ETH zürich

Machine Learning and the Real-Space Renormalization Group

Maciej Koch-Janusz

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Zohar Ringel

``Mutual Information, Neural Networks and the Renormalization Group''
MKJ and Zohar Ringel, *Nature Physics* ***14****, 578-582 (2018)*

ETH zürich

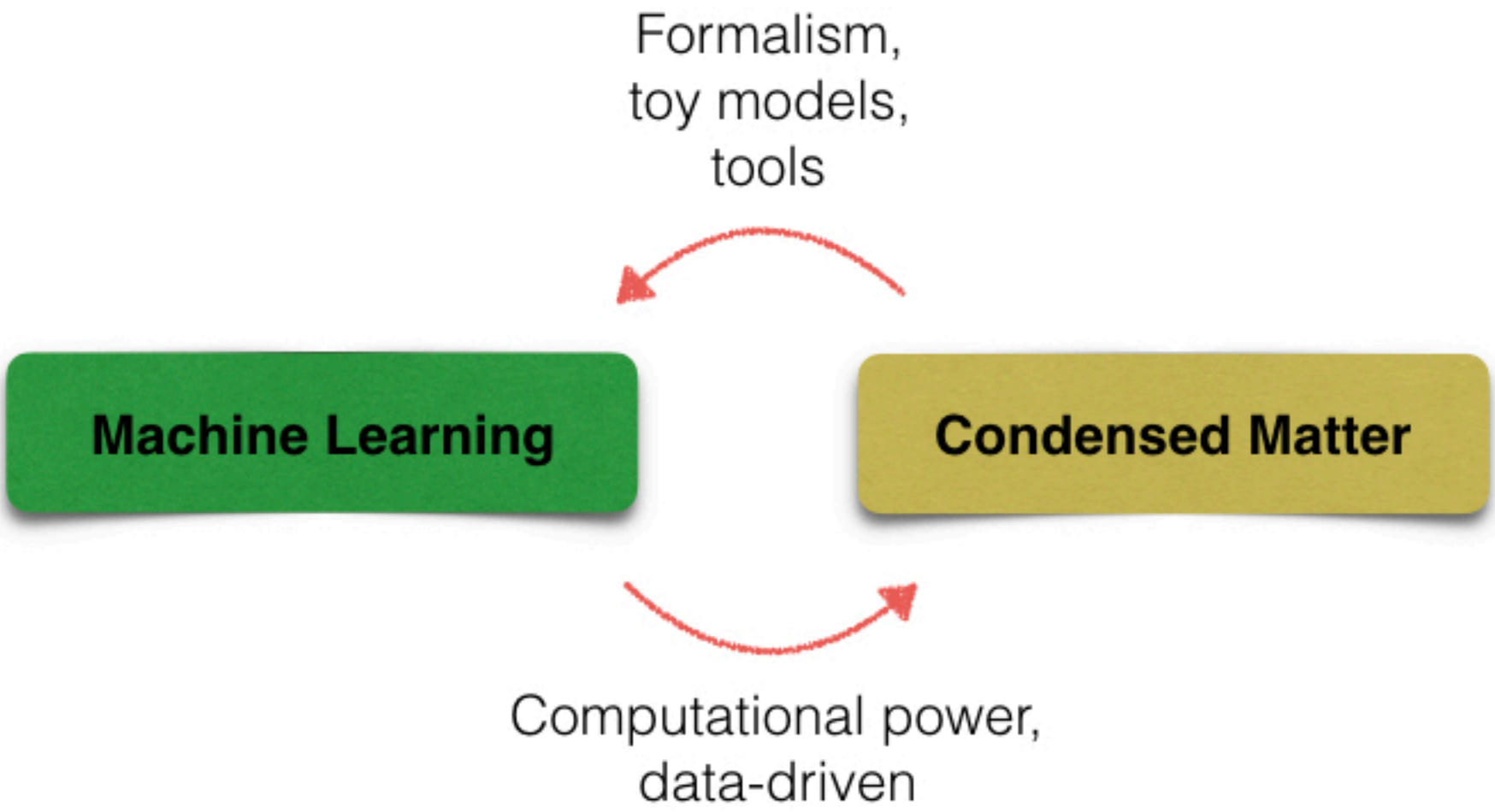
Sebastian D. Huber

Patrick Lenggenhager

``Optimal Renormalization Group Transformation from Information Theory''
Lenngenhager, Ringel, Huber and MKJ, *arXiv:1809.09632*

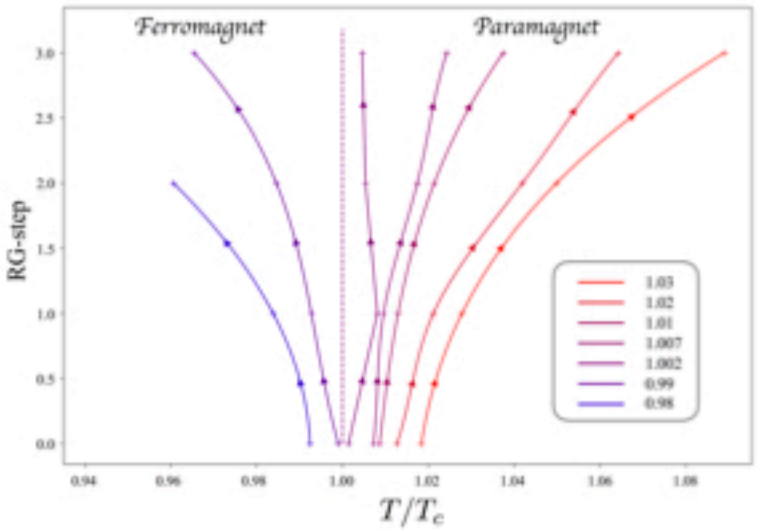
Outline

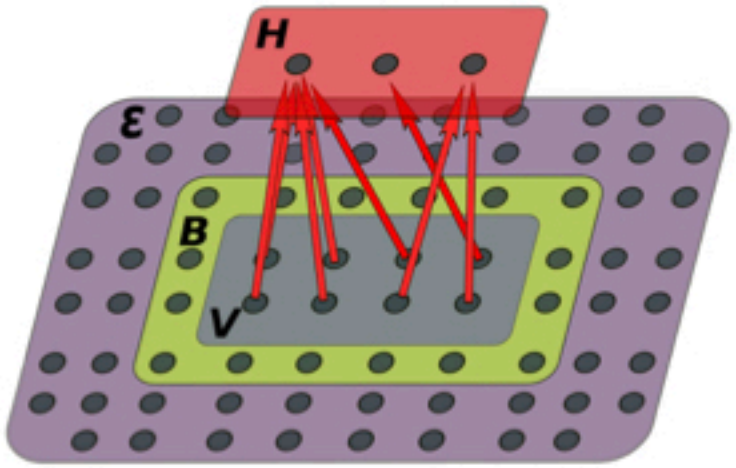
- Machine learning in condensed matter
- Information-theoretic approach to real-space RG:
  - The Real Space Mutual Information algorithm
  - Numerical results
  - Optimality of the RSMI prescription


Formalism,
toy models,
tools
Machine Learning
Condensed Matter
Computational power,
data-driven

The punchline

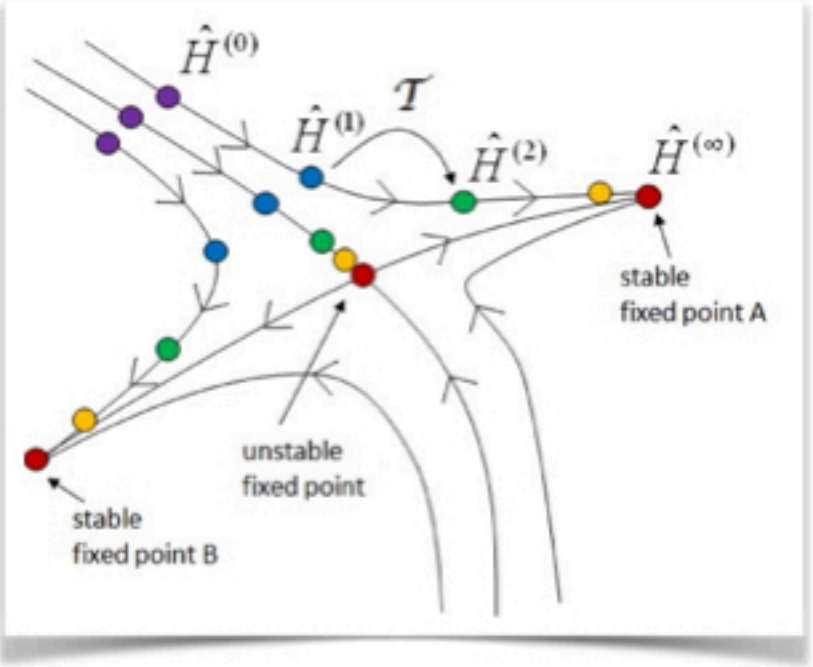
An information-theoretic approach and an *unsupervised* machine learning algorithm performing real-space RG of classical stat. mech. systems. Optimality results.

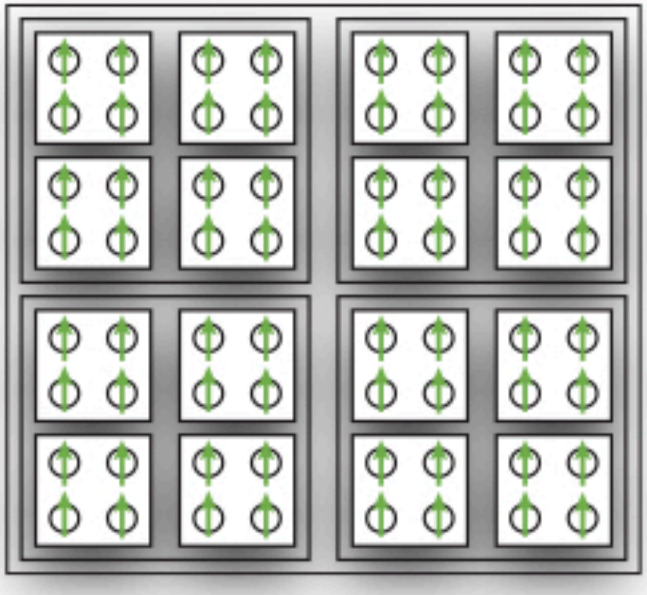




Renormalization Group

- Conceptually important: formalizes the notion of separation of scales
- Universality
- Many flavors exist: Wilsonian, DMRG, real-space, ….





Leitmotiv: integrate out some ‘fast’ degrees of freedom to obtain effective theory of the ‘slow’ ones

In a sense the relation of RG to information theory is obvious as averaging loses information

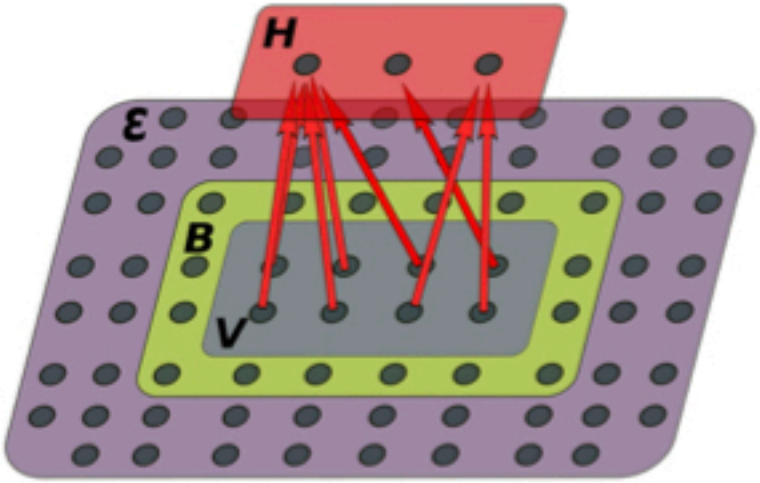
- Can it be formalized?
- Is it useful in practice?

Real-space RG from Information Theory perspective

$$e^{\mathcal{K}'(\mathcal{X}')} = \sum_{\mathcal{X}} e^{\mathcal{K}(\mathcal{X})} P_\Lambda(\mathcal{X}'|\mathcal{X})$$

Task: Learn $P_\Lambda(\mathcal{H}|\mathcal{V})$ such that $\mathcal{H}$ tracks the *slow degrees of freedom* within region $\mathcal{V}$

$$\boxed{\Lambda = (a_i, b_j, \lambda_i^j)}$$

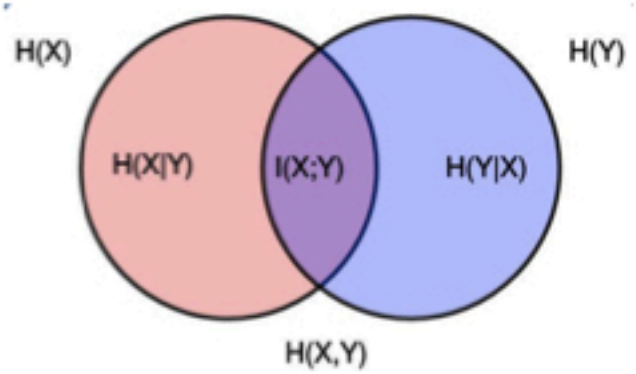


$$P(\mathcal{X}) = \frac{1}{Z} e^{\mathcal{K}(\mathcal{X})}$$

Method: Require that *slow degrees of freedom* maximize spatial mutual information

Formally: find $\max[I_\Lambda(\mathcal{H}:\mathcal{E})]$ over parameters Λ

Mutual Information



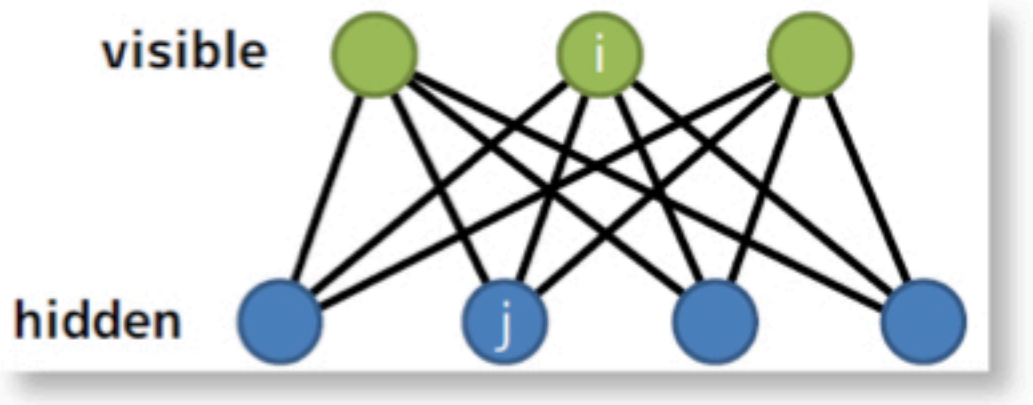
$$I_\Lambda(\mathcal{H}:\mathcal{E}) = \sum_{\mathcal{H},\mathcal{E}} P_\Lambda(\mathcal{E},\mathcal{H}) \log\left(\frac{P_\Lambda(\mathcal{E},\mathcal{H})}{P_\Lambda(\mathcal{H})P(\mathcal{E})}\right)$$

- Vanishes for independent variables
- Bounded from above by information entropy
- More general than correlation functions

$$I_\Lambda(\mathcal{H}:\mathcal{E}) = H(\mathcal{H}) - H(\mathcal{H}|\mathcal{E})$$

(Restricted) Boltzmann Machines

- Stochastic networks
- Model probability distributions
- Can be used for efficient sampling



$$E_\Theta \equiv E_{a,b,\theta}(\mathcal{V},\mathcal{H}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{ij} v_i \theta_{ij} h_j$$

$$P_\Theta(\mathcal{V},\mathcal{H}) = \frac{1}{\mathcal{Z}} e^{-E_{a,b,\theta}(\mathcal{V},\mathcal{H})}$$

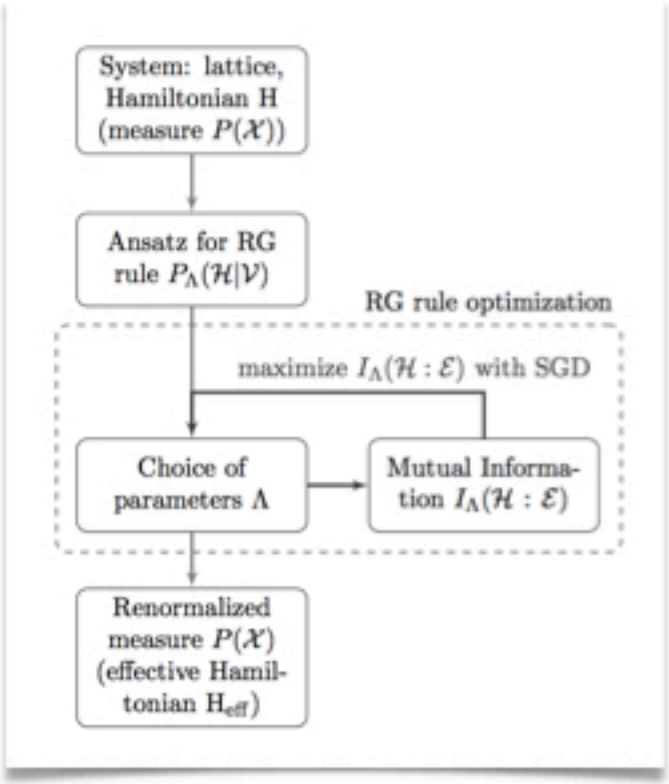
$$P_\Theta(\mathcal{V}) \qquad P_\Theta(\mathcal{H}|\mathcal{V})$$

- How to chose the parameters?

Stage I. - Train RBMs to reproduce P(V,E) and P(V) via contrastive divergence

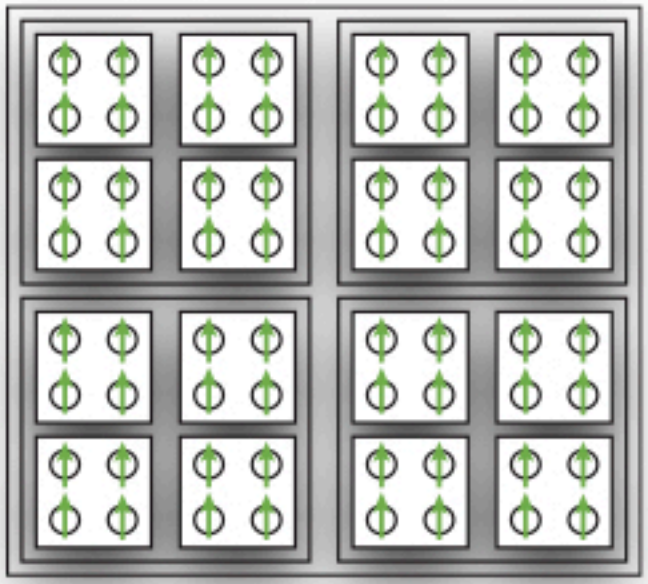
$$P_\Theta(\mathcal{V},\mathcal{H}) = \frac{1}{\mathcal{Z}} e^{-E_{a,b,\theta}(\mathcal{V},\mathcal{H})}$$

Stage II. - Model $P_\lambda(H \mid V)$ as an RBM, obtain $P_\lambda(H,E)$, do SGD to maximize I(H:E) (involves Monte Carlo)

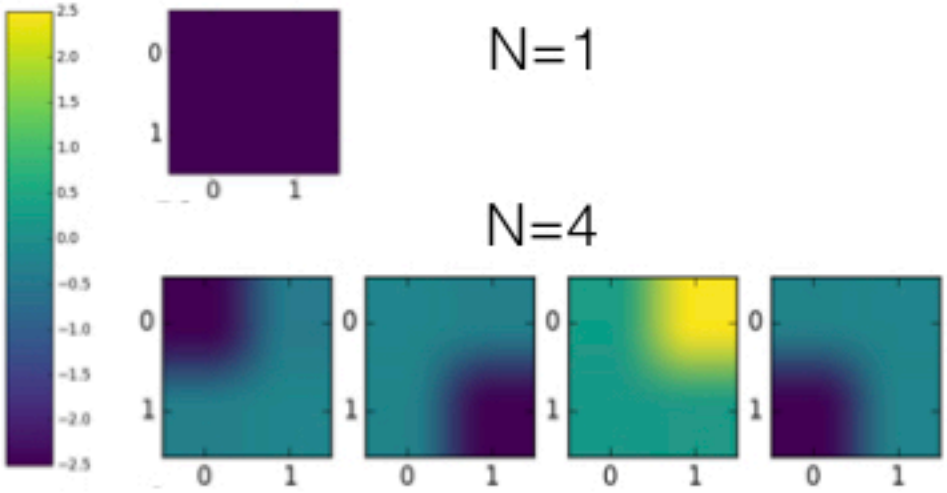


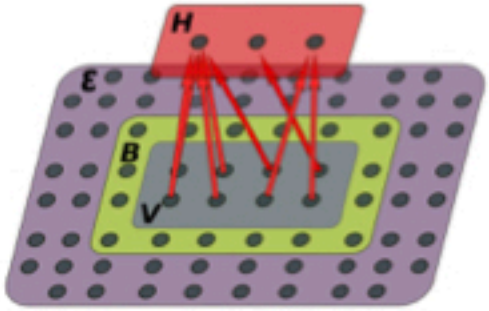
Test case 1: the 2D Ising model

$$H_I = -\sum_{<i,j>} s_i s_j$$



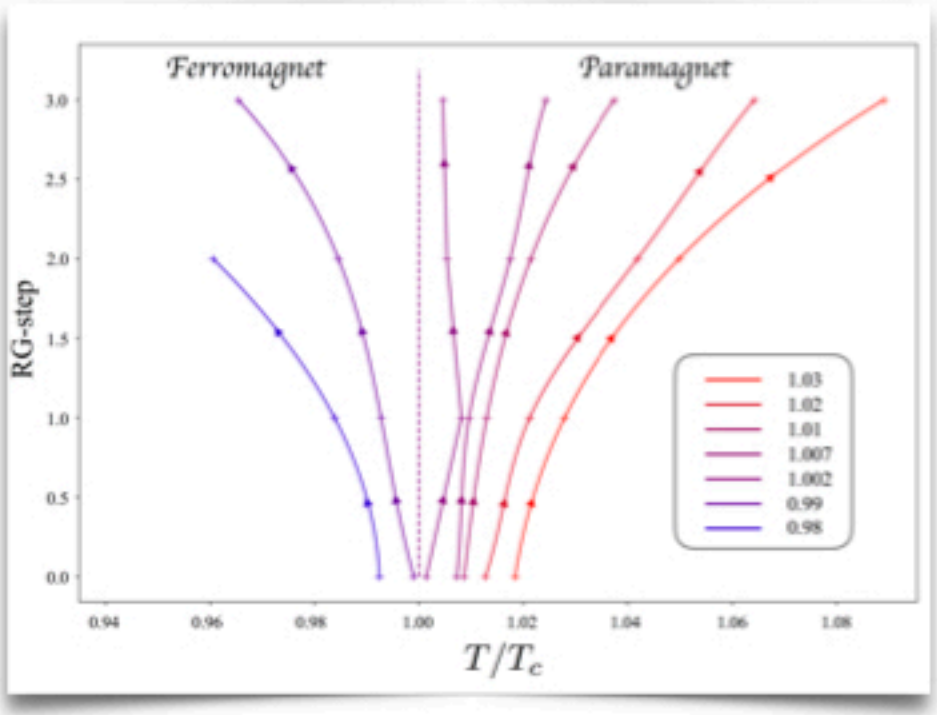
Migdal-Kadanoff block-spins:



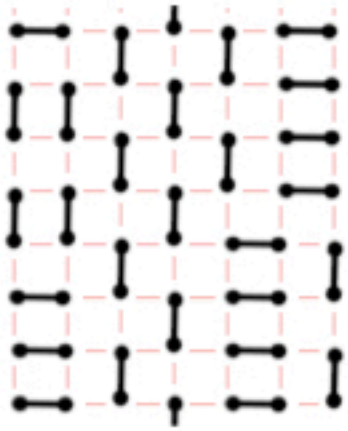


RG flow and critical exponents

- RG flow reconstruction
- Position and type of critical points
- Critical exponents



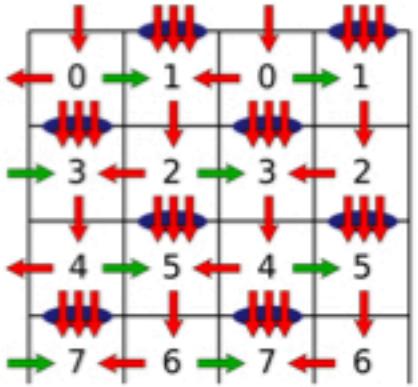
Test case 2: the dimer model

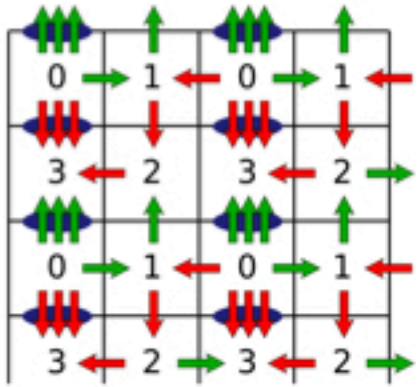


- Defined by local constraints
- Partition function counts configurations

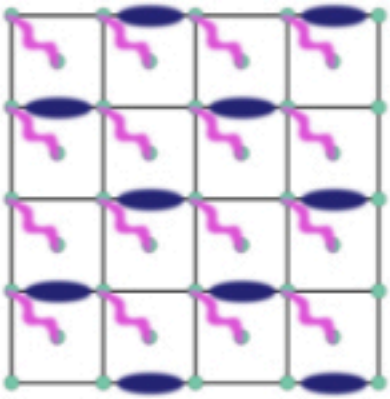
RG of dimer model:
mapping to height field h(x)

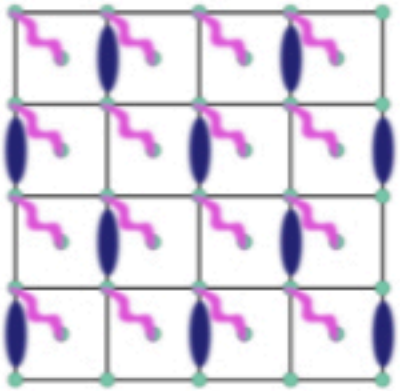
$$S_{dim}[h] = \int d^2x \ (\nabla h(\vec{x}))^2 \equiv \int d^2x \ \vec{E}^2(\vec{x})$$

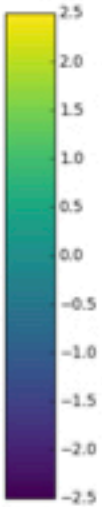




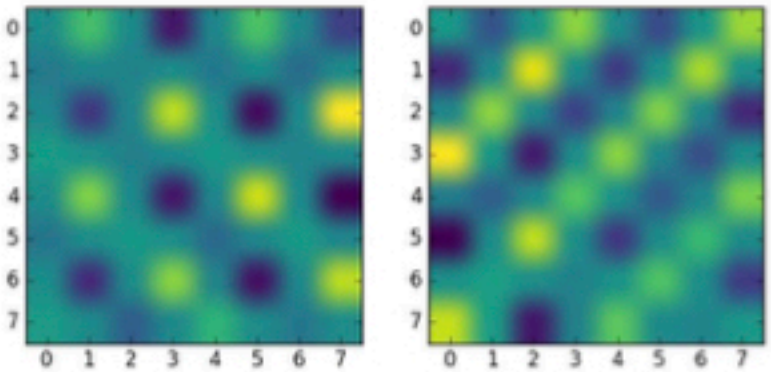
- Let's add noise!
- Physically irrelevant, but strong pattern

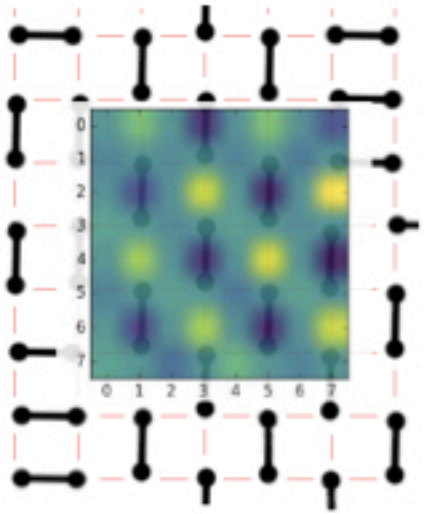






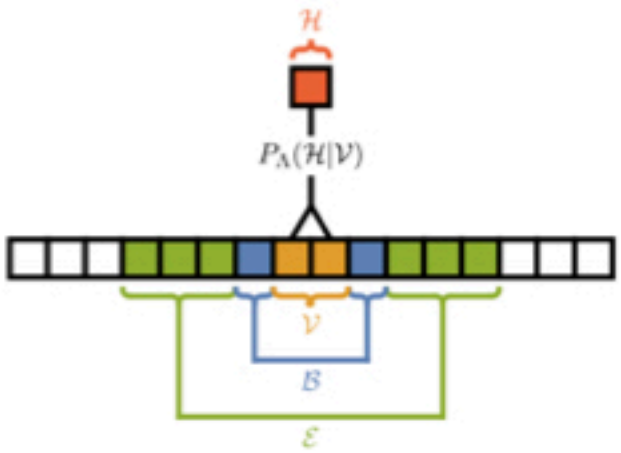
N=2





Optimality of the RSMI approach

$$I_\Lambda(\mathcal{H} : \mathcal{E}) \leq I(\mathcal{V} : \mathcal{E})$$



- Assume first perfect RSMI capture

$$P(\mathcal{H}_{j\leq 2}, \mathcal{H}_0^*, \mathcal{H}_{j\geq 2}) = P(\mathcal{H}_{j\leq -2}, \mathcal{H}_0^*) \cdot P(\mathcal{H}_0^*, \mathcal{H}_{j\geq 2})$$

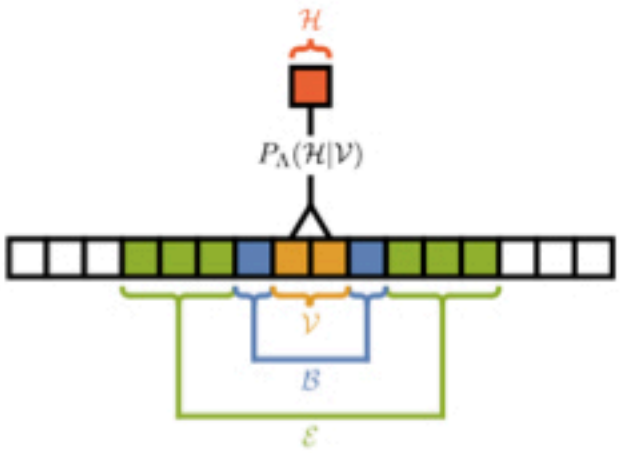
→ The range of the Hamiltonian does not increase!

BUT: Realistically, this assumption is rarely true

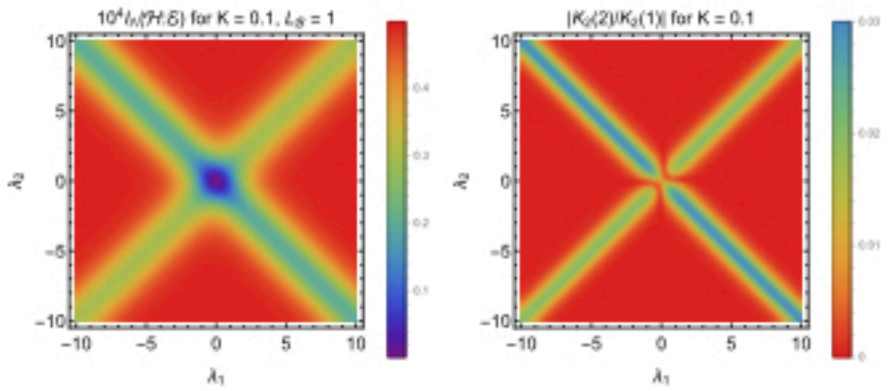
Optimality: 1D Ising model

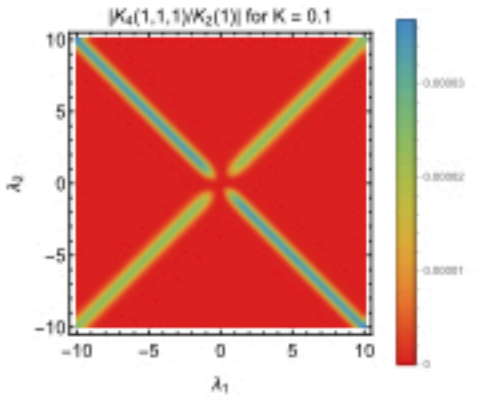
- Effective Hamiltonian:

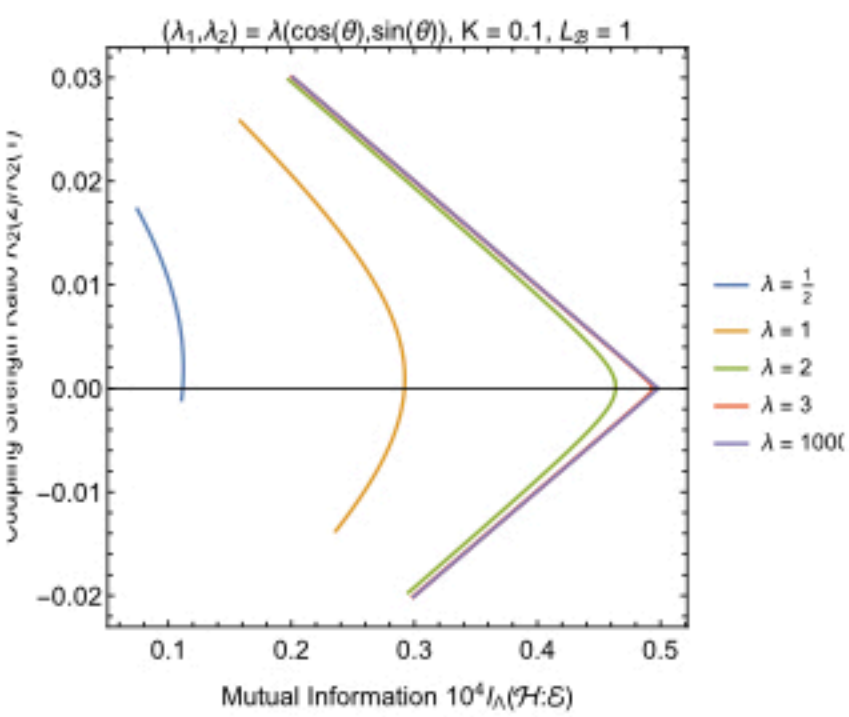
$$\mathcal{K}'[\mathcal{X}'] = \log(Z_{\Lambda,0}[\mathcal{X}']) + \sum_{k=0}^{\infty} \frac{1}{k!} C_k[\mathcal{X}']$$

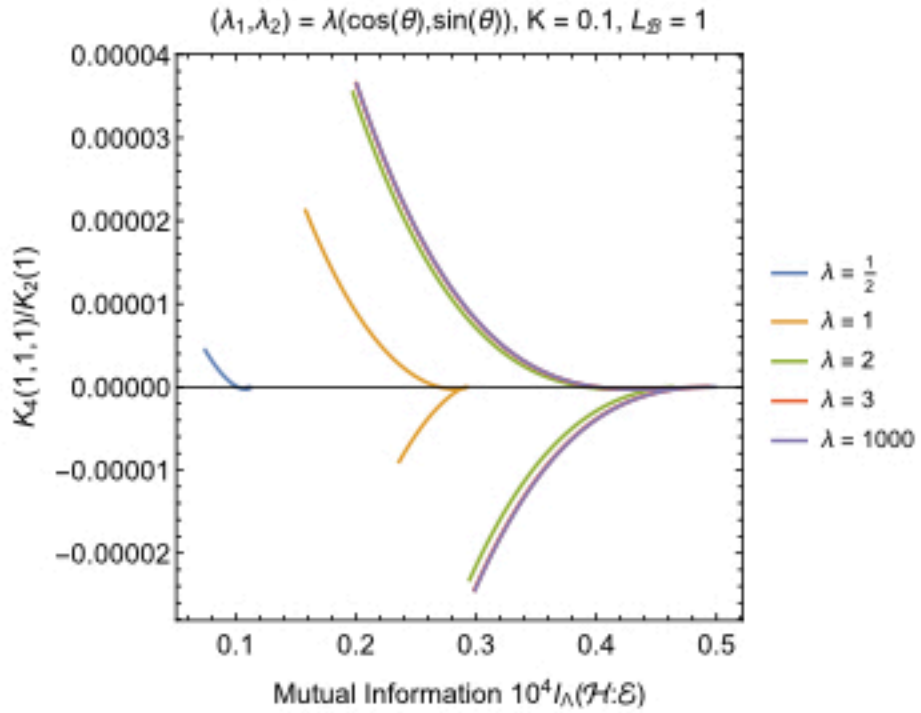


- The “rangeness” and “n-body-ness”



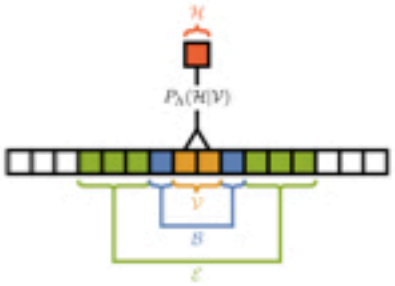



$(\lambda_1,\lambda_2) = \lambda(\cos(\theta),\sin(\theta))$, K = 0.1, $L_{\mathcal{B}}$ = 1
0.03
0.02
0.01
0.00
−0.01
−0.02
0.1
0.2
0.3
0.4
0.5
Mutual Information $10^4 I_\Lambda(\mathcal{H}:\mathcal{E})$
$\lambda = \frac{1}{2}$
$\lambda = 1$
$\lambda = 2$
$\lambda = 3$
$\lambda = 1000$

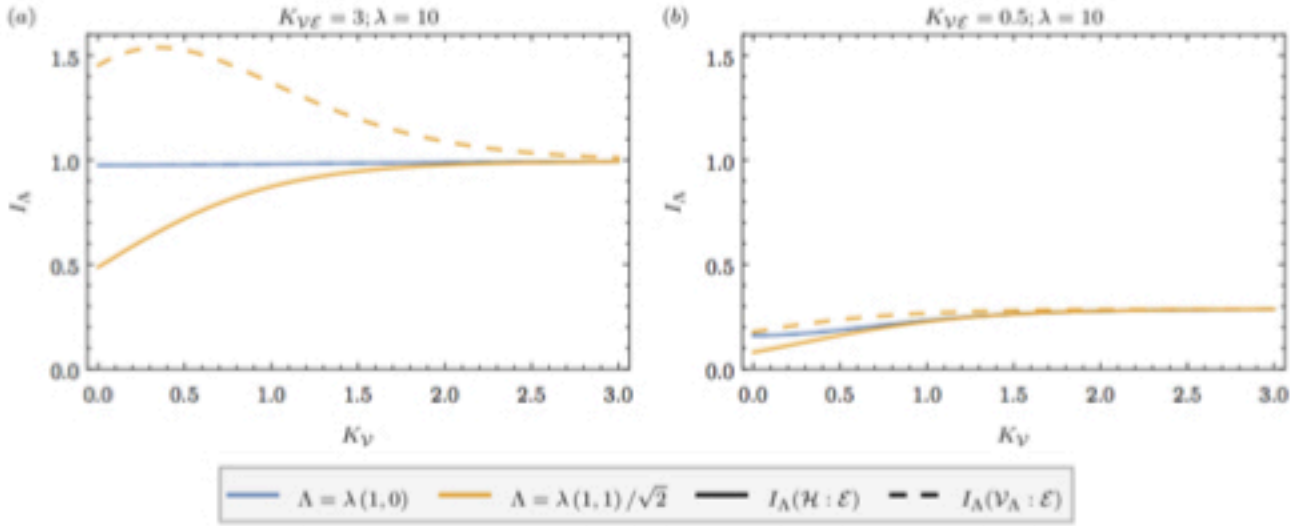

$(\lambda_1,\lambda_2) = \lambda(\cos(\theta),\sin(\theta))$, K = 0.1, $L_{\mathcal{B}}$ = 1
$K_4(1,1,1)/K_2(1)$
0.00004
0.00003
0.00002
0.00001
0.00000
−0.00001
−0.00002
0.1
0.2
0.3
0.4
0.5
Mutual Information $10^4 I_\Lambda(\mathcal{H}:\mathcal{E})$
$\lambda = \frac{1}{2}$
$\lambda = 1$
$\lambda = 2$
$\lambda = 3$
$\lambda = 1000$

The "shape" of coarse-grained variables

- The hidden couples to: $\mathcal{V}_{\Lambda_i} = \frac{1}{||\Lambda_i||} \sum_j \lambda_{ij} v_j$

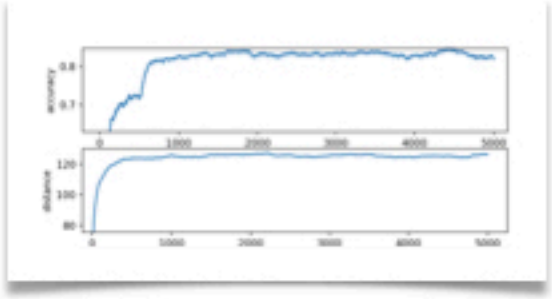


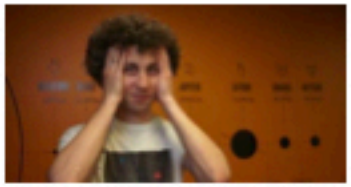
$$I_\Lambda(\mathcal{H}:\mathcal{E}) \leq I(\mathcal{V}_\Lambda:\mathcal{E}) \leq I(\mathcal{V}:\mathcal{E})$$



$$I_\Lambda(\mathcal{H}:\mathcal{E}) = I(\mathcal{V}_\Lambda:\mathcal{E}) - I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$$

- Binary Network training





M. Maksymenko
(SoftServe, Ukraine)

- Properties of dropout

Y. Efron, V. Pushkarov
(Technion, Israel)

M. Maksymenko
(SoftServe, Ukraine)

- Quantum ML

Mark H. Fisher
(ETH/Uni. Zurich)

Sebastian D. Huber
(ETH)

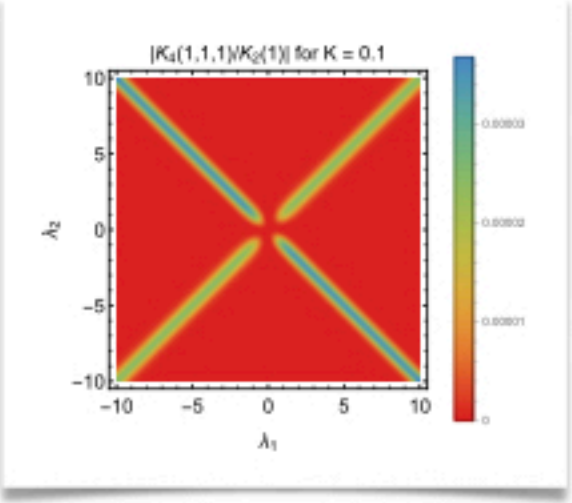
- Quantum state modeling with non-Markovian noise

E. van Nieuwenburg
(Caltech)

Alireza Seif
(JQI)

Conclusions

- Information-theoretic view of RG
  - The Real Space Mutual Information algorithm
  - Optimality of the RSMI prescription



Lenggenhager, Ringel, Huber, MKJ,
arXiv:1809.09632

MKJ and Z. Ringel
Nature Physics ***14****, 578-582 (2018)*

ETH zürich

Thank you!

(R)BM training

$$P_\Theta(\mathcal{V},\mathcal{H}) = \frac{1}{\mathcal{Z}} e^{-E_{a,b,\theta}(\mathcal{V},\mathcal{H})}$$

- First attempt: Maximal Likelihood (ML)

$$\sum_{\text{all } \mathcal{V}} P_{data}(\mathcal{V}) \log P_\Theta(\mathcal{V}) = \sum_{\mathcal{V}\in \text{data}} \log P_\Theta(\mathcal{V}) P_{data}(\mathcal{V}) = \sum_{\text{all } \mathcal{V}} \log P_\Theta(\mathcal{V}) P_{data}(\mathcal{V}) = \sum_{\text{all } \mathcal{V}} P_{data}(\mathcal{V}) \log \left(\frac{P_{data}(\mathcal{V})}{P_\Theta(\mathcal{V})} \right)$$

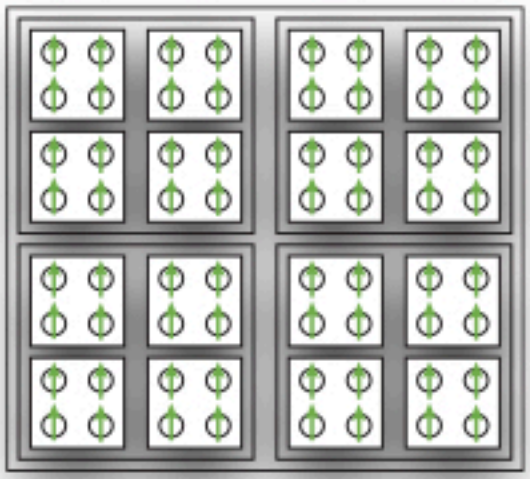
- Better solution: Contrastive Divergence (CD)

$$KL(P_{data}||P_\Theta) - KL(P_n||P_\Theta)$$

Effective hamiltonian

$$\mathcal{K}[\mathcal{X}] = \mathcal{K}_0[\mathcal{X}] + \mathcal{K}_1[\mathcal{X}] \qquad \mathcal{K}_0[\mathcal{X}] = \sum_{j=1}^{n} \mathcal{K}_\mathrm{b}[\mathcal{V}_j]$$

$$Z_0 = \sum_{\mathcal{X}} e^{\mathcal{K}_0[\mathcal{X}]} = \prod_{j=1}^{n} Z_\mathrm{b}, \qquad Z_\mathrm{b} = \sum_{\mathcal{V}} e^{\mathcal{K}_\mathrm{b}[\mathcal{V}]}$$

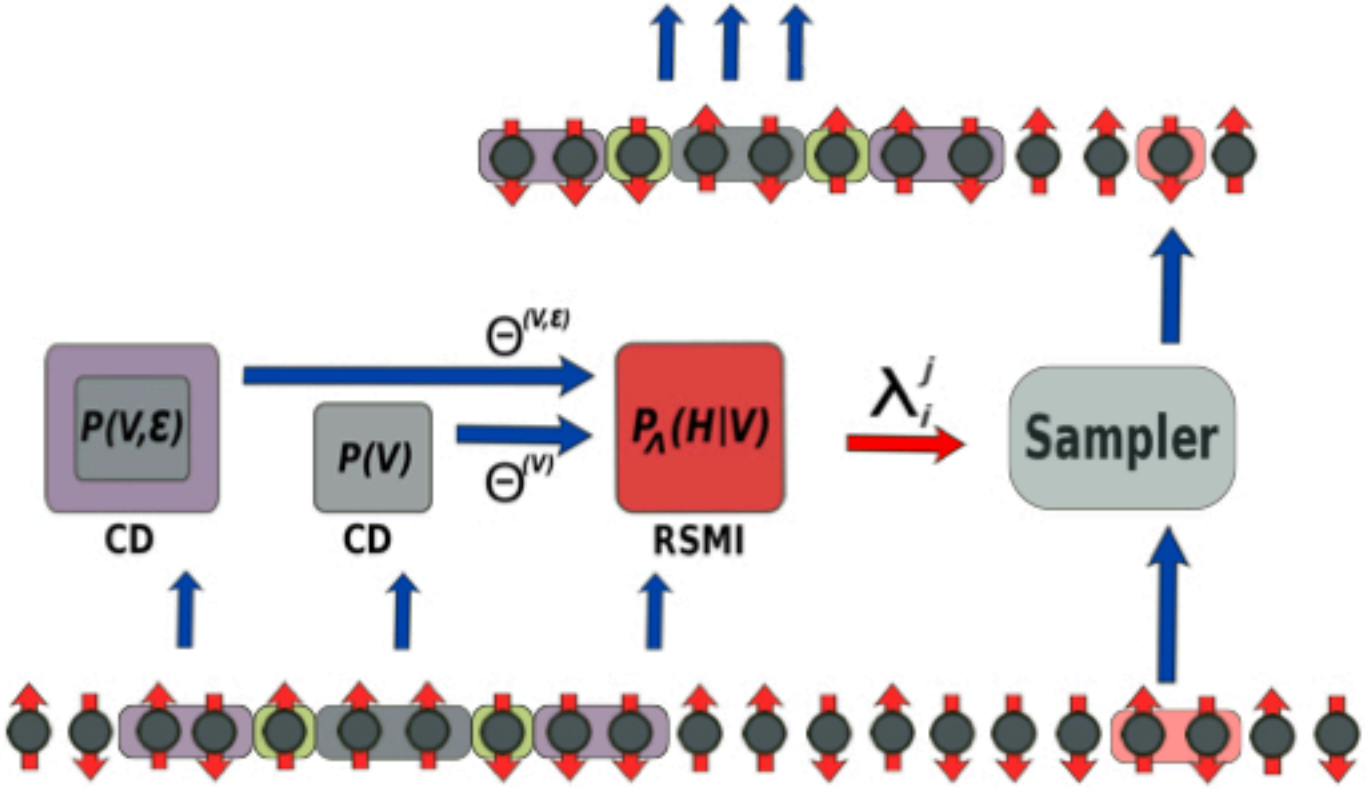


$$\begin{aligned}
e^{\mathcal{K}'[\mathcal{X}']} &= Z_0 \sum_{\mathcal{X}} e^{\mathcal{K}_1[\mathcal{X}]} \prod_{j=1}^{n} \underbrace{\frac{e^{\mathcal{K}_\mathrm{b}[\mathcal{V}_j]}}{Z_\mathrm{b}} P_\Lambda(\mathcal{H}_j|\mathcal{V}_j)}_{=:P_{\Lambda,\mathrm{b}}(\mathcal{H}_j,\mathcal{V}_j)} \\
&= Z_0 \sum_{\mathcal{X}} e^{\mathcal{K}_1[\mathcal{X}]} \prod_{j=1}^{n} P_{\Lambda,\mathrm{b}}(\mathcal{V}_j|\mathcal{H}_j) P_{\Lambda,\mathrm{b}}(\mathcal{H}_j) \\
&= Z_0 \underbrace{\prod_{j=1}^{n} P_{\Lambda,\mathrm{b}}(\mathcal{H}_j)}_{P_{\Lambda,0}(\mathcal{X}')} \sum_{\mathcal{X}} e^{\mathcal{K}_1[\mathcal{X}]} \underbrace{\prod_{j=1}^{n} P_{\Lambda,\mathrm{b}}(\mathcal{V}_j|\mathcal{H}_j)}_{=:P_{\Lambda,0}(\mathcal{X}|\mathcal{X}')} \\
&= Z_0 P_{\Lambda,0}(\mathcal{X}') \left\langle e^{\mathcal{K}_1[\mathcal{X}]} \right\rangle_{\Lambda,0} [\mathcal{X}'],
\end{aligned}$$

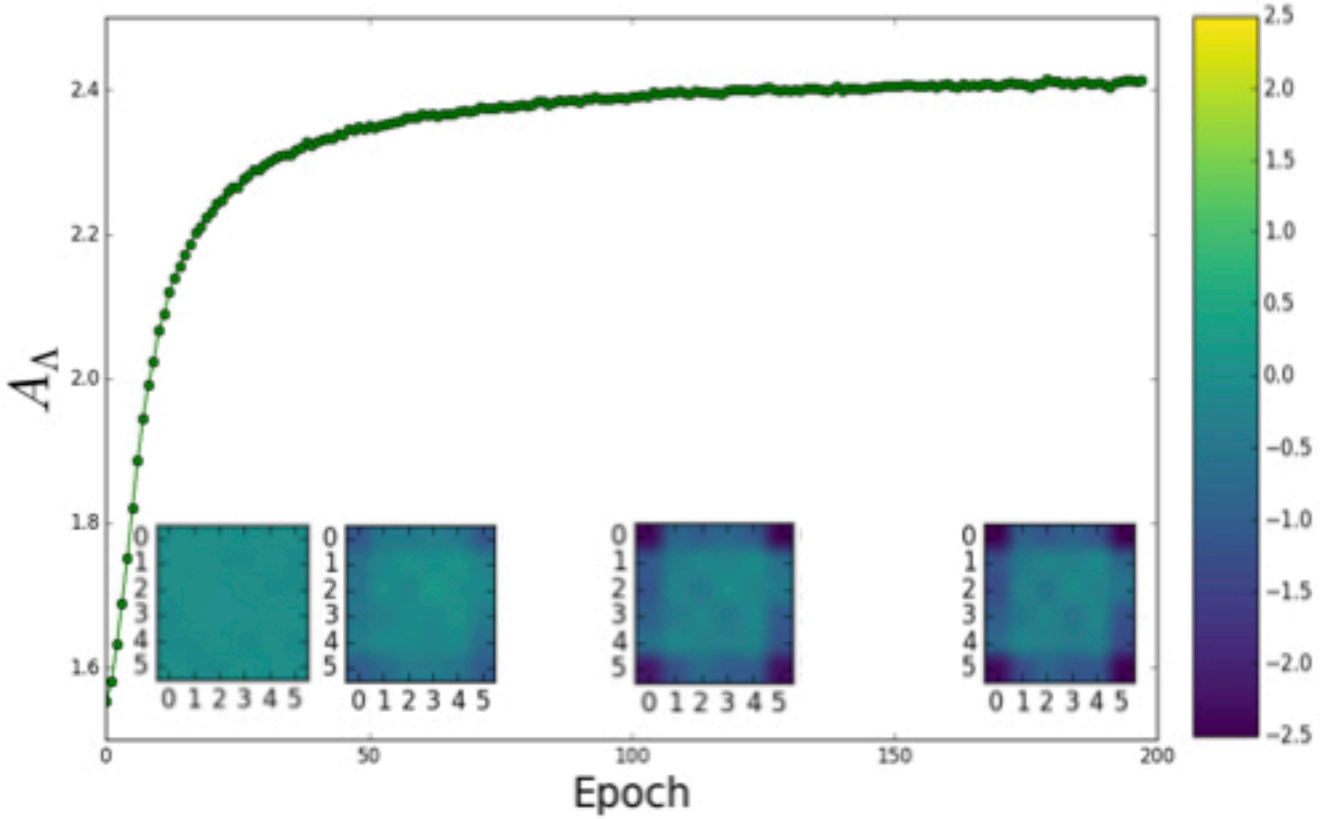
$$\mathcal{K}'[\mathcal{X}'] = \log(Z_{\Lambda,0}[\mathcal{X}']) + \sum_{k=0}^{\infty} \frac{1}{k!} C_k[\mathcal{X}']$$

$$\begin{aligned}
C_1 &= \langle \mathcal{K}_1 \rangle_{\Lambda,0}, \\
C_2 &= \langle \mathcal{K}_1^2 \rangle_{\Lambda,0} - \langle \mathcal{K}_1 \rangle_{\Lambda,0}^2, \\
C_3 &= \langle \mathcal{K}_1^3 \rangle_{\Lambda,0} - 3 \langle \mathcal{K}_1^2 \rangle_{\Lambda,0} \langle \mathcal{K}_1 \rangle_{\Lambda,0} + 2 \langle \mathcal{K}_1 \rangle_{\Lambda,0}^3
\end{aligned}$$

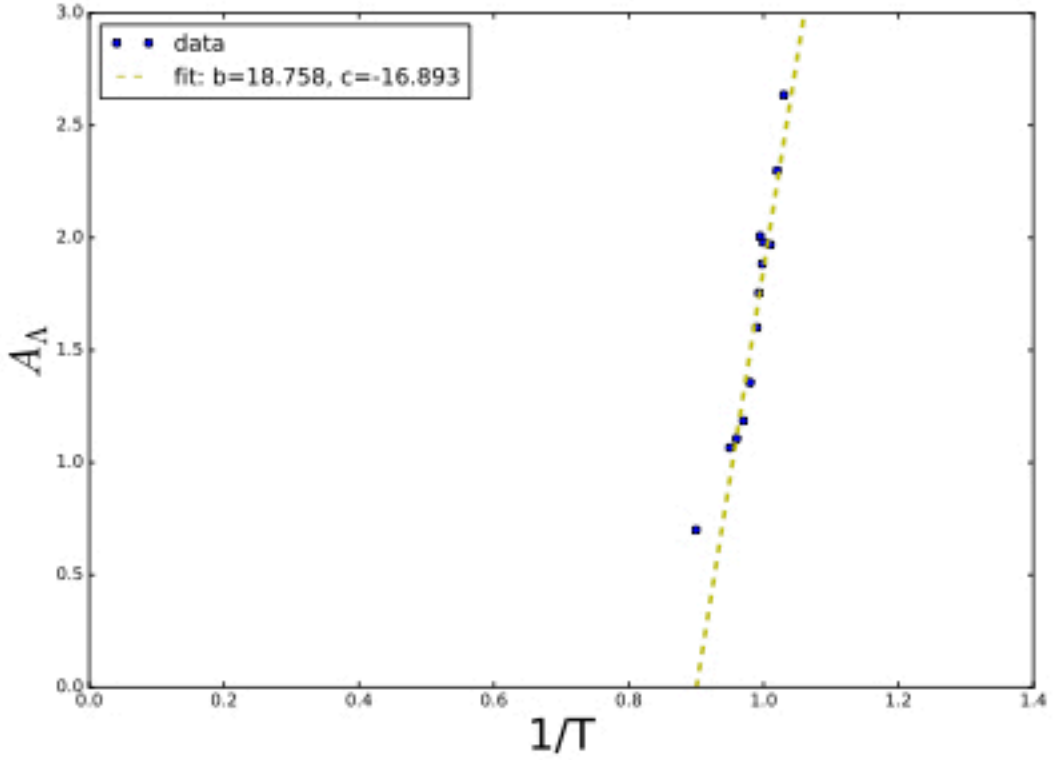
Multiple RG steps



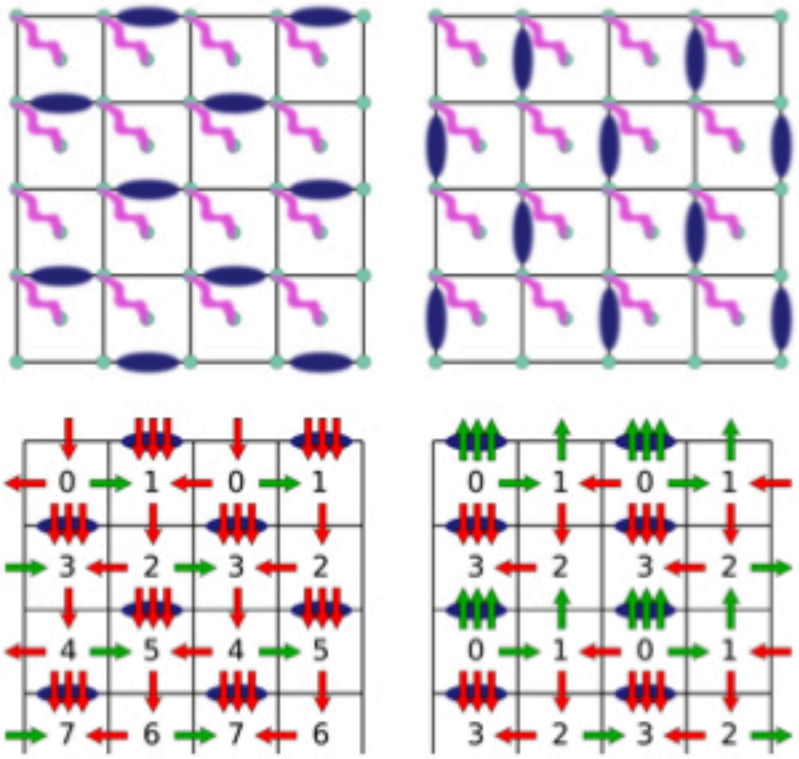
MI - Training


A_Λ
2.4
2.2
2.0
1.8
1.6
0
50
100
150
200
Epoch
0 1 2 3 4 5
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5

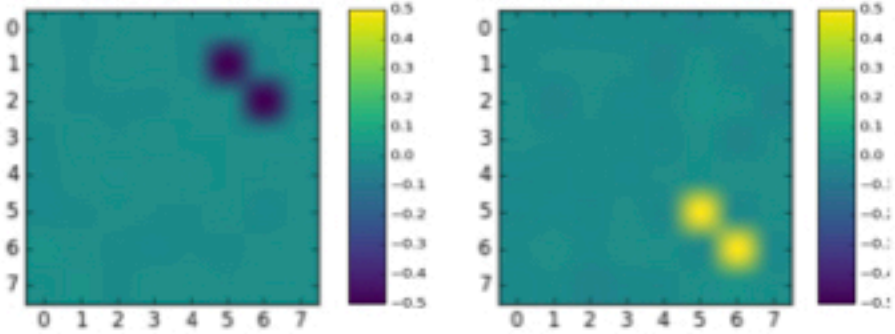
The MI “thermometer”



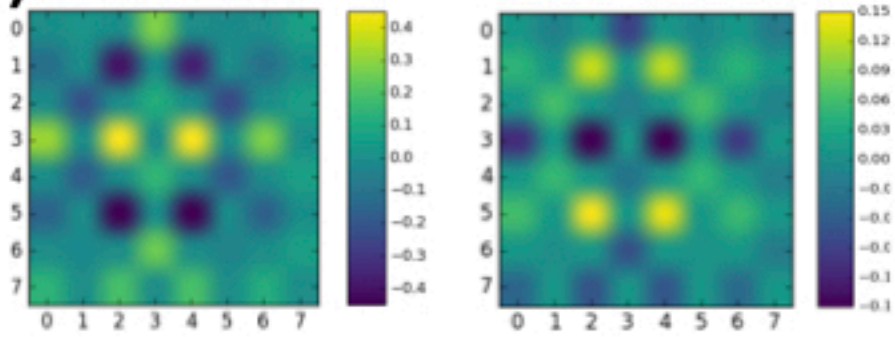
KL-training failure



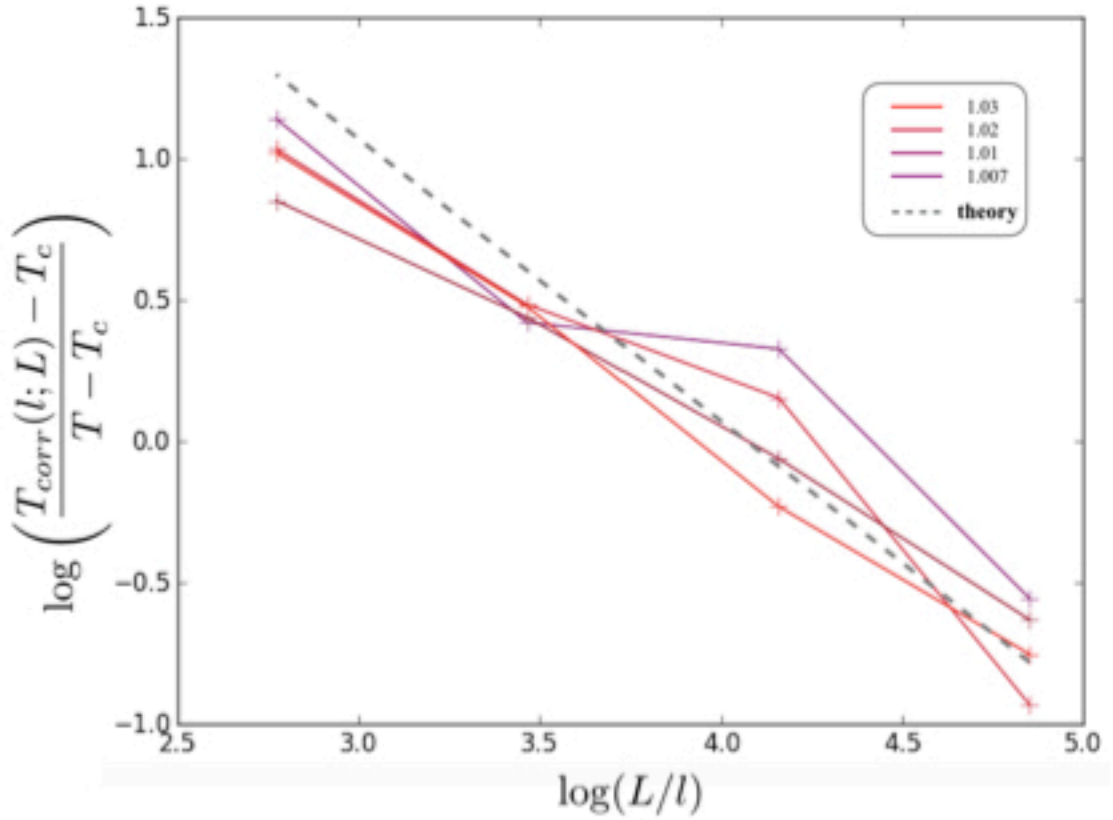
A)



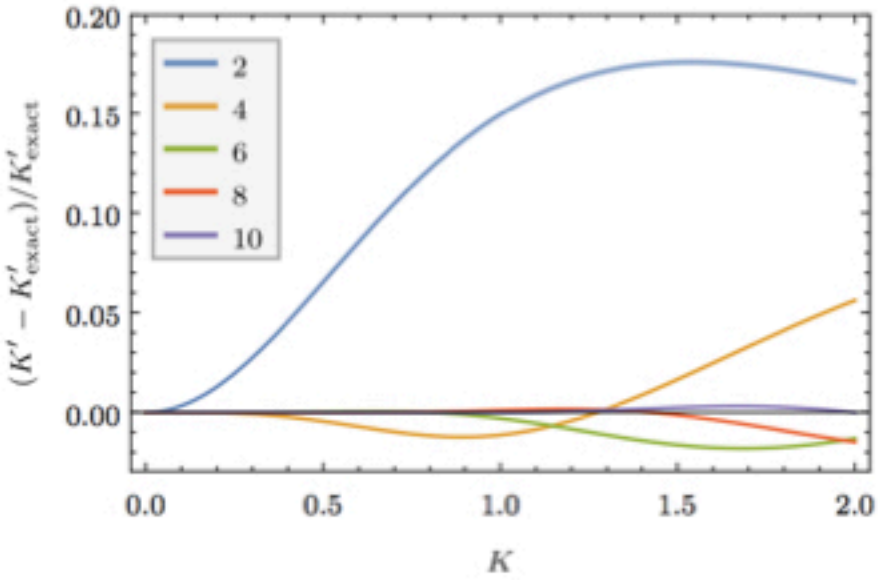
B)

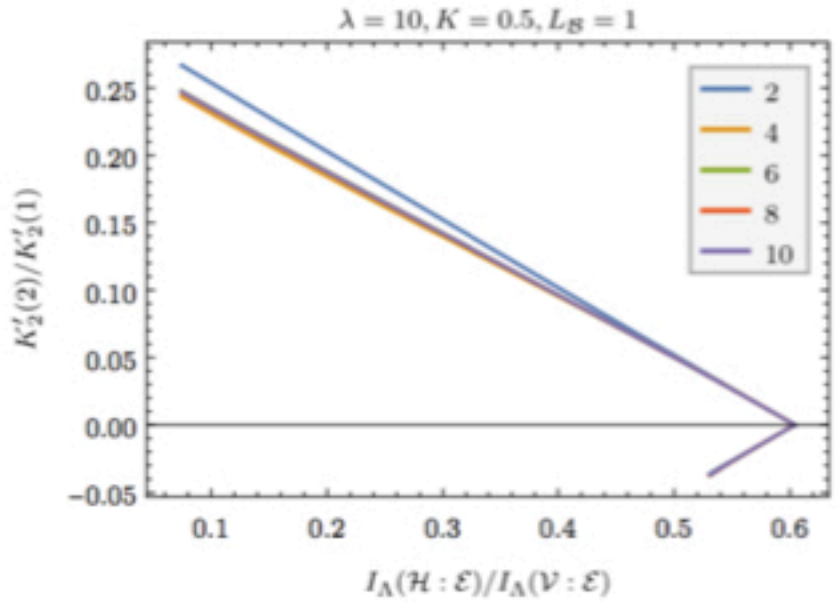


Critical exponent

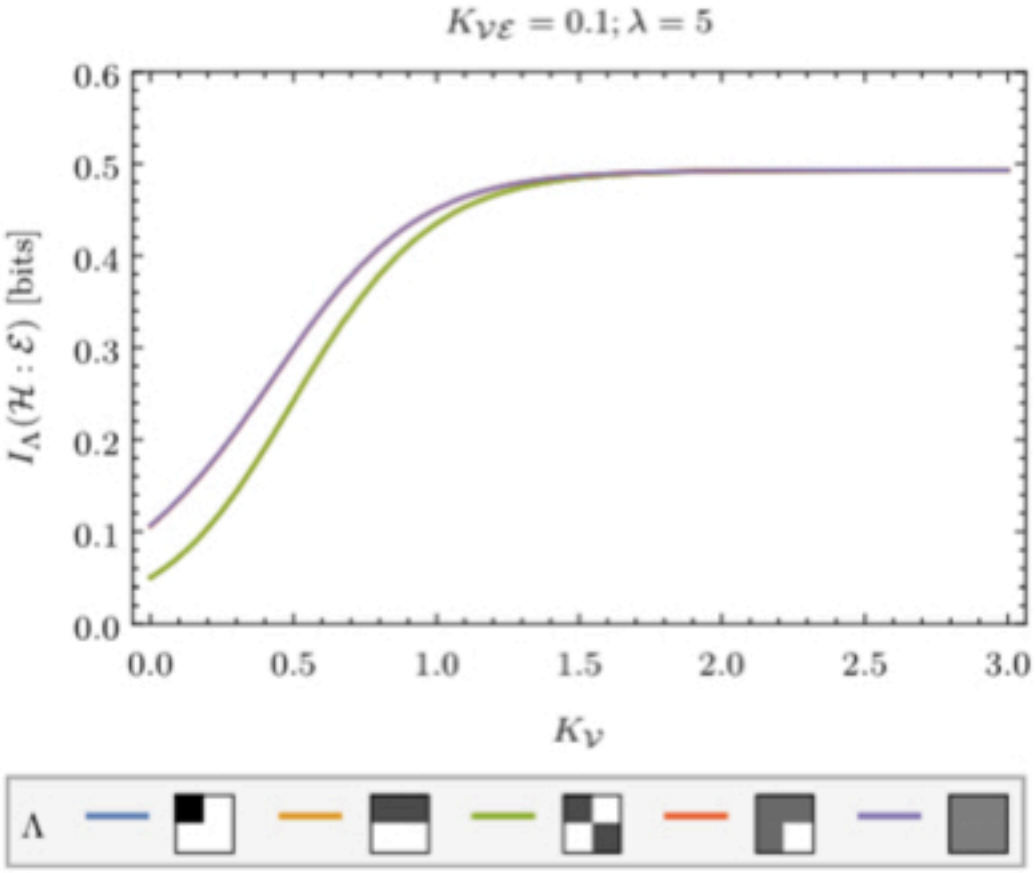


Convergence





2D Toy model


$K_{\mathcal{V}\mathcal{E}} = 0.1; \lambda = 5$
$I_\Lambda(\mathcal{H} : \mathcal{E})$ [bits]
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
$K_{\mathcal{V}}$
Λ

1D Toy model with PBC