140B Statistical Physics

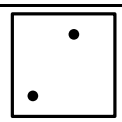
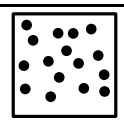
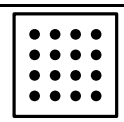
Part 3. Phase Transitions

Liquid-Gas Transition

■ Phases and Phase Transitions

■ States of Matter

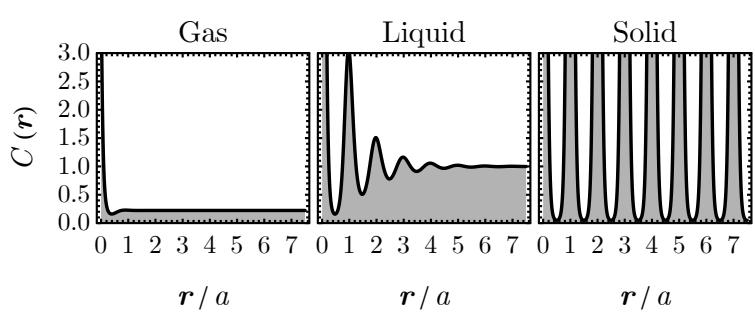
Gas, **liquid**, and **solid** are *states of matter* in our everyday life. They are all made of atoms (or molecules), yet their macroscopic behaviors are very different.

	Fluid		Solid
	Gas	Liquid	
			
Atom arrangement	Random	Random	Periodic
Traslation symmetry	Preserved	Preserved	Broken
Correlation	Negligible	Short-range	Long-range
Density	Low	High	High
Compressibility	High	Low	Low
Fluidity	✓	✓	×
Sound modes	1	1	3

- The key difference lies in the **density correlation**

$$C(\boldsymbol{r}) = \langle n(\boldsymbol{r})\, n(\boldsymbol{0})\rangle, \tag{1}$$

which quantifies how the presence of an atom at the origin $\boldsymbol{0}$ affects the likelihood of finding another atom at some point $\boldsymbol{r}$ away from the origin.

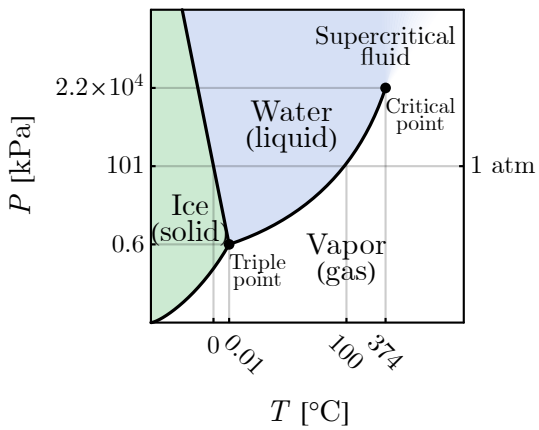


- **Gas**: atoms are far apart, almost independent, no regular arrangement. Correlation is *negligible* (very weak and short-ranged).
- **Liquid**: atoms are close together, sliding around each other, no regular arrangement. Correlation is *strong* but *short-ranged*.
- **Solid**: atoms are closely packed, arranging in *regular patterns* (crystal lattices). Correlation is *strong* and *long-ranged*.

- Different *microscopic order* of atoms leads to different *macroscopic properties* of the matter.

Phase Diagram

Phase diagram show the macroscopic conditions under which distinct states of matter occur at equilibrium. The phase diagram of water (H_2O) looks like



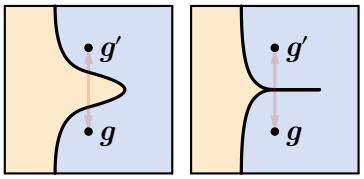
- Each **macrostate** of a system is labeled by a set of coordinates $\boldsymbol{g} = (g_1, g_2, \ldots)$ in the phase diagram. For example, $\boldsymbol{g} = (T, P)$ in the above phase diagram.
- **Phase transition** happens across the **phase boundary**, where the free energy density

$$f(\boldsymbol{g}) = \lim_{V\to\infty} \frac{F(\boldsymbol{g})}{V} \tag{2}$$

  becomes *non-analytic* with respect to $\boldsymbol{g}$ in the thermodynamic limit, i.e. its derivatives become *singular* or *discontinuous*.

- **Phase**: Two macrostates $\boldsymbol{g}$ and $\boldsymbol{g}'$ belongs to the *same phase* as long as there exists a continuous path from $\boldsymbol{g}$ to $\boldsymbol{g}'$ without encountering any phase transition.
  - **Liquid** and **gas** are in the *same* phase (the **fluid** phase).
    - Even though liquid and gas are separated by a phase transition, it does not make them different phases of matter, because the phase transition can be circumvented through the **supercritical fluid** regime.

- This is the only logically consistent way of defining phases: they should be *unified* by the *absence* of phase transitions, other than *separated* by the *presence* of phase transitions.



- **Solid** is in a *different* phase from liquid and gas.

  There can be many different solid phases characterized by different ways of breaking the *translation* and *rotation* symmetry of the space.

Lattice Gas Model

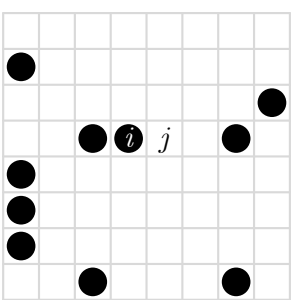
Model Hamiltonian

Interaction plays an essential role to bind atoms together to form the liquid state at low temperature. To model the liquid-gas transition, we must go beyond the non-interacting limit of ideal gas, and include the *interaction effects*.

The **lattice gas model** is toy model to describe *interacting* atoms on a lattice. Its energy function is

$$E(\boldsymbol{n}) = -g \sum_{\langle ij \rangle} n_i \, n_j - \mu \sum_i n_i. \tag{3}$$

- Assuming the space is partitioned into many small cells (or lattice **sites**), each labeled by an index, denoted as i, j.



- The **microstate** of the system is described by a collection of binary random variables $\boldsymbol{n} = \{n_i\}$, where

$$n_i = \begin{cases} 0 & \text{empty}, \\ 1 & \text{occupied}, \end{cases} \tag{4}$$

describes the number of atoms on site i, assuming each site can not host more than one atom

(the **hard-core** condition).

- $g > 0$ is the **attractive interaction strength** between atoms: it provides an *energy reward* if atoms are *neighboring* to each other.
  - $\langle ij \rangle$ denotes the summation over pairs of neighboring sites i and j.
  - The energy $-g\, n_i\, n_j$ under different scenarios:

n_i	n_j	$-g\, n_i\, n_j$	
0	0	0	
0	1	0	(no energy reward)
1	0	0	
1	1	$-g$	(attractive energy reward)

(5)

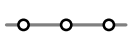
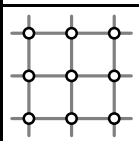
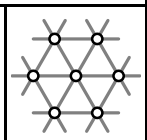
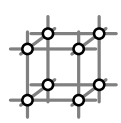
- $\mu \in \mathbb{R}$ is the **chemical potential**, s.t. each atom experiences a potential energy of $-\mu$ to occupy a site.
  - It is introduced to tune the total **particle number**

$$\boxed{N(\boldsymbol{n}) = \sum_i n_i.} \tag{6}$$

To lower the energy, N tends to increase with increasing μ.

■ Lattice Structure

The lattice model can be implemented on various lattices in different dimensions.

1D	2D		3D
			
chain	sqaure	triangle	cubic
$q = 2$	$q = 4$	$q = 6$	$q = 6$

- **Coordination number** q: the number of *nearest-neighbor sites* around each given site, reflecting how connected or coordinated each site is within the lattice.
- **Volume**: Number of **sites** on the lattice.

$$\boxed{\sum_i 1 = V} \tag{7}$$

  - Number of nearest-neighbor **bonds** (links) on the lattice.

$$\boxed{\sum_{\langle ij \rangle} 1 = \frac{q}{2}\, V.} \tag{8}$$

■ Probability Distribution

The probability distribution of the lattice gas configuration $\boldsymbol{n}$ is

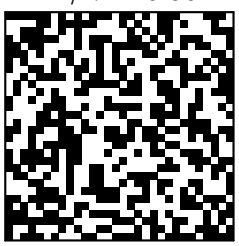
$$\boxed{p(\boldsymbol{n}) = \frac{1}{Z}\, e^{-\beta\, E(\boldsymbol{n})},} \tag{9}$$

where $\beta = 1/T$ is the inverse temperature, and Z is the partition function

$$Z = \sum_{\boldsymbol{n}} e^{-\beta\, E(\boldsymbol{n})}. \tag{10}$$

- **High temperature limit** ($T \to \infty$, $\beta \to 0$): $p(\boldsymbol{n}) = 1/Z$ reduces to a uniform distribution of $\boldsymbol{n}$. In this case, each site will be occupied or empty with $1/2$ to $1/2$ probability ⇒ **Supercritical fluid** (an uniform mixture of liquid and gas).

$N/V = 0.50$

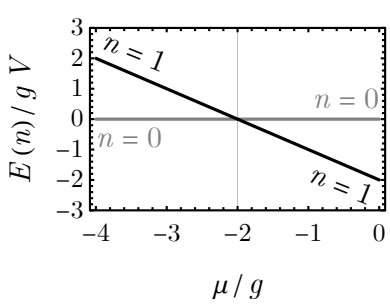


- **Low temperature limit** ($T \to 0$, $\beta \to \infty$): $p(\boldsymbol{n})$ will be dominated by the lowest-energy configuration $\boldsymbol{n}$.
  - A over-simplified estimation: assuming the same particle number on every site, i.e. $\forall\, i : n_i = n$, the energy function becomes

$$\boxed{E(n) = -V\, n\left(\frac{q}{2}\, g\, n + \mu\right).} \tag{11}$$

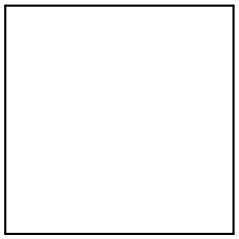
where V is the lattice volume and q is the coordination number (assuming square lattice with $q = 4$).

Exc 1 Derive Eq. (11).



- When $\mu < -2\,g$, the configuration $n = 0$ is lower in energy. ⇒ **Gas** (in the dilute limit). Atoms are far apart from each other — so *dilute* that we can't see any atom on a finite lattice.

$N / V = 0$



- When $\mu > -2\,g$, the configuration $n = 1$ is lower in energy. ⇒ **Liquid** (in the dense limit). Atoms are closely packed, next to each other — so *dense* that we can't see any bubble on a finite lattice.

$N / V = 1$

Markov Chain Monte Carlo

Monte Carlo Sampling

Monte Carlo sampling is an important numerical algorithm to draw random variables $\boldsymbol{x}$ from the Boltzmann distribution $p(\boldsymbol{x}) \propto e^{-\beta\, E(\boldsymbol{x})}$, given the energy function $E(\boldsymbol{x})$.

The main idea is to

- start with an *arbitrary* initial sample $\boldsymbol{x}^{(0)} \approx p_0\big(\boldsymbol{x}^{(0)}\big)$,
- generate a new sample $\boldsymbol{x}^{(1)}$ following the **transition probability** $p\big(\boldsymbol{x}^{(1)} \,|\, \boldsymbol{x}^{(0)}\big)$ (to be designed later)

$$\boldsymbol{x}^{(0)} \to \boldsymbol{x}^{(1)} \approx p\big(\boldsymbol{x}^{(1)} \,|\, \boldsymbol{x}^{(0)}\big), \tag{12}$$

- iterate the process to construct a chain of samples

$$\boldsymbol{x}^{(0)} \to \boldsymbol{x}^{(1)} \to \boldsymbol{x}^{(2)} \to \ldots \to \boldsymbol{x}^{(n)} \to \ldots. \tag{13}$$

The hope is that after sufficient number of steps ($n \gg 1$), the probability distribution $p_n(\boldsymbol{x}^{(n)})$ of the last generated sample $\boldsymbol{x}^{(n)}$ will converge to the desired distribution $p(\boldsymbol{x}) \propto e^{-\beta\, E(\boldsymbol{x})}$.

Markov Chain

The iterative sampling process is a **Markov process**. "Markov" means the probability distribution of the *next* sample $\boldsymbol{x}^{(n+1)}$ depends *only* on the *current* sample $\boldsymbol{x}^{(n)}$ but *not* any *previous*

samples $\boldsymbol{x}^{(n-1)}$, $\boldsymbol{x}^{(n-2)}$, ... In short, the Markov process is a stochastic process that has *no memory* of *history*.

- The chain of samples Eq. (13) generated by the Markov process is a **Markov chain**, along which the sample probability evolves as

$$p_n\left(\boldsymbol{x}^{(n)}\right) = \sum_{\boldsymbol{x}^{(n-1)}} p\left(\boldsymbol{x}^{(n)} \,\middle|\, \boldsymbol{x}^{(n-1)}\right) p_{n-1}\left(\boldsymbol{x}^{(n-1)}\right). \tag{14}$$

  The evolution is uniquely determined by the transition probability $p(\boldsymbol{x}' \mid \boldsymbol{x})$.

- One may start with any initial distribution $p_0(\boldsymbol{x})$. Under the Markov process, the distribution $p_n(\boldsymbol{x})$ might evolve and relax to a **stationary distribution** $p(\boldsymbol{x})$

$$\lim_{n\to\infty} p_n(\boldsymbol{x}) = p(\boldsymbol{x}). \tag{15}$$

The goal is to design the *transition* probability $p(\boldsymbol{x}' \mid \boldsymbol{x})$ wisely, such that the *stationary* distribution $p(\boldsymbol{x})$ matches our desired *target* distribution $e^{-\beta\, E(\boldsymbol{x})}$.

Detailed Balance

Detailed balance is a sufficient (but not necessary) condition for a stationary distribution $p(\boldsymbol{x})$ to exist, which requires

$$\boxed{p(\boldsymbol{x}' \mid \boldsymbol{x})\, p(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{x}')\, p(\boldsymbol{x}').} \tag{16}$$

- Argument: detailed balance implies that, for every pair of states , the flow of probability is perfectly balanced between the *forward* sampling $\boldsymbol{x} \to \boldsymbol{x}'$ and the *reversed* sampling $\boldsymbol{x}' \to \boldsymbol{x}$, such that the probability will not evolve in time, i.e. stationary.

- Alternatively, Eq. (16) can be written as

$$\frac{p(\boldsymbol{x}' \mid \boldsymbol{x})}{p(\boldsymbol{x} \mid \boldsymbol{x}')} = \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x})}. \tag{17}$$

  Note: for $\boldsymbol{x}' = \boldsymbol{x}$, Eq. (17) is trivially satisfied, thus one only need to check the detailed balance for the case of $\boldsymbol{x}' \neq \boldsymbol{x}$.

Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm is a systematic approach to realize the Markov process $\boldsymbol{x} \to \boldsymbol{x}'$ that is detailed balanced with respect to the desired stationary distribution $p(\boldsymbol{x})$.

- Given the current sample $\boldsymbol{x}$.
- Propose a new sample $\tilde{\boldsymbol{x}}$ by drawing from the **proposal probability** $q(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})$.
- Calculate the **acceptance probability**

$$\boxed{\alpha(\tilde{\boldsymbol{x}}, \boldsymbol{x}) = \min\left(1, \frac{p(\tilde{\boldsymbol{x}})\, q(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})}{p(\boldsymbol{x})\, q(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})}\right) \in [0, 1].} \tag{18}$$

- **Accept** or **Reject**:
  - with probability $\alpha(\tilde{\boldsymbol{x}}, \boldsymbol{x})$, *accept* the new sample and set $\boldsymbol{x}' = \tilde{\boldsymbol{x}}$,
  - with probability $1 - \alpha(\tilde{\boldsymbol{x}}, \boldsymbol{x})$, *reject* the new sample and set $\boldsymbol{x}' = \boldsymbol{x}$.
- Return $\boldsymbol{x}'$ as the updated sample.

For $\boldsymbol{x}' \neq \boldsymbol{x}$, the transition probability is then given by

$$p(\boldsymbol{x}' \mid \boldsymbol{x}) = q(\boldsymbol{x}' \mid \boldsymbol{x})\, \alpha(\boldsymbol{x}', \boldsymbol{x}), \tag{19}$$

which satisfies detailed balance by design.

Exc 2 Show that the transition probability in Eq. (19) satisfies the detailed balance condition in Eq. (17).

If the proposal probability is *symmetric*,

$$q(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) = q(\tilde{\boldsymbol{x}} \mid \boldsymbol{x}), \tag{20}$$

the acceptance probability will be simplified to

$$\boxed{\alpha(\tilde{\boldsymbol{x}}, \boldsymbol{x}) = \min\left(1, \frac{p(\tilde{\boldsymbol{x}})}{p(\boldsymbol{x})}\right).} \tag{21}$$

If the desired stationary probability distribution $p(\boldsymbol{x})$ is further modeled by the energy function $E(\boldsymbol{x})$ as $p(\boldsymbol{x}) = e^{-E(\boldsymbol{x})} / Z$, the acceptance probability will be further reduced to

$$\boxed{\begin{aligned} \alpha(\tilde{\boldsymbol{x}}, \boldsymbol{x}) &= \min\left(1, e^{E(\boldsymbol{x}) - E(\tilde{\boldsymbol{x}})}\right) \\ &= \begin{cases} 1 & E(\tilde{\boldsymbol{x}}) \leq E(\boldsymbol{x}), \\ e^{E(\boldsymbol{x}) - E(\tilde{\boldsymbol{x}})} & E(\tilde{\boldsymbol{x}}) > E(\boldsymbol{x}). \end{cases} \end{aligned}} \tag{22}$$

In this case, the accept-or-reject rule is simple:

- if the proposed sample has a lower (or equal) energy, i.e. $E(\tilde{\boldsymbol{x}}) \leq E(\boldsymbol{x})$, the proposal is always accepted,
- if the proposed sample has a higher energy, i.e. $E(\tilde{\boldsymbol{x}}) > E(\boldsymbol{x})$, the proposal is accepted with probability $e^{E(\boldsymbol{x}) - E(\tilde{\boldsymbol{x}})}$ that decays exponentially with the energy difference.

Monte Carlo Simulation

Monte Carlo sampling is an important numerical algorithm to draw samples from the Boltzmann distribution $p(\boldsymbol{n}) \propto e^{-\beta E(\boldsymbol{n})}$, given the energy function $E(\boldsymbol{n})$:

$$E(\boldsymbol{n}) = -g \sum_{\langle ij \rangle} n_i\, n_j - \mu \sum_i n_i. \tag{23}$$

The algorithm has the following steps:

- Start with (arbitrary) initial configuration $\boldsymbol{n}$.

- Randomly choose a site i.
- Propose a new configuration $\boldsymbol{n}'$ by flipping the occupation number on site i, i.e.

$$\begin{aligned} n'_i &= 1 - n_i, \\ n'_j &= n_j \text{ (for all } j \neq i), \end{aligned} \tag{24}$$

- Compute the energy difference

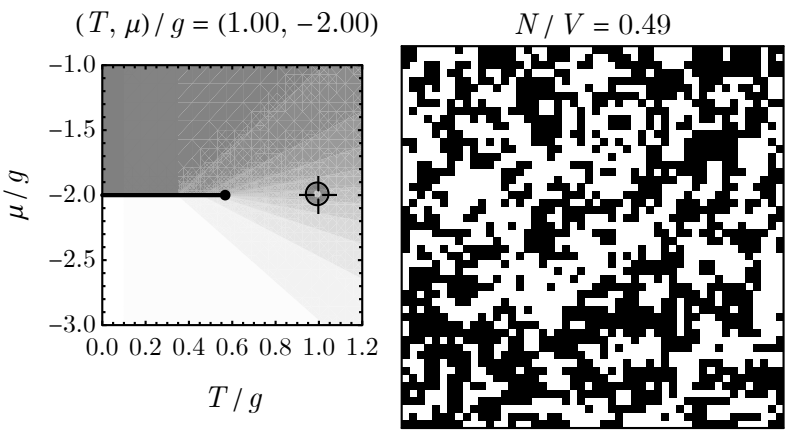
$$\Delta E = E(\boldsymbol{n}') - E(\boldsymbol{n}). \tag{25}$$

  - If the energy lowers, $\Delta E < 0$, *accept* the update $\boldsymbol{n} \to \boldsymbol{n}'$,
  - If the energy raises, $\Delta E > 0$, *accept* the update $\boldsymbol{n} \to \boldsymbol{n}'$ with probability

$$\alpha = e^{-\beta \Delta E}, \tag{26}$$

  and *reject* with probability $1 - \alpha$ (if rejected, configuration will remain in $\boldsymbol{n}$).
- Repeat to collect a series of configurations $\boldsymbol{n}$.

Update rate: 0 1 10 200 5000



Magnetic Transition

Ising Model

Model Hamiltonian

Ising model is another lattice model in statistical mechanics that describes the behavior of **magnetic spins** on a lattice, originally introduced to study the order-disorder **magnetic transition** in magnetic materials. It is described by the energy function

$$\boxed{E(\boldsymbol{s}) = -J \sum_{\langle ij \rangle} s_i\, s_j - h \sum_i s_i.} \tag{27}$$

- The **microscopic state** is describes by a set of Ising spins $\boldsymbol{s} = \{s_i\}$, where

$$\boxed{s_i = \begin{cases} +1 & \text{spin up,} \\ -1 & \text{spin down,} \end{cases}} \tag{28}$$

  describes the spin state on the lattice site i.

- $J > 0$ is the **coupling constant**, representing the *interaction strength* between neighboring spins.
  - $\langle ij \rangle$ denotes the summation over pairs of neighboring sites i and j.
  - The energy $-J\, s_i\, s_j$ under different scenarios:

spins	s_i	s_j	$-J\, s_i\, s_j$	energy
aligned	$+1$	$+1$	$-J$	gain
anti-aligned	$+1$	-1	$+J$	cost
anti-aligned	-1	$+1$	$+J$	cost
aligned	-1	-1	$-J$	gain

(29)

- $h \in \mathbb{R}$ is the **external magnetic field**, s.t. the spin *gains* energy h by *aligning* with the external magnetic field, and *costs* energy h to *anti-align*.
  - It is introduced to tune the total **magnetization**

$$\boxed{M = \sum_i s_i.} \tag{30}$$

  The magnetization M is energetically favored to *align* with the external magnetic field h (i.e. to be of the same sign).

Ising Symmetry

The **Ising model** and the **lattice gas model** are closely related.

- Mapping between on-site binary variables.

Lattice gas	$n_i = 0$	$n_i = 1$
	☐	●
	empty	occupied
Ising	$s_i = +1$	$s_i = -1$
	↑	↓
	spin up	spin down

$$\boxed{n_i = \frac{1 + s_i}{2}.} \tag{31}$$

- Using Eq. (31), the lattice gas model

$$E(\boldsymbol{n}) = -g \sum_{\langle ij \rangle} n_i\, n_j - \mu \sum_i n_i, \tag{32}$$

is equivalent to the Ising model (up to a constant energy shift)

$$E(\boldsymbol{s}) = -J \sum_{\langle ij \rangle} s_i\, s_j - h \sum_i s_i. \tag{33}$$

with the following relations

$$\boxed{J = \frac{g}{4},\ \ h = \frac{q\, g + 2\, \mu}{4}.} \tag{34}$$

Exc 3 Derive the relations in Eq. (34).

However, there is a key conceptual difference between the two models.

- The **Ising model** has an explicit $\mathbb{Z}_2$ **Ising symmetry** at $h = 0$ (in the absence of external magnetic field).
  - The symmetry acts by flipping all Ising spins together,

$$\boxed{\mathbb{Z}_2 : s_i \to -s_i.} \tag{35}$$

The symmetry transformations a $\mathbb{Z}_2$ **group** — the two-fold cyclic group containing two elements: *identity* (do nothing) and *spin flip*, s.t.

$$\text{flip} + \text{flip} = \text{identity}. \tag{36}$$

  - The energy function (at $h = 0$) is invariant under the symmetry transformation

$$E(\boldsymbol{s}) = -J \sum_{\langle ij \rangle} s_i\, s_j. \tag{37}$$

There is no such $\mathbb{Z}_2$ symmetry in the lattice gas model in general. (Although one may argue that at $\mu = -q\, g / 2$, there is a $n_i \to 1 - n_i$ symmetry, but this requires *fine-tuning* and is *not generic* if the hard-core assumption is relaxed.)

■ Probability Distribution

The probability distribution of the **Ising configuration** $\boldsymbol{s}$ is

$$\boxed{p(\boldsymbol{s}) = \frac{1}{Z}\, e^{-\beta\, E(\boldsymbol{s})},} \tag{38}$$

where $\beta = 1/T$ is the **inverse temperature**, and Z is the **partition function**

$$Z = \sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}. \tag{39}$$

Solving a statistical mechanics problem amounts to computing the *partition function* Z as a function of model parameters (such as $T = 1/\beta$, J and h). However, it is not easy to solve Ising model, because

- **Combinatorial complexity**: With V spins, there are 2^V possible configurations. This *exponential growth* in configurations makes exact summation in Eq. (39) *intractable* for large systems, especially in higher dimensions.
- **Correlation complexity**: The *correlated* nature of spins rooted in the *interactions* among them, where the state of one spin can influence the probable states of others. This preclude simplifying the analysis to isolated spins, necessitating a *collective analysis*.

Given these challenges, exact solutions to the Ising model are only possible in limited cases. Our current status of knowledge:

	$h = 0$	$h \neq 0$	Approach
$d = 1$	✓	✓	Transfer matrix
$d = 2$	✓	?	Fermionization
$d = 3$	?	?	–
$d \geq 4$	(✓)	(✓)	Mean field theory

- d - spatial dimension of lattice,
- h - external magnetic field.

Mean Field Theory

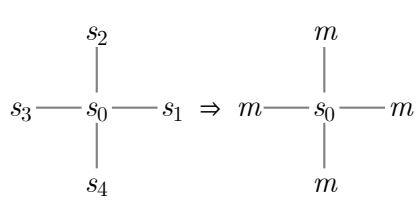
General Idea

Mean field (MF) theory is a powerful *approximation* method to solve many-body problems.

- **Key assumption**: *neglect* the *correlation* between Ising spins, assuming each spin only feels an *average* interaction effect from other spins, hence the name "mean field".

  Introduce average magnetization density m,

$$\boxed{m = \langle s_i \rangle = \sum_{\boldsymbol{s}} s_i \, p_{\mathrm{MF}}(\boldsymbol{s}).} \tag{40}$$



This enables two crucial simplifications:

- **Independence**: Each spin s_0 is treated as if it is independent of the others $s_{i\neq 0}$, under the influence of the mean field m only.
- **Factorization**: The joint probability distribution $p(\boldsymbol{s})$ is *approximated* by a *product* of independent distributions for each spin.

$$\boxed{p_{\mathrm{MF}}(\boldsymbol{s}) = \prod_i p_{\mathrm{MF}}(s_i).} \tag{41}$$

- **Key steps**:
  - **Probability Factorization**: propose a factorized *probability model* $p_{\mathrm{MF}}(\boldsymbol{s})$ to approximate the ground truth $p(\boldsymbol{s})$.
  - **Objective Function**: formulate an *objective function* that measure the *divergence* between $p_{\mathrm{MF}}(\boldsymbol{s})$ and $p(\boldsymbol{s})$.
  - **Parameter Optimization**: optimize *model parameters* in $p_{\mathrm{MF}}(\boldsymbol{s})$ to minimize its distance to $p(\boldsymbol{s})$, such that $p_{\mathrm{MF}}(\boldsymbol{s})$ approximates $p(\boldsymbol{s})$ as much as possible (within the model capacity).
  - **Model Prediction**: use the model $p_{\mathrm{MF}}(\boldsymbol{s})$ to predict whatever thermodynamic properties of interest.

■ Probability Factorization

Starting with a factorized probability model Eq. (41),

$$p_{\mathrm{MF}}(\boldsymbol{s}) = \prod_i p_{\mathrm{MF}}(s_i). \tag{42}$$

For each Ising spin $s_i = \pm 1$ on site i, we can parameterize its probability distribution by the **model parameter** $m \in [-1, 1]$ in the following form

$$\boxed{p_{\mathrm{MF}}(s_i) = \frac{1 + m\, s_i}{2} = \begin{cases} \frac{1+m}{2} & s_i = +1, \\ \frac{1-m}{2} & s_i = -1. \end{cases}} \tag{43}$$

- Normalization is automatically satisfied

$$\sum_{s_i=\pm 1} p_{\mathrm{MF}}(s_i) = 1. \tag{44}$$

- The model parameter m has a clear physical meaning — the expectation value of s_i (i.e. the **magnetization density**)

$$\langle s_i \rangle = \sum_{s_i=\pm 1} s_i\, p_{\mathrm{MF}}(s_i) = m. \tag{45}$$

Exc 4 Show that Eq. (43) is the unique solution of Eq. (45).

Objective Function

Kullback-Leibler (KL) Divergence

In statistical machine learning, the **Kullback-Leibler (KL) divergence** $D_{\mathrm{KL}}(p \,\|\, q)$, also known as the **relative entropy**, is often used to characterize the *deviation* of one probability distribution $q(x)$ with respect to the other $p(x)$ for $x \in \mathcal{X}$ in the same support space.

$$\boxed{D_{\mathrm{KL}}(p \,\|\, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.} \tag{46}$$

- The term "divergence" is used instead of "distance" because D_{KL} is not symmetric in general, i.e. $D_{\mathrm{KL}}(p \,\|\, q) \neq D_{\mathrm{KL}}(q \,\|\, p)$, therefore not qualified as a distance.
- **Non-negativity** of KL divergence: for any two distributions $p(x)$ and $q(x)$,

$$D_{\mathrm{KL}}(p \,\|\, q) \geq 0. \tag{47}$$

  The equality is achieved iff $p(x)$ and $q(x)$ are identical distributions, i.e.

$$D_{\mathrm{KL}}(p \,\|\, q) = 0 \Leftrightarrow \forall\, x : p(x) = q(x). \tag{48}$$

Exc 5 Prove Eq. (47) that the KL divergence is non-negative.

$D_{\mathrm{KL}}(p \,\|\, q)$ quantifies the **average relative surprise** we experienced when our *belief* $q(x)$ deviates from the *reality* $p(x)$.

- **Reality** $p(x)$: The world is full of randomness. Let $p(x)$ describe the probability of event x happening in reality.
- **Belief** $q(x)$: When we observe reality, our brain unconsciously construct a (different) probability model $q(x)$ in our mind, representing our belief about the world.
- **Surprise**: Sometimes, we witness an event x that, according to our belief $q(x)$, was highly *unlikely*, yet it actually *occurs* in reality. This discrepancy results in *surprise*:

$$\text{surprise} = -\log q(x). \tag{49}$$

  If an event x happens:

$q(x)$	$-\log q(x)$	Level of suprise
1	0 bit	Not at all.
1 / 2	1 bit	As if flipping a coins, landing heads up.
1 / 1024	10 bit	As if flipping 10 coins, all landing heads up!

- **Relative surprise**: Now, I observe the world with my friend. We hold different beliefs $q(x)$ (me) and $q'(x)$ (my friend). Thus, even if we witness the same event x, we might experience different levels of surprise. My surprise *relative* to my friend:

$$\begin{aligned}&\text{relative surprise} = (-\log q(x)) - (-\log q'(x))\\&= \log \frac{q'(x)}{q(x)}.\end{aligned} \tag{50}$$

- **Compare belief with reality**: It turns out that my friend is omniscient, that their belief equals the reality, i.e. $q'(x) = p(x)$, then relative to them, I would feel *more surprised*.

  This is because

  - when I am *more surprised* than my friend, it indicates that I *underestimated* the likelihood of an event occurring,
  - when I am *less surprised*, it means I *overestimated* it.

  Consequently, I tend to be more surprised than my friend more frequently, leading to a higher average relative surprise for me.

  $$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{51}$$

  The **KL divergence** D_{KL} precisely measures this *relative*, *comparative* level of *surprise*, quantifying the *deviation* of my *belief* $q(x)$ from *reality* $p(x)$.

Example: asymmetry of KL divergence

Who will be the next US president?

x	Biden	Trump
$q_{\text{Pro-Biden}}(x)$	0.999	0.001
$q_{\text{Neutral}}(x)$	0.5	0.5
$q_{\text{Pro-Trump}}(x)$	0.001	0.999

(52)

The KL divergence between each pair of probability distributions:

$D_{\mathrm{KL}}(p \,\|\, q)$ [bit]		belief $q(x)$		
		$q_{\text{Pro-Biden}}$	q_{Neutral}	$q_{\text{Pro-Trump}}$
reality $p(x)$	$p_{\text{Biden wins}}$	0	0.99	9.94
	$p_{\text{Election ties}}$	3.98	0	3.98
	$p_{\text{Trump wins}}$	9.94	0.99	0

(53)

▫ Variational Free Energy

We want to use the KL divergence to quantify the deviation between $p_{\mathrm{FM}}(\boldsymbol{s})$ and $p(\boldsymbol{s})$.

- In practice, it does not matter which distribution is treated as belief and which is reality. Both objectives are fine:

  $$D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p) \quad \text{or} \quad D_{\mathrm{KL}}(p \,\|\, p_{\mathrm{MF}}). \tag{54}$$

  However, $D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p)$ is much *easier* to evaluate analytically, because sampling $\boldsymbol{s}$ from the uncorrelated (factorized) distribution p_{MF} is tractable.

- So we choose the **objective function** to be

$$\boxed{D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p) = \sum_{\boldsymbol{s}} p_{\mathrm{MF}}(\boldsymbol{s}) \log \frac{p_{\mathrm{MF}}(\boldsymbol{s})}{p(\boldsymbol{s})}.} \tag{55}$$

  By minimizing the objective function, we expect to bring $p_{\mathrm{MF}}(\boldsymbol{s})$ close to $p(\boldsymbol{s})$ as much as possible.

Given Eq. (27), Eq. (38), Eq. (41), and Eq. (43),

$$\begin{aligned} p(\boldsymbol{s}) &= \frac{1}{Z}\, e^{-\beta\, E(\boldsymbol{s})} = \frac{1}{Z} \exp\left(\frac{1}{T}\left(J \sum_{\langle ij \rangle} s_i\, s_j + h \sum_i s_i\right)\right), \\ p_{\mathrm{MF}}(\boldsymbol{s}) &= \prod_i p_{\mathrm{MF}}(s_i) = \prod_i \frac{1 + m\, s_i}{2}, \end{aligned} \tag{56}$$

the objective function can be written as

$$D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p) = \beta\, F + \log Z, \tag{57}$$

where

- F is the **variational free energy**, given by

$$\boxed{\begin{aligned} F &= E - T\, S, \\ E &= \left(-\frac{q\, J}{2}\, m^2 - h\, m\right) V, \\ S &= \left(-\frac{1+m}{2} \log \frac{1+m}{2} - \frac{1-m}{2} \log \frac{1-m}{2}\right) V, \end{aligned}} \tag{58}$$

  - V - lattice volume
  - q - coordination number (assuming square lattice with $q = 4$)

Exc 6 Derive Eq. (58) by substituting Eq. (56) into Eq. (55).

- Note that $F_{\mathrm{true}} := -T \log Z$ is the **true free energy** of the system, Eq. (57) can be written as

$$\boxed{D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p) = \beta\, (F - F_{\mathrm{true}}),} \tag{59}$$

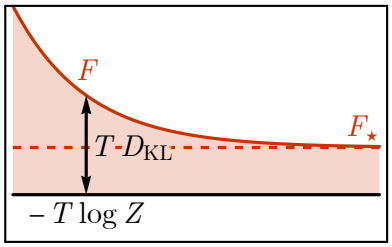
  - Since F_{ture} (or Z) is *independent* of the *model parameter* m, it can be *dropped* from the objective function without affecting the optimization of m.

    $\Rightarrow$ To *minimize* $D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p)$, we only need to *minimize* F.

  - Given $D_{\mathrm{KL}}(p_{\mathrm{MF}} \,\|\, p) \geq 0$,

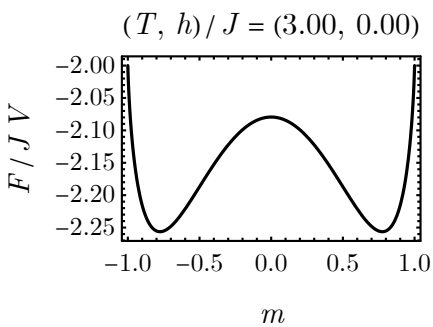
$$\boxed{F \geq F_{\mathrm{true}} = -T \log Z,} \tag{60}$$

i.e. the variational free energy F serves as an *upper bound* of the true free energy.



Optimization steps

The **variational free energy** $F(V, T, h; m)$ is a function of the **magnetization density** m (as model parameter), given the volume V, temperature T and external field h (with the Ising coupling J taken as the energy unit).



The goal of bringing $p_{\mathrm{MF}}(\boldsymbol{s})$ close to $p(\boldsymbol{s})$ boils down to optimizing the model parameter m to minimize the objective function $F(V, T, h; m)$.

■ Parameter Optimization

The optimal parameter m can be found by solving the **saddle point equation** (the point where gradient vanishes)

$$\boxed{\frac{\partial F(V, T, h; m)}{\partial m} = 0,} \tag{61}$$

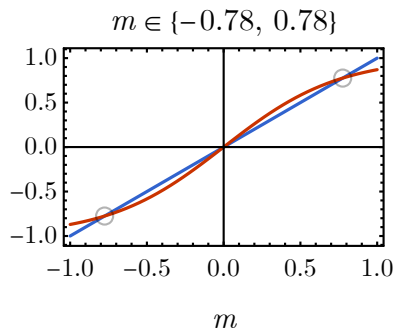
which results in

$$-h - q\,J\,m + T \operatorname{arctanh} m = 0. \tag{62}$$

Exc 7 Verify Eq. (62).

Or equivalently written as the **mean-field equation** (to determine m self-consistently)

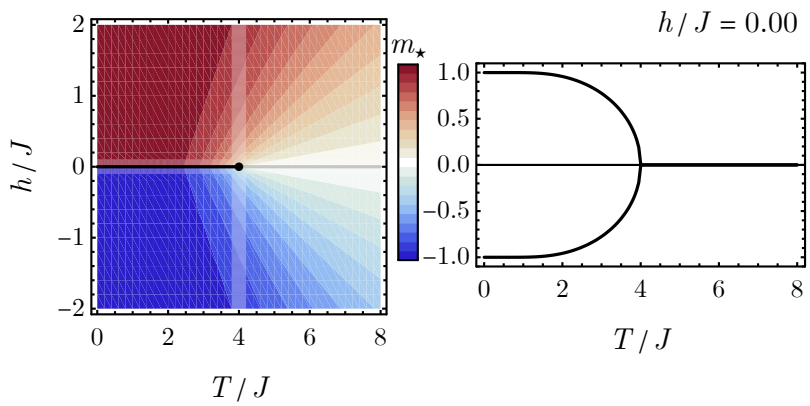
$$\boxed{m = \tanh\left(\frac{q\,J\,m + h}{T}\right).} \tag{63}$$



By solving the mean-field equation, the **stable solution** (the free energy *minimum* solution) of the magnetization density

$$\boxed{m_\star(T, h) := \underset{m}{\operatorname{argmin}}\, F(V, T, h; m),} \tag{64}$$

can be found as a function of the temperature T and the external magnetic field h (but not the volume V, because is m is *intensive*).



- $h = 0$: there is a **phase transition**, with the **critical temperature**

$$\boxed{T_c = q\, J.} \tag{65}$$

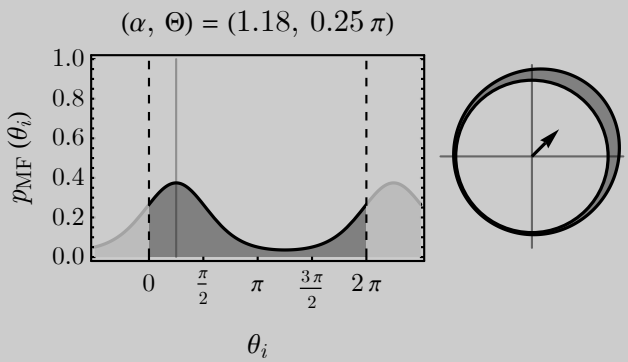
  - $T \geq T_c$: Only one *stable* solution $m = 0$. ⇒ Ising spins are **disordered**, resulting in zero magnetization on average.
  - $T < T_c$: $m = 0$ is *unstable*, two *stable* solutions emerge at $m \neq 0$ (as $T \to 0$, $m \to \pm 1$). ⇒ Ising spins are **ordered** (spontaneously align in the same direction), resulting in finite magnetization.
  - m is also called the **order parameter** — a local observable that indicates the ordering of Ising spins.
- $h \neq 0$: there is no sharp transition, only a **crossover**.

- $T \gg q\,J$: $m \sim h\,/\,T$ varies *smoothly*.
- $T \lesssim q\,J$: The anti-aligned solution ($h\,m < 0$) will become *meta-stable* or *unstable*. The aligned solution ($h\,m > 0$) is always *stable*.

- **Magnetization curve** shows how magnetization $m_\star$ responding to external magnetic field h.
  - $T > T_c$: **Paramagnetic** (spin disordered) phase. Magnetization curve is continuous with finite slope near $h \to 0$.
  - $T < T_c$: **Ferromagnetic** (spin ordered) phase. Magnetization curve is discontinuous across $h \to 0$.
  - $T = T_c$: **Critical point** of the magnetic transition (order-disorder transition). Magnetization curve is continuous but with divergent slope near $h \to 0$.

HW 1

Mean field theory of XY model. The XY model is a lattice model of planar spins, where each spin can rotate continuously in a plane. The model is defined on a lattice with a spin at each site i is represented by an angle $\theta_i \in [0, 2\pi)$, describing the spin orientation with respect to a reference axis (say, the horizontal axis). The spin configuration is denoted as $\boldsymbol{\theta} = \{\theta_i\}$. The energy function is given by
$E(\boldsymbol{\theta}) = -J \sum_{\langle ij \rangle} \cos(\theta_i - \theta_j)$,
where J is the coupling constant, and the sum runs over all pairs of nearest neighboring sites. Assume the lattice coordination number is q. Consider a mean field distribution that is factorizable
$p_{\mathrm{MF}}(\boldsymbol{\theta}) = \prod_i p_{\mathrm{MF}}(\theta_i)$,
where the model has two parameters α and Θ:
$p_{\mathrm{MF}}(\theta_i) = \exp(\alpha \cos(\theta_i - \Theta)) / Z$.



(i) Determine the normalization constant Z in $p_{\mathrm{MF}}(\theta_i)$.
(ii) Compute the average magnetization density in the horizontal $\langle \cos\theta_i \rangle$ and the vertical $\langle \sin\theta_i \rangle$ directions based on the mean field distribution $p_{\mathrm{MF}}(\theta_i)$, express the result as a function of α and Θ.
(iii) Evaluate the average energy E of the system (assuming volume V).
(iv) Evaluate the entropy S associated with the mean field distribution $p_{\mathrm{MF}}(\boldsymbol{\theta})$.
(v) Compute the variational free energy $F = E - T\,S$. Verify that F is only a function of α but not Θ. Provide a symmetry argument for the Θ-independence.
(vi) Determine the magnetic transition temperature T_c given the coupling strength J and coordination number q within the mean field theory.
(vii)* Find the optimal α as a function of temperature T numerically. Define the magnitude of the magnetization density as $m = \left(\langle \cos\theta_i \rangle^2 + \langle \sin\theta_i \rangle^2\right)^{1/2}$, plot the optimal m as a function of T.

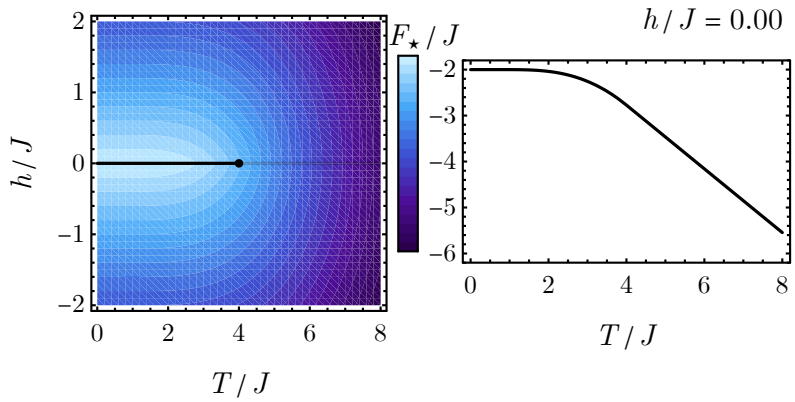
Hint: Useful facts about the Bessel function
$\int_0^{2\pi} \cos(n\,\theta)\, e^{\alpha \cos\theta}\, d\theta = 2\pi\, I_n(\alpha)$,
$I_n(\alpha) = \left(\frac{\alpha}{2}\right)^n \sum_{k=0}^{\infty} \frac{1}{k!\,(n+k)!} \left(\frac{\alpha^2}{4}\right)^k$.

■ Model Prediction

Free energy is all we need! Define the **mean-field free energy** $F_\star(V,\ T,\ h)$ as the minimum of the variational free energy $F(V,\ T,\ h;\ m)$:

$$\boxed{\begin{aligned} &F(V,\ T,\ h;\ m) \xrightarrow[\text{minimize}]{m=m_\star(T,h)} \\ &\quad F_\star(V,\ T,\ h) := F(V,\ T,\ h;\ m_\star(T,\ h)). \end{aligned}} \tag{66}$$

Recall Eq. (60), it is an upper bound of the true free energy.



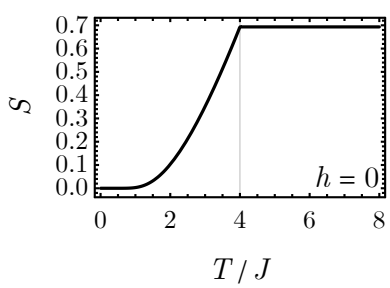
It obeys the following thermodynamic identity

$$d\,F_\star = -P_\star\,d\,V - S_\star\,d\,T - M_\star\,d\,h, \tag{67}$$

which enables us to compute the following thermodynamic quantities.

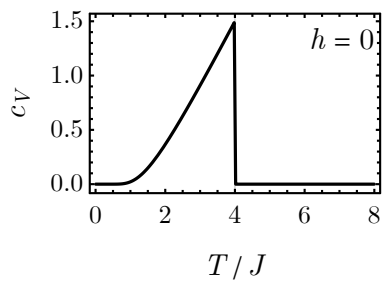
- **Entropy**

$$S_\star = -\left(\frac{\partial F_\star}{\partial T}\right)_{V,h}. \tag{68}$$



- **Specific heat**

$$c_V = \frac{T}{V}\left(\frac{\partial S_\star}{\partial T}\right)_{V,h} = -\frac{T}{V}\left(\frac{\partial^2 F_\star}{\partial T^2}\right)_{V,h}. \tag{69}$$



- Mean field theory: c_V becomes discontinuous at $T = T_c$.
- Beyond mean field: c_V diverges at $T = T_c$ as

$$\boxed{c_V \sim \frac{1}{|T - T_c|^{\alpha}},} \tag{70}$$

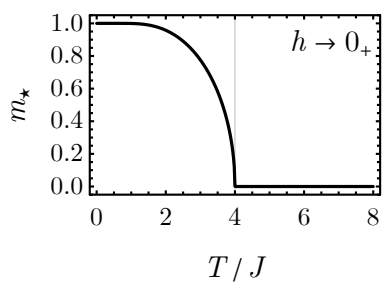
with a critical exponent α.

Along $h = 0$: the mean-field free energy $F_\star$ becomes discontinuous in its 2nd order derivative with respect to T at $T = T_c \Rightarrow$ indicating a **2nd order phase transition**.

- **Magnetization**

$$M_\star = -\left(\frac{\partial F_\star}{\partial h}\right)_{V,T}. \tag{71}$$

Magnetization density $m_\star = M_\star / V$.



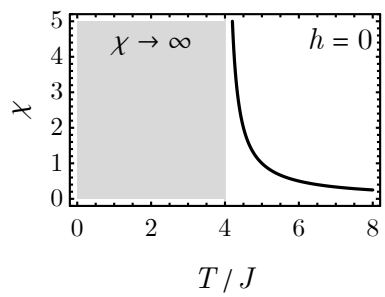
- Mean field theory: $m_\star$ drops to zero following $m_\star \sim (T_c - T)^{1/2}$ as $T \to T_c$ (from below).
- Beyond mean field: $m_\star$ drops to zero as $T \to T_c$ following

$$\boxed{m_\star \sim (T_c - T)^{\beta},} \tag{72}$$

with a critical exponent β.

- **Magnetic susceptibility**: in the paramagnetic phase, $m_\star$ responses to h linearly (for small h), a behavior known as **linear response**. The linear coefficient is defined as susceptibility

$$\chi = \frac{\partial m_\star}{\partial h} = -\frac{1}{V}\left(\frac{\partial^2 F_\star}{\partial h^2}\right)_{V,T}. \tag{73}$$

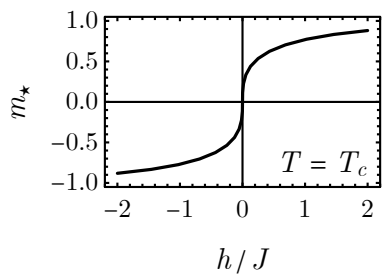


- Mean field theory: χ diverges as $\chi \sim 1/(T - T_c)$ as $T \to T_c$ (from above).
- Beyond mean field: χ diverges as $T \to T_c$ as

$$\boxed{\chi \sim \frac{1}{(T - T_c)^\gamma},} \tag{74}$$

with a critical exponent γ.

At the critical point $T = T_c$, $\chi \to \infty$ at $h = 0$, indicating that $m_\star$ is no longer linear in h.



- Mean field theory: $m_\star$ responses to h as $m_\star \sim h^{1/3}$ at $T = T_c$.
- Beyond mean field: $m_\star$ responses to h as

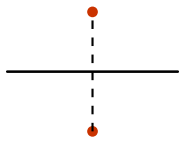
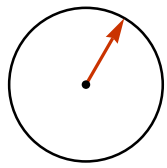
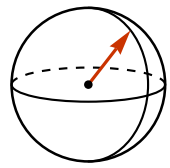
$$\boxed{m_\star \sim h^{1/\delta},} \tag{75}$$

with a critical exponent δ.

Landau Theory of Phase Transition

Spontaneous Symmetry Breaking

In physics, **symmetry** refers to the invariance of a system under certain transformations. A important class of symmetries are the O(n) symmetries.

Name	$\mathbb{Z}_2 \cong \mathrm{O}(1)$	$\mathrm{O}(2)$	$\mathrm{O}(3)$
Meaning	2-fold reflection	Circular rotation	Spherical rotation
			
Order parameter	m	$\boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$	$\boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}$
Transform	$m \to \pm m$	$\boldsymbol{m} \to e^{i\,\theta\,\sigma^y}\,\boldsymbol{m}$	$\boldsymbol{m} \to e^{\frac{i}{2}\,\boldsymbol{\theta}\cdot\boldsymbol{\sigma}}\,\boldsymbol{m}$

It is logically plausible that if the *cause* has a certain symmetry, its *effect* will also respect the same symmetry.

- For example, if the microscopic energy model $E(\boldsymbol{s}) = E(-\boldsymbol{s})$ is invariant under the Ising symmetry $\boldsymbol{s} \to -\boldsymbol{s}$, the probability distribution $p(\boldsymbol{s}) \propto e^{-\beta\,E(\boldsymbol{s})}$ will inherit the same symmetry, hence any macroscopic properties $\langle X \rangle$ must also be symmetric,

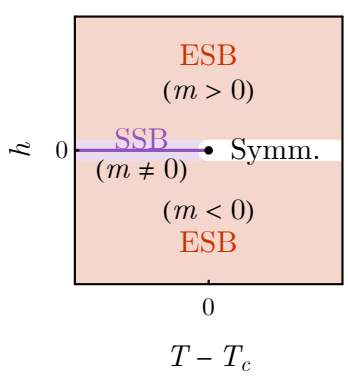
$$\begin{aligned} &\forall\, X : \langle X(\boldsymbol{s}) \rangle \\ &= \sum_{\boldsymbol{s}} X(\boldsymbol{s})\, p(\boldsymbol{s}) \overset{\boldsymbol{s}\to-\boldsymbol{s}}{=} \sum_{\boldsymbol{s}} X(-\boldsymbol{s})\, p(-\boldsymbol{s}) \\ &= \sum_{\boldsymbol{s}} X(-\boldsymbol{s})\, p(\boldsymbol{s}) = \langle X(-\boldsymbol{s}) \rangle. \end{aligned} \tag{76}$$

- This would imply the magnetization density $m := \langle m(\boldsymbol{s}) \rangle$ to vanish, because it is odd under the Ising symmetry,

$$m(\boldsymbol{s}) = \frac{1}{V} \sum_i s_i = -\,m(-\boldsymbol{s}), \tag{77}$$

thus $\langle m(\boldsymbol{s}) \rangle = -\langle m(-\boldsymbol{s}) \rangle = -\langle m(\boldsymbol{s}) \rangle \Rightarrow m = 0$.

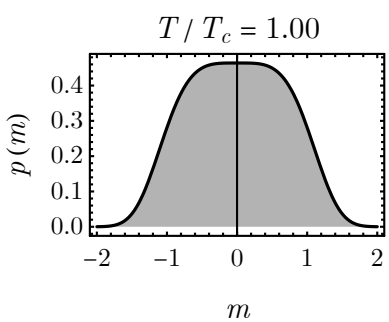
However, this is not the case we have seen in the phase diagram. ⇒ There can be **symmetry breaking**:



- **Explicit symmetry breaking (ESB)**: When $h \neq 0$, the external magnetic field *explicitly* breaks the Ising symmetry, s.t. $E(\boldsymbol{s})$ is no longer invariant under $\boldsymbol{s} \to -\boldsymbol{s}$. If the cause doesn't have the symmetry, so doesn't its effect.
- **Spontaneous symmetry breaking (SSB)**: When $h = 0$ and $T < T_c$, while the underlying *microscopic model* $E(\boldsymbol{s})$ has the Ising symmetry, the system settles into a *macroscopic state* that does not exhibit this symmetry.

  SSB presents a paradox where a *symmetric cause* leads to *asymmetric effects*. There are several interpretations:

  - **Collective choice**: At low temperature, the Ising spins collectively choose a particular *symmetry breaking order* to lower the free energy. SSB is a collective effect driven by **interaction**.
  - **Historical contingency**: At the critical point, the Ising spins fluctuates strongly. Once a *small*, *random* fluctuation biases towards a symmetry breaking pattern, the bias can be *amplified* by interaction and gets *frozen* in the system at low temperature. SSB is a historical contingency driven by **critical fluctuations**.
  - **Ergodicity breaking**: Ergodicity — the assumption that a system can sample all feasible microstate in its equilibrium distribution — can break down when the distribution *splits* into disconnected *modes* (peaks). Once settled in one mode, the system can hardly explore other modes via *local fluctuations*. The symmetry appears to be broken despite that the distribution is still symmetric.



  SSB is an asymmetric illusion in a symmetric probability distribution due to **mode collapse**.

Classification of Phases

Symmetry (and its breaking) serves as one central principles in *classifying* **phases of matter**: two phases *differed* by symmetry must be *separated* by phase transitions, because the symmetry can not change continuously.

Phase	Order	Broken symmetry
Ferromagnet	Spin alignment	Spin rotational symmetry
Liquid crystal (nematic)	Molecule alignment	Spatial rotation symmetry
Crystalline solid	Atom arrangement (periodic lattice)	Translation and rotation symmetries
Superfluid	Phase coherence of particles	U (1) symmetry (particle number conservation)
Electricmagnetic field	Phase coherence of strings	Magnetic 1-form symmetry

■ Landau Free Energy

Landau theory provides *phenomenological* approach to phase transition, focusing on the *order parameter* (e.g. m) that characterizes symmetry-breaking phases.

- Given a **symmetry** G of the system, an **order parameter** is a *local observable* that transforms *non-trivially* under the symmetry action.
  - For example, $G = \mathbb{Z}_2$ in Ising model is the **Ising** (spin-flip) **symmetry**. Given $m = \langle s_i \rangle$,

$$\mathbb{Z}_2 : m \to -m, \tag{78}$$

thus m qualifies as an order parameter, as it can indicate symmetry breaking:

$$m \begin{cases} = 0 & \mathbb{Z}_2 \text{ symmetric,} \\ \neq 0 & \mathbb{Z}_2 \text{ symmetry breaking.} \end{cases} \tag{79}$$

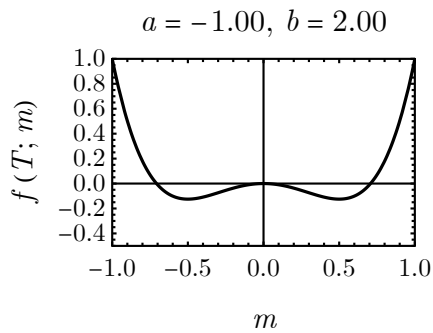
- Landau theory employs a *Taylor series* expansion of the **free energy** (density) in terms of the *order parameter*, without considering the microscopic details of the system.
  - For example, expanding the variational free energy $F(V, T, h; m)$ in Eq. (58) with respect to the order parameter m gives

$$f := \frac{F}{V} = -T \log 2 - h\, m + \frac{T - T_c}{2}\, m^2 + \frac{T\, m^4}{12} + \ldots. \tag{80}$$

Exc 8 Verify Eq. (80).

  - Landau argues that, in the absence of external field (i.e. $h = 0$), the **free energy density** f should take the following general form

$$\boxed{f(T; m) = f_0(T) + a(T)\, m^2 + b(T)\, m^4 + \ldots} \tag{81}$$



- **Symmetry argument**: With $h = 0$, the system is $\mathbb{Z}_2$ symmetric $\Rightarrow f(T; m)$ must be an *even* function of m as

  $$f(T; m) = f(T; -m), \tag{82}$$

  So the expansion Eq. (81) must only contain *even* powers of m.

- To include an **external magnetic field** h, simply add $-h\,m$ to the free energy density:

  $$\boxed{\begin{aligned} &f(T, h; m) = f(T; m) - h\,m \\ &= f_0(T) - h\,m + a(T)\,m^2 + b(T)\,m^4 + \ldots \end{aligned}} \tag{83}$$

- **Symmetry argument**: $h \neq 0$ *explicitly* breaks the $\mathbb{Z}_2$ symmetry, such that $f(T, h; m)$ is no longer an even function of m, therefore a linear term $-h\,m$ is allowed.

■ Phase Diagram

When m is small (near the critical point), it will be sufficient to consider the first several leading terms in the expansion:

- The m^0 term $f_0(T)$ is unimportant (unrelated to m).
- The m^2 coefficient $a(T)$, called the **tuning parameter**, tunes the phase transition:

  $$\boxed{a(T) \approx a_0(T - T_c).} \tag{84}$$

  Phases along $h = 0$:

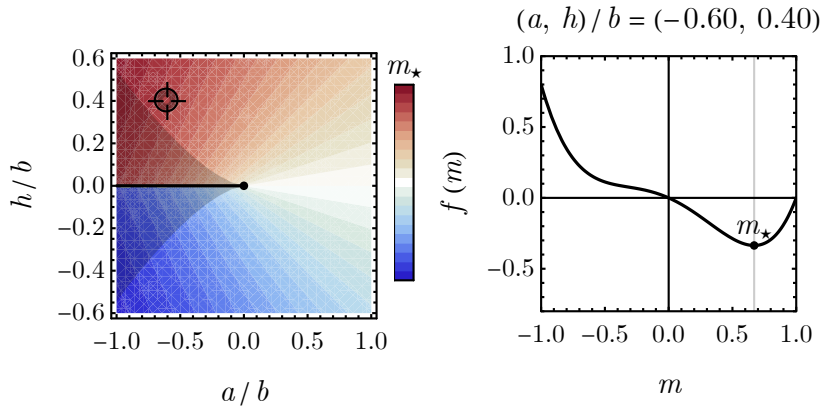
Tuning parameter	Order parameter	Phase	$\mathbb{Z}_2$ symmetry
$a(T) > 0$	$m_\star = 0$	Disorder	Preserved
$a(T) < 0$	$m_\star \neq 0$	Ordered	Broken (spontaneously)

- The m^4 coefficient $b(T)$ should be positive to ensure the stability (if not, m^6 term must be considered). We may assume

$$b(T) \approx b_0 > 0, \tag{85}$$

for the temperature range of interest (around $T \sim T_c$).

Phase diagram in the (a, h) plane (treating b as the energy unit).



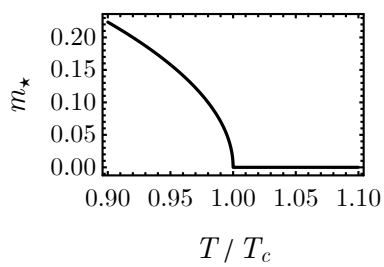
Critical Behaviors

By minimizing the Landau free energy Eq. (81), we can compute:

- the stable solution of **order parameter**

$$m_\star(T) = \begin{cases} \pm\left(\frac{a_0}{2\, b_0}\right)^{1/2} (T_c - T)^{1/2} & T < T_c, \\ 0 & T \geq T_c, \end{cases} \tag{86}$$

which exhibits $m_\star \sim (T_c - T)^{1/2}$ behavior near the critical point;



- the optimal **free energy** density

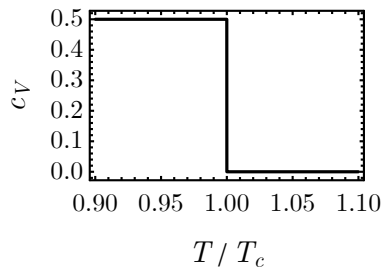
$$f_\star(T) = \begin{cases} f_0 - \frac{a_0^2}{4\, b_0} (T - T_c)^2 & T < T_c, \\ f_0 & T \geq T_c. \end{cases} \tag{87}$$

Exc 9 Derive Eq. (86) and Eq. (87).

Based on the optimal free energy density, the specific heat can be evaluated

$$c_V = -T\,\frac{\partial^2 f_\star}{\partial T^2} = \begin{cases} \frac{a_0^2}{2\,b_0}\,T_c & T < T_c, \\ 0 & T \geq T_c, \end{cases} \tag{88}$$

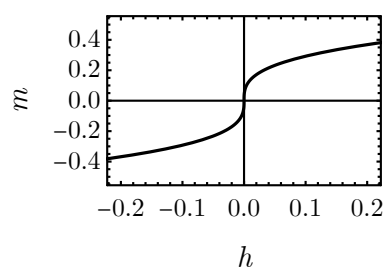
which jumps across the critical point ⇒ a manifestation of **2nd order phase transition**.



Putting back the external magnetic field h, the Landau free energy density takes the form of Eq. (83), for which the saddle point equation $\partial f / \partial m = 0$ reads

$$\boxed{h = 2\,a_0(T - T_c)\,m + 4\,b_0\,m^3.} \tag{89}$$

Exc 10 Derive Eq. (89).



- In the disorder phase ($T > T_c$), the magnetic susceptibility at $h = 0$ is given by

$$\chi = \left(\frac{\partial m_\star}{\partial h}\right)_{h=0} = \frac{1}{2\,a_0(T - T_c)}, \tag{90}$$

  which diverges as $\chi \sim (T - T_c)^{-1}$ as $T \to T_c$ from above, known as the **Curie-Weiss law**.

- At the critical point $T = T_c$, $a(T) = 0$ vanishes, then Eq. (89) reduces to

$$h = 4\,b_0\,m^3, \tag{91}$$

  indicating that $m_\star$ responses to h as $m_\star \sim h^{1/3}$.

Summary: **critical exponents** predicted by the **Landau theory**.

Exponent	Definition	Condition	Prediction
α	$c_V \sim \lvert T - T_c\rvert^{-\alpha}$		0
β	$m_\star \sim (T_c - T)^\beta$	$T < T_c$	1/2
γ	$\chi \sim (T - T_c)^{-\gamma}$	$T > T_c$	1

(92)

δ	$m_\star \sim h^{1/\delta}$	$T = T_c$	3

Limitation of Landau Theory

Landau theory, in its basic form, is a **mean field theory**, because it only focus on the **order parameter** as an macroscopic *average* of spin states *without* microscopic *fluctuations*. — The spin correlations across the space is ignored.

- **Limitations** of Landau theory:
  - **Absence of critical fluctuations**: Fluctuations are long-ranged and highly correlated near the critical point. Ignoring them leads to inaccurate predictions of critical exponents.

Exponent	Definition	$D = 2$	$D = 3$	$D = 4$	Mean field
α	$c_V \sim \lvert T - T_c\rvert^{-\alpha}$	0	0.110	0	0
β	$m_\star \sim (T_c - T)^\beta$	$1/8$	0.336	$1/2$	$1/2$
γ	$\chi \sim (T - T_c)^{-\gamma}$	$7/4$	1.237	1	1
δ	$m_\star \sim h^{1/\delta}$	15	4.790	3	3

(93)

  - **Ignorance of dimensionality**: Dimensionality of the system significantly influences critical phenomena, which is not account for in mean field theories.
- Approaches beyond Landau theory:
  - **Field theory**: Promote the order parameter m to a fluctuating field $m(x)$, e.g.

$$F = \int d^D x \left(\frac{1}{2} (\partial_\mu m)^2 + a\, m^2 + b\, m^4 + \ldots \right), \tag{94}$$

    such that the spatial dimension D can also be incorporated.
  - **Numerical simulations**: Use computer to simulate the system at large scales (e.g. Monte Carlo simulation).
  - **Renormalization group**: A powerful framework to analyze how physics systems behaves at different length scales, from which critical exponents can be more accurately estimated.
  - **Conformal bootstrap**: A non-perturbative approach leveraging conformal symmetries to classify universality classes and solve for critical exponents.

Renormalization Group

Overview

Renormalization group (RG) is

- An elegant **conceptual framework** to understand *phases* and *phase transitions*,
- A powerful **computational approach** to identify *critical points* and estimate *critical exponents*.

It iteratively *coarsening* the *local* degrees of freedom in physical systems, extracting **representative features** at every *scale*, and tracking the *flow* of **effective theories** from scale to scale.

- How to identify *representative features*? - **Representation learning**.
- How to establish the *effective theory* from scale to scale? - **Coarse graining**.
- What is the usage of tracking the *flow* of effective theories? - **Renormalization group flow**.

Representation Learning

Multivariate Probabilities

A **multivariate random variable** $(X, Y, \ldots)$ is made up of a *collection* of random variables, each *possible value* $(x, y, \ldots) \in \mathcal{X} \times \mathcal{Y} \times \ldots$ is assigned a **joint probability** $p(x, y, \ldots) = \Pr(X = x, Y = y, \ldots)$. The joint probability also satisfies

- Positivity

$$p(x, y, \ldots) \geq 0. \tag{95}$$

- Normalization

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \ldots p(x, y, \ldots) = 1. \tag{96}$$

For two random variables X, Y

- $p(x, y)$: **joint probability**. The probability of observing $X = x$ *and* $Y = y$ jointly.
- $p(x)$: **marginal probability**. The probability of observing $X = x$ regardless of Y (before any observation is made on Y).

$$\boxed{p(x) = \sum_{y \in \mathcal{Y}} p(x, y).} \tag{97}$$

- $p(x \mid y)$: **conditional probability**. The probability of observing $X = x$ after observing $Y = y$.

$$\boxed{p(x \mid y) = \frac{p(x, y)}{p(y)}.} \tag{98}$$

The conditional probability is simply proportional to the joint probability, and is normalized by the marginal probability for each given condition separately.

Example:

- Joint and marginal distributions:

$p(x, y)$	y			$p(x)$
	1	2	3	
1	0	$\frac{1}{2}$	0	$\frac{1}{2}$

x	2	0	0	$\frac{1}{6}$	$\frac{1}{6}$
	3	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{3}$
$p(y)$		$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	

- Conditional distributions:
  - $p(x \mid y) = p(x, y) / p(y)$

		$p(x \mid Y = 1)$	$p(x \mid Y = 2)$	$p(x \mid Y = 3)$
x	1	0	1	0
	2	0	0	$\frac{1}{2}$
	3	1	0	$\frac{1}{2}$

(100)

  - $p(y \mid x) = p(x, y) / p(x)$

	y		
	1	2	3
$p(y \mid X = 1)$	0	1	0
$p(y \mid X = 2)$	0	0	1
$p(y \mid X = 3)$	$\frac{1}{2}$	0	$\frac{1}{2}$

(101)

Action and Probability

In statistical physics, for a set of random variables $\boldsymbol{x}$ with an **energy function** $E(\boldsymbol{x})$, we can define an equilibrium **probability distribution** known as the *Boltzmann distribution*:

$$p(\boldsymbol{x}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{x})} \tag{102}$$

Conversely, for any given **probability distribution** $p(\boldsymbol{x})$, one can derive an underlying (dimensionless) energy function, called the **action** in *physics*, or the **negative log likelihood** (NLL) in *math*,

$$\boxed{S(\boldsymbol{x}) = -\log p(\boldsymbol{x}) = \beta E(\boldsymbol{x}) + \log Z.} \tag{103}$$

- This concept of "action" parallels that in *classical mechanics*, where the **principle of least action** simply states that the most-likely configuration of $\boldsymbol{x}$ is the one that minimizes the action $S(\boldsymbol{x})$ — a straightforward interpretation of Eq. (103).
  - $\log Z$ is often *dropped* in describing the action, as it is just a constant shift, not affecting how $S(\boldsymbol{x})$ depends on $\boldsymbol{x}$, and Z can always be recovered by normalizing $p(\boldsymbol{x})$.
- Interestingly, $S(\boldsymbol{x})$ can also be interpreted as the **entropy** reduction, or the **information** gain, associated with observing the specific state $\boldsymbol{x}$, because the average entropy of the distribution $p(\boldsymbol{x})$ is precisely given by

$$S = \sum_{\boldsymbol{x}} p(\boldsymbol{x})\, S(\boldsymbol{x}) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}). \tag{104}$$

What a nice coincidence in the use of the symbol "S"!

The action model $S(\boldsymbol{x})$ and the probability distribution $p(\boldsymbol{x})$ are *equivalent* descriptions of random variables

$$\boxed{p(\boldsymbol{x}) = e^{-S(\boldsymbol{x})}.} \tag{105}$$

Which one is more fundamental? - There are two perspectives:

$$S(\boldsymbol{x}) \underset{\text{machine learning}}{\overset{\text{statistical physics}}{\rightleftharpoons}} p(\boldsymbol{x}). \tag{106}$$

- **Statistical physics** starts with the theoretical model of action (or energy function) $S(\boldsymbol{x}) = \beta\, E(\boldsymbol{x})$ and derive the probability distribution $p(\boldsymbol{x})$ to characterize the equilibrium thermodynamic properties.
- **Machine learning** starts with the empirical data distribution $p(\boldsymbol{x})$ and infer or learn a underlying action model $S(\boldsymbol{x})$ that describe or generate the observed data distribution.

■ Effective Action

The idea of **effective action** originates from *marginalization* of probability distributions. Let $\boldsymbol{x}$ and $\boldsymbol{z}$ be two sets of random variables. Given a *joint* distribution $p(\boldsymbol{x}, \boldsymbol{z})$ described by an *joint* action $S(\boldsymbol{x}, \boldsymbol{z})$:

$$p(\boldsymbol{x}, \boldsymbol{z}) = e^{-S(\boldsymbol{x},\boldsymbol{z})}, \tag{107}$$

if we only care about the distribution of $\boldsymbol{x}$ (regardless of $\boldsymbol{z}$), we should consider the *marginal* distribution

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z}} e^{-S(\boldsymbol{x},\boldsymbol{z})}, \tag{108}$$

which defines an **effective action** (the *marginal* action) $S(\boldsymbol{x})$ for $\boldsymbol{x}$ only:

$$S(\boldsymbol{x}) = -\log p(\boldsymbol{x}) = -\log \sum_{\boldsymbol{z}} e^{-S(\boldsymbol{x},\boldsymbol{z})}. \tag{109}$$

The effective action $S(\boldsymbol{x})$ is related to the joint action $S(\boldsymbol{x}, \boldsymbol{z})$ by

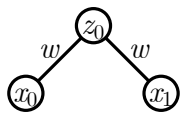
$$\boxed{e^{-S(\boldsymbol{x})} = \sum_{\boldsymbol{z}} e^{-S(\boldsymbol{x},\boldsymbol{z})}.} \tag{110}$$

The idea can be summarized by the diagram:

$$\begin{array}{ccc} p(\boldsymbol{x}, \boldsymbol{z}) & \xrightarrow{\text{marginal}} & p(\boldsymbol{x}) \\ \downarrow & & \downarrow \\ S(\boldsymbol{x}, \boldsymbol{z}) & \xrightarrow[\text{effective}]{} & S(\boldsymbol{x}) \end{array} \tag{111}$$

Example: interaction mediated by a latent spin

- Consider an Ising model of three Ising variables x_0, x_1, z_0 (grouped into $\boldsymbol{x} = (x_0, x_1)$ and $\boldsymbol{z} = (z_0)$). Suppose that they interact as



  with $w = \beta J > 0$ being the *dimensionless* **coupling strength** (or call the **weight** in machine learning).

- The joint action is

$$\boxed{S(\boldsymbol{x}, \boldsymbol{z}) = -w\,(x_0\, z_0 + x_1\, z_0).} \tag{112}$$

  There is no *direct interaction* between x_0, x_1.

- However, z_0 can mediate an *effective interaction* between $x_0$and x_1,

$$\begin{aligned} S(\boldsymbol{x}) &= -\log \sum_{\boldsymbol{z}} e^{-S(\boldsymbol{x},\boldsymbol{z})} \\ &= -\log \sum_{z_0=\pm 1} e^{w\,(x_0+x_1)\,z_0} \\ &= -\log\,(2\cosh(w\,(x_0+x_1))). \end{aligned} \tag{113}$$

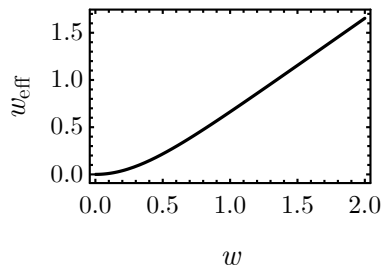
  Take $w = 1$ for example, the values of $S(\boldsymbol{x})$ for different configurations of $\boldsymbol{x}$ are as follows

		x_1	
		$+1$	-1
x_0	$+1$	-2.01815	-0.693147
	-1	-0.693147	-2.01815

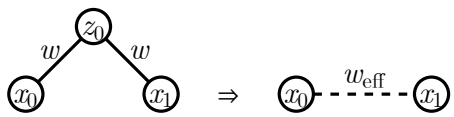
  - The action is lower if x_0 and x_1 are aligned ⇒ meaning that there is an *effective* ferromagnetic interaction between x_0 and x_1, even if there was *no direct interaction* between them in the original model.
  - In fact, for Ising variables x_0, x_1, Eq. (113) can be rewritten as

$$\boxed{\begin{aligned} &S(\boldsymbol{x}) = -w_{\text{eff}}\, x_0\, x_1 + \text{const.}, \\ &w_{\text{eff}} = \frac{1}{2}\log\cosh(2\,w) > 0. \end{aligned}} \tag{114}$$

Exc 11 Show that Eq. (114) is consistent with Eq. (113).



- In conclusion, the latent spin z_0 can mediate an effective interaction between x_0 and x_1 with an effective interaction strength w_{eff} that increases with w.



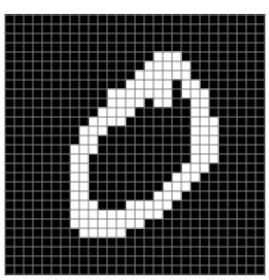
■ Boltzmann Machine

A **Boltzmann machine** is a *generative model* that learns to model the **data distribution** $p_{\text{dat}}(\boldsymbol{x})$ over input variables $\boldsymbol{x}$ by approximating $p_{\text{dat}}(\boldsymbol{x})$ as a Boltzmann distribution of some effective action:

$$\boxed{\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \frac{1}{Z_{\boldsymbol{\theta}}} \sum_{\boldsymbol{z}} e^{-S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}, \\ Z_{\boldsymbol{\theta}} &= \sum_{\boldsymbol{x},\boldsymbol{z}} e^{-S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}. \end{aligned}} \tag{118}$$

- Random variables:
  - $\boldsymbol{x} = \{x_i\}$ - **input** variables (describing a sample of data, such as an image).
  - $\boldsymbol{z} = \{z_i\}$ - **latent** variables (introduced to mediate effective interactions/correlations among input variables).

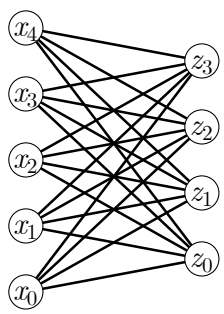
  Consider x_i, $z_i = 0, 1$ as *binary* variables (like occupation numbers in a lattice gas model), such that each sample of $\boldsymbol{x}$ can represent a black-white image (as particles arranged into a pattern on the lattice).

$\boldsymbol{x} =$ 

- $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ - the **model distribution** to optimize.

- $S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ - a joint action parametrized by a set of **parameters** $\boldsymbol{\theta}$ (such as the coupling strengths). For example,

$$\begin{aligned} S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) &= -\sum_{i,j} W_{ij}\, x_i\, z_j - \sum_i b_i^x\, x_i - \sum_j b_j^z\, z_j \\ &= -\boldsymbol{x} \cdot \boldsymbol{W} \cdot \boldsymbol{z} - \boldsymbol{b}^x \cdot \boldsymbol{x} - \boldsymbol{b}^z \cdot \boldsymbol{z}. \end{aligned} \tag{119}$$



  - The weights w_{ij} and biases b_i^x, b_j^z are the model parameters, collectively denoted as $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{b}^x, \boldsymbol{b}^z)$.
  - The setting that the $\boldsymbol{x}$ and $\boldsymbol{z}$ only interact with each other, but *not* within themselves, is to make the training more *efficient* by Gibbs sampling. This special variant is called the **restricted Boltzmann machine** (RBM).

Objective: to optimize $\boldsymbol{\theta}$ to align $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ as close as possible with $p_{\text{dat}}(\boldsymbol{x})$, which can be achieved by minimizing the **KL divergence**, or (in practice) the **cross entropy**.

$$\boxed{\begin{aligned} \mathcal{L}_{\boldsymbol{\theta}} &= D_{\text{KL}}(p_{\text{dat}} \,\|\, p_{\boldsymbol{\theta}}) \\ &= -\sum_{\boldsymbol{x}} p_{\text{dat}}(\boldsymbol{x}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \text{const.} \end{aligned}} \tag{120}$$

Training: the training algorithm iteratively samples $\boldsymbol{x}$ from the dataset, computes the loss function $\mathcal{L}_{\boldsymbol{\theta}}$, adjusts parameters $\boldsymbol{\theta}$ to minimize $\mathcal{L}_{\boldsymbol{\theta}}$ by gradient descent, until convergence (thermal equilibrium) is reached.

Usage: training yields an optimal action model $S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$, which, as a *generative model*, is capable of:

- **Sample generation**: generate new samples $\boldsymbol{x}$ by drawing from the distribution

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\theta}}} \sum_{\boldsymbol{z}} e^{-S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}. \tag{121}$$

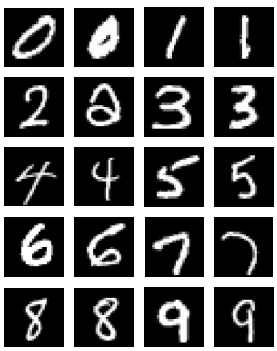
- Other task-specific applications
  - Anomaly detection
  - Data imputation
  - Feature extraction (representation learning)

- Classification

Demonstration

The MNIST (Modified National Institute of Standards and Technology) dataset is a famous dataset of hand-written digits. Each digit is represented as a 28×28 gray-scale image:

```
SeedRandom[42];
data = Map[···] +;
Grid[Partition[data, 4]]
```



Can an lattice gas model learns to write digits?

- Step 1: write a code to define the restricted Boltzmann machine (RBM) model using PyTorch package.
  - There are $28 \times 28 = 784$ input variables in $\boldsymbol{x}$.
  - We choose to use 32 latent variables for $\boldsymbol{z}$.

In[ ]:=

```
import torch
import numpy
torch.manual_seed(12) # set random seed
# define the Boltzmann machine model
class RBM(torch.nn.Module):
    def __init__(self, nx=784, nz=32):
        super().__init__()
        self.nx = nx
        self.nz = nz
        self.w = torch.nn.Parameter(5*torch.randn((nx, nz)))  #
adjusted weight initialization
        self.bx = torch.nn.Parameter(torch.zeros(nx))  # input
variables bias
        self.bz = torch.nn.Parameter(torch.zeros(nz))  # latent
variables bias

    def x_to_z(self, x, beta=1):
```

```
        logit_z = torch.nn.functional.linear(x, self.w.T, self.bz)
        return torch.bernoulli(torch.sigmoid(beta*logit_z))

    def z_to_x(self, z, beta=1):
        logit_x = torch.nn.functional.linear(z, self.w, self.bx)
        return torch.bernoulli(torch.sigmoid(beta*logit_x))

    def forward(self, x, steps=5, beta=1):
        for _ in range(steps):
            z = self.x_to_z(x, beta=beta)
            x = self.z_to_x(z, beta=beta)
        return x

    def eff_action(self, x):
        bxx = x.matmul(self.bx)
        logit_z = torch.nn.functional.linear(x, self.w.T, self.bz)
        logZ = -torch.nn.functional.logsigmoid(-logit_z).sum(-1)
        return (- bxx - logZ).mean()

    def loss(self, x, steps=5):
        x1 = self(x, steps=steps)
        return self.eff_action(x) - self.eff_action(x1)
```

ExternalFunction[System: Python Arguments: {self, nx, nz } Command: RBM]

- Step 2: setup the model, prepare the data, and connect the model to an optimizer.

In[ ]:=

```
rbm = RBM()
data = torch.tensor(numpy.array(<*ImageData/@data*>),
dtype=torch.float).view(-1, 28*28)
optimizer = torch.optim.Adam(rbm.parameters(), lr=0.1,
weight_decay=0.002)
```

- Step 3: train the model on the data for 1000 steps.

In[ ]:=

```
loss_sum = 0.
for k in range(1000):
    optimizer.zero_grad()
    loss = rbm.loss(data, steps=5)
    loss.backward()
    optimizer.step()
```

```
    loss_sum += loss.item()
    if (k+1)%100==0:
        print(f'step {k+1:4}: loss = {loss_sum/100:6.2f}')
        loss_sum = 0.
```

```
step  100: loss = 948.41
step  200: loss =  33.70
step  300: loss = -13.40
step  400: loss = -22.80
step  500: loss = -26.63
step  600: loss = -27.51
step  700: loss = -28.19
step  800: loss = -28.58
step  900: loss = -28.65
step 1000: loss = -28.34
```

The loss function $\mathcal{L}_{\boldsymbol{\theta}}$ decreases significantly, then gradually converges to the minimum within the error range of stochastic fluctuations.

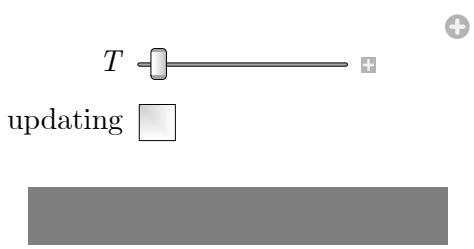
After training, we can use the the model to generate new samples.

- Start with an initial input $\boldsymbol{x}$, follow the **Markov chain Monte Carlo** (MCMC) approach to obtain a sequence of new samples

$$\boldsymbol{x} \to \boldsymbol{x}' \to \boldsymbol{x}'' \to \ldots \tag{122}$$

according to the model distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

- Recall that every *action* $S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = \beta\, E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ can be viewed as an rescaling of its *energy* function by an overall scale β. Tuning this **temperature** parameter $T = 1/\beta$ controls the strength of thermal fluctuations in the probability model.



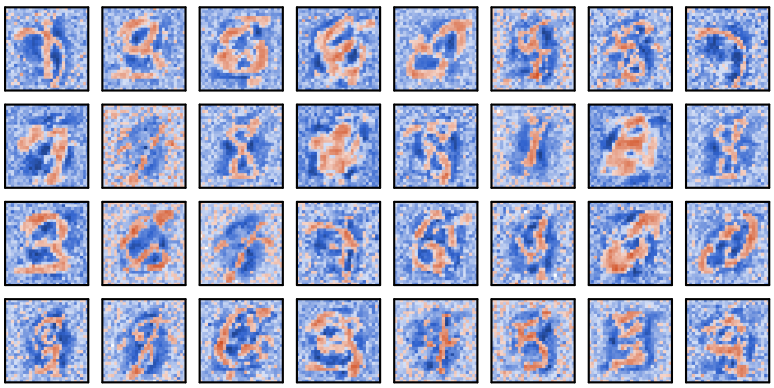
- Each digits is a (meta)-stable phase of the lattice gas model.
- Raising and lowering the temperature can anneal the model to its stable phase.

We can open up the model to inspect the learned parameters:

$$S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = -\boldsymbol{x} \cdot \boldsymbol{W} \cdot \boldsymbol{z} - \boldsymbol{b}^x \cdot \boldsymbol{x} - \boldsymbol{b}^z \cdot \boldsymbol{z}. \tag{123}$$

- $\boldsymbol{b}^x$ - the chemical potential for the lattice gas.
- $\boldsymbol{W}$ - the pattern of $\boldsymbol{x}$ interacting with each of the latent variables (there are 32 of them) in $\boldsymbol{z}$.

$\boldsymbol{b}^x =$

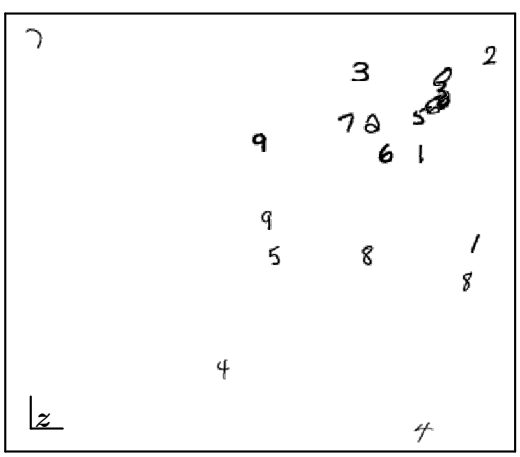
$\boldsymbol{W} =$ 

These *emergent* patterns are the key *features* automatically extracted by the model from the data, demonstrating the power of *generative models* to perform **unsupervised representation learning**.

Given an input $\boldsymbol{x}$, the model can infer the latent variables $\boldsymbol{z}$ according to the *conditional distribution*

$$p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{e^{-S_\theta(\boldsymbol{x},\boldsymbol{z})}}{\sum_{\boldsymbol{x}} e^{-S_\theta(\boldsymbol{x},\boldsymbol{z})}}. \tag{124}$$

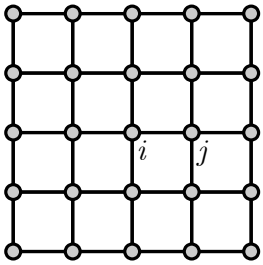
- $\boldsymbol{z}$ serves as an encoding of $\boldsymbol{x}$ in the latent (feature) space.
- By projecting the high-dimensional latent space to a two-dimensional plane, we can visualize the arrangement of hand-written digits $\boldsymbol{x}$ in the latent space: we expect *similar* digits to be *close* to each other.



Coarse Graining

Problem Setup

Consider a 2D Ising model, defined on a square lattice:



- Degree of freedom: $\boldsymbol{s} = \{s_i\}$, each Ising spin $s_i = \pm 1$ is a binary variable on site i.
- The **energy** model (without external magnetic field)

$$E(\boldsymbol{s}) = -J \sum_{\langle ij \rangle} s_i \, s_j. \tag{125}$$

Or the **action** model

$$\boxed{S(\boldsymbol{s}) = \beta \, E(\boldsymbol{s}) = -w \sum_{\langle ij \rangle} s_i \, s_j,} \tag{126}$$

with $w = \beta \, J = J / T$ being the *dimensionless* coupling constant.

- Small w: weak coupling, high temperature;
- Large w: strong coupling, low temperature.

- **Probability distribution** (Boltzmann distribution)

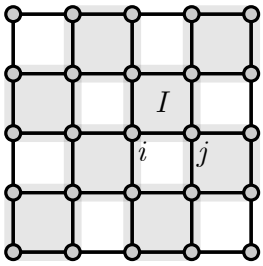
$$\boxed{\begin{aligned} p(\boldsymbol{s}) &= \frac{1}{Z} \, e^{-S(\boldsymbol{s})}, \\ Z &= \sum_{\boldsymbol{s}} e^{-S(\boldsymbol{s})}. \end{aligned}} \tag{127}$$

Key idea: **Coarse-graining** the spin configuration to understand their interactions at a higher level (larger scale). This contains three steps:

- **Block Partition**: partitioning spins into blocks.
- **Representation Learning**: emergence of coarse-grained (higher-level) spins.
- **Decimation**: summing over (marginalize) fine-grained (lower-level) spins to obtain an *effective theory* for coarse-grained (higher-level) spins.

Block Partition

Partition the square lattice into 2×2 corner sharing blocks.



- I - block index (larger scale).
- i, j - site index (smaller scale).

Due to the corner sharing structure, every **site** is shared by *two* adjacent blocks. But the partition is such designed that every **link** $\langle ij \rangle$ *uniquely* belongs to only *one* block, which enables the link summation to be split

$$\sum_{\langle ij \rangle} = \sum_{I} \sum_{\langle ij \rangle \in I} . \tag{128}$$

Apply Eq. (128) to the action in Eq. (126),

$$\boxed{S(\boldsymbol{s}) = \sum_{I} S_I(\boldsymbol{s}),} \tag{129}$$

with the ***local* action** model within each block I:

$$\boxed{S_I(\boldsymbol{s}) = -w \sum_{\langle ij \rangle \in I} s_i \, s_j.} \tag{130}$$

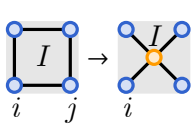
Correspondingly, the (joint) probability distribution can be written as a product of local distributions,

$$p(\boldsymbol{s}) \propto e^{-S(\boldsymbol{s})} = \prod_{I} e^{-S_I(\boldsymbol{s})} \propto \prod_{I} p_I(\boldsymbol{s}). \tag{131}$$

- Note: Eq. (131) does *not* mean the distribution can be factored into independent distributions in each block, because the spins are *shared* between blocks — not independent variables.

Representation Learning

We can focus on each block and try to learn a **local representative spin** for the 4 individual spins in a block.



- Spin degrees of freedom:
  - $\boldsymbol{s} = \{s_i\}$ with $s_i = \pm 1$ - the original *fine-grained* spin (in blue),
  - $\boldsymbol{s}' = \{s'_I\}$ with $s'_I = \pm 1$ - the emergent *coarse-grained* spin (in orange), as a *representative* of the fine-grained spins.
- The *local action*

$$S_I(\boldsymbol{s}) = -w \sum_{\langle ij \rangle \in I} s_i \, s_j \,, \tag{132}$$

defines a *local distribution* as a **target distribution** for a *generative model* to learn

$$\boxed{p_I(\boldsymbol{s}) \propto e^{-S_I(\boldsymbol{s})},} \tag{133}$$

in which s_i interact with each other.

Goal: extract the representative features by generative modeling.

- Introduce a **generative model** with the following action

$$\boxed{\tilde{S}_I(\boldsymbol{s}, \boldsymbol{s}') = -\tilde{w} \sum_{i \in I} s_i \, s'_I,} \tag{134}$$

where

  - s_i no longer interacts with each other directly, but they all interact with an emergent *representative spin* s'_I at the block center. — the architecture of a *restricted Boltzmann machine*.
  - $\tilde{w}$ is the **model parameter** to be optimized. It controls how strong the representative spin $\boldsymbol{s}'$ interacts with the underlying spins $\boldsymbol{s}$, described by the conditional distribution

$$\tilde{p}_I(\boldsymbol{s}' \mid \boldsymbol{s}) = \frac{e^{-\tilde{S}_I(\boldsymbol{s}, \boldsymbol{s}')}}{\sum_{\boldsymbol{s}} e^{-\tilde{S}_I(\boldsymbol{s}, \boldsymbol{s}')}} \,. \tag{135}$$

$\tilde{w} = 0.50$

$\boldsymbol{s}$	$\tilde{p}_I(\boldsymbol{s}' \mid \boldsymbol{s})$
−−−−	− |+
−−−+	− +
−−+−	− +
−+−−	− +
+−−−	− +
−−++	− +
−+−+	− +
−++−	− +
+−−+	− +
+−+−	− +
++−−	− +
−+++	− +
+−++	− +
++−+	− +
+++−	− +
++++	− +

- Given $\tilde{S}_I(\boldsymbol{s}, \boldsymbol{s}')$, the **model distribution** $\tilde{p}_I(\boldsymbol{s})$ of the original spins $\boldsymbol{s}$ is

$$\boxed{\tilde{p}_I(\boldsymbol{s}) \propto \sum_{\boldsymbol{s}'} e^{-\tilde{S}_I(\boldsymbol{s},\boldsymbol{s}')}.} \tag{136}$$

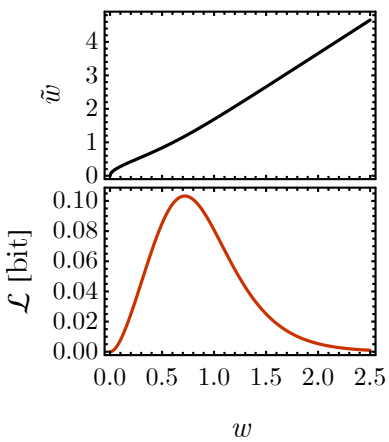
Comparing the target distribution $p_I(\boldsymbol{s})$ (parametrized by w) and the model distribution $\tilde{p}_I(\boldsymbol{s})$ (parametrized by $\tilde{w}$) for all configurations of $\boldsymbol{s}$:

$\boldsymbol{s}$	$w = 0.72$ $p_I(\boldsymbol{s})$	$\tilde{w} = 1.18$ $\tilde{p}_I(\boldsymbol{s})$
−−−−	0.373	0.349
−−−+	0.021	0.033
−−+−	0.021	0.033
−−++	0.021	0.006
−+−−	0.021	0.033
−+−+	0.001	0.006
−++−	0.021	0.006
−+++	0.021	0.033
+−−−	0.021	0.033
+−−+	0.021	0.006
+−+−	0.001	0.006
+−++	0.021	0.033
++−−	0.021	0.006
++−+	0.021	0.033
+++−	0.021	0.033
++++	0.373	0.349

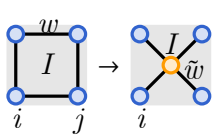
- **Loss function**: the KL divergence

$$\boxed{\begin{aligned}\mathcal{L} &= D_{\mathrm{KL}}(p_I(\boldsymbol{s}) \,\|\, \tilde{p}_I(\boldsymbol{s})) \\ &= \sum_{\boldsymbol{s}} p_I(\boldsymbol{s}) \log \frac{p_I(\boldsymbol{s})}{\tilde{p}_I(\boldsymbol{s})},\end{aligned}} \tag{137}$$

which is a function of $\tilde{w}$ given w. For each w, find the optimal $\tilde{w}$ to minimize the KL divergence $\mathcal{L}$.



- The optimization is *not perfect* given the non-vanishing KL divergence, indicating that the model distribution $\tilde{p}_I(\boldsymbol{s})$ *approximates* rather than precisely matches the target distribution $p_I(\boldsymbol{s})$.
  - However, the small residual KL divergence (< 0.1 bit) suggests that the approximation is sufficiently accurate.
  - The approximation can be improved as more representative (latent) spins are introduced, see Ref. [1].
- As a result, we obtain the **optimal model parameter** $\tilde{w}(w)$ as a *function* of the original parameter w.

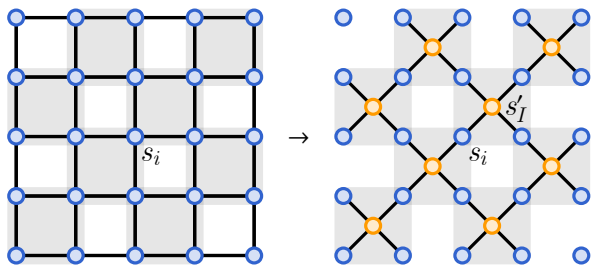


The optimization result tells us how to treat the local action $S_I(\boldsymbol{s})$ as an effective action of $\tilde{S}_I(\boldsymbol{s}, \boldsymbol{s}')$ approximately.

[1] Wanda Hou, Yi-Zhuang You. Machine Learning Renormalization Group for Statistical Physics. arXiv:2306.11054.

▪ Decimation

Within each block, replace the local action $S_I(\boldsymbol{s})$ by its optimal approximation $\tilde{S}_I(\boldsymbol{s}, \boldsymbol{s}')$,



the joint action becomes

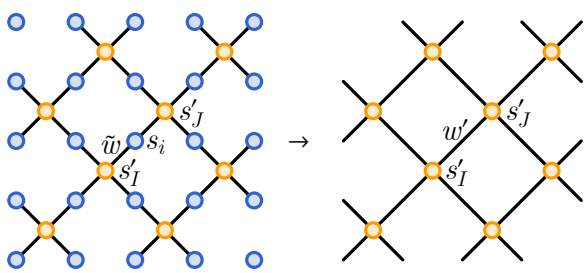
$$\boxed{\tilde{S}(\boldsymbol{s},\, \boldsymbol{s}') = \sum_I \tilde{S}_I(\boldsymbol{s},\, \boldsymbol{s}') = -\,\tilde{w} \sum_I \sum_{i \in I} s_i\, s'_I.} \tag{138}$$

- If we *marginalize* the emergent coarse-grained spins $\boldsymbol{s}'$, we fall back to the original action for the fine-grained spins $\boldsymbol{s}$ approximately

$$\sum_{\boldsymbol{s}'} e^{-\tilde{S}(\boldsymbol{s},\boldsymbol{s}')} \simeq e^{-S(\boldsymbol{s})}. \tag{139}$$

- However, if we *marginalize* the fine-grained spins $\boldsymbol{s}$, we will obtain a new effective action $S'(\boldsymbol{s}')$ for the coarse-grained spins $\boldsymbol{s}'$

$$\boxed{e^{-S'(\boldsymbol{s}')} = \sum_{\boldsymbol{s}} e^{-\tilde{S}(\boldsymbol{s},\boldsymbol{s}')}.} \tag{140}$$



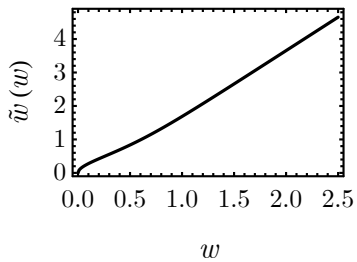
- The fine-grained spins $\boldsymbol{s}$ mediate effective interactions between the emergent coarse-grained spins $\boldsymbol{s}'$,

$$\boxed{S'(\boldsymbol{s}') = -\,w' \sum_{\langle I\,J \rangle} s'_I\, s'_J,} \tag{141}$$

with an effective coupling w' as a function of the original coupling w by (recall Eq. (114))

$$\boxed{w' = \frac{1}{2} \log \cosh(2\, \tilde{w}(w)),} \tag{142}$$

where $\tilde{w}(w)$ is given by the following curve.



- If we rotate the lattice by 45 ° and rescale the lattice spacing by $\sqrt{2}$, the effective action $S'(\boldsymbol{s}')$ in Eq. (141) maps to the original action $S(\boldsymbol{s})$ in Eq. (126), with the coupling modified by $w \to w'$ in Eq. (142).

The coarse graining procedure can be summarized as the following diagram

$$\begin{array}{ccccc} p(\boldsymbol{s}) & & \tilde{p}(\boldsymbol{s},\, \boldsymbol{s}') & \to & p'(\boldsymbol{s}') \\ \downarrow & & \uparrow & & \downarrow \\ S(\boldsymbol{s}) & & \tilde{S}(\boldsymbol{s},\, \boldsymbol{s}') & & S'(\boldsymbol{s}') \\ = \sum_I S_I(\boldsymbol{s}) & \to & = \sum_I \tilde{S}_I(\boldsymbol{s},\, \boldsymbol{s}') & & \ldots \end{array} \tag{143}$$

- Every step is exact except for the *representation learning*, where the KL divergence is non-vanishing (i.e. the learning is not perfect).
- Under coarse graining,
  - the **lattice** up-scales by $\sqrt{2}$, i.e. the lattice spacing rescaled by $\ell \to \ell' = \sqrt{2}\, \ell$,
  - the **action** restores the same form (this doesn't need to be so in general, but let us assume it for now),
  - the **coupling constant** is updated by $w \to w'$ (more generally, it is the generative model that gets updated).

Renormalization Group Flow

RG Flow Equation

Iteratively coarse graining the system will generate a renormalization group (RG) flow of the coupling constant,

$$w^{(0)} \to w^{(1)} \to w^{(2)} \to w^{(3)} \to \ldots, \tag{144}$$

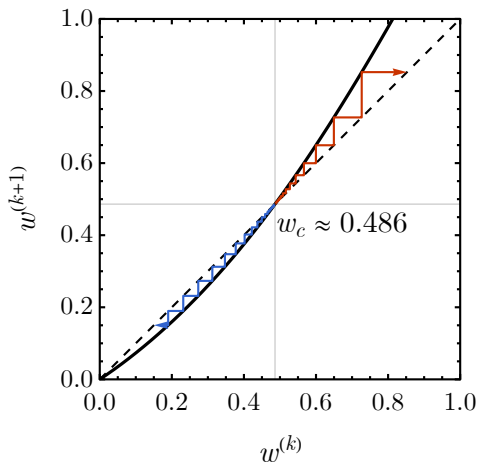
described by a *recurrent* equation, called the **RG flow equation**:

$$\boxed{\begin{aligned} & w^{(k+1)} = \mathcal{R}\big(w^{(k)}\big), \\ & \mathcal{R}(w) := \frac{1}{2} \log \cosh(2\, \tilde{w}(w)). \end{aligned}} \tag{145}$$

- $\mathcal{R}$ denotes the **RG transformation**, mapping the coupling constant w from one scale to another.

- Its specific form in Eq. (145) is derived for 2D Ising model under the given RG scheme (i.e., block size, number of representative spins, parameter space dimension, etc.). $\mathcal{R}$ is expected to be different for different models and RG schemes.

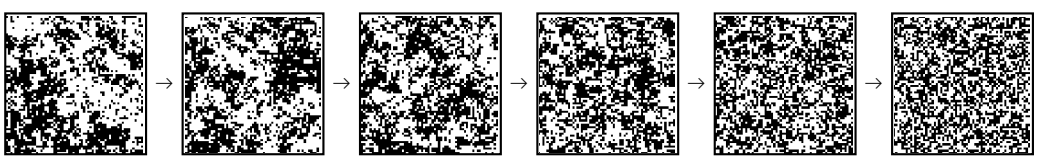
The RG transformation looks like



w - the dimensionless coupling strength (with respect to the temperature)

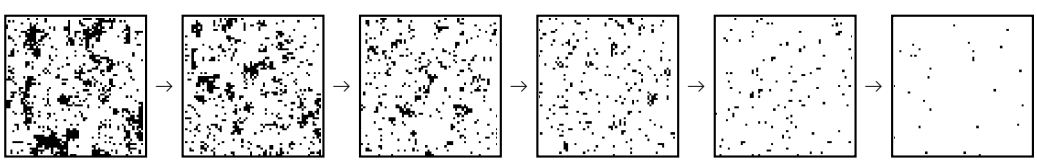
$$w = \beta J = J / T. \tag{146}$$

- *Small* w (weak coupling, high temperature) - **disordered phase**. Under coarse graining:



Spins are randomly oriented due to strong thermal fluctuations. Their orientations averages out, allowing the representative spin to fluctuate more freely, weakening spin correlation at larger scales.

- *Large* w (strong coupling, low temperature) - **ordered phase**. Under coarse graining:



Spins are locally aligned, reinforcing the orientation of local representative spins, mediating in a mutual alignment of representative spins, enhancing spin correlation at larger scales.

- The two phases are separated by a **critical point** at $w = w_c$, where the **phase transition** happens.
  - At the critical point, w stops flowing — realizing a **RG fixed point**. The critical value w_c can be found by solving the *fixed point equation*

$$\boxed{w_c = \mathcal{R}(w_c),} \tag{147}$$

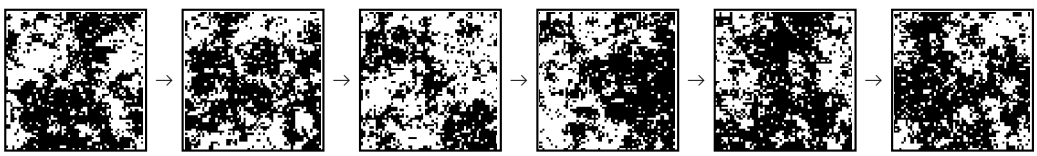
where the RG transformation $\mathcal{R}$ was defined in Eq. (145). One solution of Eq. (147) is

$$w_c \approx 0.486, \tag{148}$$

which is close to the known exact solution of 2D Ising model

$$w_c = \frac{1}{2}\log\left(1+\sqrt{2}\right) \approx 0.4407. \tag{149}$$

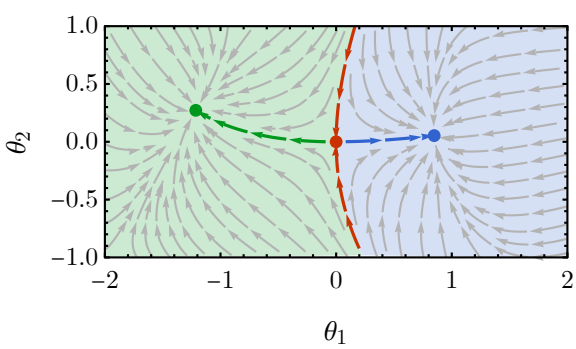
- At the RG fixed point, the probability model $p(\boldsymbol{s})$ looks the same at different scales — an *emergent* **scale invariance**. Under coarse graining:



The spin configuration looks (statistically) the same as we zoom out (or in) — exhibiting a **self-similarity**. As a consequence, all thermodynamic properties must be **scale-free** at the *critical point*.

■ RG Flow Diagram

An **RG flow diagram** illustrates how the **parameters** $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots\}$ of a *physical system* (e.g. temperature, pressure, coupling constants) or a *generative model* (e.g. model parameters) evolves under the RG flow.



- **Flow lines**: the trajectories of $\boldsymbol{\theta}$ under coarse graining. In the continuum limit, the flow of $\boldsymbol{\theta}$ can be described by a *differential* RG flow equation

$$\boxed{\frac{d\boldsymbol{\theta}}{d\log\ell} = -\boldsymbol{\beta}(\boldsymbol{\theta}),} \tag{150}$$

where $\boldsymbol{\beta}(\boldsymbol{\theta})$ describes the flow velocity at each $\boldsymbol{\theta}$ point, and $\log\ell$ measures the RG step (or RG "time") with ℓ being the **cutoff length scale** (such as the lattice spacing).

- **Fixed points**: the points $\boldsymbol{\theta}_\star$ where the flow lines *converge* or *diverge*, i.e. where the flow velocity vanishes

$$\boldsymbol{\beta}(\boldsymbol{\theta}_\star) = \mathbf{0}. \tag{151}$$

  Solutions of $\boldsymbol{\theta}_\star$ represent *scale-invariant* states of the system, where the physical behavior does not change under renormalization.

  - **Stable fixed points** (the green, blue points): attracting flow lines from all directions in their vicinity, representing stable **phases** of the system that are robust against perturbations.
  - **Unstable fixed points** (the red point): repelling flow lines in at least one directions, marking the **phase transition** between adjacent phases.
    - **Critical points**: unstable fixed points with one relevant perturbation (the small perturbation under which $\boldsymbol{\theta}$ flows away from the fixed point).
    - **Multi-critical points**: unstable fixed points with more than one relevant perturbations.
- **Critical surface** (the red flow line): the line (or surface, more generally) on which the RG flow converges to a critical point. It separates the RG flow towards different stable fixed points, and serves as the **phase boundary**.

Perturbations at RG Fixed Points

The RG equation Eq. (150) can be *linearized* for small **perturbations** $\delta\boldsymbol{\theta} := \boldsymbol{\theta} - \boldsymbol{\theta}_\star$ *near* any RG *fixed point* $\boldsymbol{\theta}_\star$,

$$\boxed{\frac{d\,\delta\boldsymbol{\theta}}{d\log\ell} = -\delta\boldsymbol{\theta}\cdot\nabla_{\boldsymbol{\theta}}\boldsymbol{\beta}\left(\boldsymbol{\theta}_\star\right) + O\left(\delta\boldsymbol{\theta}^2\right).} \tag{152}$$

Exc 12 Derive Eq. (152) from Eq. (150).

Any perturbations can be decomposed as a linear combination of **eigen directions** $\boldsymbol{n}_a$,

$$\delta\boldsymbol{\theta} = \sum_a \delta\theta_a\,\boldsymbol{n}_a, \tag{153}$$

such that each **eigenvector** $\boldsymbol{n}_a$ is a solution of the eigen equation

$$\boldsymbol{n}_a\cdot\nabla_{\boldsymbol{\theta}}\boldsymbol{\beta}\left(\boldsymbol{\theta}_\star\right) = \lambda_a\,\boldsymbol{n}_a, \tag{154}$$

with the corresponding **eigenvalue** λ_a.

The eigen perturbation (perturbation along an *eigen direction*) flows as

$$\boxed{\frac{d\,\delta\theta_a}{d\log\ell} = -\lambda_a\,\delta\theta_a + O\left(\delta\boldsymbol{\theta}^2\right),} \tag{155}$$

Exc 13 Derive Eq. (155).

whose solution is (with some integration constant A)

$$\boxed{\delta\theta_a = A\, e^{-\lambda_a \log \ell} = A\, \ell^{-\lambda_a},} \tag{156}$$

meaning that the eigen perturbation always scales with the length scale ℓ in a **power-law** manner with the *exponent* set by the *eigenvalue* λ_a.

- **Relevant perturbation**: an eigen perturbation $\delta\theta_a$ with $\lambda_a < 0$, that *increases* under RG and becomes important at large scale, serving as a **driving parameter** of the phase transition, *perpendicular* to the *critical surface*.
- **Irrelevant perturbation**: an eigen perturbation $\delta\theta_a$ with $\lambda_a > 0$, that *diminishes* under RG and become insignificant at large scale. Irrelevant perturbations span the *tangent space* of the *critical surface*.
- **Marginal perturbation**: an eigen perturbations $\delta\theta_a$ with $\lambda_a = 0$. There is no RG flow under such perturbation — a delicate balance at the threshold between relevance and irrelevance.
  - **Exact marginal perturbation**: all higher order terms in the RG flow equation also vanish, such that $d\delta\theta_a / d\log \ell = 0$ exactly.
  - **Marginally relevant/irrelevant perturbation**: if the higher order terms do not vanish, they will decide whether the perturbation is relevant (increasing) or irrelevant (diminishing) under RG flow.

Universal Scaling Behaviors

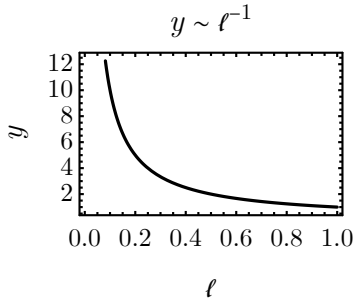
Scaling Hypothesis

The *scale invariance* at the **critical point** implies that all thermodynamic properties of the system must be **scale-free**, i.e. not depending on any specific length scale.

- The only function that is *scale-free* is the **power-law function**. For example

$$y = A\, \ell^{-\lambda}, \tag{157}$$

  any change in the length scale $\ell \to b\,\ell$ can be absorbed into the coefficient A, leaving the form of the function unchanged.



- All thermodynamic properties are derivatives of **free energy density** $f = F / V$ (the free energy per site) with respect to *model parameters* $\boldsymbol{\theta}$ (or Lagrangian multipliers). At the critical point, they scale with the length scale ℓ as

$$\boxed{\begin{aligned} &f \sim \ell^d, \\ &\delta\theta_a \sim \ell^{-\lambda_a}, \end{aligned}} \tag{158}$$

  where d is the **spatial dimension**, and λ_a is the RG eigenvalue associated with the eigen perturbation $\delta\theta_a$.

  - The free energy density scaling $f \sim \ell^d$ follows from the fact that a single site at scale ℓ is a representative of ℓ^d microscopic spins, which has a free energy that is ℓ^d times the average free energy of each microscopic spin.
  - The parameter scaling $\delta\theta_a \sim \ell^{-\lambda_a}$ was given by the solution of the linearized RG equation near the critical point, see Eq. (156).

- To make all thermodynamic properties scale invariant, the free energy density f as a function of $\delta\theta_a$ must take a *scale-free* form, described by the **scaling hypothesis**

$$\boxed{f(\delta\boldsymbol{\theta}) = \ell^d\, \phi\left(\ell^{\lambda_1}\, \delta\theta_1,\, \ell^{\lambda_2}\, \delta\theta_2,\, \ldots\right).} \tag{159}$$

  - ϕ is a **universal scaling function** relating the dimensionless parameters $\delta\theta_a \big/ \ell^{-\lambda_a}$ with the dimensionless free energy $f \big/ \ell^d$, such that the scale ℓ is canceled everywhere, and Eq. (159) is scale-free.

■ Scaling Laws

Take the Ising model for example, the free energy density $f(T, h)$ is a function of the **temperature** T and the **external magnetic field** h

- The *critical point* is at $(T, h) = (T_c, 0)$.
- Eigen perturbations scale as

$$\begin{aligned} &\delta T = T - T_c \sim \ell^{-\lambda_T}, \\ &\delta h = h \sim \ell^{-\lambda_h}. \end{aligned} \tag{160}$$

with two exponents λ_T and λ_h to be determined. We may call λ_T the **temperature exponent** and λ_h the **magnetic exponent**.

- The scaling form Eq. (159) implies

$$\boxed{f(\delta T, h) = \ell^d \, \phi\left(\ell^{\lambda_T} \, \delta T, \ell^{\lambda_h} \, h\right).} \tag{161}$$

Given the free energy in Eq. (161), the scaling behavior of thermodynamic properties can be determined at the critical point

- **Specific heat**

$$c_V = -T \frac{\partial^2 f}{\partial T^2} \sim \ell^{d+2\lambda_T}. \tag{162}$$

- **Magnetization density**

$$m = -\frac{\partial f}{\partial h} \sim \ell^{d+\lambda_h}. \tag{163}$$

- **Spin susceptibility**

$$\chi = -\frac{\partial^2 f}{\partial h^2} \sim \ell^{d+2\lambda_h}. \tag{164}$$

Exc 14 Derive Eq. (162), Eq. (163), Eq. (164).

Instead of scaling with ℓ, we are more interested in how these properties scales with the perturbations $\delta T = T - T_c$ and h (that can be directly tuned). Using Eq. (160) to eliminate ℓ, one arrives at the following **scaling laws**:

$$\boxed{\begin{aligned} &c_V \sim \delta T^{-(d+2\lambda_T)/\lambda_T}, \\ &m \sim \delta T^{-(d+\lambda_h)/\lambda_T}, \\ &\chi \sim \delta T^{-(d+2\lambda_h)/\lambda_T}, \\ &m \sim h^{-(d+\lambda_h)/\lambda_h}. \end{aligned}} \tag{165}$$

Recall the definition of critical exponents, they can all be expressed in terms of λ_T and λ_h:

Definition	Exponent
$c_V \sim \lvert T - T_c \rvert^{-\alpha}$	$\alpha = (d + 2\lambda_T)/\lambda_T$
$m_\star \sim (T_c - T)^\beta$	$\beta = -(d + \lambda_h)/\lambda_T$
$\chi \sim (T - T_c)^{-\gamma}$	$\gamma = (d + 2\lambda_h)/\lambda_T$
$m_\star \sim h^{1/\delta}$	$\delta = -\lambda_h/(d + \lambda_h)$

(166)

As there are only two independent exponents λ_T and λ_h to start with, the four derived exponents α, β, γ, δ must be related by two relations, known as **scaling relations**.

- Rushbrooke's relation [2]

$$\alpha + 2\,\beta + \gamma = 2. \tag{167}$$

- Widom's relation [3]

$$\gamma = \beta\,(\delta - 1). \tag{168}$$

[2] G. S. Rushbrooke, J. Chem. Phys., **39**, 842, (1963).

[3] B. Widom, J. Chem. Phys. **41**, 1633 (1964).

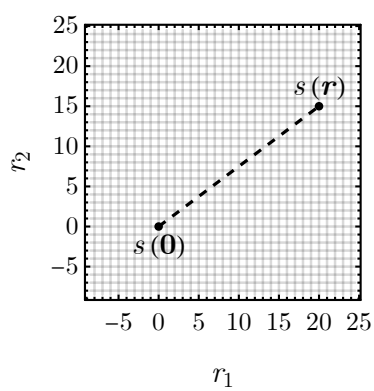
■ Correlation Function

Although scaling laws for c_V, m, χ are intuitively accessible for *experiments*, these quantities are not conveniently calculable in *theory* or *numerics*. Instead, the correlation function often presents a more direct route for calculation.

The *connected two-point* **correlation function** $G(\boldsymbol{r})$ is defined as

$$G(\boldsymbol{r}) = \langle s(\boldsymbol{r})\, s(\boldsymbol{0})\rangle - \langle s(\boldsymbol{r})\rangle\,\langle s(\boldsymbol{0})\rangle, \tag{169}$$

where $s(\boldsymbol{r})$ denotes the spin variable at position $\boldsymbol{r}$ in the space (as the continuum limit of the spin s_i on a site i), distance $|\boldsymbol{r}|$ away from another spin variable $s(\boldsymbol{0})$ at the origin (arbitrarily chosen).



- *At the critical point*, the *self-similar* and *scale-free* properties necessitate a **power-law** correlation

$$G(\boldsymbol{r}) \sim \frac{1}{|\boldsymbol{r}|^{d-2+\eta}}, \tag{170}$$

  thereby defining an exponent η.

- *Away from the critical point*, a characteristic length scale — the **correlation length** ξ — emerges, and the correlation function decays exponentially as

$$G(\boldsymbol{r}) \sim \frac{1}{|\boldsymbol{r}|^{d-2+\eta}}\, e^{-|\boldsymbol{r}|/\xi}. \tag{171}$$

The correlation length is expected to *diverge* ($\xi \to \infty$) as the system approaches the *critical point* (for Eq. (171) to reduce to Eq. (170)). The divergence is also a power-law,

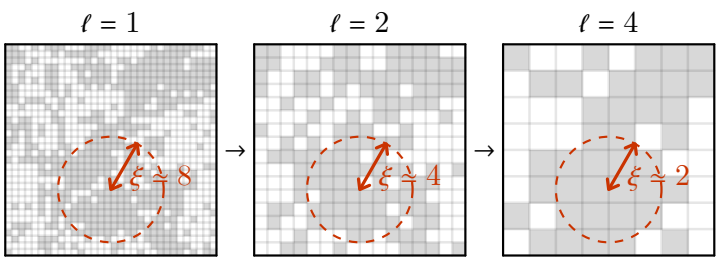
$$\boxed{\xi \sim |T - T_c|^{-\nu},} \tag{172}$$

with another exponent ν.

■ Hyperscaling Relations

The exponents ν and η are also related to λ_T and λ_h. The relations can be analyzed as follows.

- The exponent ν regards how the correlation length ξ scales.
  - The correlation length ξ describes the *typical size* of a **magnetic domain** (a region in which spins are predominantly aligned) in the spin configuration.



  - It worth noting that ξ is measured in *units* of the *lattice spacing* ℓ. So as ℓ increases under coarse graining, ξ will decrease inversely,

$$\xi \sim \ell^{-1}. \tag{173}$$

  - Using the scaling $\delta T \sim \ell^{-\lambda_T}$ in Eq. (160), one finds

$$\xi \sim (\delta T)^{1/\lambda_T}. \tag{174}$$

    Comparing with the definition of ν in Eq. (172), $\xi \sim (\delta T)^{-\nu}$, one concludes

$$\nu = -1/\lambda_T. \tag{175}$$

- The exponent η concerns how fast the spin correlation decays with distance (larger $\eta \to$ more rapid decay).
  - Using the **fluctuation-response relation**, the variance of magnetization $M = \int d^d\boldsymbol{r}\, s(\boldsymbol{r})$ and the spin susceptibility $\chi = -\frac{1}{V}\,\partial_h^2 F$ are related

$$\operatorname{var} M = -T\,\frac{\partial^2 F}{\partial h^2} = V\,T\,\chi. \tag{176}$$

  - Given that $\operatorname{var} M := \langle M^2\rangle - \langle M\rangle^2$, the variance can be expressed as an integration over the correlation function $G(\boldsymbol{r})$,

$$\text{var } M = V \int d^d \boldsymbol{r}\, G(\boldsymbol{r}). \tag{177}$$

Exc 15 Derive Eq. (177).

Using the form of the correlation function in Eq. (171), $G(\boldsymbol{r}) \sim |\boldsymbol{r}|^{-d+2-\eta}\, e^{-|\boldsymbol{r}|/\xi}$, the integration scales as

$$\text{var } M = V \int d^d \boldsymbol{r}\, |\boldsymbol{r}|^{-d+2-\eta}\, e^{-|\boldsymbol{r}|/\xi} \propto V\, \xi^{2-\eta}. \tag{178}$$

Exc 16 Show the scaling of the integration in Eq. (178) by nondimentionalization.

- Comparing Eq. (176) and Eq. (178), one arrives at

$$\chi \sim T^{-1}\, \xi^{2-\eta}. \tag{179}$$

At the critical point, $T = T_c$ is a constant, and by Eq. (165),

$$\chi \sim \delta T^{-(d+2\lambda_h)/\lambda_T} \sim \xi^{-(d+2\lambda_h)}, \tag{180}$$

where the scaling $\delta T \sim \xi^{\lambda_T}$ in Eq. (174) was used. Comparing Eq. (179) and Eq. (180),

$$\eta = 2 + d + 2\lambda_h. \tag{181}$$

In conclusion, ν and η can be expressed in terms of λ_T and λ_h:

$$\boxed{\begin{aligned} \nu &= -1/\lambda_T, \\ \eta &= 2 + d + 2\lambda_h. \end{aligned}} \tag{182}$$

- Just by computing the behavior of the *correlation function* $G(\boldsymbol{r})$ near the critical point, λ_T and λ_h can be determined.

Combining Eq. (166) and Eq. (182), more scaling relations can be discovered

- Josephson's relation [4]

$$\boxed{2 - \alpha = d\,\nu.} \tag{183}$$

- Stell's relation [5]

$$\boxed{2 - \eta = d\,\frac{\delta - 1}{\delta + 1}.} \tag{184}$$

- Fisher's relation [6]

$$\boxed{\gamma = (2 - \eta)\,\nu.} \tag{185}$$

Here is a table of critical exponents for O(n) models in d dimensional space. [$n = 0$: polymers; $n = 1$: Ising universality, liquid-gas transition; $n = 2$: XY universality, superfluid transition;

$n = 3$: Heisenberg universality.] (results from Ref. [7])

		α	β	γ	δ	ν	η
$d = 2$	$n = 1$	0	1/8	7/4	15	1	1/4
$d = 3$	$n = 0$	0.24	0.30	1.16	4.83	0.59	0.028
	$n = 1$	0.11	0.33	1.24	4.81	0.63	0.034
	$n = 2$	0.00	0.35	1.32	4.79	0.67	0.035
	$n = 3$	−0.12	0.37	1.39	4.79	0.71	0.036
$d \geq 4$		0	1/2	1	3	1/2	0

(186)

- Each distinct set of *critical exponents* specifies a **universality class** of critical systems (or continuous phase transitions). Systems within a universality class share the same *critical exponents* and *universal scaling functions*.
- Universality classes are affected by **spatial dimensions** d and **symmetry groups** (such as $\mathrm{O}(n)$).
- For $\mathrm{O}(n)$ models, above the upper critical dimension ($d \geq 4$), all models are governed by the *mean-field* universality class, as fluctuations becomes less significant in higher dimensions.
  - There is a special class of scaling relations, often referred to as **hyperscaling relations**, in which the **spatial dimension** d appears explicitly, such as Eq. (183) and Eq. (184). These relations break down for $d > 4$, where the mean-field exponents prevails.

[4] B. D. Josephson, Phys. Lett. **21**, 608 (1966).

[5] G. Stell, Phys. Rev. Lett. **20**, 533 (1968).

[6] M. E. Fisher, J. Math. Phys. **5**, 944 (1964).

[7] J. Zinn-Justin, Scholarpedia, **5**(5):8346 (2010).

Critical Exponent Estimations

Critical Exponent ν

Can we estimate the critical exponents based on the RG flow equation Eq. (145) derived for the 2D Ising model?

$$\begin{aligned} w^{(k+1)} &= \mathcal{R}\big(w^{(k)}\big), \\ \mathcal{R}(w) &:= \frac{1}{2} \log \cosh(2\, \tilde{w}(w)). \end{aligned} \tag{187}$$

- Introduce the perturbation of the coupling constant w near the critical point w_c,

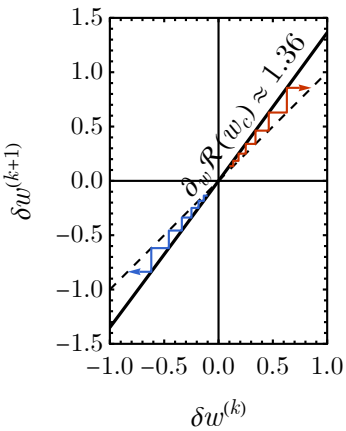
$$\boxed{\delta w := w - w_c,} \tag{188}$$

to linearize the RG flow equation

$$\boxed{\delta w^{(k+1)} = \partial_w \mathcal{R}(w_c)\, \delta w^{(k)} + \ldots.} \tag{189}$$

Exc 17 Derive Eq. (189) based on Eq. (187).

- In each RG step, the perturbation δw will be multiplied by a factor $\partial_w \mathcal{R}(w_c)$ (the slop of $\mathcal{R}(w)$ at $w = w_c$). The fact that $\partial_w \mathcal{R}(w_c) > 1$ makes δw a relevant perturbation.



- Following Eq. (189), starting with an initial perturbation $\delta w^{(0)}$, after k steps of iteration, the perturbation grows to

$$\delta w^{(k)} = (\partial_w \mathcal{R}(w_c))^k \, \delta w^{(0)}. \tag{190}$$

Meanwhile, the correlation length ξ shrinks to

$$\xi^{(k)} = \xi^{(0)} \Big/ \left(\sqrt{2}\right)^k, \tag{191}$$

since the lattice scale ℓ enlarges by $\sqrt{2}$ in each RG step, and the correlation length $\xi \sim \ell^{-1}$ scales inversely with ℓ.

- Eliminate k from Eq. (190) and Eq. (191), one obtains

$$\xi \propto \delta w^{-\nu}, \tag{192}$$

Exc 18 Eliminate k and show Eq. (192).

with the exponent ν given by

$$\boxed{\nu = \frac{\log \sqrt{2}}{\log \partial_w \mathcal{R}(w_c)} \approx 1.1.} \tag{193}$$

The RG estimation is close to the exact result of $\nu = 1$ for 2D Ising universality class. The estimation can be further improved by introducing more latent variables to improved the accuracy of representation learning.

■ Critical Exponent η

The critical exponent η can be estimated by *fitting* the *power-law* decay of the spin **correla-**

tion function at the critical point.

$$G(\boldsymbol{r}) \sim \frac{1}{|\boldsymbol{r}|^{d-2+\eta}}. \tag{194}$$

For 2D Ising model, $d = 2$, Eq. (194) simply reduces to

$$\boxed{G_{ij} \sim \frac{1}{r_{ij}^{\eta}},} \tag{195}$$

where

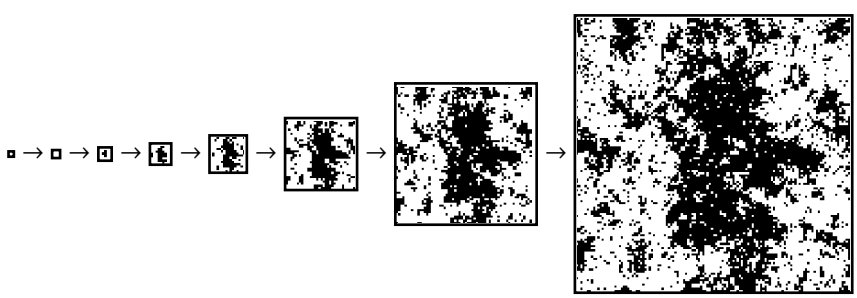
- the correlation function G has been redefined on the *lattice*, indexed by two sites i and j

$$\begin{aligned} G_{ij} &= \langle s_i\, s_j \rangle - \langle s_i \rangle \langle s_j \rangle \\ &= \sum_{\boldsymbol{s}} s_i\, s_j\, p(\boldsymbol{s}) - \left(\sum_{\boldsymbol{s}} s_i\, p(\boldsymbol{s})\right)\left(\sum_{\boldsymbol{s}} s_j\, p(\boldsymbol{s})\right). \end{aligned} \tag{196}$$

- $r_{ij} := |\boldsymbol{r}_i - \boldsymbol{r}_j|$ denotes the distance between sites i and j on a lattice, in unit of the lattice spacing.

How to evaluate the correlation function? - Sample spin configurations $\boldsymbol{s}$ from the distribution $p(\boldsymbol{s})$, calculate correlation by averaging over many samples. Some sampling methods are:

- **Monte Carlo sampling**: unbiased, but suffers from *critical slowdown* (i.e. local update becomes slow to equilibrate the system at criticality).
- **Inverse RG sampling**: biased (approximate), but permits fast direct sampling at the RG fixed point.
  - Key idea: starting from a single spin, progressively fine-graining, regenerate large spin configurations in logarithmic steps.



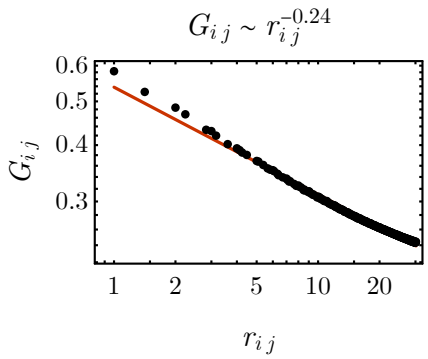
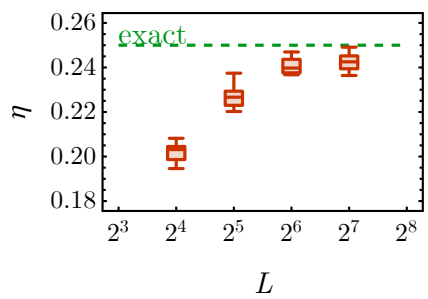
  - The fine-graining is realized by sampling from the conditional distribution

$$\begin{aligned} p(\boldsymbol{s} \mid \boldsymbol{s}') &= \frac{e^{-\tilde{S}(\boldsymbol{s},\boldsymbol{s}')}}{\sum_{\boldsymbol{s}} e^{-\tilde{S}(\boldsymbol{s},\boldsymbol{s}')}}, \\ \tilde{S}(\boldsymbol{s},\, \boldsymbol{s}') &= -\tilde{w}_c \sum_I \sum_{i \in I} s_i\, s'_I. \end{aligned} \tag{197}$$

    - $\boldsymbol{s}$ - fine-grained spin configuration.

- $\boldsymbol{s}'$ - coarse-grained spin configuration.
- $\tilde{w}_c$ - fixed the coupling at the critical value (where the variance of magnetization is maximal).
- The sampling is efficient because given $\boldsymbol{s}'$, the components of $\boldsymbol{s}$ are *independent*, and can be sampled *independently* (no equilibrium process required).

The correlation indeed decays in a power law.



By fitting the power-law exponent, η can be estimated. The calculation can be performed on systems of different sizes L, the estimated η is converging to $\eta \approx 0.24$, close to the exact value $\eta = 1/4$.