



THE UNIVERSITY *of* EDINBURGH

# Machine Learning Dualities in Statistical Physics

---

Andrea E. V. Ferrari

based on an ICML25 paper with P. Gupta and N. Iqbal

Data Intensive Science Seminar, DAMTP, 11.11.2025

[1] University of Edinburgh, School of Mathematics & DESY Hamburg & DAMTP

*andrea.e.v.ferrari@gmail.com, andreaferrari.ch*

# ICML Link

Here is a linke to the ICML25 paper:

ICML25

There's also a longer but older arxiv preprint (2411.04838), with complementary results I won't talk about today.

# Motivation – Dualities

What is a duality?

*Duality* is a key idea of modern theoretical physics: the same physical system may have two (or more!) surprisingly different descriptions. Biased set of examples:

- Bosonisation, Sine-Gordon/Thirring [Coleman, Mandelstam]
- Particle/Vortex duality in 2+1 dimensions (3d mirror symmetry)
  [Dasgupta/Halperin, Peskin, Song/Tarch/Tong, Intriligator/Seiberg,..]
- 2d Ising Model Kramers-Wannier [Kramers/Wannier]
- AdS/CFT (string theory/conformal field theory) [Maldacena, Witten,...]:
- ...

These different descriptions, or *frames*, are often related at different regimes: strong coupling/weak coupling, high temperature/low temperature (hence "surprising").

# Motivation – Dualities

In all cases we have a "dictionary" between *observable quantities* in the two frames.

**Slogan**

Whatever can in principle be computed in one frame, may be computed in the other frame.

But the strong/weak relation implies that computational efficiency may be different in the two frames.

# Motivation – Dualities

In all cases we have a "dictionary" between *observable quantities* in the two frames.

**Slogan**

Whatever can in principle be computed in one frame, may be computed in the other frame.

But the strong/weak relation implies that computational efficiency may be different in the two frames.

**Algorithm analogy**

Embracing an extended Church-Turing-like thesis, frames can be thought of as algorithms that compute observables (outputs) based on different initial conditions (inputs). Dualities are algorithms taking different inputs, but producing the same outputs.

Can we make this precise in some settings?

# Motivation – Statistical physics tasks

Let's try to do so in statistical physics!

**Proto-task**

Given a distribution of states $\mathbf{s}_\alpha$, $\alpha \in C$, guess an *energy function* (Hamiltonian) determining Boltzmann weights

$$\{\mathbf{s}_\alpha\}_{\alpha \in C} \mapsto \{\mathbf{s}_\alpha \mapsto e^{-H(\mathbf{s}_\alpha)}\}$$

such that the observed data is "close" to the respective probability distribution

$$p(\mathbf{s}_\alpha) = \frac{e^{-H(\mathbf{s}_\alpha)}}{Z}, \qquad Z = \sum_\beta e^{-H(\mathbf{s}_\beta)} \ .$$

# Motivation – Statistical physics tasks

Let's try to do so in statistical physics!

**Proto-task**

Given a distribution of states $\mathbf{s}_\alpha$, $\alpha \in C$, guess an *energy function* (Hamiltonian) determining Boltzmann weights

$$\{\mathbf{s}_\alpha\}_{\alpha \in C} \mapsto \{\mathbf{s}_\alpha \mapsto e^{-H(\mathbf{s}_\alpha)}\}$$

such that the observed data is "close" to the respective probability distribution

$$p(\mathbf{s}_\alpha) = \frac{e^{-H(\mathbf{s}_\alpha)}}{Z}, \qquad Z = \sum_\beta e^{-H(\mathbf{s}_\beta)} \ .$$

Machines can perform some of these tasks very well, and fast.

# Motivation – Statistical physics tasks

Example: Fully Visible Boltzmann Machines (FVBM). Let $\mathbf{s} = \{s_i\}_{i \in L}$, $s_i \in \{\pm 1\}$. Let

$$H(\mathbf{s}) = -\frac{1}{2} \sum_{i \neq j} w_{ij} s_i s_j \ .$$

The proto-task corresponds to finding $w_{ij}$ such that the negative log-likelihood

$$\mathcal{L}^{FVBM} = \log(Z) + \frac{1}{|C|} \sum_{\alpha \in C} H(\mathbf{s}_\alpha)$$

is minimised. Gradients:

$$\frac{\partial \mathcal{L}^{FVBM}}{\partial w_{ij}} = \langle s_i s_j \rangle_H - \langle s_i s_j \rangle_{data} \ .$$

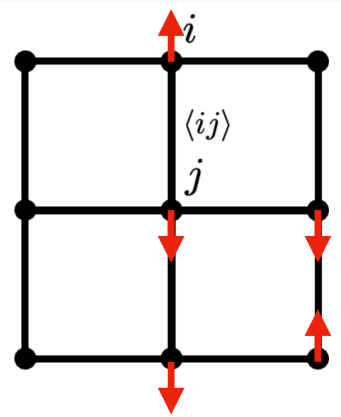
Note:

- It is very expensive to compute $\langle s_i s_j \rangle_H$, because it is very expensive to compute $Z$.
- Standard algorithms involve a Monte Carlo estimation of $\langle s_i s_j \rangle_H$
- See e.g. Contrastive Divergence [Hinton]

# Motivation – Statistical physics tasks

Special case: The *2d Ising model* on a square lattice is an example. Let $i \in L$ label the sites of the lattice, and

$$w_{ij} = \begin{cases} \beta = J/k_B T & i,j \text{ neighbouring, } \langle ij \rangle \\ 0 & \text{otherwise ,} \end{cases}$$




that is

$$H(\mathbf{s}) = -\beta \sum_{\langle ij \rangle} s_i s_j \ .$$

In this case, the problem of finding the parameters is called *Inverse Ising Problem*.

There is a huge literature that I won't be able to use (see arxiv paper 2411.04838 [Ferrari et al.] for some nice physics-based exploitations).

# Motivation – Statistical physics tasks

We will be concerned with a different, more general problem. Consider a more sophisticated statistical physicist that is *agnostic* about:

- Parametrisation of the underlying configuration of states. Spins on a lattice? Which lattice?
- Precise form of the Hamiltonian/energy
- The actual *observable* that is actually being measured.

**Guiding principle**

The less is assumed the more a potential description may be surprising/useful! Still, one needs to find well-defined training tasks...

Look at dualities in statistical physics.

# 2d Ising and Kramers-Wannier

Today we will consider one of the most elementary examples of a *duality*, Kramers-Wannier duality in the 2d Ising model [Kramers-Wannier, ...]. Recall $s_i \in \{\pm 1\}$, and

$$H(\mathbf{s}) = -\beta \sum_{\langle ij \rangle} s_i s_j \ .$$

Note that there is a $\mathbb{Z}_2$ symmetry

$$s_i \mapsto -s_i \ , \ \beta \mapsto -\beta \ .$$

The energy is lowest ("ground state") when spins are "aligned", $s_i = +1$ or $s_i = -1$.

Kramers-Wannier duality is in this case a *self*-duality relating different $\beta$ regimes.[1]

[1] Physically, the duality is related to *gauging* the $\mathbb{Z}_2$ symmetry.

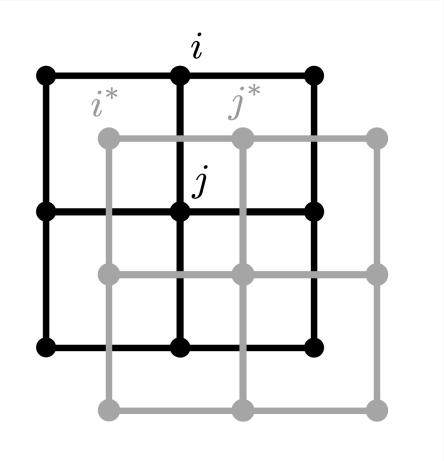
# 2d Ising and Kramers-Wannier

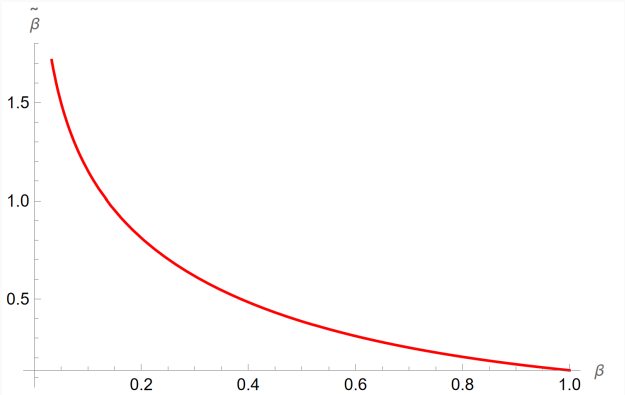
More in details, the *self*-duality relates models

$$H(\mathbf{s}) = -\beta \sum_{\langle ij \rangle} s_i s_j \;\leftrightarrow \tilde{H}(\mathbf{s}^*) = -\tilde{\beta} \sum_{\langle i^* j^* \rangle} s_{i^*} s_{j^*}$$

where $\tilde{H}$ is defined on the *dual lattice*[2] $L^\vee$ and

$$\sinh(2\beta)\sinh(2\tilde{\beta}) = 1 \ .$$



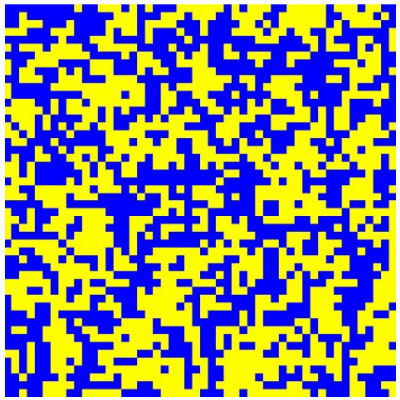




The partition functions agree up to an overall constant.

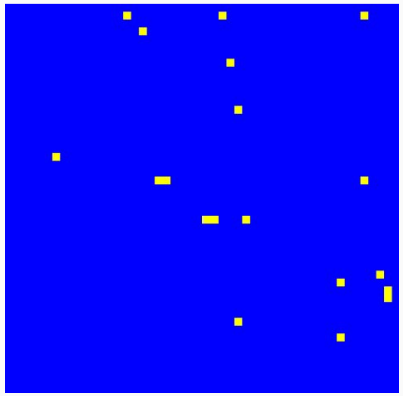
[2] In general, $L^\vee := \{y \in \mathbb{R}^n \mid \forall x \in L,\ (y,x) \in \mathbb{Z}\}$ .

# 2d Ising and Kramers-Wannier

Remark: this is a large-temperature/low-temperature duality.

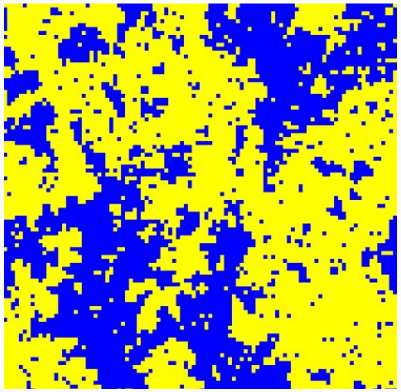


$\leftrightarrow$



$\beta << 0.44$ $\beta >> 0.44$

with a self-dual point at $\beta \simeq 0.44$



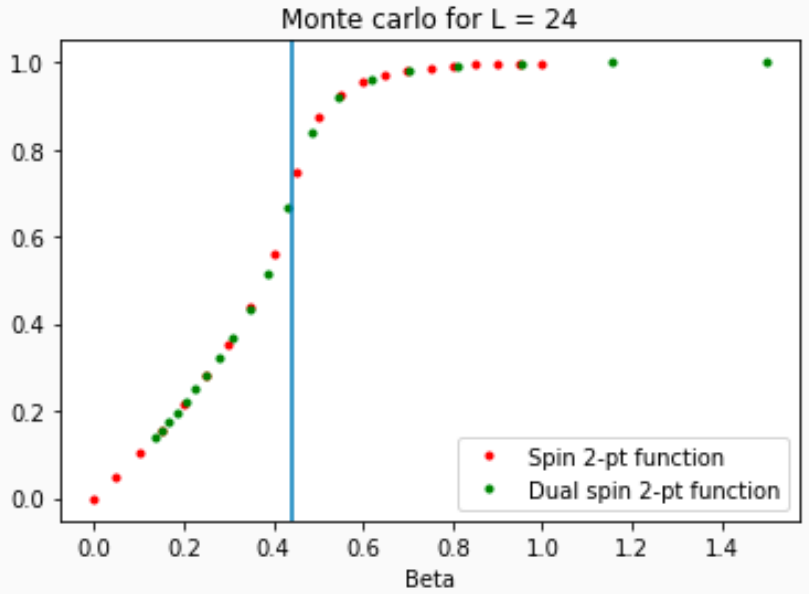
Important: the map between spin configurations is *not* 1:1!

# 2d Ising and Kramers-Wannier

Correlation functions, however, need to match

$$\langle s_i s_j \cdots \rangle_H = \langle e^{-2\tilde{\beta}\tilde{s}_{i*}\tilde{s}_{j*}} \cdots \rangle_{\tilde{H}}$$

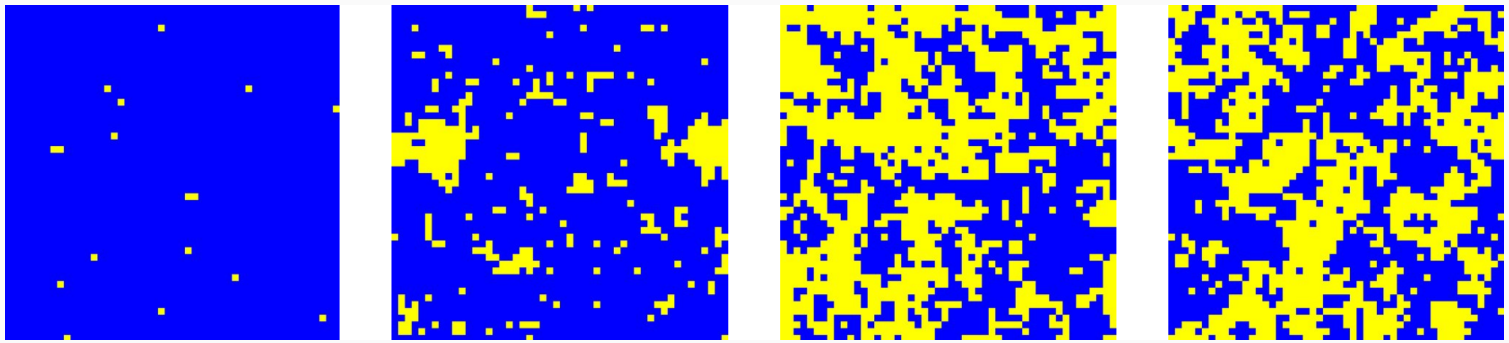
Example: Implement Monte Carlo in *e.g. Mathematica* to compute the correlations on a $24 \times 24$ lattice:




Can this matching of data be used to *learn* the duality?

# Duality as an optimisation problem – Data generation

Yes we can! Start from the theory at a given $\beta$, and generate lattice spin configurations $C = \{L_\alpha\}_\alpha$ using Markov Chain Monte Carlo



We can the compute

$$\langle \mathcal{O}_{ij}(\mathbf{s})\mathcal{O}_{kl}(\mathbf{s})...\mathcal{O}_{mn}(\mathbf{s})\rangle_{data} := \frac{1}{|C|}\sum_{\alpha\in C}\mathcal{O}_{ij}(\mathbf{s}_\alpha)\mathcal{O}_{kl}(\mathbf{s}_\alpha)...\mathcal{O}_{mn}(\mathbf{s}_\alpha)\ .$$

which is an estimate for

$$\langle \mathcal{O}_{ij}(\mathbf{s})\mathcal{O}_{kl}(\mathbf{s})...\mathcal{O}_{mn}(\mathbf{s})\rangle_H\ .$$

# Duality as an optimisation problem – Duality learning

Once the data has thus been prepared, we can define the following problem:

**Kramers-Wannier Duality Learning Task - Rough sketch**

Given *some* lattice $L^?$ with spins $s_{i'} \in \{\pm 1\}$, as well as

- A reasonably general Hamiltonian

$$H^t(\mathbf{s}') = -t_0 \sum_{\langle i'j' \rangle} s_{i'} s_{j'} - t_1 \sum_{(i'j'k'l')} s_{i'} s_{j'} s_{k'} s_{l'} - \cdots$$

- A (sufficiently local) observable $\mathcal{O}^\theta_{ij} : \mathbf{s}' \to \mathbb{R}$ parametrised by $\theta$

find $(t = t^\star, \theta = \theta^\star)$ such that all distances between allcorrelation functions

$$|\langle \mathcal{O}_{ij}(\mathbf{s}) \mathcal{O}_{kl}(\mathbf{s}) ... \mathcal{O}_{mn}(\mathbf{s}) \rangle_H - \langle \mathcal{O}^\theta_{ij}(\mathbf{s}') \mathcal{O}^\theta_{kl}(\mathbf{s}') ... \mathcal{O}^\theta_{mn}(\mathbf{s}') \rangle_{H^t}|^2$$

are minimised.

Note: The above strategy may be thought of as a decoding/encoding process.

See also [Betzler/Krippendorf]

# Duality as an optimisation problem – Expected solutions

Provided the problem is set up correctly, one should recover two solutions:

The original model:

- $L^? = L$
- $H^{t^*} = H(\beta)$
- $\mathcal{O}_{ij}^{\theta^\star} = \mathcal{O}_{ij} = s_i s_j$.

The dual model:

- $L^? = L^\vee$
- $H^{t^*} = H(\tilde{\beta})$
- $\mathcal{O}_{ij}^{\theta^\star} = e^{-\tilde{\beta} s_{i^*} s_{j^*}}$.

# Duality as an optimisation problem – Expected solutions

Provided the problem is set up correctly, one should recover two solutions:

The original model:

- $L^? = L$
- $H^{t^*} = H(\beta)$
- $\mathcal{O}_{ij}^{\theta^\star} = \mathcal{O}_{ij} = s_i s_j$.

The dual model:

- $L^? = L^\vee$
- $H^{t^*} = H(\tilde{\beta})$
- $\mathcal{O}_{ij}^{\theta^\star} = e^{-\tilde{\beta} s_{i*} s_{j*}}$.

To achieve this result, one needs to:

- identify a suitable parametrisation of the Hamiltonian $H^t$ (as before) and observable $\mathcal{O}_{ij}^{\theta}$
- define an appropriate *loss function*
- develop an *optimisation scheme*.

# Parametrisation of the observables

We assume that the lattice $L^?$ is also square and so $\mathbf{s}^* \cong \mathbf{s}$.

In principle, one could parametrise $\mathcal{O}^\theta_{ij}$ as any NN with a large input (the spins) and one output (the value of the observable).

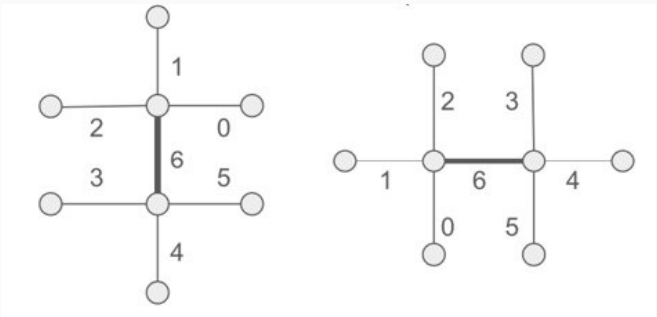
**Observable Intuition**

Focus attention on some neighbouring links.

Reasonable example:

$$\mathcal{O}^\theta(\mathbf{s}) = \theta_2 \text{Gumbel-Softmax}(\mathbf{s})(\theta_1)^T \delta^{ij}(\mathbf{s}) + \theta_3$$

where $\delta^{ij}(\mathbf{s})$ is a vector with the values of the *neighbouring links* of $\langle ij \rangle$ (=6 below).



# Loss function

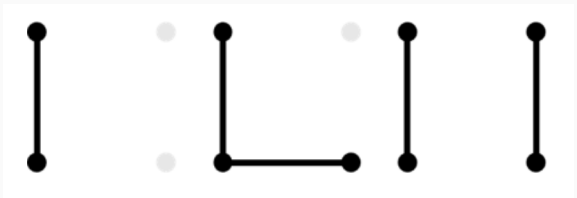
Recall desired matches (with $\mathbf{s} = \mathbf{s}^*$):

$$|\langle \mathcal{O}_{ij}(\mathbf{s})\mathcal{O}_{kl}(\mathbf{s})...\mathcal{O}_{mn}(\mathbf{s})\rangle_H - \langle \mathcal{O}^\theta_{ij}(\mathbf{s})\mathcal{O}^\theta_{kl}(\mathbf{s})...\mathcal{O}^\theta_{mn}(\mathbf{s})\rangle_{H^t}|^2$$

Standard approaches involve *kernel methods* [Li/Zwersky/Zemel][3], however:

- Care is needed because $s_i s_i = 1$. The issue can be solved by modifying kernels.
- Correlation functions involving far-away spins carry little information, noise-to-signal problem

Solution: just focus on a handful of these correlators! For instance:



[3] Cf. *Maximum Mean Discrepancy*.

# Loss function

More formally, let

- $f_{ij} : \mathbf{s} \to \mathbb{R}$ a function labelled by a link $\langle ij \rangle$ (such as the observables);
- $A$ be a set of subsets of links.

and for $a \in A$ define the "product of the funcion values over links"

$$\psi^a(f) = \prod_{\langle ij \rangle \in a} f_{ij} \ .$$

**Loss function**

We define the loss $\mathcal{L}$ as the Mean-Squared-Error

$$\mathcal{L} = \sum_{a \in A} l_a^2 \ .$$

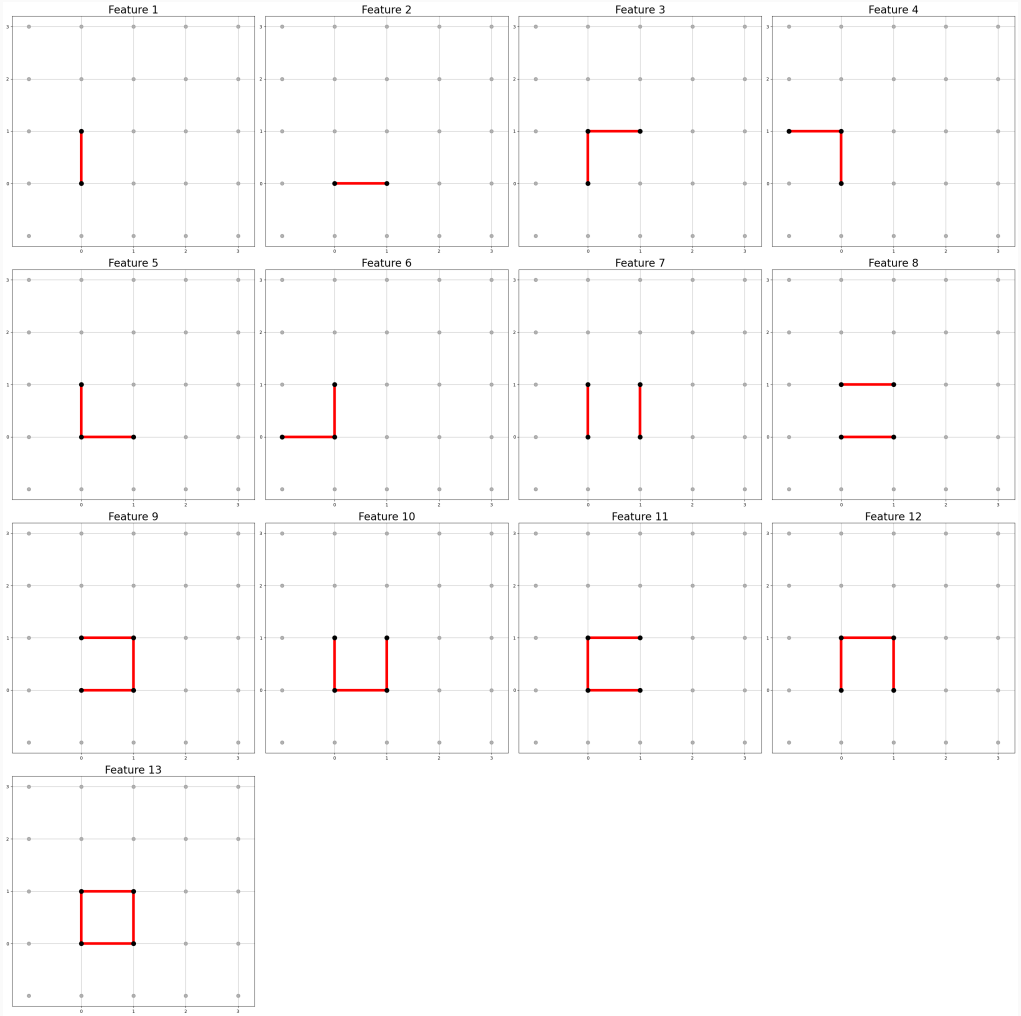
with

$$l_a = \langle \psi^a(\mathcal{O}_{ij}) \rangle_H - \left\langle \psi^a \left( \mathcal{O}_{ij}^\theta \right) \right\rangle_{H^t}$$

for a suitable subset $A$ (see next slide).

# Loss function

Here are the selected features (subset $A$) in a little more details:



# Optimisation

To compute the loss, we need to compute gradients with respect to the parameters $t$ and $\theta$ entering the pair $(H^t, \mathcal{O}_{ij}^{\theta})$. This can be done as follows:

- $\partial_\theta \mathcal{L}$ is easy to compute: simply use backpropagation `loss.backward()`
- $\partial_t \mathcal{L}$ is computationally hard, as $t$ directly affects the probability measure of the sampling.
- As far as I can tell, classic methods such as CD cannot be used in this case.

However, a simple computation shows that the $\partial_t \mathcal{L}$ gradients can be expressed in terms of expecation values:[4]

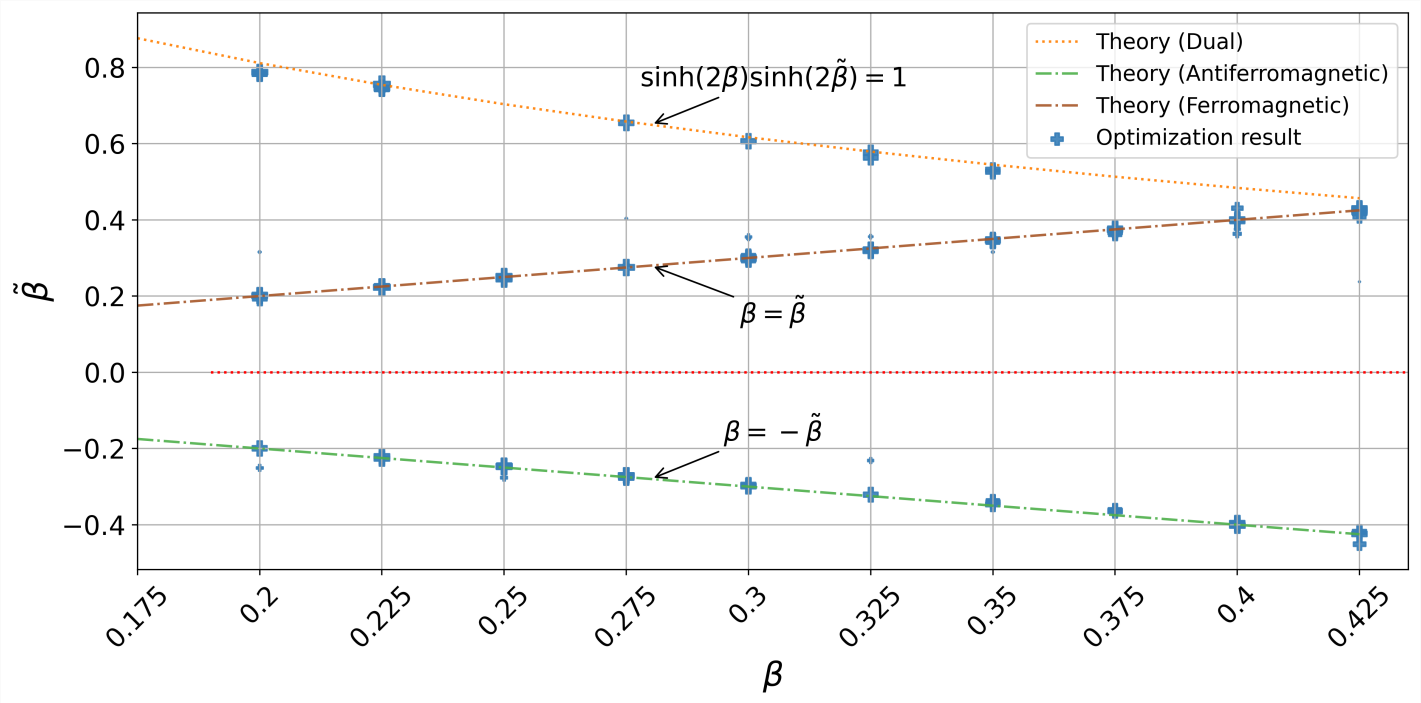
$$\partial_t \mathcal{L} = 2 \left\langle \sum_{a \in A} l_a \left( \sum_{\langle ij \rangle} \langle s_i s_j \rangle_{H^t} - \sum_{\langle ij \rangle} s_i s_j \right) \psi^a \right\rangle_{H^t}$$

Thus, they can be estimated using MCMC! Slow but feasible...

[4] Similar to REINFORCE

# First results – KW rediscovery

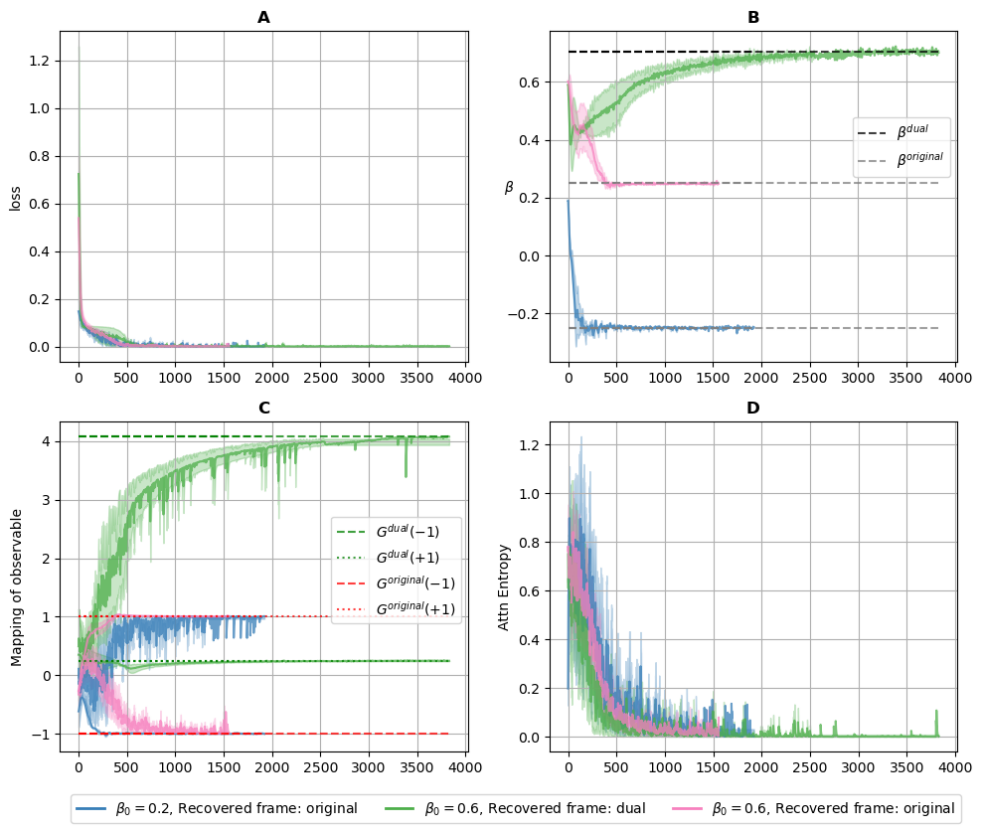
Run the optimisation problem on a $8 \times 8$ lattice at different temperatures, first with $H^t$ truncated to the Ising Hamiltonian (that is $t = (t_0)$, in the graph $\tilde{\beta}$).




Rejoice, the duality is rediscovered! As well as the antiferromagnetic image of the original model $\beta \mapsto -\beta$.

# First results – KW rediscovery

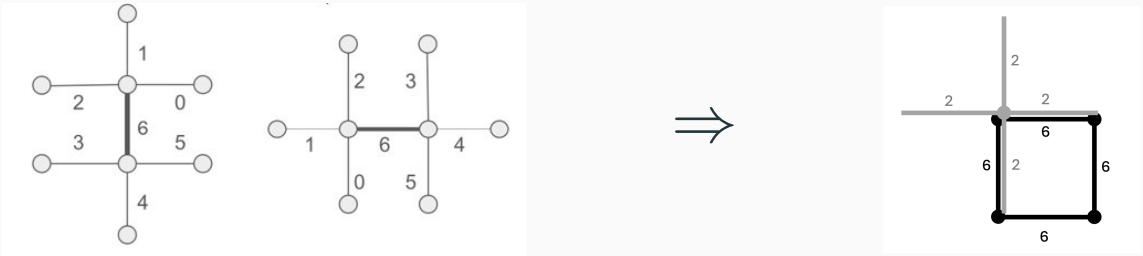
Let's have a look at the training dynamics as well as the discovered observables



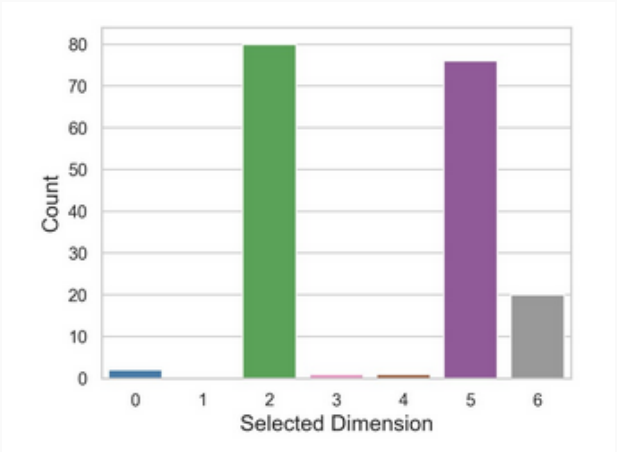

All looks pretty good. Low value of attention entropy indicates one link is mostly picked. Expressions $\mathcal{O}_{ij}^{\theta^\star} = s_i s_j$, $\mathcal{O}_{ij}^{\theta^\star} = e^{-2\tilde{\beta} s_{i*} s_{j*}}$ are found!

# First results – KW rediscovery

It is interesting to see whether it is *really* the original links $\langle ij \rangle$ or the dual links $\langle j^* i^* \rangle$ that are discovered. Recall:




That is, link 6 recovers the original lattice, 2 or 5 recover the dual. Let us look at the links selected by the optimisation problem:



# First results – KW rediscovery

Some details and problems:

- Computing gradiends takes time, as MCMC chains need to be equilibrated
- The MCMC-based gradients are also noisy. It is advisable to use variance reduction techniques such as *control variates*.
- The training rarely converges when the original model is in the ordered phase. This is similar to observations in the Inverse Ising problem –deep?
- $8 \times 8$ is very small, still it takes hours to do these runs.
- Adding more couplings complicates things...

# Generalisations

In any case, given the success with KW in the standard 2d Ising model, one can consider generalisations. Obvious candidates:

- Next-to-nearest neighbour 2d Ising
- Plaquette (see below).

The former has interesting dualities to brick-wall or honeycomb lattices (explicit expressions in e.g. [Strycharski(Koza]). The latter has not been studied in great details, it probably requires "non-planar" lattices.

**Plaquette generalisation task**

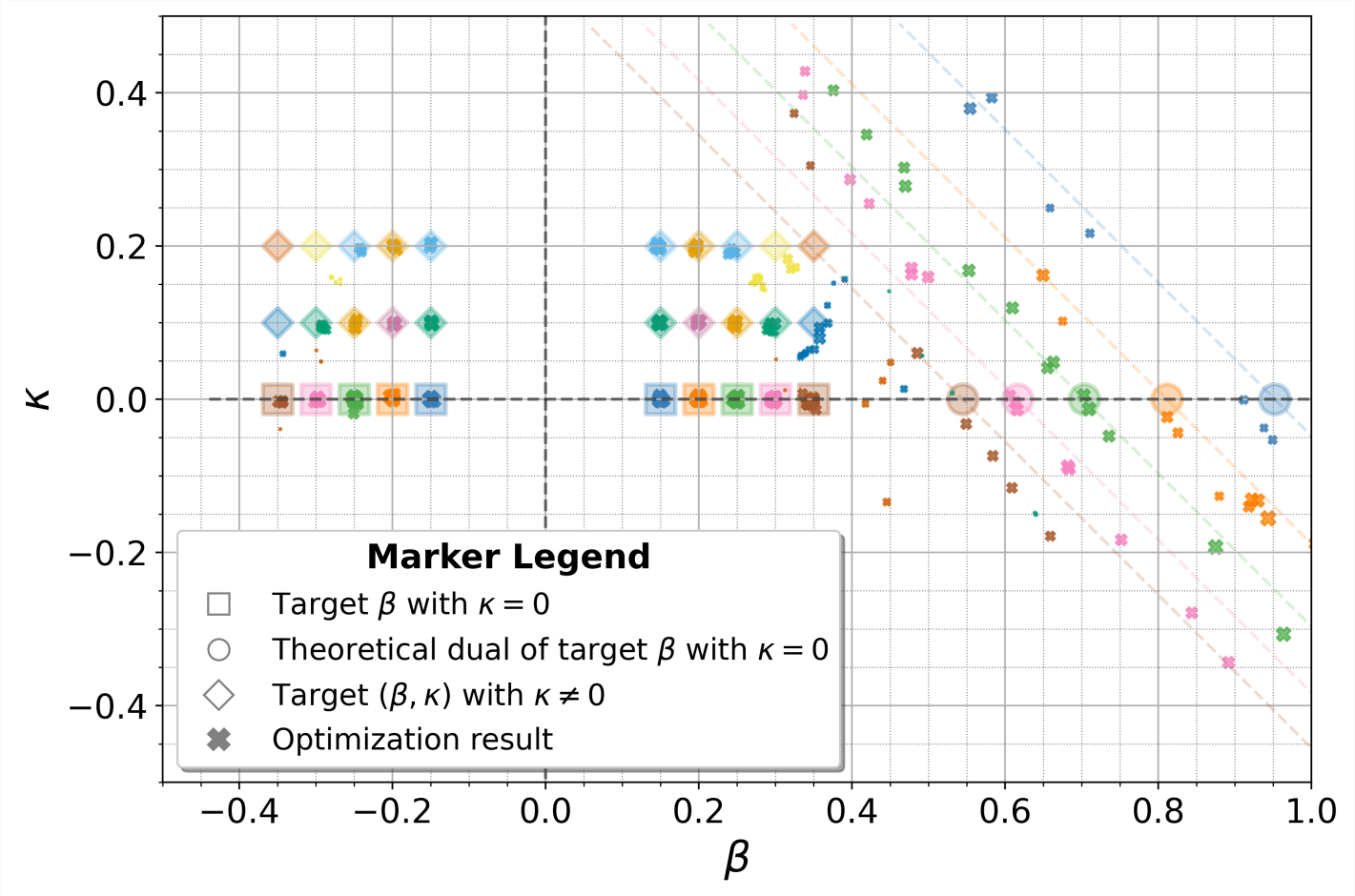
What does our set-up outputs for the plaquette model

$$H(\mathbf{s}) = -\beta \sum_{\langle ij \rangle} s_i s_j - \kappa \sum_{(ijkl)} s_i s_j s_k s_l \ ,$$

where $(ijkl)$ label *plaquettes*?

# Generalisation – Plaquette results

Some results for the plaquette model, square lattice:



# Generalisation – Plaquette results

Some observations on the emergence of anti-diagonal "clusters":

- When models are deep in the ordered phase (almost all spins aligned), an excellent approximation for the model is given in terms of 'rare, isolated spin flips
- These happen with probability $p = p(\tilde{\beta} + \tilde{\kappa})$
- The ML model learned that the model becomes effectively a one-parameter model in this regime
- The loss is very low, 0 up to noise!

**Approximate duals.**

Even on a square lattice, one can find "approximate duals" deep in the ordered phase.

# Conclusions and future directions

To conclude:

- The standard problem of fitting parameters in a Hamiltonian of a statistical model/energy of an energy-based model has interesting generalisations
- It is possible to define duality discovery in statistical physics as an instance of such more general problems
- Solutions lead to known dualities as well as hints about novel physics

Several future directions:

- Study the ML problem better! (New ways to optimise etc.)
- Generalise duality discovery! New dualities?
- More modestly, apply our framework to more statistical physics models!

# The End