CS 229br Lecture 5: Inference and Statistical Physics

Boaz Barak

Yamini Bansal
Official TF

Javin Pombra
Official TF

Dimitris Kalimeris
Unofficial TF

Gal Kaplun
Unofficial TF

Preetum Nakkiran
Unofficial TF

Note: $\#TF \ll \#students$

Digression: Frequentism vs Bayesianism

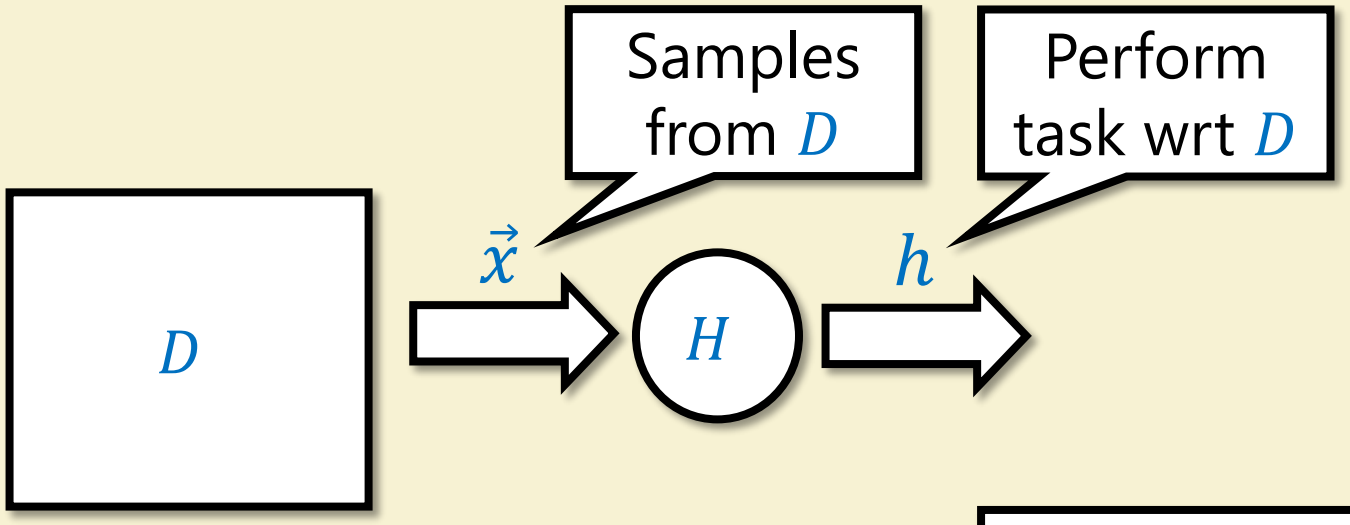
"The probability of winning a battle has no place in our theory … any more than the physical concept of work can be applied to the 'work' done by an actor reciting his part.", Richard von Mises, 1928

I am unable to see why 'objectivity' requires us to interpret every probability as a frequency … in most problems probabilities are frequencies only in an imaginary universe invented just for the purpose of allowing a frequency interpretation.", E.T. Jaynes, 1976

"To the statistician probability appears simply as the ratio which a part bears to the whole … Mr. Keynes adopts a psychological definition. It measures the degree of rational belief", *R. Fischer 1923*

Digression: Frequentism vs Bayesianism

Setup:

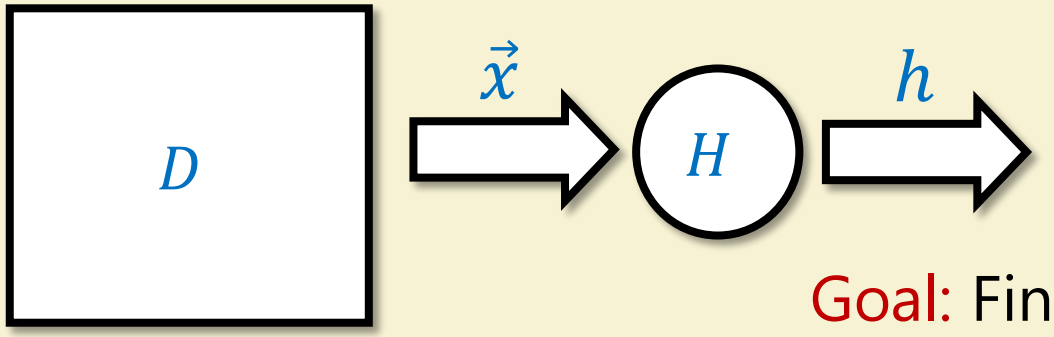


Can measure with samples from D

Goal: Find $h(\vec{x})$ that minimizes $\mathcal{L}_D(h)$

For now: assume no computational limitations

Setup:



Goal: Find $h(\vec{x})$ that minimizes $\mathcal{L}_D(h)$

Frequentist:

Define family $\mathcal{D}$ of potential distributions

Find transformation $\vec{x} \mapsto h(\vec{x})$ minimizing cost if $D \in \mathcal{D}$

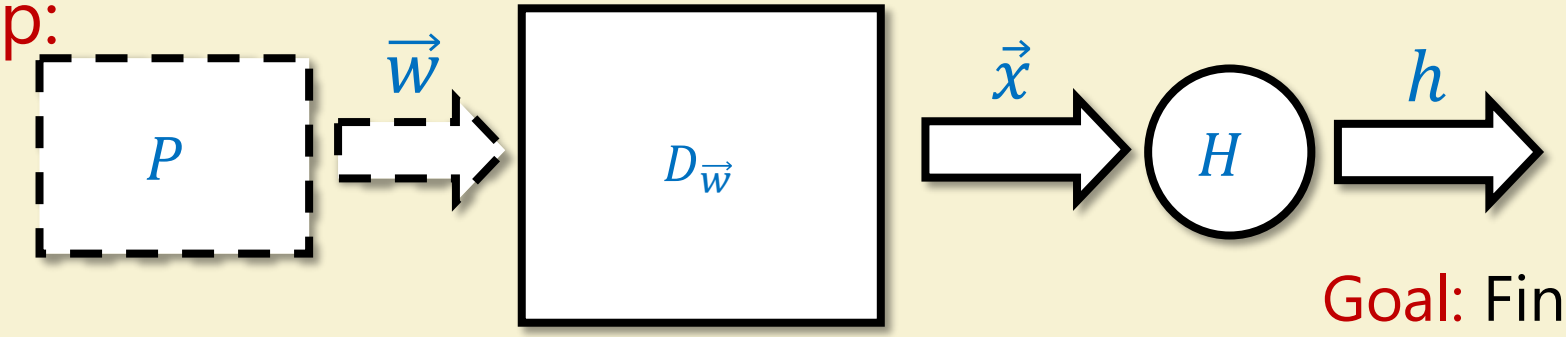
$$h(\cdot) = \arg \min_{\vec{x} \mapsto h(\vec{x})} \max_{D \in \mathcal{D}} \underbrace{\mathbb{E}_{\vec{x} \sim D} \mathcal{L}_D(h(\vec{x}))}$$

Probability over D: "forward looking"

Estimable with sampling access to D

* *Makes no sense if D is deterministic process*

Setup:



Goal: Find $h(\vec{x})$ that minimizes $\mathcal{L}_D(h)$

Bayesian: Assume $D = D_{\vec{w}}$ where $\vec{w} \sim P$ are latent variables

Prior

Let $Q_{\vec{x}}(\vec{w}) = \mathrm{P}(\mathrm{w} \mid \mathrm{D}_{\vec{\mathrm{w}}} = \mathrm{x})$

Posterior

$$h(\vec{x}) = \arg\min_{\mathrm{h}} \mathbb{E}_{\vec{w} \sim Q_{\vec{x}}} \mathcal{L}_{D_{\vec{w}}}(h)$$

Probability over posterior: "backward looking"

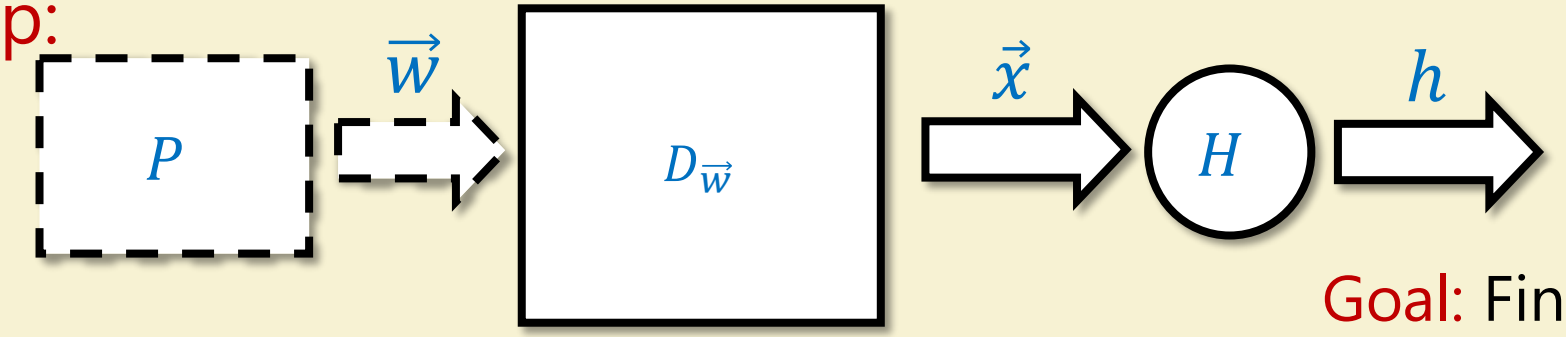
Probability over inaccessible latent variables

* *Makes sense even if* $D_{\vec{w}}$

** *Can do more than min*

How to choose prior?
One approach – maximum entropy principle.

Setup:



Goal: Find $h(\vec{x})$ that minimizes $\mathcal{L}_D(h)$

Computational Constraints

Frequentist: $h(\cdot) = \arg \min_{\vec{x} \mapsto h(\vec{x})} \max_{D \in \mathcal{D}} \mathbb{E}_{\vec{x} \sim D} \mathcal{L}_D(h(\vec{x}))$

Minimize over smaller set of transformations

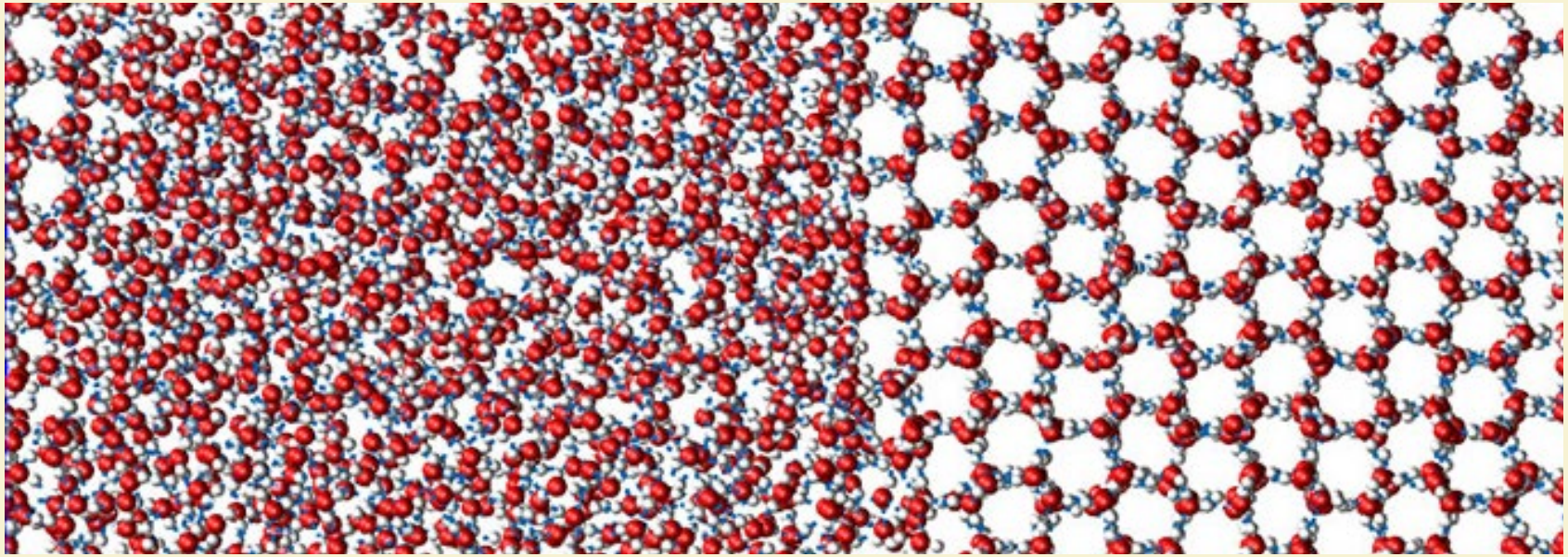
Bayesian: $h(\vec{x}) = \arg \min_{\mathrm{h}} \mathbb{E}_{\vec{w} \sim Q_{\vec{x}}} \mathcal{L}_{D_{\vec{w}}}(h)$

Approximate posterior by simpler distribution

Part I: Intro to statistical physics

Statistical physics

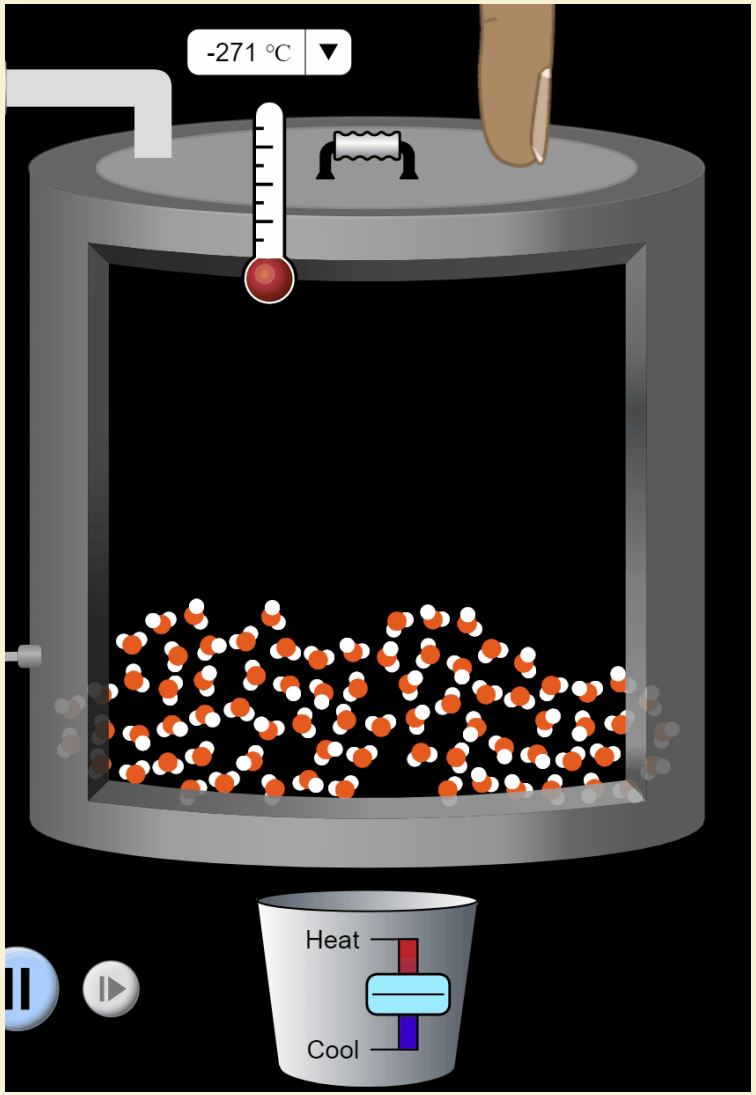
water ice



energy

temperature

Statistical physics



U of Colorado PhET https://phet.colorado.edu/

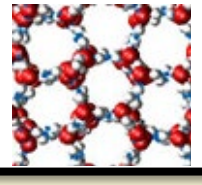
Statistical physics 101

Atomic state: $x \in \{0,1\}^n$

System state: p distribution over $\{0,1\}^n$

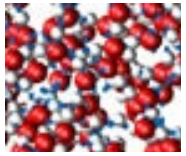
"Utility"/ negative energy function $W: \{0,1\}^n \to \mathbb{R}$

System "wants" to have large $\mathbb{E}_{x\sim p}\, W(x)$



τ = temperature

High temperature –
system "wants" to have large $H(p)$



Equilibrium: $p = \arg\max_p \mathbb{E}_{x\sim p} W(x) + \tau \cdot H(p)$

Variational principle: $p(x) \propto \exp(\tau^{-1} \cdot W(x)) = \exp(\tau^{-1} \cdot W(x) - A_\tau(W))$

"Boltzman distribution"

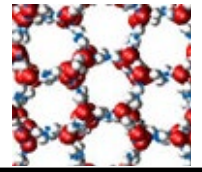
$A_\tau(W) = \log \int \exp(\tau^{-1} \cdot W(x))$

Atomic state: $x \in \{0,1\}^n$

System state: p distribution over $\{0,1\}^n$

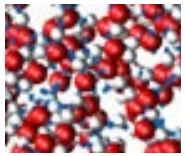
"Utility"/ negative energy function $W: \{0,1\}^n \to \mathbb{R}$

System "wants" to have large $\mathbb{E}_{x\sim p}\, W(x)$



τ = temperature

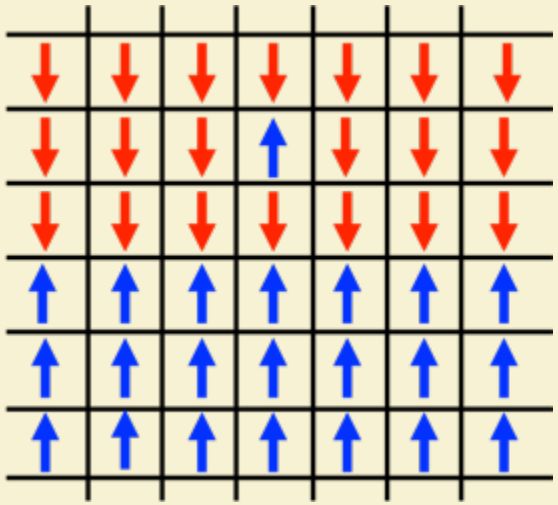
High temperature –
system "wants" to have large $H(p)$



Variational principle: $p(x) \propto \exp(\tau^{-1} \cdot W(x))$

"Boltzman distribution"

Example 1: Ising model



$x \in \{\pm 1\}^n$ represents "spin"

$W(x) = J \sum_{i\sim j} x_i x_j + J' \sum_i x_i$

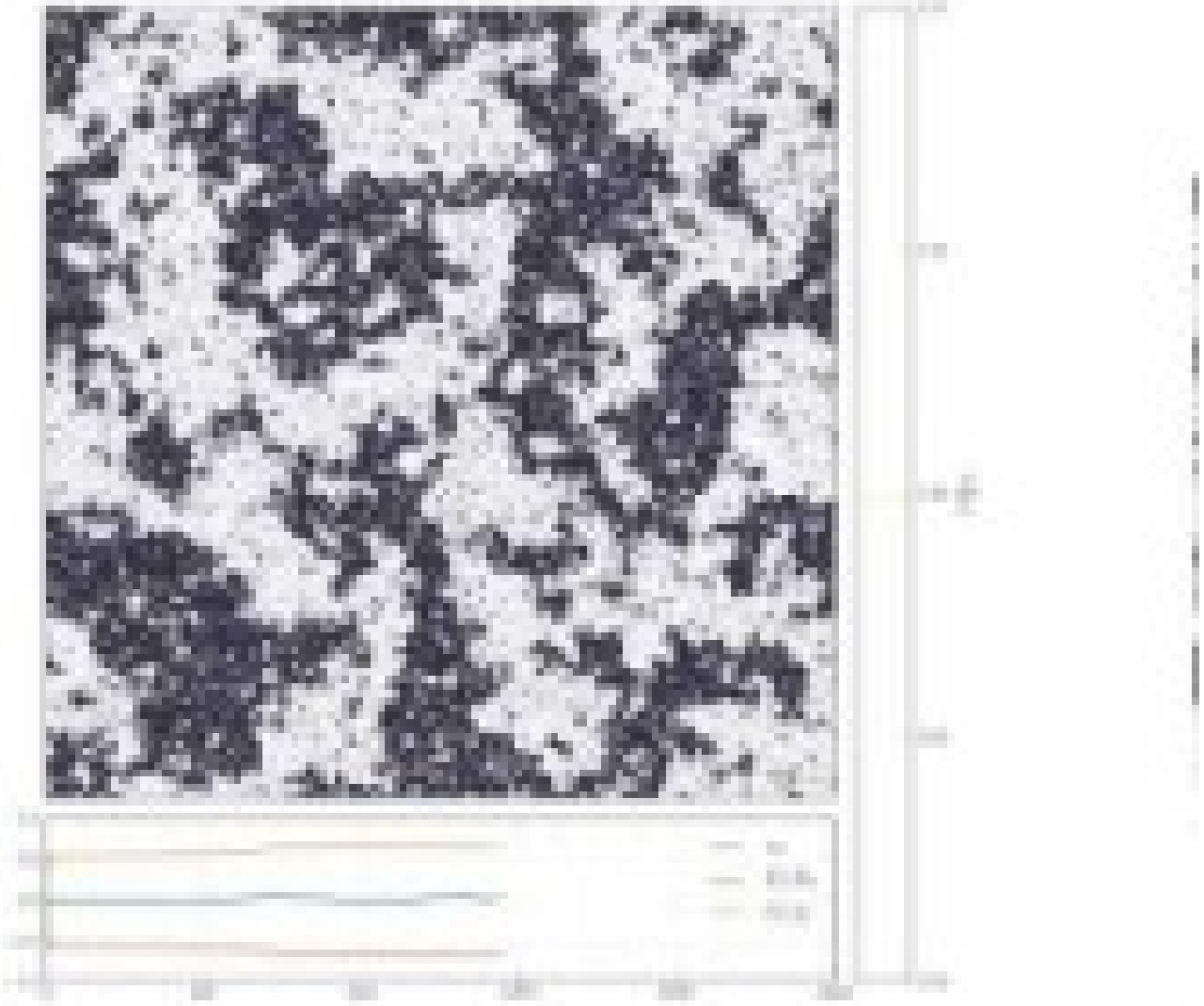
$\{x_i, x_i x_j\}_{i,j}$ are "sufficient statistics"

Variational principle: $p(x) \propto \exp(\tau^{-1} \cdot W(x))$

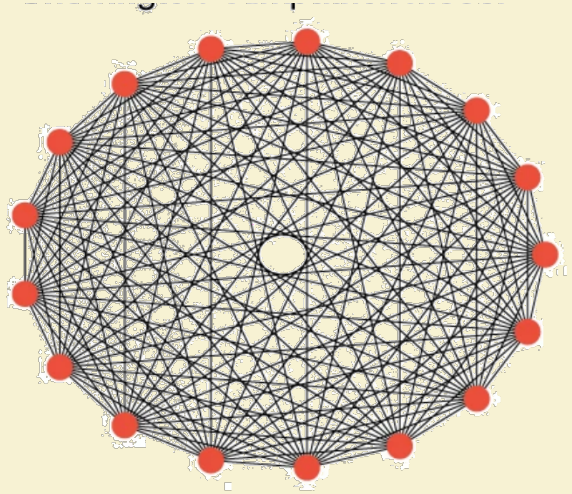
Example 1: Ising model

$x \in \{\pm 1\}^n$ represents "spin"

$W(x) = J \sum_{i \sim j} x_i x_j + J' \sum_i x_i$



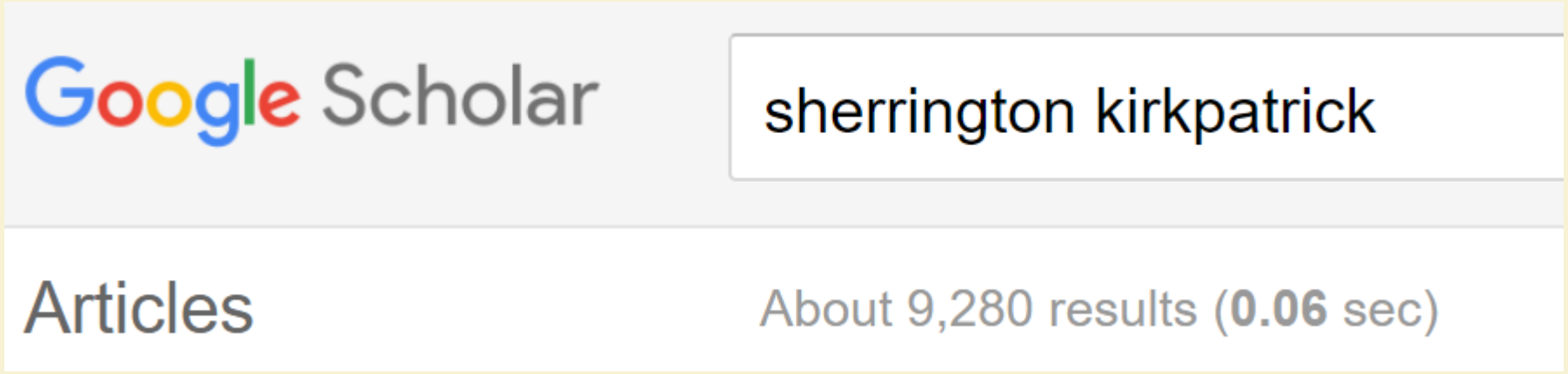
Variational principle: $p(x) \propto \exp(\tau^{-1} \cdot W(x))$



Example 2: Sherrington Kirpatrick Model

Ising on random graph – disordered mean field model

$x \in \{\pm 1\}^n$ represents "spin" $W(x) = \sum w_{i,j} x_i x_j$ $w_{i,j} \sim N(0,1)$



Variational principle: $p(x) \propto \exp(\tau^{-1} \cdot W(x))$

Example 2: Posterior distribution

x = hidden variable with uniform prior

Make k independent observations $O_1, \ldots, O_k$ about x

$$p(x) \propto \Pr[x \text{ satisfies } O_1] \cdot \Pr[x \text{ satisfies } O_2] \cdots \Pr[x \text{ satisfies } O_k]$$

$$= \exp(-\log \sum \Pr[x \text{ satisfies } O_i])$$

"High temperature" = "low learning rate"

Proof of variational principle

$A_\tau(W) = \log \int \exp(\tau^{-1} \cdot W(x))$

Thm: Let $p(x) \propto \exp(\tau^{-1} \cdot W(x)) \quad = \exp(\tau^{-1} \cdot W(x) - A_\tau(W))$

Then $\quad p = \arg\max_p \mathbb{E}_{x\sim p} W(x) + \tau \cdot H(p)$

PF:

$$H(p) = -\int \overbrace{(\tau^{-1} \cdot W(x) - A_\tau(W))}^{\log p(x)} p(x) dx$$

$$= -\tau^{-1} \cdot \mathbb{E}_{x\sim p} W(x) + \mathbb{E}_{x\sim p}\, A_\tau(W)$$

Independent of x

$\Rightarrow$ Claim: $\tau \cdot A_\tau(W) = \tau \cdot H(p) + \mathbb{E}_{x\sim p}\, W(x)$

Free Energy

Canonical entropy

(neg) internal energy

Proof of variational principle

Thm: Let $p(x) \propto \exp(\tau^{-1} \cdot W(x)) \quad = \exp(\tau^{-1} \cdot W(x) - A_\tau(W))$

Then $\quad p = \arg\max_p \mathbb{E}_{x\sim p} W(x) + \tau \cdot H(p)$

PF: Let q be other dist

$$0 \le \tau \cdot \Delta_{KL}(q \parallel p) = \tau \cdot \mathbb{E}_q \log q \; - \tau \cdot \mathbb{E}_q \log p$$

$$= -\tau \cdot H(q) \quad - \tau \cdot \mathbb{E}_{x\sim q} \tau^{-1} \cdot W(x) + \tau \cdot A_\tau(\mathrm{W})$$

$$\Rightarrow \tau \cdot H(p) + \mathbb{E}_{x\sim p} W(x) \quad -\tau \cdot H(q) - \mathbb{E}_{x\sim q} W(x) \ge 0$$

Claim: $\tau \cdot A_\tau(W) = \quad \tau \cdot H(p) + \mathbb{E}_{x\sim p}\, W(x)$

■

Proof of variational principle

Thm: Let $p(x) \propto \exp(\tau^{-1} \cdot W(x)) \quad = \exp(\tau^{-1} \cdot W(x) - A_\tau(W))$

Then $\quad p = \arg\max_p \mathbb{E}_{x\sim p} W(x) + \tau \cdot H(p)$

Claim: $\tau \cdot A_\tau(W) = \quad \tau \cdot H(p) + \mathbb{E}_{x\sim p}\, W(x)$

Recall: VAE

$\min \Delta_{KL}(E(x) \parallel N(0,I))$

$\min \|x - D(E(x))\|^2$

Cor ($\tau = 1$): $A(W) = \max_q H(q) + \mathbb{E}_{x\sim q} W(x)$

divergence

Energy/reconstruction

In particular, for every q, $A(W) \geq H(q) + \mathbb{E}_{x\sim q} W(x)$

log likelihood

ELBO

Computable for "tractable" q

Sampling from Boltzman distribution

$$p(x) = \exp(\tau^{-1} \cdot W(x) - A_\tau(W))$$

$$A_\tau(W) = \log \int \exp(\tau^{-1} \cdot W(x))$$

Typically: Can compute $W(x)$, can't compute $A_\tau(W)$

$\Rightarrow$ Can compute $p(x)/p(x')$!

Gibbs /MH sampling: Let x_0 random and for $i = 1,2,$

Choose x' "near" x_{i-1}

e.g. x' agrees with x_{i-1} in all but one coordinate

$$x_i = \begin{cases} x', & w.p. \min\{1, p(x')/p(x)\} \\ x_{i-1}, & otherwise \end{cases}$$

Rejection sampling

Gibbs /MH sampling: Let x_0 random and for $i = 1,2, \ldots$

Choose x' "near" x_{i-1}

$$x_i = \begin{cases} x', & w.p. \min\{1, p(x')/p(x)\} \\ x_{i-1}, & otherwise \end{cases}$$

Claim: p is stationary distribution of MH

Detailed balance

PF: $p(x'|x)p(x) = p(x|x')p(x') = \min\{p(x), p(x')\}$

$\Rightarrow$If $x_{i-1} \sim p$, probability ratios same for x_i

Under certain connectivity conditions, can prove stationary distribution is unique

Main question is time to converge

Optimization: Simulated annealing

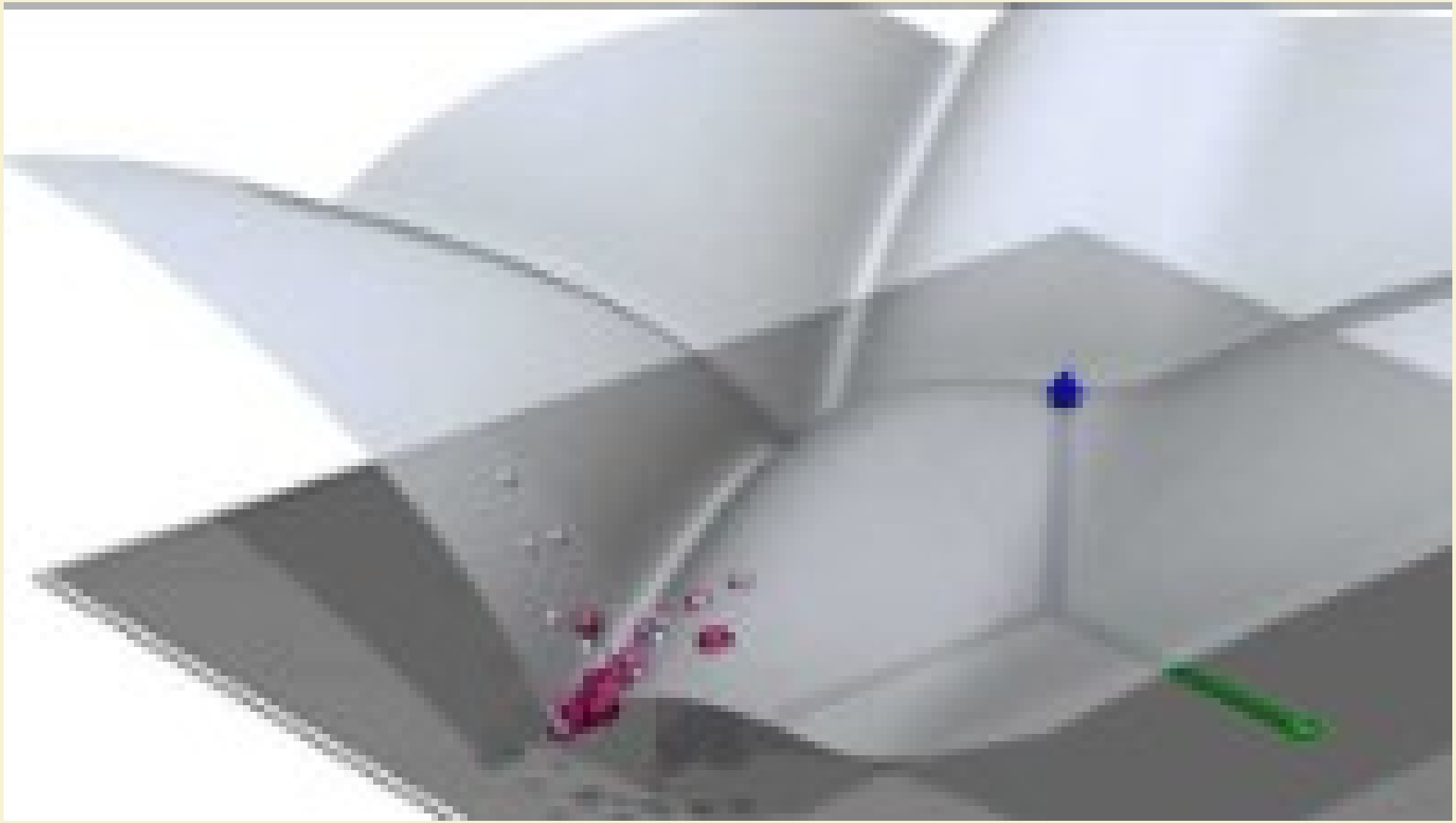
Input: $W: \{0,1\}^n \to \mathbb{R}$

Goal: Find $x = \arg\min W(x)$

Idea/Hope: For $\tau = \infty, \ldots, 0$: sample $x \sim p_\tau$

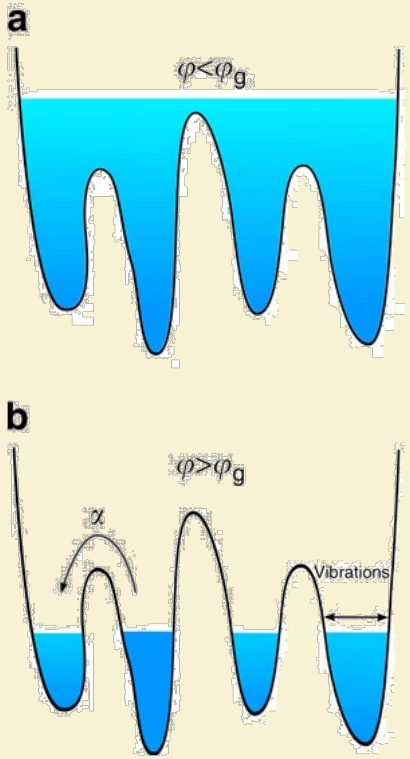
Sampling $x \sim p_\infty$ easy

At each stage, move from x to x' with probability $\exp(\tau^{-1} \cdot (W(x') - W(x)))$

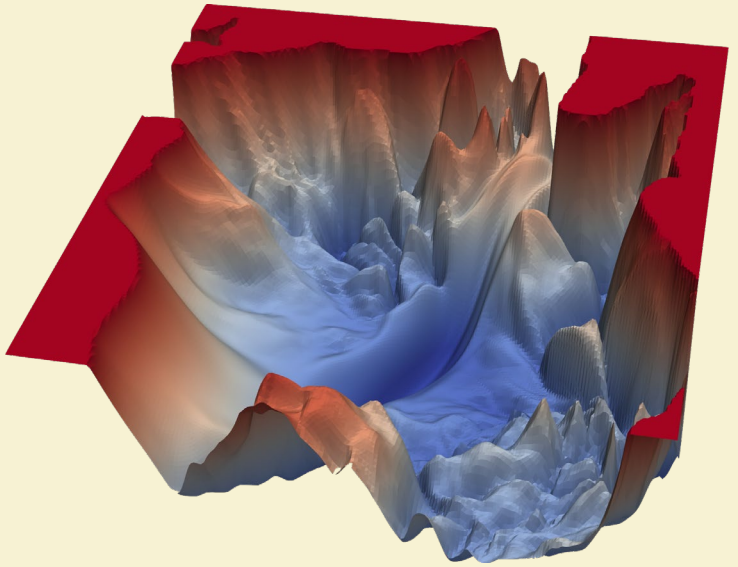


Andreas Müller https://www.youtube.com/watch?v=iaq_Fpr4KZc

Barriers in simulated annealing



Charbonneau et al Nature Comms 2014

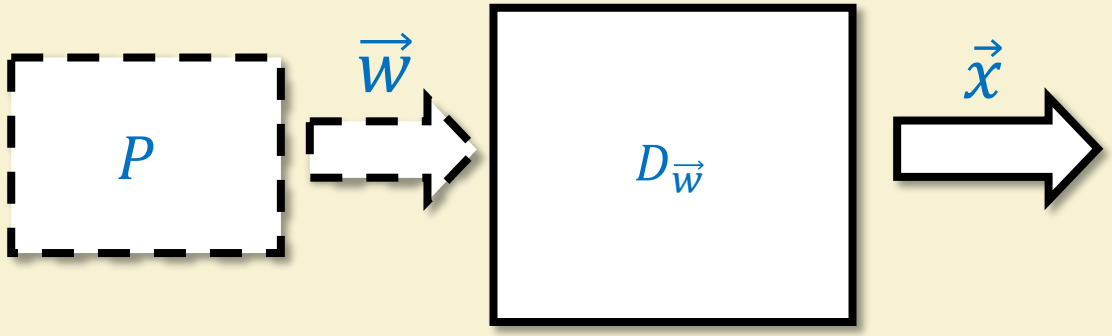


Li, Xu, Taylor, Studer, Goldstein ‘18

https://www.cs.umd.edu/~tomg/projects/landscapes/

Part II: From physics to learning

Bayesian Analysis



$\vec{x} = (x_1, \ldots, x_n)$

Easy to compute / constant

$X_i(\vec{w}) := \log \Pr[x_i|\vec{w}]$

$$p(\vec{w} \mid x_1 \ldots x_n) = \frac{p(\vec{w})}{p(\vec{x})} \cdot p(x_1|\vec{w})p(x_2|\vec{w}) \cdots p(x_n|\vec{w}) \quad \propto \exp(-\sum X_i(\vec{w}))$$

hard to compute

For fixed $\vec{x}$, probability on $\vec{w}$ is Boltzmann with roles flipped

$\vec{x}$ defines energy function ⇒ inference = sampling from posterior

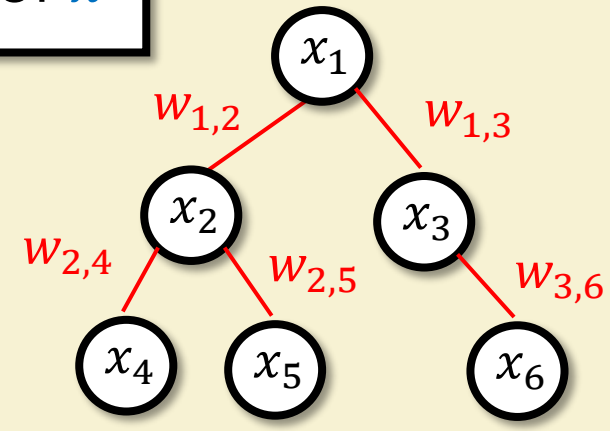
Exponential distributions

Assume $\tau = 1$

$$p_W(x) = \exp(W(x) - A(W)) = \exp(\langle w, \hat{x} \rangle - A(w))$$

Assume $W(x) = \langle w, \hat{x} \rangle$

$\hat{x} \in \mathbb{R}^m$ are sufficient statistics of x



Example: $W(x) = \sum_{(i,j)\in E} w_{i,j} x_i x_j$

$\langle w, \hat{x} \rangle$

$$\hat{x} = (x_1x_2, x_1x_3, x_2x_4, x_2x_5, x_3x_6)$$

To know $\mathbb{E}_{x\sim p_W} W(x)$ enough to know $\mu = \mathbb{E}_{x\sim p_W} \hat{x}$

In example: $\mu_6 = \mu_3 \cdot \frac{\exp(w_{3,6})}{1+\exp(w_{3,6})} + (1 - \mu_3) \cdot \frac{1}{1+\exp(w_{3,6})}$

n linear equations on n marginals

Given marginal μ_6 can sample x_6 & recursively sample $x_1 \dots x_5$

Exponential distributions

Assume $\tau = 1$

$$p_W(x) = \exp(\langle w, \hat{x} \rangle - A(w))$$

Average statistics: Let $\mu = \mathbb{E}_{x \sim p_w} \hat{x}$

By variational principle:

$$p_w = \arg \max_{q \ s.t. \mathbb{E}_q \hat{x} = \mu} H(q)$$

Recall: $A(W) = H(p_W) + \mathbb{E}_{x \sim p_W} W(x) = \mathrm{H}(\mathrm{p_w}) + \langle w, \mu \rangle$

Facts: $\nabla \mathrm{A}(\mathrm{w}) = \mathbb{E}_{p_w} \hat{x} = \mu$

$\mathbf{H}A(w) = Cov_{p_w}(\hat{x}) \succcurlyeq 0$

A is convex

Exponential distributions

$$A(w) = H(p_w) + \langle w, \mu \rangle$$

$$p_w(x) = \exp(\langle w, \hat{x} \rangle - A(w))$$

Average statistics: Let $\mu = \mathbb{E}_{x \sim p_w} \hat{x}$

$$p_w = \arg \max_{q \ s.t. \mathbb{E}_q \hat{x} = \mu} H(q)$$

$$H(\mu) = \max_{\text{p s.t.} \mathbb{E}_p \hat{x} = \mu} H(p)$$

$\{\mu | H(\mu) \geq \alpha\}$ convex set

$$\mu = \arg \max_{H(\mu) \geq H(p_w)} \langle w, \mu \rangle$$

p_w is the maximum entropy distribution consistent with observations μ

Parameter representation

w

Sampling*

(compute $A(w)$)

Mean representation

μ

Inference*

(compute $A^*(\mu) = -H(\mu)$)

* When p is posterior then roles flip

Examples (see Wainright&Jordan)

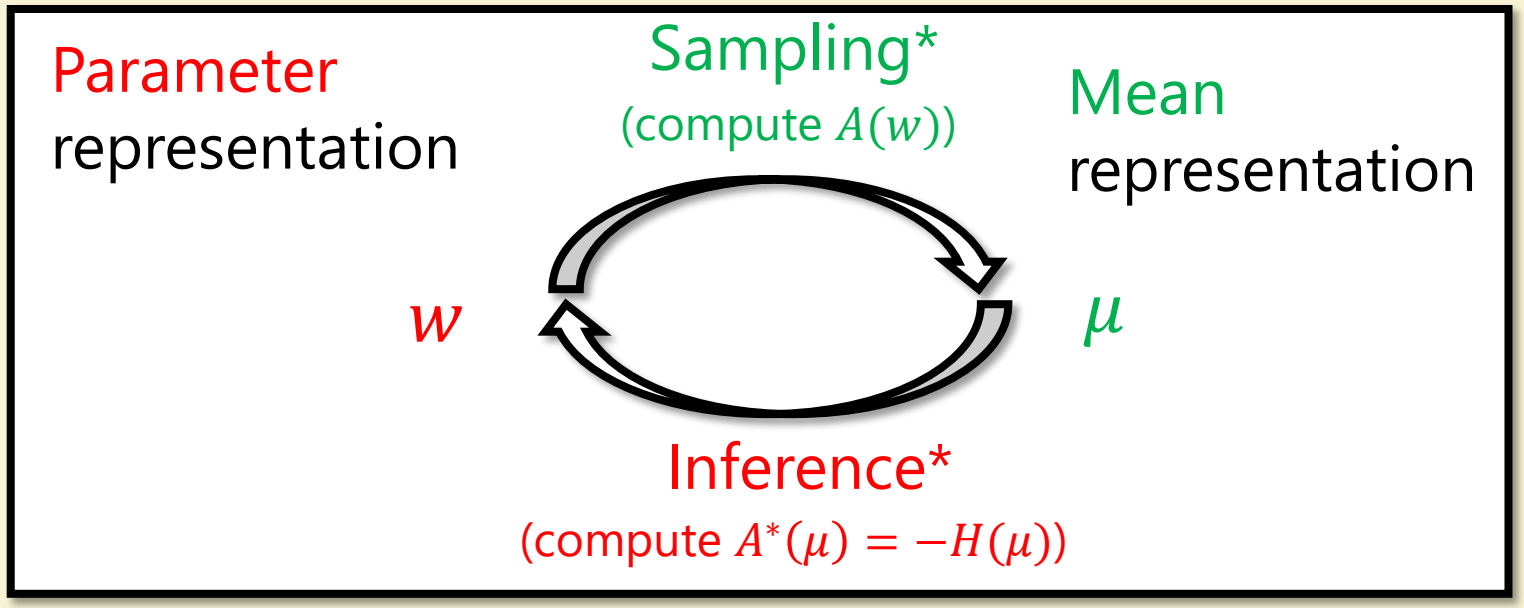
High dim normal: $x \in \mathbb{R}^d$, $W(x) = -(x-\mu)^\top \Sigma^{-1}(x-\mu)$

Ising model: $x \in \{0,1\}^d$, $W(x) = \sum w_i x_i + \sum_{(i,j)\in E} w_{i,j} x_i x_j$

Only case we will deal today

$$\hat{x} = (x, xx^\top) = (x_1, \dots, x_n, x_1^2, x_1 x_2, \dots, x_{n-1} x_n, x_n^2)$$

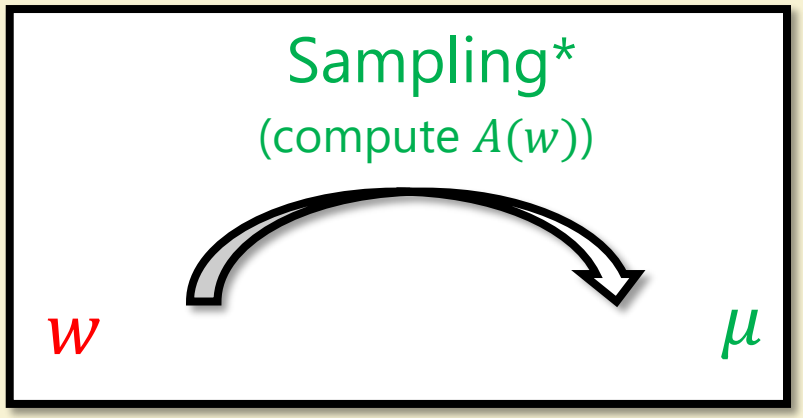
Gaussian MRF, Mixture of Gaussians, latent Dirichlet allocation,...



Mean Field Approximation

$$p_W(x) = \exp(\langle w, \hat{x} \rangle - A(w))$$

$$A(w) = \max_q \langle w, \mathbb{E}_q \hat{x} \rangle + H(q)$$



$\mu_{i,j} = \mu_i \mu_j$

Restrict to q which is product distribution over $x_1 \ldots . x_d$ $(\in \{0,1\}^d)$

$H(\alpha) = -\alpha \log \alpha - (1-\alpha) \log(1-\alpha)$

$$A(w) \geq \max_{\mu \in \mathbb{R}^d} \sum w_i \mu_i + \sum w_{i,j} \mu_i \mu_j + \sum H(\mu_i)$$

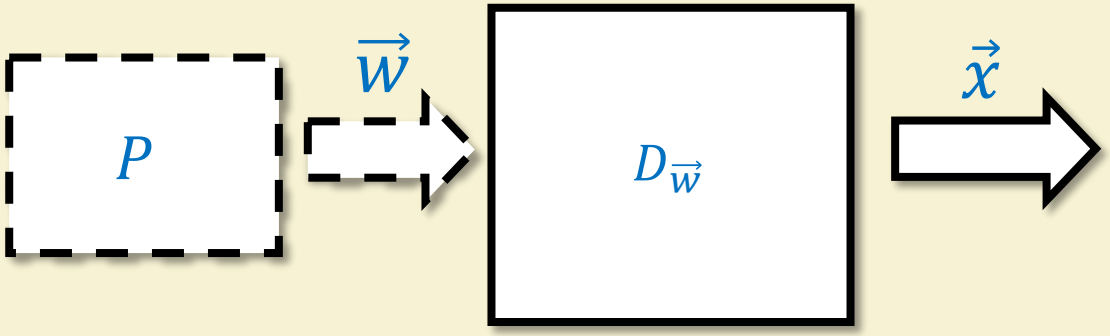
Not concave ☹

Concave in every coordinate ☺

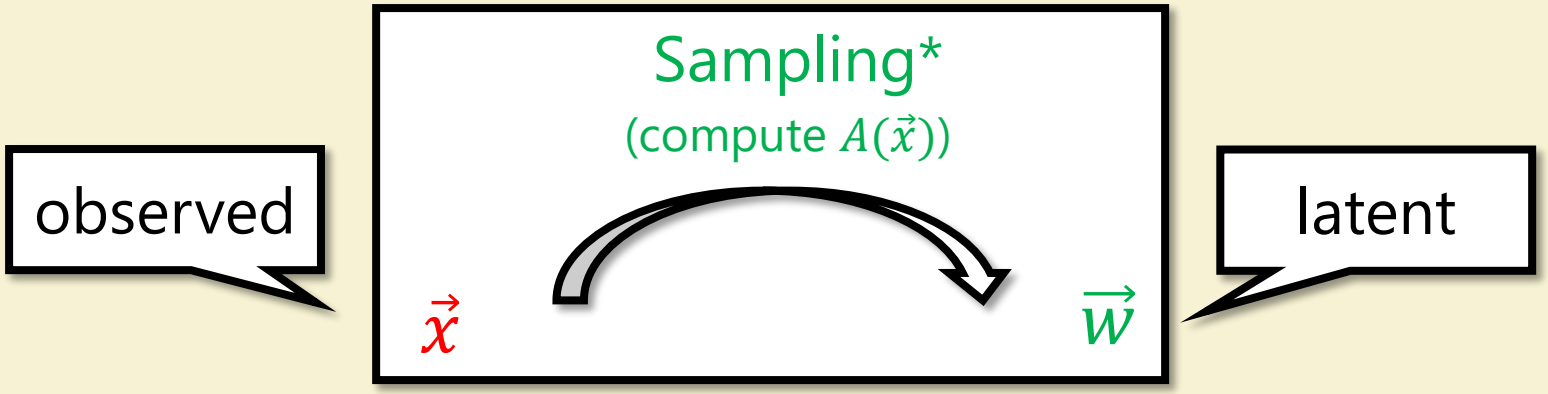
Coordinate Ascent mean-field Variational Inference (CAVI)

Generalizations: Other tractable q

Bayesian Analysis



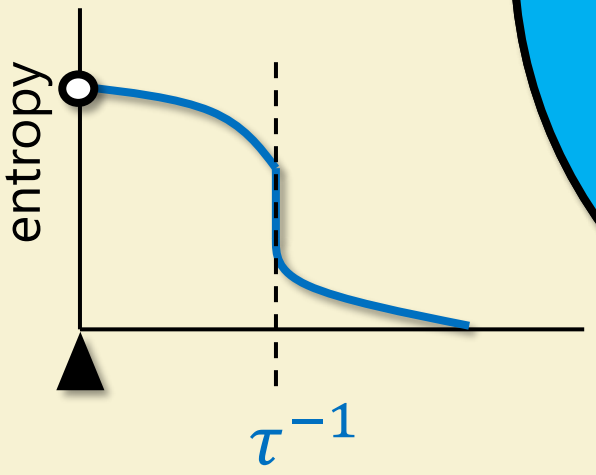
$$p(\vec{w} \mid x_1 \ldots . x_n) \propto \exp(-\sum X_i(\vec{w}))$$

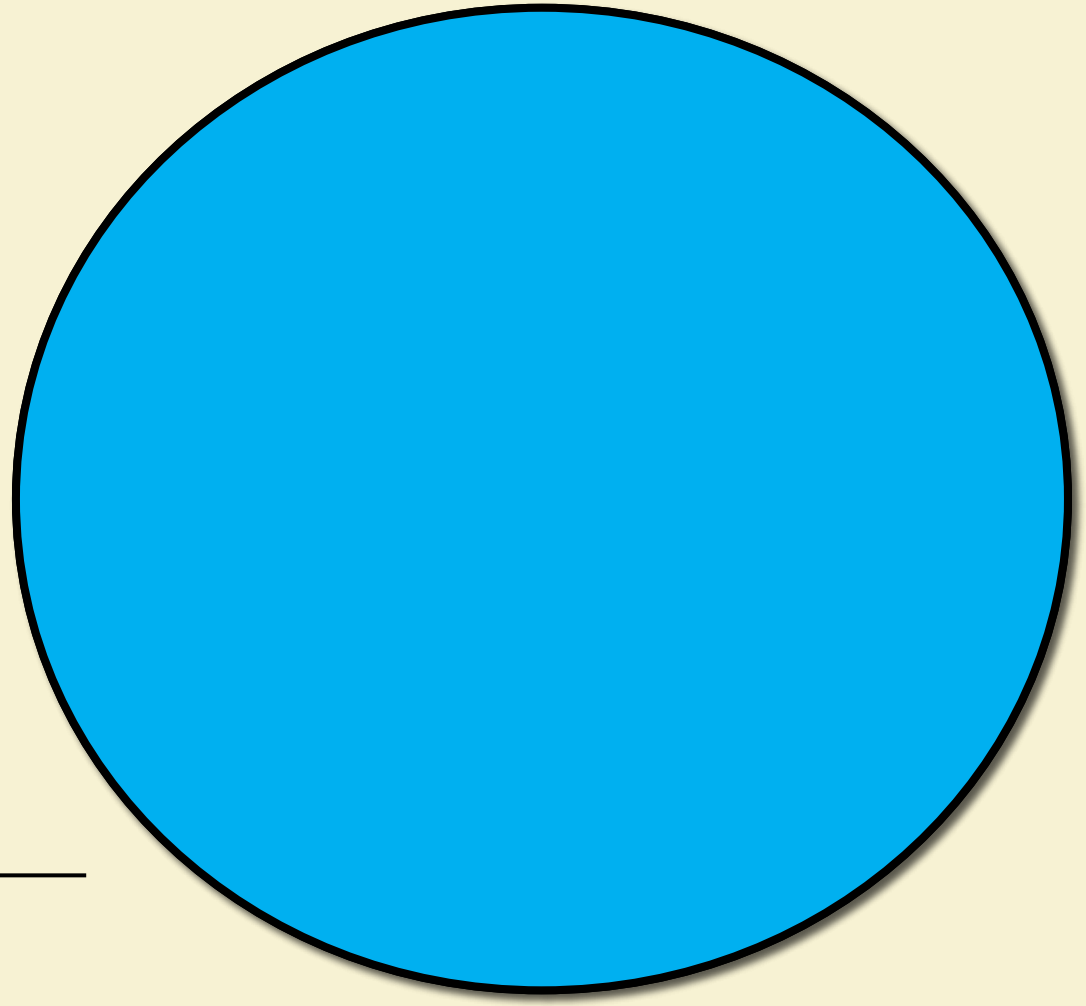


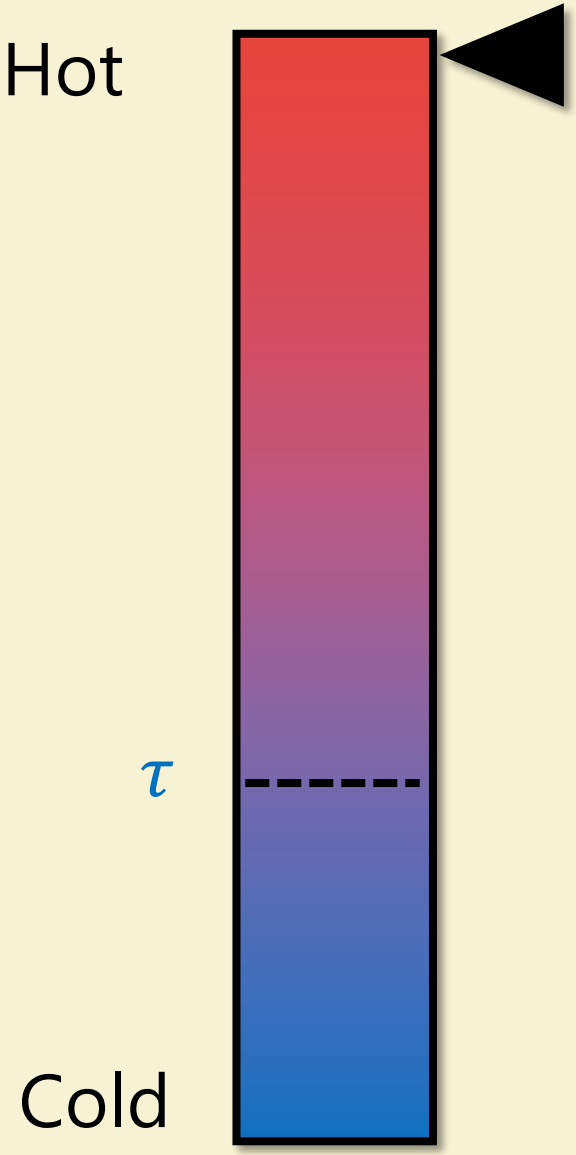
Mean-field approximation: for fixed $\vec{x}$, probability on $\vec{w}$ is product (with parameters depending on $\vec{x}$)

Part III: Solution landscape & replica method

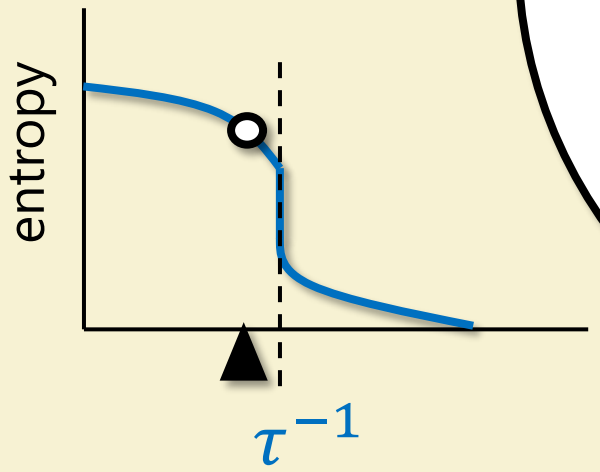
Support of distribution

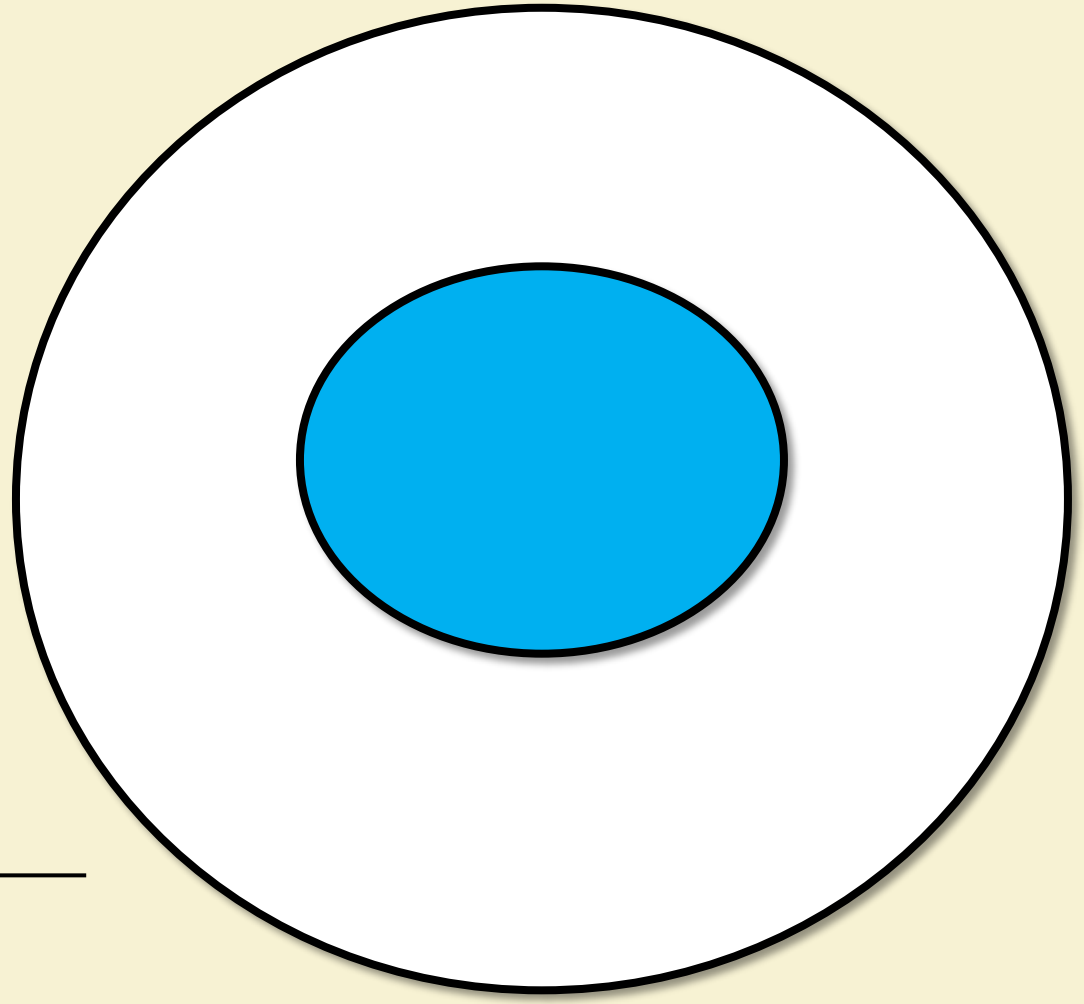


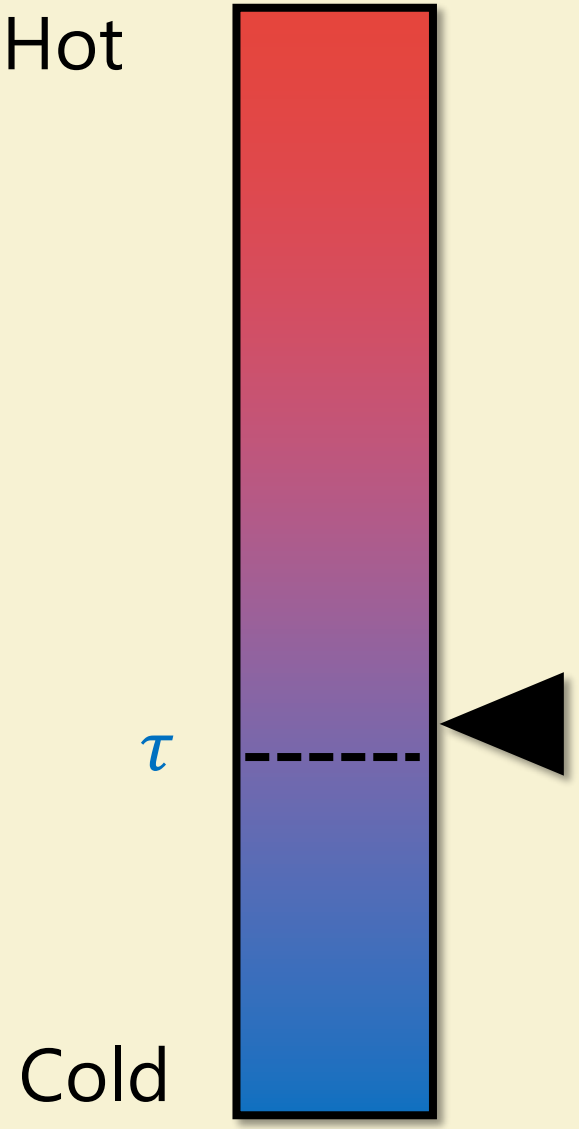




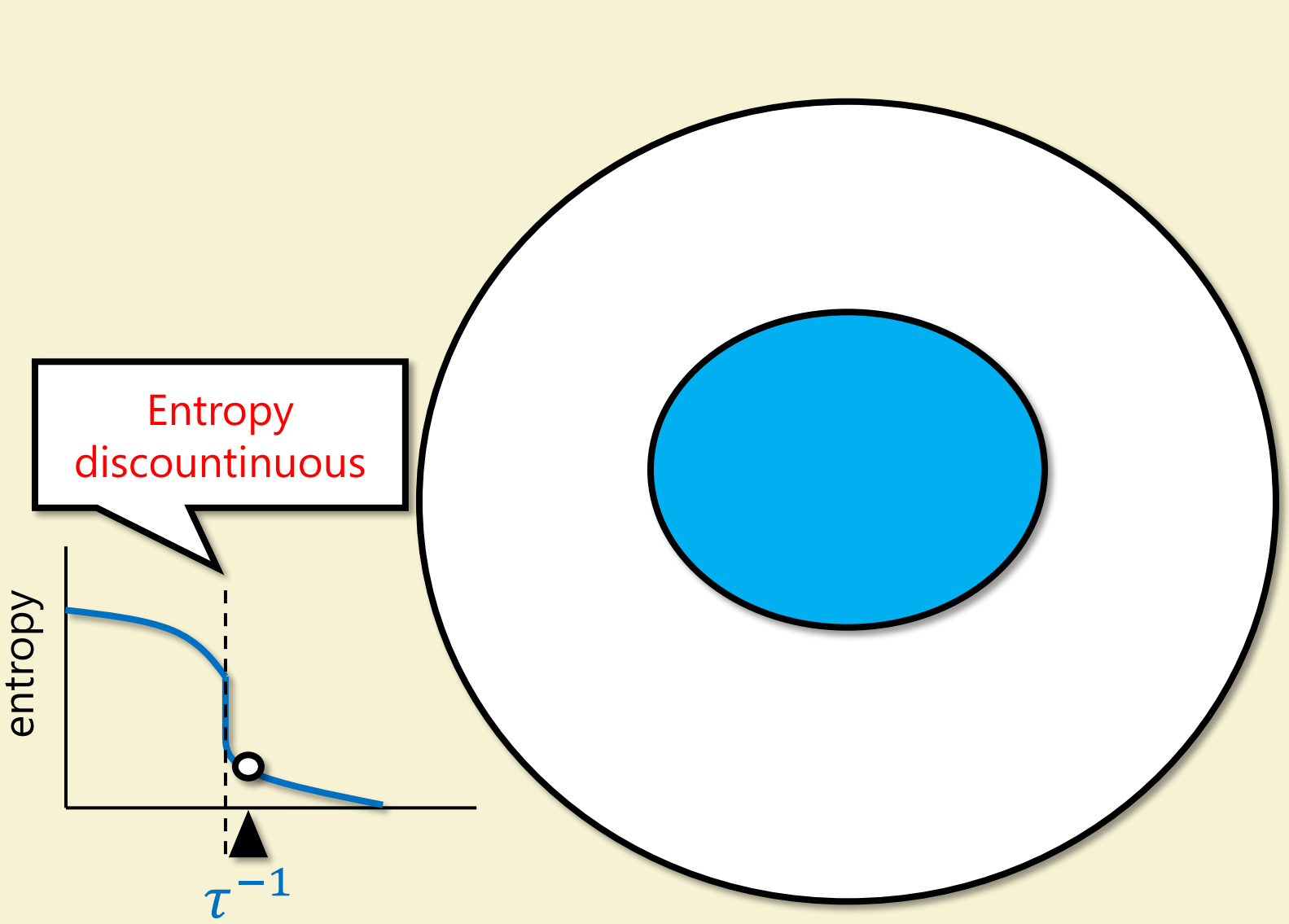
Support of distribution

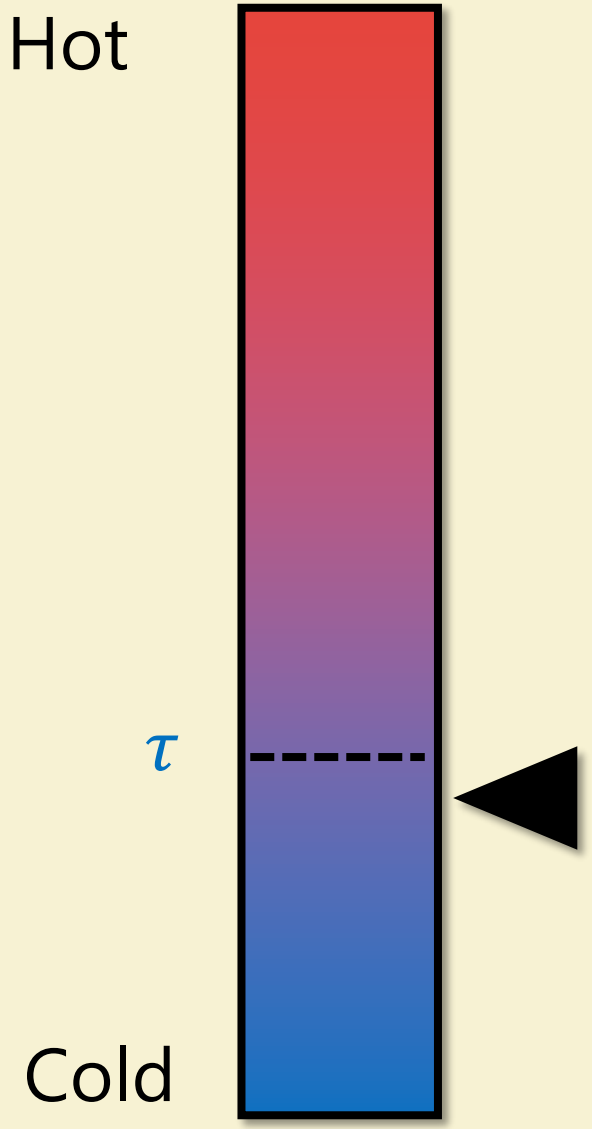




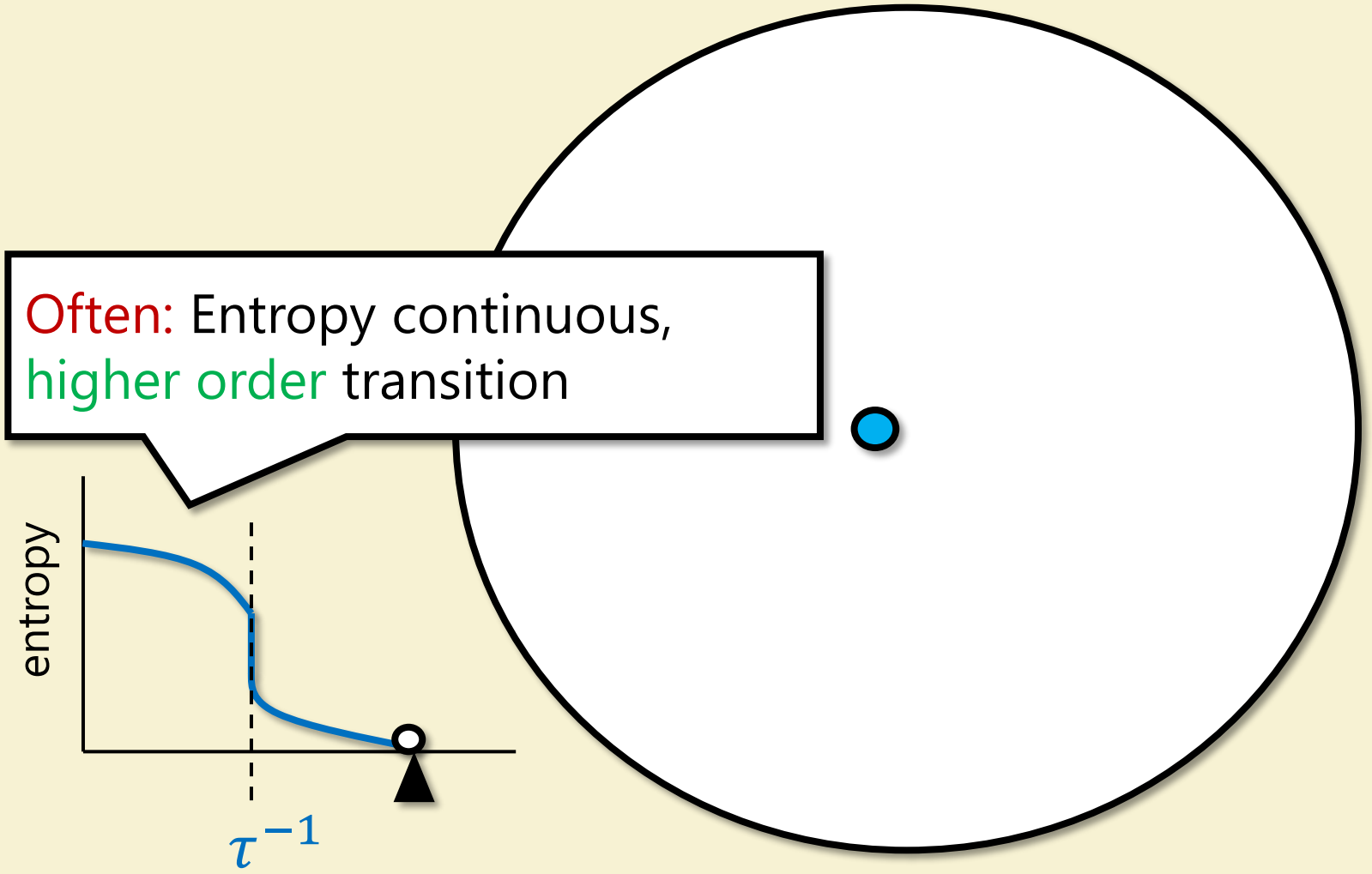


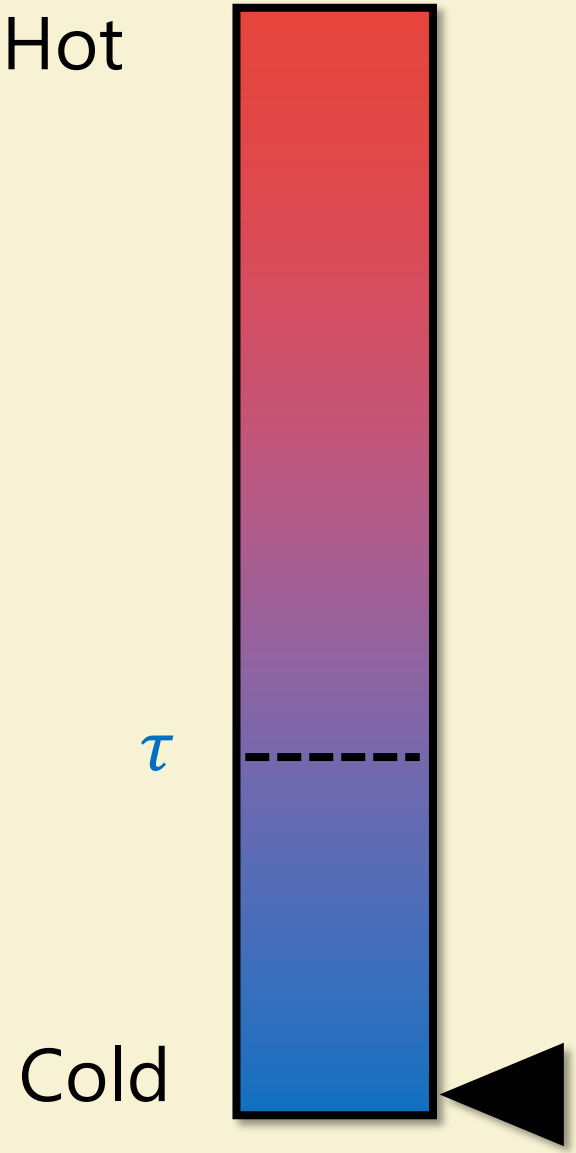
Support of distribution





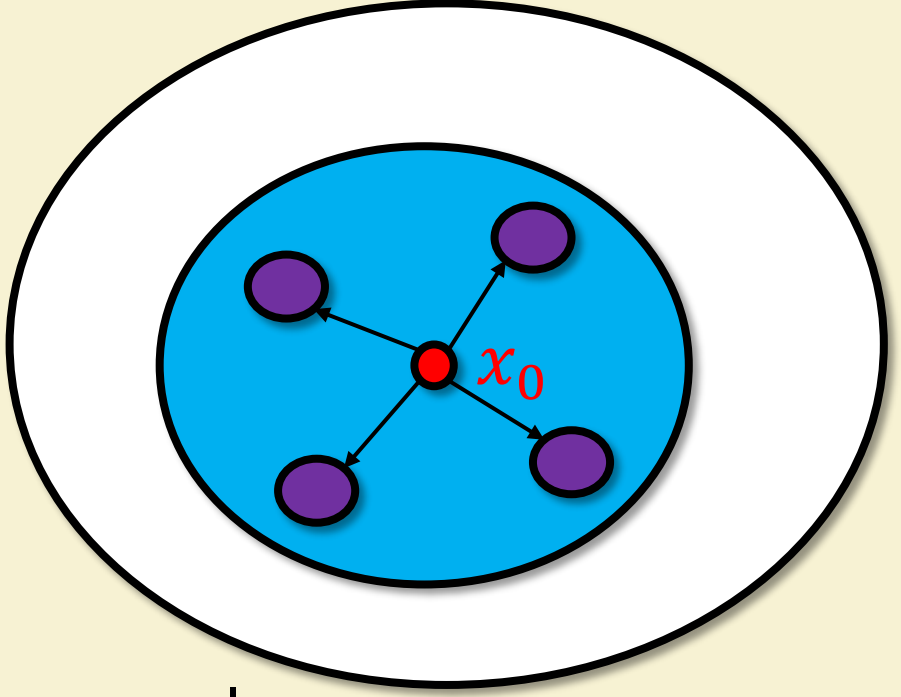
Support of distribution





Replicas

Suppose p_w looks like "ball" around x_0



Pick $x_1, \ldots, x_n$ i.i.d from p_w

High dimension $\Rightarrow x_i - x_0$, $x_j - x_0$ roughly orthogonal

Overlap matrix $$Q = \begin{pmatrix} \langle x_1, x_1 \rangle & \cdots & \langle x_1, x_n \rangle \\ \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix} \approx \begin{pmatrix} 1 & \cdots & q \\ \vdots & \ddots & \vdots \\ q & \cdots & 1 \end{pmatrix}$$

Replica method

Setup: w comes from prob distribution W

Goal: Compute $\mathbb{E}_w A(w) = \mathbb{E}_w \log \int \exp(\langle w, \hat{x} \rangle)$

Easy: Compute $\log \mathbb{E}_w \int \exp(\langle w, \hat{x} \rangle) = \log \int \mathbb{E}_w \exp(\langle w, \hat{x} \rangle)$

$$\mathbb{E}_w A(w) = \lim_{n \to 0} \frac{\mathbb{E}_w \exp(n \cdot A(w)) - 1}{n} = \lim_{n \to 0} \frac{\mathbb{E}_w \left(\int_x \exp(\langle w, \hat{x} \rangle) \right)^n - 1}{n}$$

Take limit of integer to 0!

$$\approx \lim_{n \to 0} \frac{\mathbb{E}_w \int_{x_1, \ldots, x_n} \exp(\langle w, \hat{x}_1 + \cdots + \hat{x}_n \rangle) - 1}{n}$$

replicas

Replica method

Goal: Compute $\mathbb{E}_w A(w) = \mathbb{E}_w \log \int \exp(\langle w, \hat{x} \rangle)$

$$\mathbb{E}_w A(w) \approxeq \lim_{n \to 0} \frac{\mathbb{E}_w \int_{x_1,\dots,x_n} \exp(\langle w, \hat{x}_1 + \cdots + \hat{x}_n \rangle) - 1}{n}$$

$$= \lim_{n \to 0} \frac{\int_{x_1,\dots,x_n} \mathbb{E}_w \exp(\langle w, \hat{x}_1 + \cdots + \hat{x}_n \rangle) - 1}{n}$$

$$Q = \begin{pmatrix} \langle x_1, x_1 \rangle & \cdots & \langle x_1, x_n \rangle \\ \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix}$$

For many natural W,

$\mathbb{E}_w \exp(\langle w, \hat{x}_1 + \cdots + \hat{x}_n \rangle)$ only depends on overlaps of $\{ x_i \}_{i=1..n}$

1) Guess shape of dominant Q (ansatz)

2) Check if makes sense

Examples:

signal

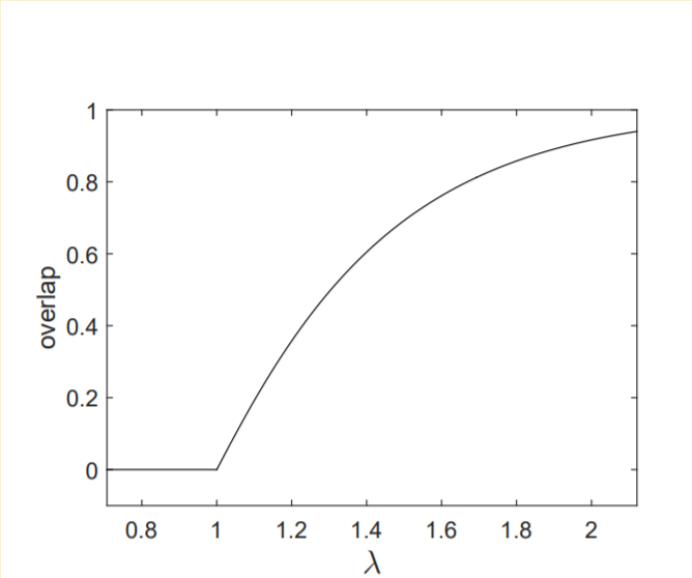
noise

Spiked matrix/tensor: $Y = \lambda S + N$

$$p(S', N'|Y) \propto \exp(\beta \langle Y - \lambda S', N' \rangle^2) \cdot \Pr[S']$$

Overlap with true signal

Want to analyze $\mathbb{E}_{S' \sim p(\cdot|Y)} \langle S, S' \rangle^2$ as function of λ



Statistical limits of spiked tensor models

Amelia Perry*†[1], Alexander S. Wein*‡[1], and Afonso S. Bandeira§[2]

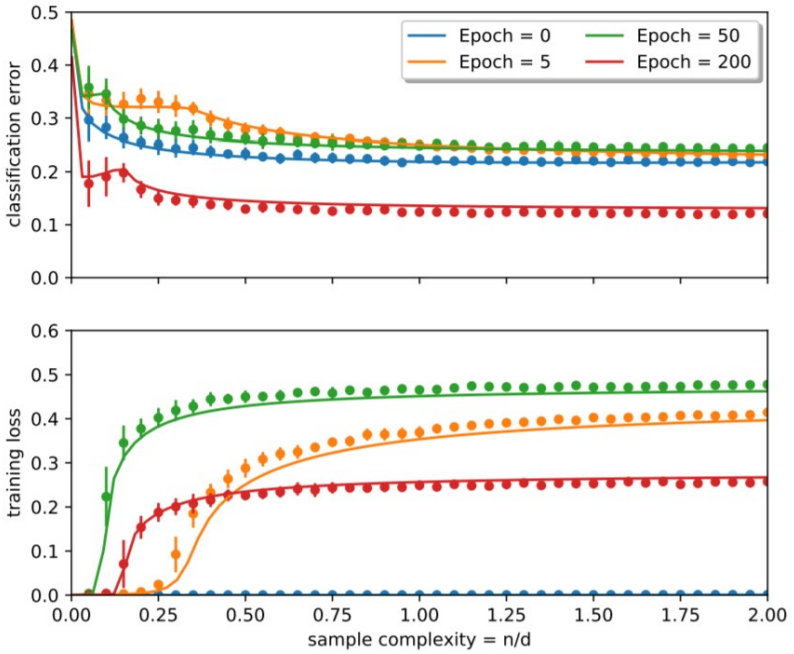
Examples:

signal

noise

Teacher student model: $X, Y = (X, f_S(X) + N)$

$$X, Y = (X, f_S(\tilde{X}) + N)$$

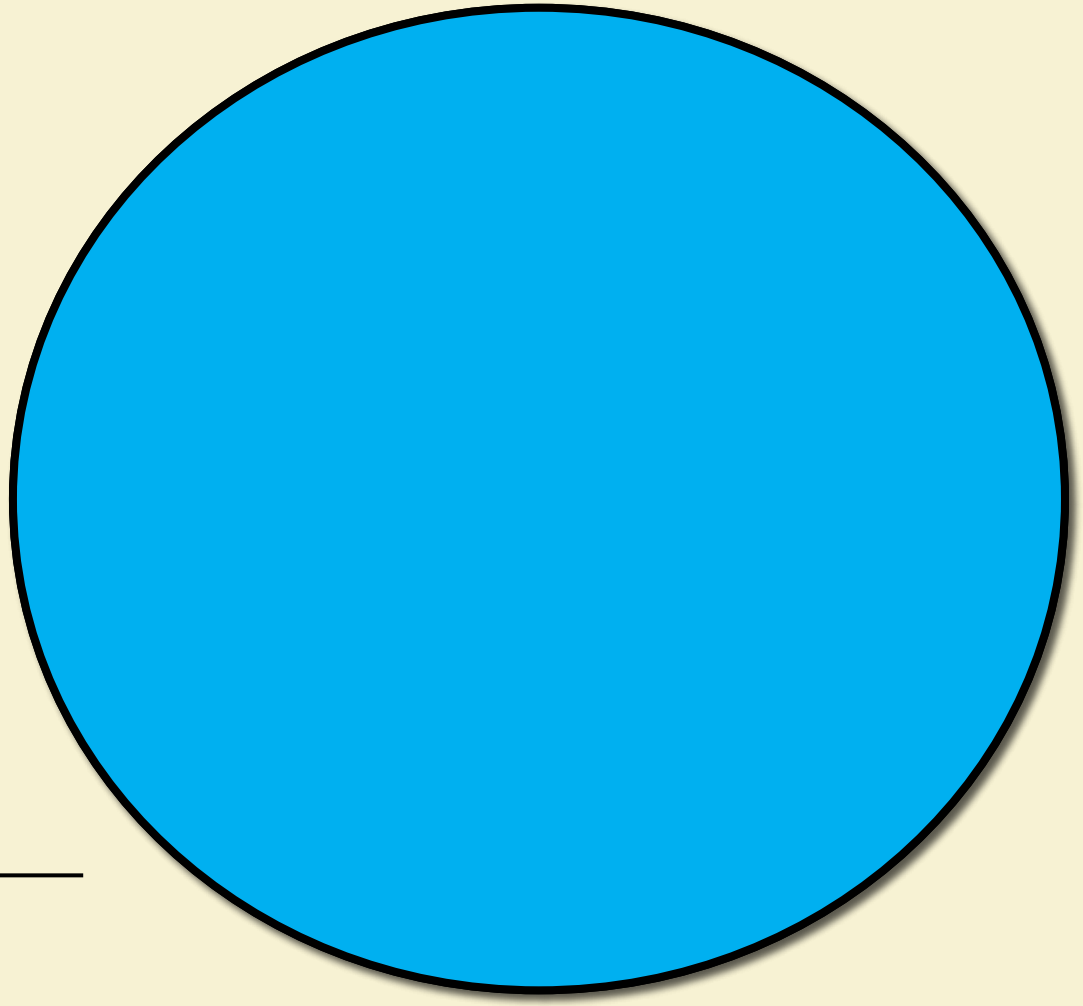


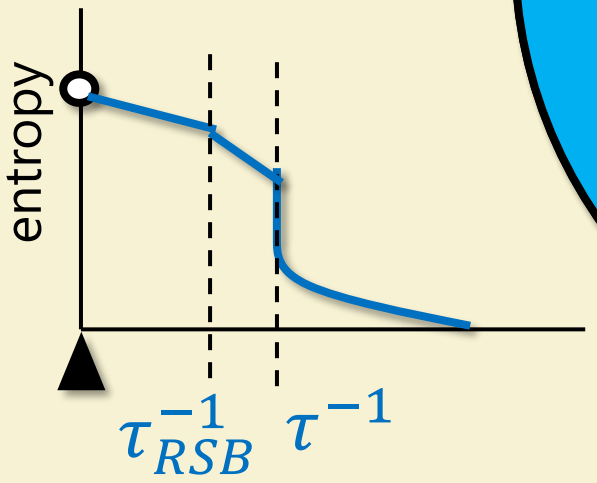
Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model

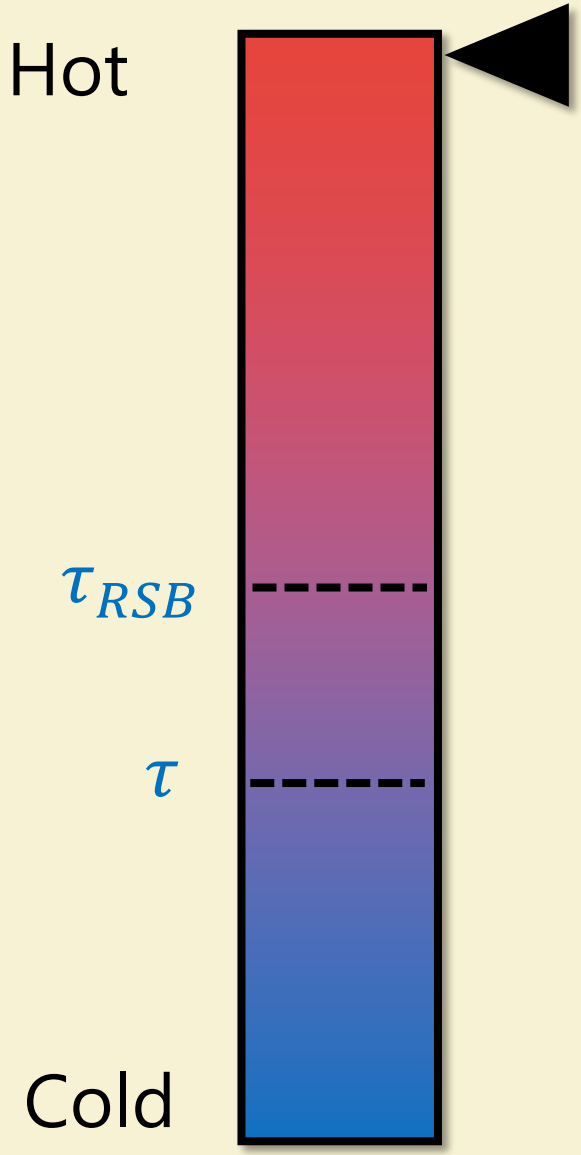
Bruno Loureiro*[1], Cédric Gerbelot†[2], Hugo Cui[3], Sebastian Goldt[4], Florent Krzakala[1], Marc Mézard[2], and Lenka Zdeborová[3]

Replica Symmetry Breaking

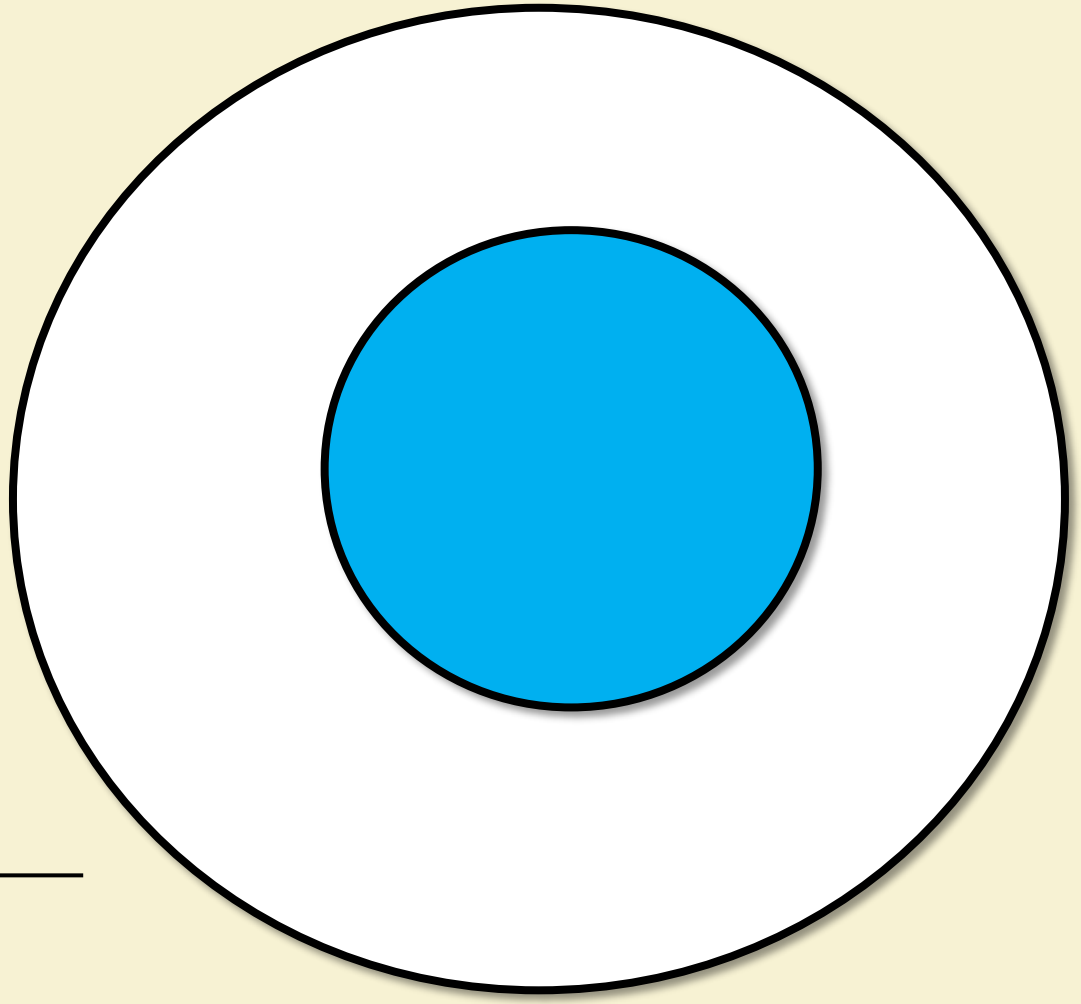
Symmetry breaking

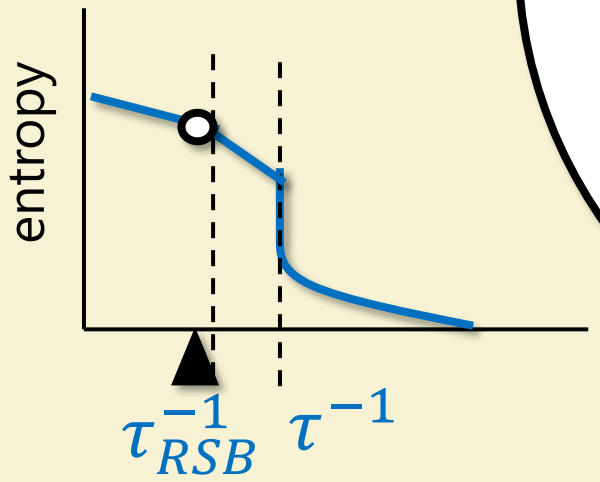


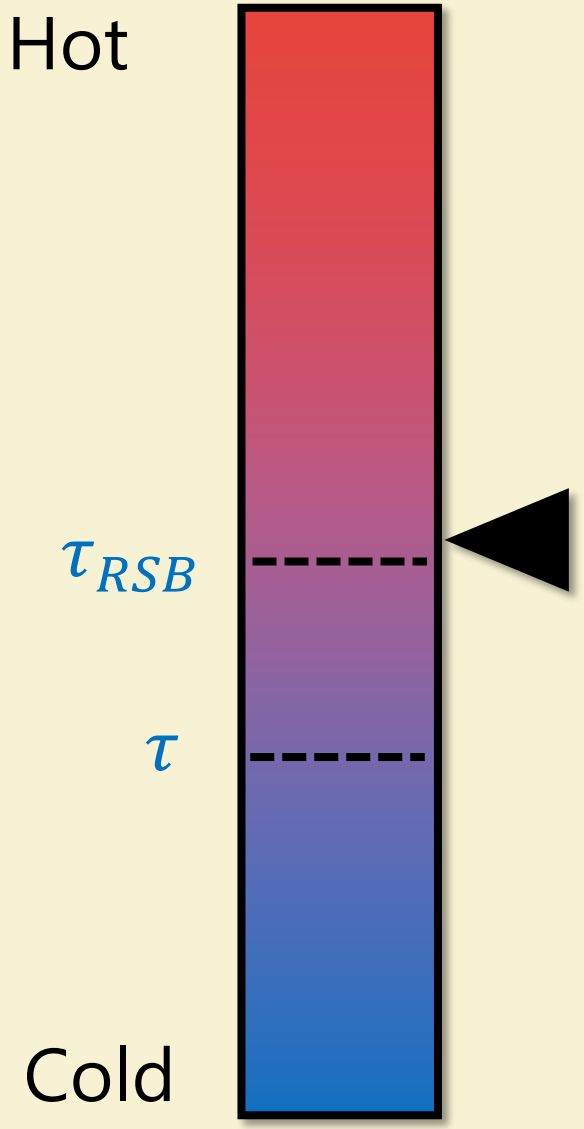




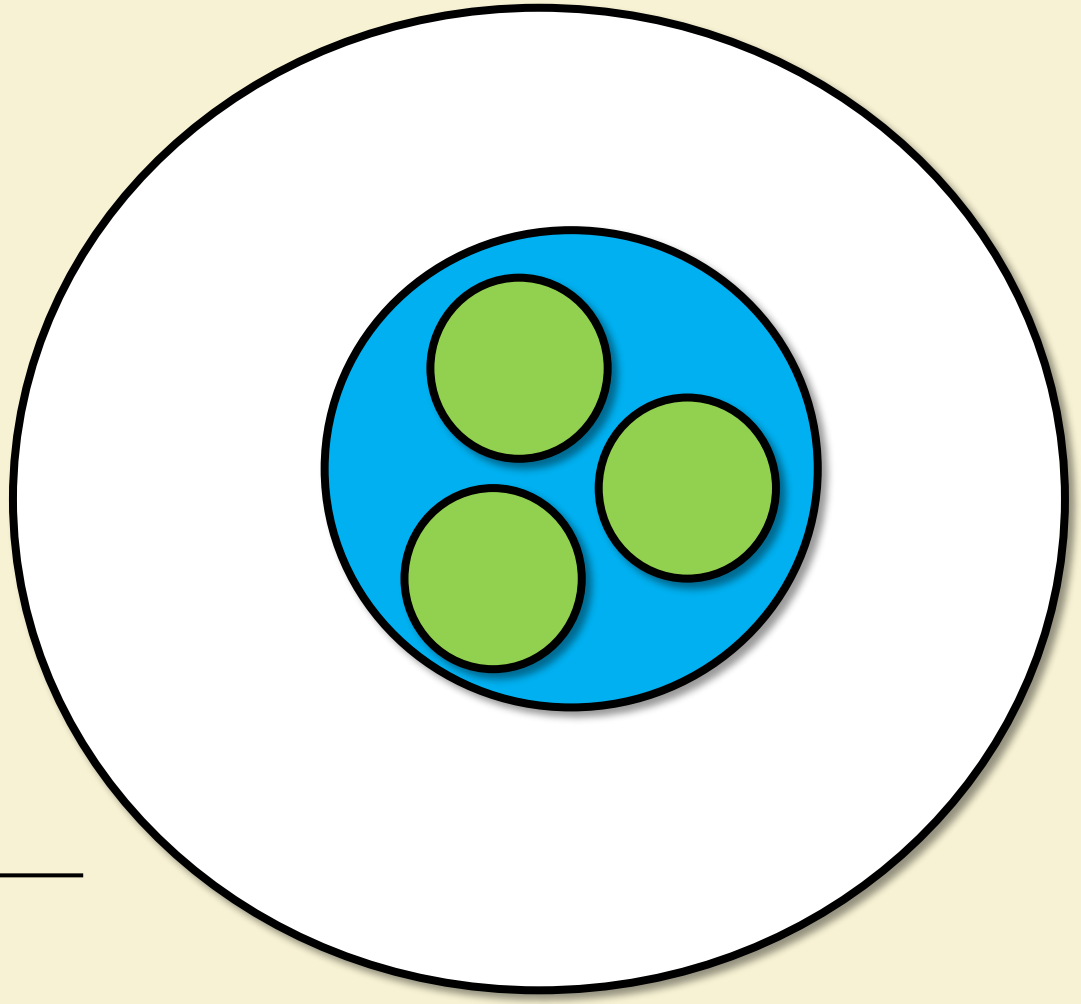
Symmetry breaking

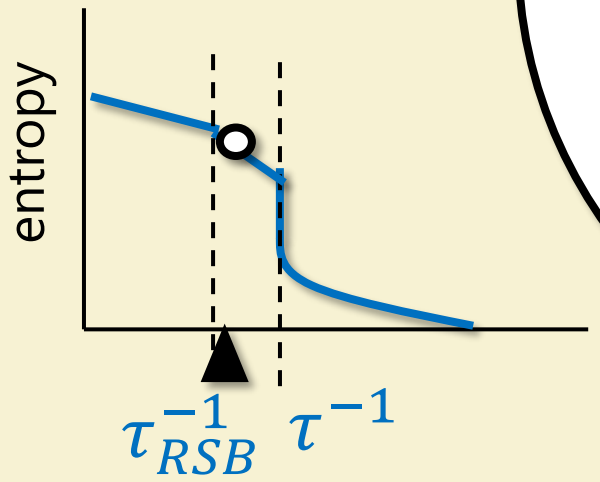


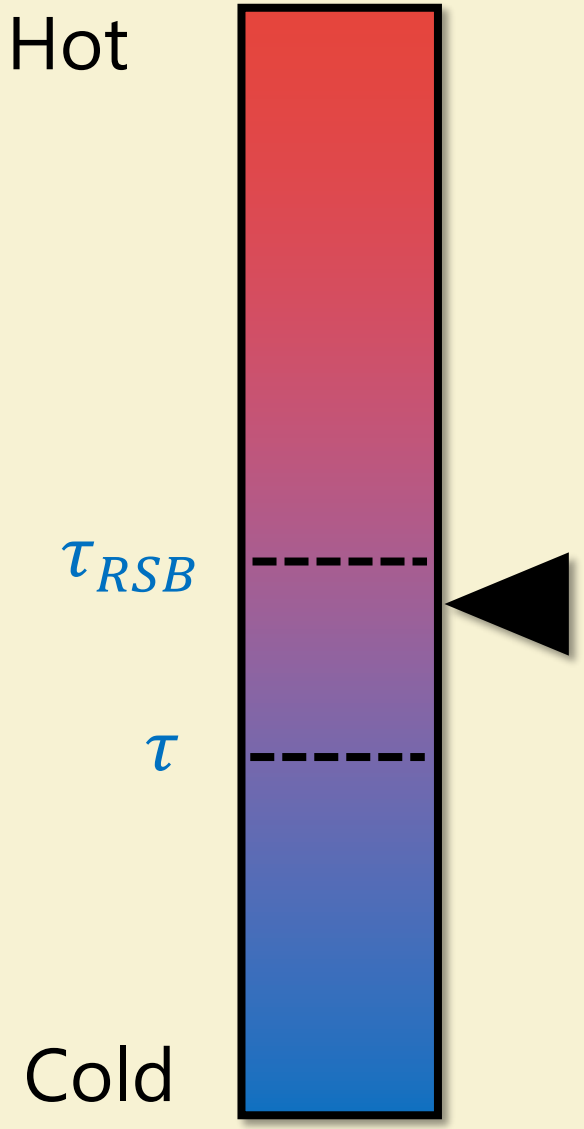




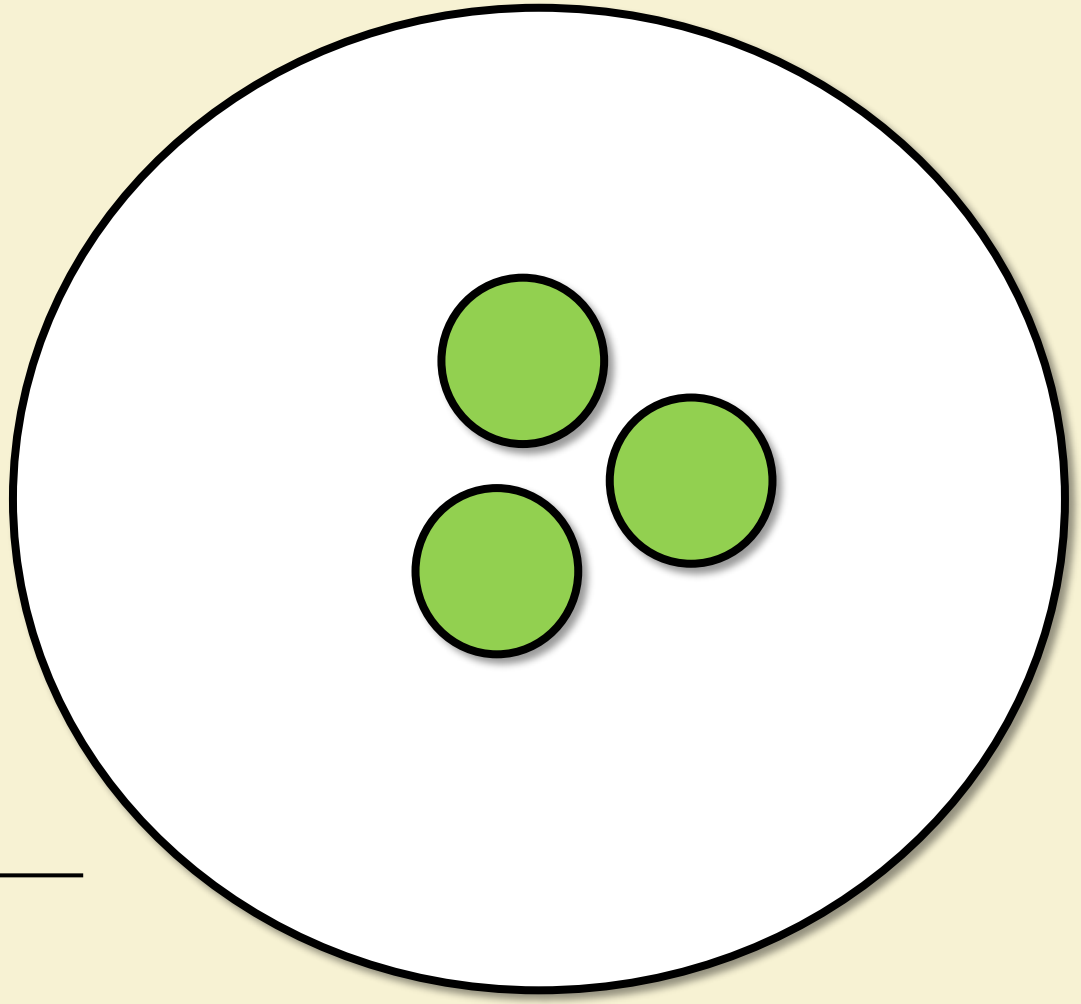
Symmetry breaking

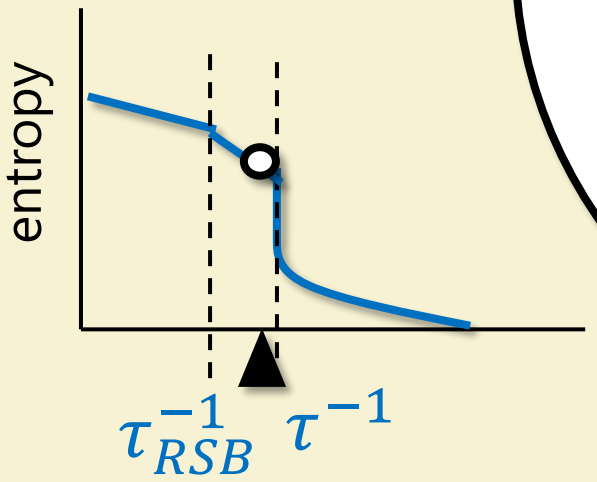


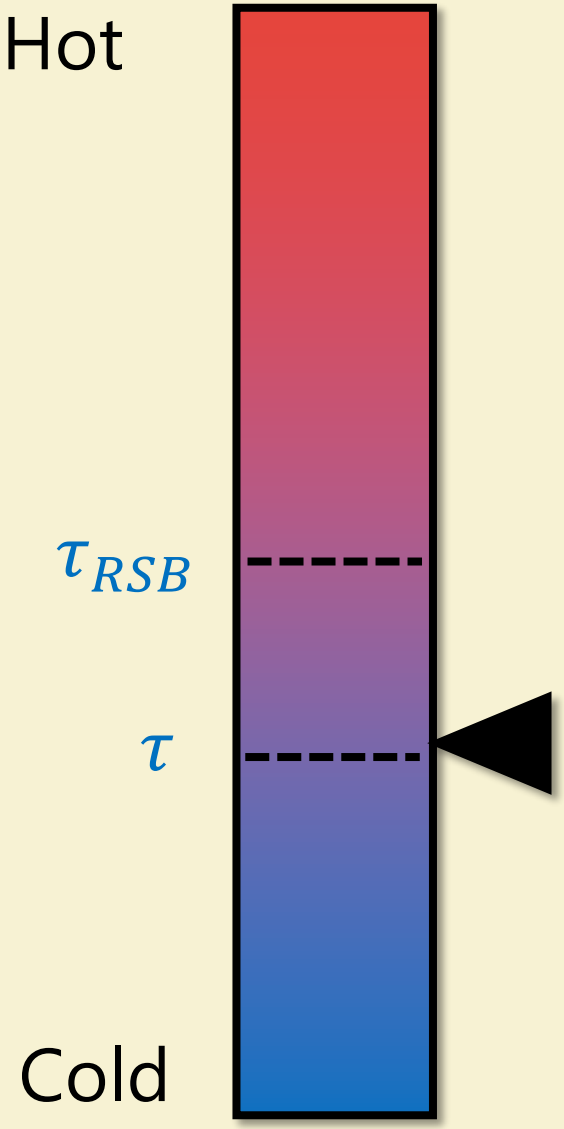




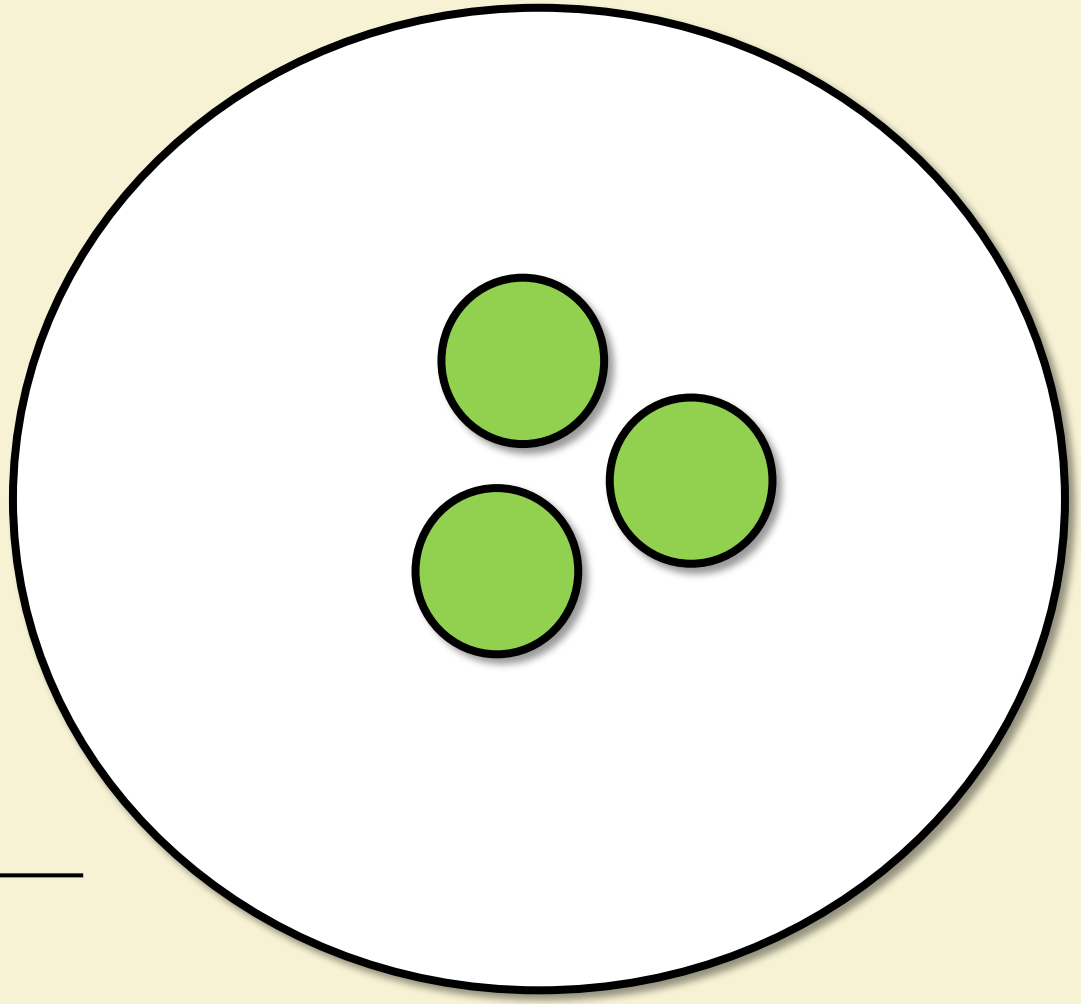
Symmetry breaking

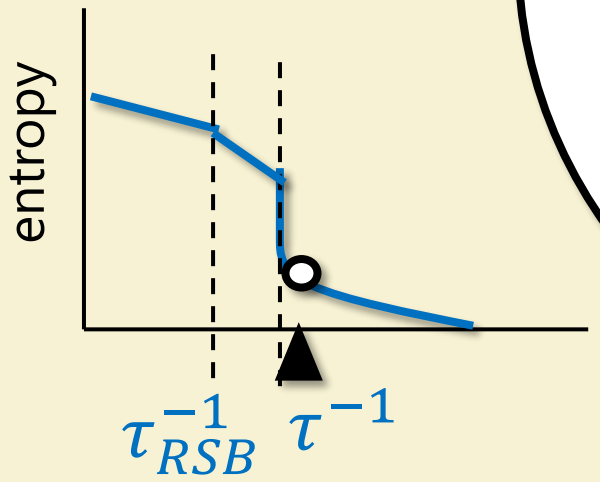


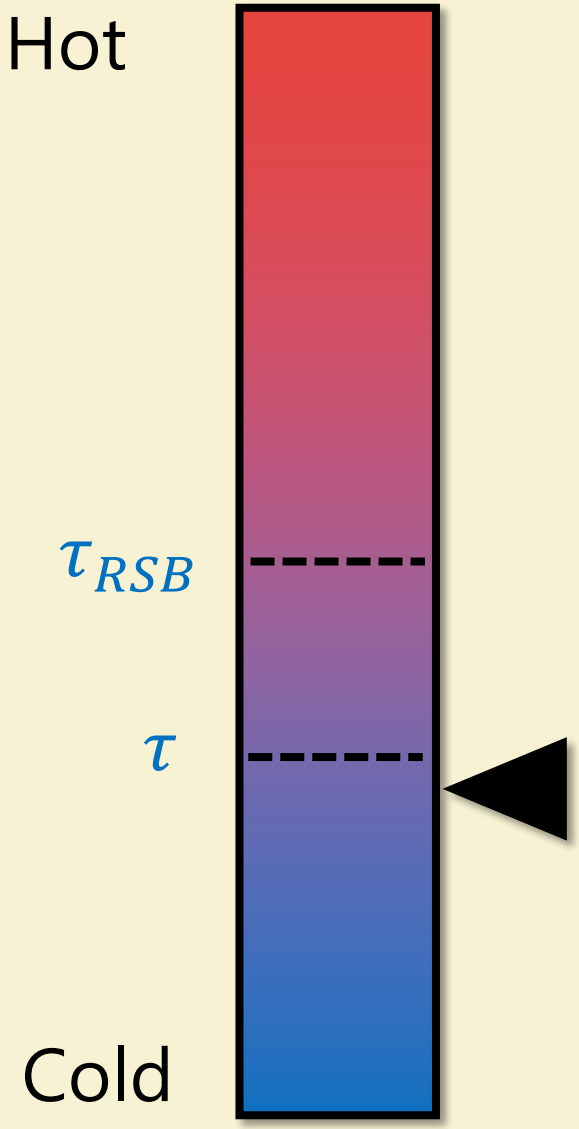




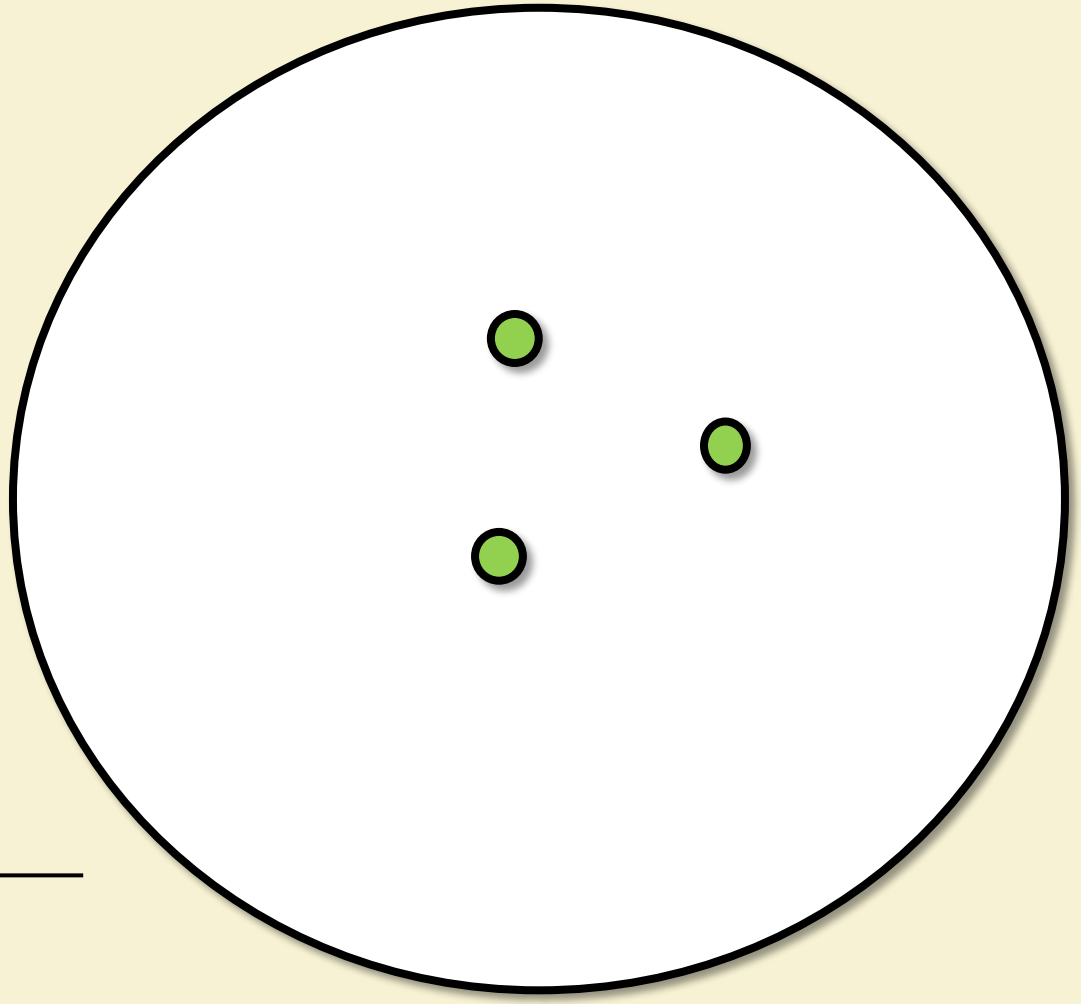
Symmetry breaking

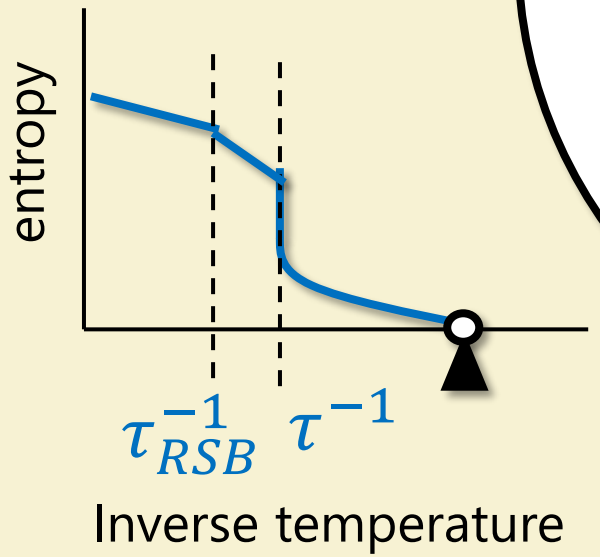


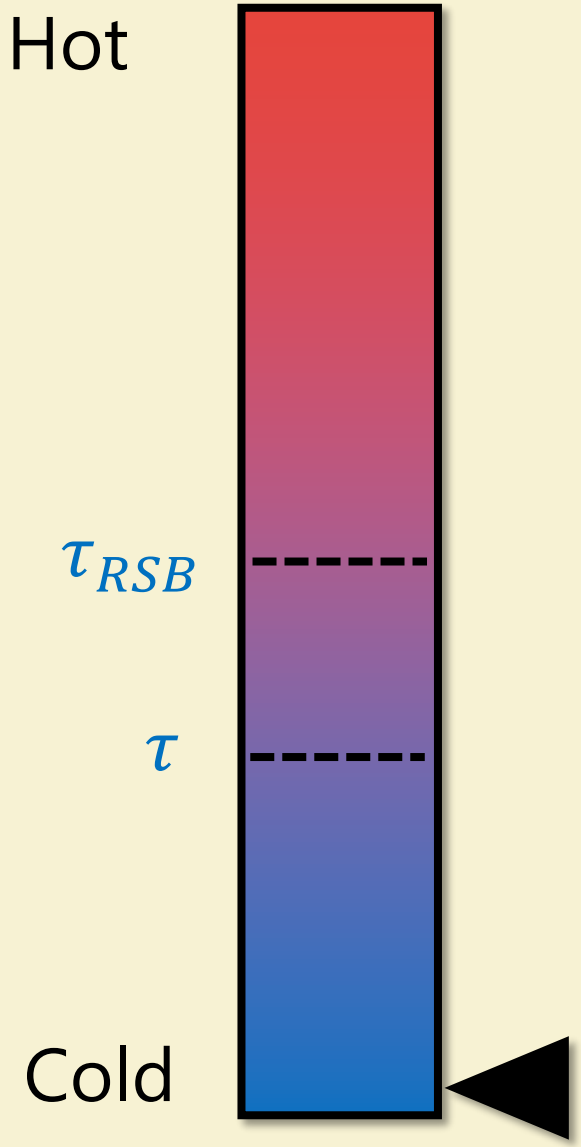




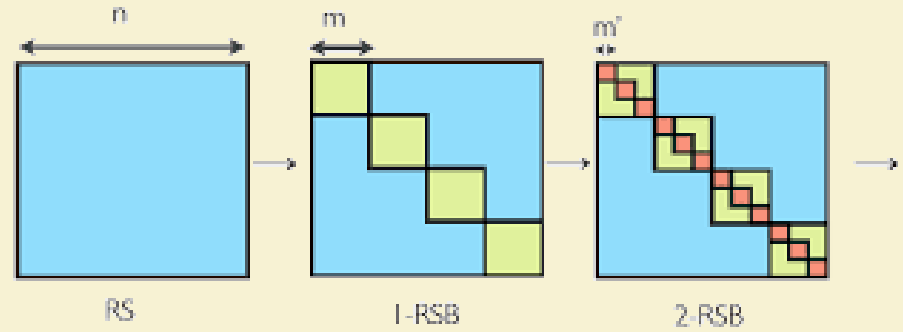
Symmetry breaking

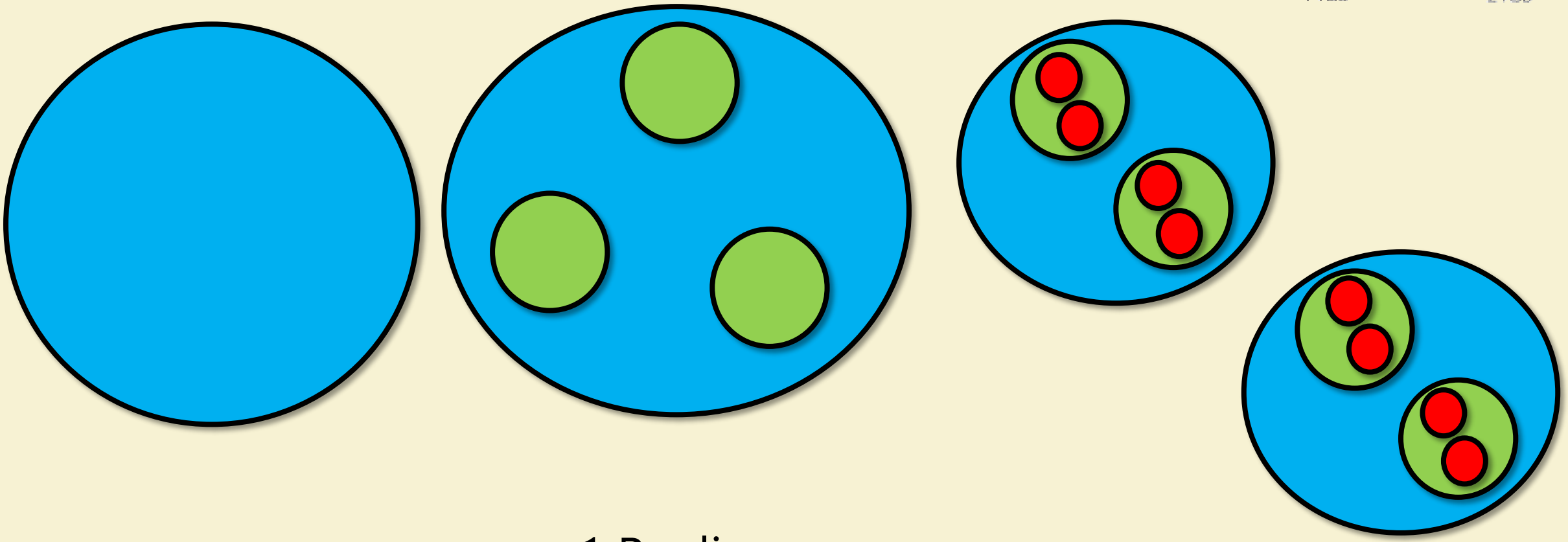






Possible scenarios





Replica Symmetry

1 Replica Symmetry Breaking

2RSB, 3RSB,…,FRSB