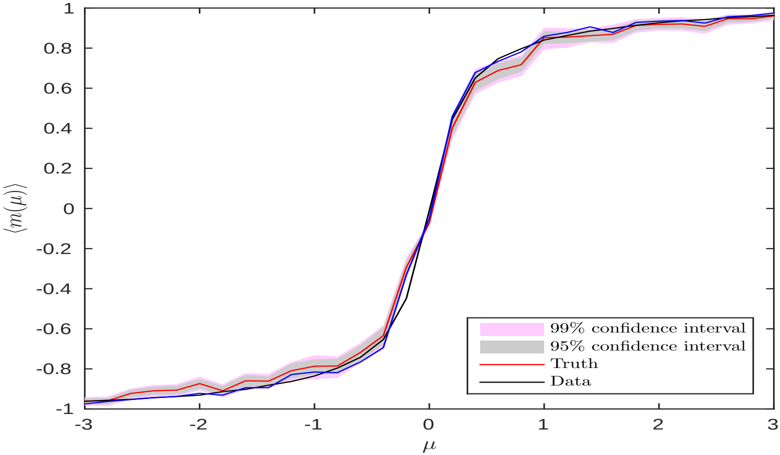
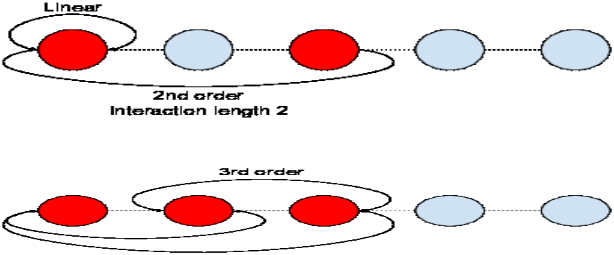


Figure : Probabilistic predictions for $dim(\boldsymbol{x})/dim(\boldsymbol{X}) = 8$

Model Selection

What is the *right* CG potential $V_c(\boldsymbol{X})$?

$$p_c(\boldsymbol{X}|\theta_c) = \exp\{\theta_c^T \phi(\boldsymbol{X}) - A(\theta_c)\}$$
$$V_c(\boldsymbol{X}) = \theta_c^T \underbrace{\phi(\boldsymbol{X})}_{\text{features}}$$

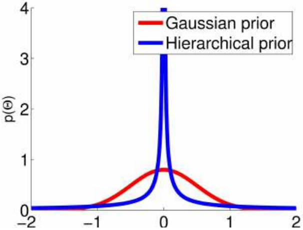


- Number of features $\phi(\boldsymbol{X})$ grows exponentially fast
- What are most important?
- Can one search *across models*?

Model Selection

- Sparsity-enforcing - Hierarchical priors (ARD, [MacKay 1994])

$$p(\boldsymbol{\theta}_c|\boldsymbol{\tau}) = \prod_j p(\boldsymbol{\theta}_{c,j}|\tau_j)$$
$$\boldsymbol{\theta}_{c,j} \sim \mathcal{N}(0, \tau_j^{-1})$$
$$\tau_j \sim Gamma(a_0, b_0)$$

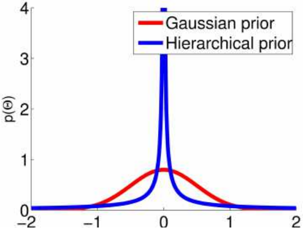


- As hyperparameters $\tau_j \to \infty$, then $\boldsymbol{\theta}_{c,j} \to 0$.
- Readily integrated in the EM-framework:
  - E-step: Compute $< \tau_j >_{p(\tau_j|\theta_{c,j})} = \frac{a_0+1/2}{b_0+\theta_{c,j}^2/2}$
  - M-step: Compute $\frac{\partial}{\partial \theta_{c,j}} = - < \tau_j > \frac{\theta_{c,j}}{2}$

Model Selection

- Sparsity-enforcing - Hierarchical priors (ARD, [MacKay 1994])

$$p(\boldsymbol{\theta}_c|\boldsymbol{\tau}) = \prod_j p(\boldsymbol{\theta}_{c,j}|\tau_j)$$
$$\boldsymbol{\theta}_{c,j} \sim \mathcal{N}(0, \tau_j^{-1})$$
$$\tau_j \sim Gamma(a_0, b_0)$$



- As hyperparameters $\tau_j \to \infty$, then $\boldsymbol{\theta}_{c,j} \to 0$.
- Readily integrated in the EM-framework:
  - E-step: Compute $< \tau_j >_{p(\tau_j|\boldsymbol{\theta}_{c,j})} = \frac{a_0+1/2}{b_0+\boldsymbol{\theta}_{c,j}^2/2}$
  - M-step: Compute $\frac{\partial}{\partial \boldsymbol{\theta}_{c,j}} = - < \tau_j > \frac{\boldsymbol{\theta}_{c,j}}{2}$

Model Selection

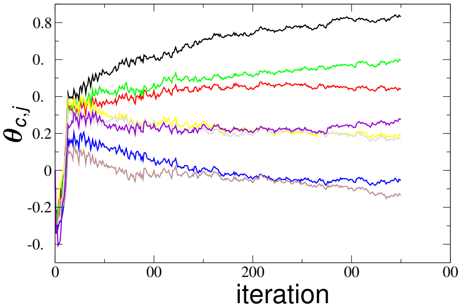


Figure : no ARD (uniform prior)

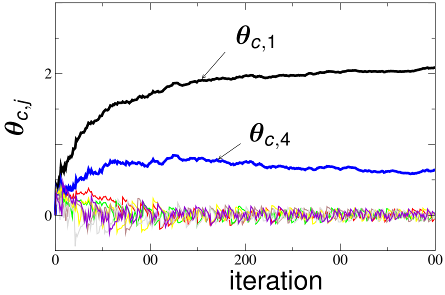


Figure : with ARD prior

Figure : 2^{nd}-order interactions - Feature functions $\phi_j(\boldsymbol{X}) = \sum_i X_i \, X_{i+j}$

Model Selection

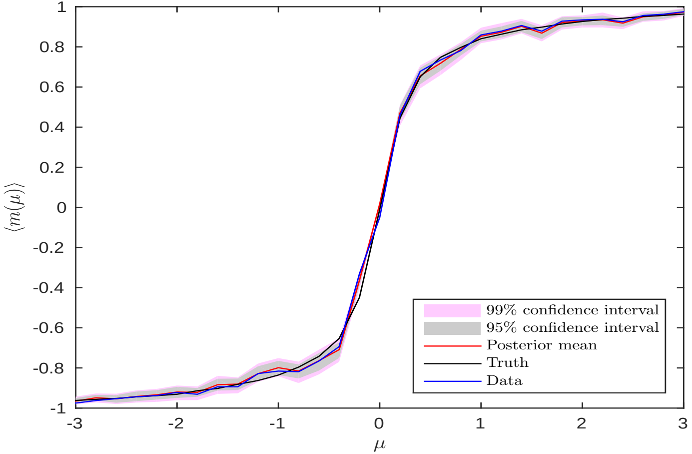


Figure : Probabilistic predictions with ARD prior for $N = 20$ data

Conclusions

Summary

- A *generative probabilistic model is proposed*
- It consists of a CG-density and a probabilistic *coarse* $\rightarrow$ *fine* map.
- Can account for information loss due to CG
- Can quantify *predictive uncertainty* in fine-scale observables.
- Can be used for *Model* selection.

Conclusions

Summary

- A *generative probabilistic model is proposed*
- It consists of a CG-density and a probabilostic *coarse* $\rightarrow$ *fine* map.
- Can account for information loss due to CG
- Can quantify *predictive uncertainty* in fine-scale observables.
- Can be used for *Model* selection.

Outlook

- Explore alternative definitions of coarse variables $\boldsymbol{X}$ and alternative *coarse* $\rightarrow$ *fine* maps p_{cf} e.g.:
  - Discrete states indicating Free-Energy wells
  - Hierachical *coarse-graining*:

$$\bar{p}_f(\boldsymbol{x}) = \int p_{cf}(\boldsymbol{x}|\boldsymbol{X}_1)\, p_c(\boldsymbol{X}_1|\boldsymbol{X}_2)\, p_c(\boldsymbol{X}_2|\boldsymbol{X}_3) \ldots p_c(\boldsymbol{X}_M)\, d\boldsymbol{X}_1 \ldots \boldsymbol{X}_M$$

- Fully Bayesian or Variational Bayesian
- Improvements in Learning by advanced sampling (instead of MCMC) and stochastic BFGS [Byrd et al. 2014, Moritz el al. 2015]