Generative Diffusion in High Dimension

Carlo Lucibello
Department of Computing Sciences
Bocconi University, Milan, Italy

Carlo Baldassi

Elizaveta Demyanenko

Davide Straziota

Luca Ambrogioni

Beatrice Achilli

Marc Mézard

Enrico Ventura

Artificial Intelligence Lab
Bocconi

Funded by
the European Union
NextGenerationEU

Generative Diffusion

Sohl-Dickstein et al ’15, Ho et al ’20, Song et al ’21, ….

Forward Process

$\mathbf{X}_0 \sim p_{data}$

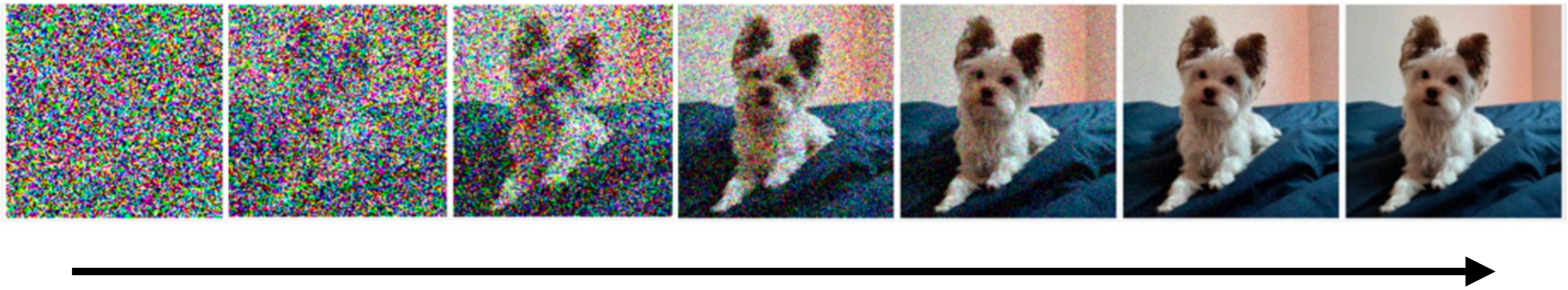
$\mathrm{d}\mathbf{X}_t = f_t(\mathbf{X}_t)\, dt + g(t)\, \mathrm{d}\mathbf{W}_t$

Reverse Process (Generative Process)

$\tilde{\mathbf{X}}_T \sim p_T \qquad T \gg 1$

$\mathrm{d}\tilde{\mathbf{X}}_t = \left(f_t(\tilde{\mathbf{X}}_t) - g^2(t)\, \nabla_x \log p_t(\tilde{\mathbf{X}}_t)\right) dt + g(t)\, \mathrm{d}\tilde{\mathbf{W}}_t$

(back in time)



Theorem **[Anderson ’82]:** under mild assumptions, the two processes have the same density $p_t(\mathbf{x})$.

If we can approximate the score $\nabla_x \log p_t(x)$, we can generate new samples (run discretized reverse)!

Score Functions

Assume for simplicity Variance Exploding process $\mathrm{d}\boldsymbol{X}_t = \mathrm{dW}_t$.

We can consider 3 types of score functions $s_t(\mathbf{x}) = \nabla_x \log p_t(\mathbf{x})$

1. True score function (needs infinite data, typically inaccessible)

$$s_t^{true}(\mathbf{x}) = \nabla_x \log p_t^{true}(\mathbf{x}) = \nabla_x \log \int p_{data}(d\boldsymbol{\xi})\ e^{-\frac{1}{2t}\|\mathbf{x}-\boldsymbol{\xi}\|^2}$$

2. Empirical score function (gives memorization)

$$s_t^{emp}(\mathbf{x}) = \nabla_x \log p_t^{emp}(\mathbf{x}) = \nabla_x \log \sum_{\mu=1}^{P} e^{-\frac{1}{2t}\|\mathbf{x}-\boldsymbol{\xi}^\mu\|^2} \qquad \boldsymbol{\xi}^\mu \sim p_{data}$$

[Ambrogioni ’23]

3. NN approximation (trained by denoising score matching objective [Vincent ’11])

$s_t^{nn}(\mathbf{x}) = NN_\theta(\mathbf{x}, t)$ trained on $\mathscr{D} = \{\boldsymbol{\xi}^\mu\}_\mu$

Sohl-Dickstein et al ’15, Ho et al ’20, Song et al ’21, ….

Empirical Score <-> Associative Memory

- Empirical time-dependent log-density for diffusion:

$$\log p_t^{emp}(\mathbf{x}) = \log \sum_{\mu=1}^{P} e^{-\frac{1}{2t}\|\mathbf{x}-\boldsymbol{\xi}^{\mu}\|^2} + const$$

- Energy of Modern Hopfield Network

[Ramsauer et al '20 "Hopfield is All You Need"] [**CL**, Mézard PRL '24]

$$E(\mathbf{x}) = -\frac{1}{\lambda}\log\left(\sum_{\mu=1}^{P} e^{\lambda\,\mathbf{x}\cdot\boldsymbol{\xi}^{\mu}}\right) + \frac{1}{2}\|\boldsymbol{x}\|^2$$

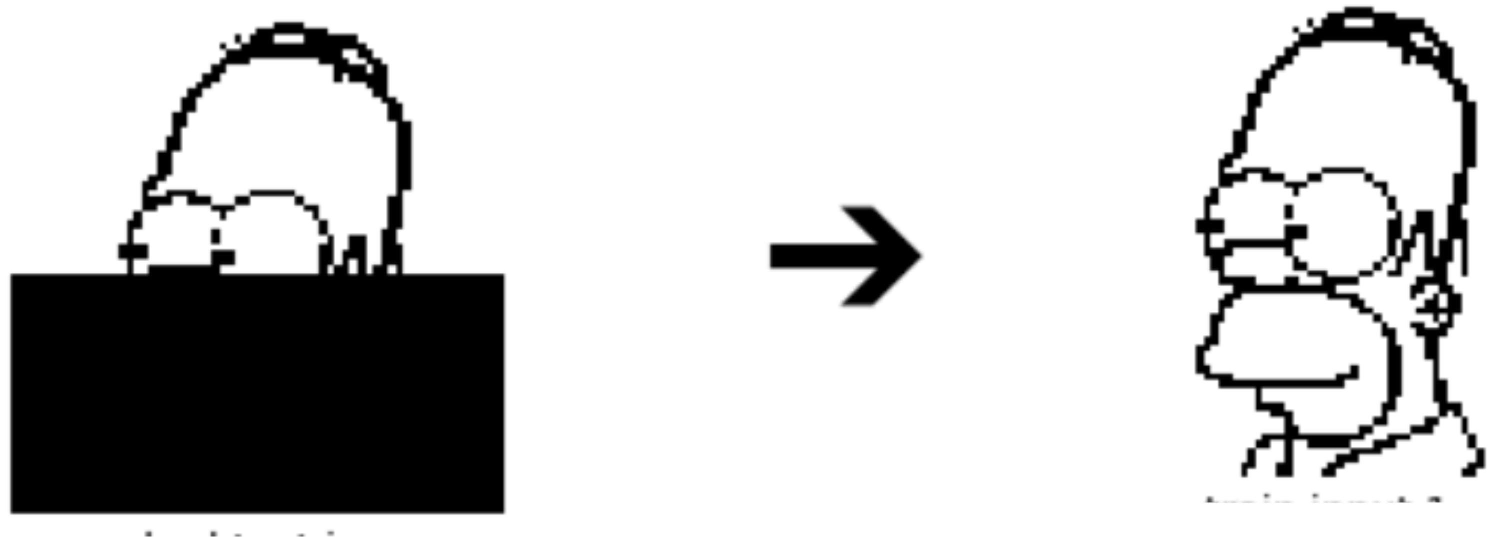
Hopfield Model

Hopfield, PNAS '82

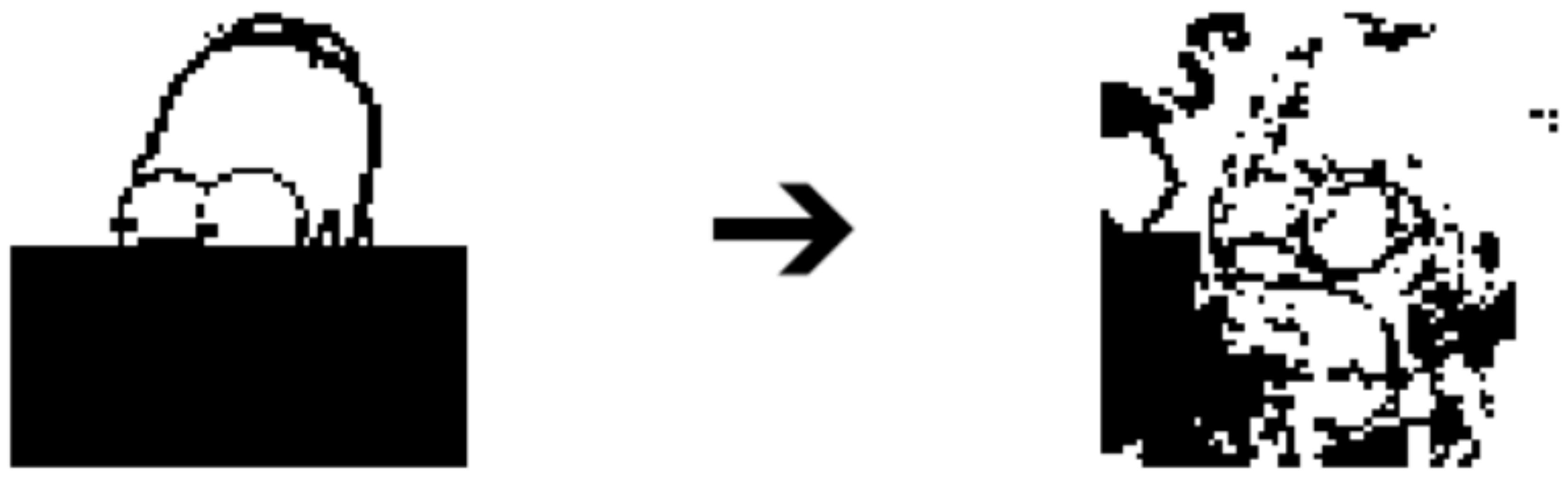
Ising spins $\sigma \in \{-1, +1\}^N$, energy $E(\sigma) = -\sum_{i,j} \sigma_i J_{ij} \sigma_j$ with $J_{ij} = \frac{1}{P} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu$

[image credit Johannes Brandstetter]

Retrieval



No Retrieval



For i.i.d.
$\xi^\mu \sim$ Unif($\{-1, +1\}^N$)
critical capacity is $P_c \approx 0.14N$

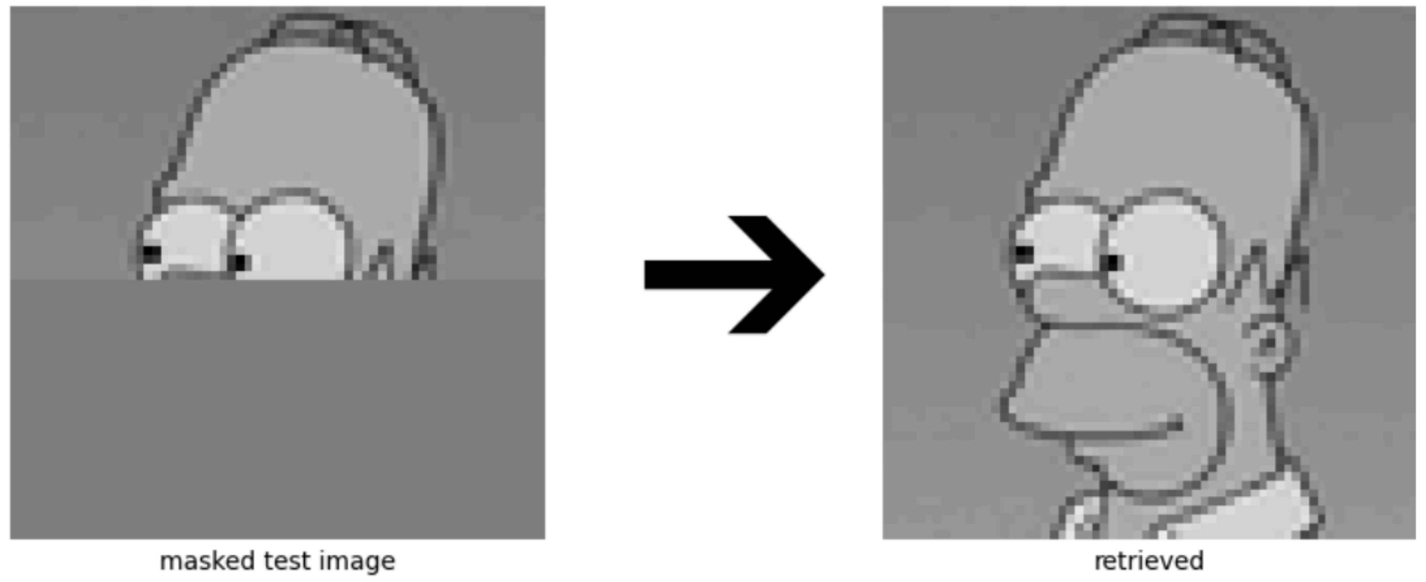
[Amit,Gutfreund,Sompolinsky '85]

Modern Hopfield Model

[Ramsauer et al ’20 “Hopfield is All You Need”]

$$E(\mathbf{x}) = -\frac{1}{\lambda}\log\left(\sum_{\mu=1}^{P} e^{\lambda\,\boldsymbol{\xi}^{\mu}\cdot\mathbf{x}}\right) + \frac{1}{2}\|\boldsymbol{x}\|^2$$

[image credit Johannes Brandstetter]



Exponential capacity!

A simple Energy Decomposition

CL, Mézard PRL '24

- We assume $\mathbf{x} \in \mathbb{R}^N$ and $P = e^{\alpha N}$ patterns i.i.d. from p_{data} (e.g. Gaussian, spherical, or from Hidden Manifold Model $\xi^\mu = \sigma(Fz^\mu)$ with intrinsic dimension D_{hidden}).
- Energy: $$E(\mathbf{x}) = -\frac{1}{\lambda}\log\left(\sum_{\mu=1}^{P} e^{\lambda\,\boldsymbol{\xi}^\mu\cdot\mathbf{x}}\right) + \frac{1}{2}\|\boldsymbol{x}\|^2$$
- Identify a signal term and a noise term: $$-\frac{1}{\lambda}\log\left(e^{\lambda\,\boldsymbol{\xi}^1\cdot\mathbf{x}} + \sum_{\mu=2}^{P} e^{\lambda\,\boldsymbol{\xi}^\mu\cdot\mathbf{x}}\right)$$
- Since the exponents are $O(N)$, for large N we can write $$E(\mathbf{x}) \approx -\max\left(\boldsymbol{\xi}^1\cdot\mathbf{x},\, \Phi(\mathbf{x})\right) + \frac{1}{2}\|\boldsymbol{x}\|^2 \qquad \text{with} \qquad \Phi(\mathbf{x}) = \frac{1}{\lambda}\log\left(\sum_{\mu=2}^{P} e^{\lambda\,\boldsymbol{\xi}^\mu\cdot\mathbf{x}}\right)$$
- If $\boldsymbol{\xi}^1$ wins the competition we have **retrieval**, since the energy becomes a quadratic form with minimum in the pattern (reached in 1 GD step).
- The noise function $\Phi(\mathbf{x})$ takes the form of the free energy of a **Random Energy Model [Derrida '81].** In fact, conditioned on (quenched) $\mathbf{x}$, we have i.i.d. energies $\epsilon^\mu = -\boldsymbol{\xi}^\mu\cdot\mathbf{x}$.

Single Pattern Retrieval Threshold

$$E(\mathbf{x}) \approx -\max\left(\boldsymbol{\xi}^1 \cdot \mathbf{x},\ \Phi\left(\mathbf{x}\right)\right) + \frac{1}{2}\|\boldsymbol{x}\|^2 \qquad \text{with} \qquad \Phi(\mathbf{x}) = \frac{1}{\lambda}\log\left(\sum_{\mu=2}^{P} e^{\lambda\,\boldsymbol{\xi}^\mu \cdot \mathbf{x}}\right)$$

Computing the energy in $\mathbf{x} = \boldsymbol{\xi}^1$, we have a simple criterium for retrieval:

$$\|\boldsymbol{\xi}^1\|^2 \ > \ \Phi\left(\boldsymbol{\xi}^1\right)$$

Condition for Retrieval

Consider $P = e^{\alpha N}$, $\mathbb{E}\|\boldsymbol{\xi}^1\|^2 = N$, and high-dimensional limit $N \to \infty$.
We can compute the REM-like noise contribution:

$$\phi_\alpha(\lambda) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Phi(\boldsymbol{\xi}^1) = \begin{cases} \dfrac{\alpha + \zeta(\lambda)}{\lambda} & \text{if } \lambda < \lambda_*(\alpha) \\ \varepsilon_*(\alpha) & \text{if } \lambda \geq \lambda_*(\alpha) \end{cases}$$

The **asymptotic threshold for single pattern retrieval** $\alpha_1(\lambda)$ is the solution of :

$$1 = \phi_{\alpha_1}(\lambda)$$

A randomly chosen pattern can be retrieved with high probability if $\alpha < \alpha_1(\lambda)$. Basins of attraction are extensive (and can compute radius). Also have bounds on all patterns retrieval threshold.

Phase Diagram

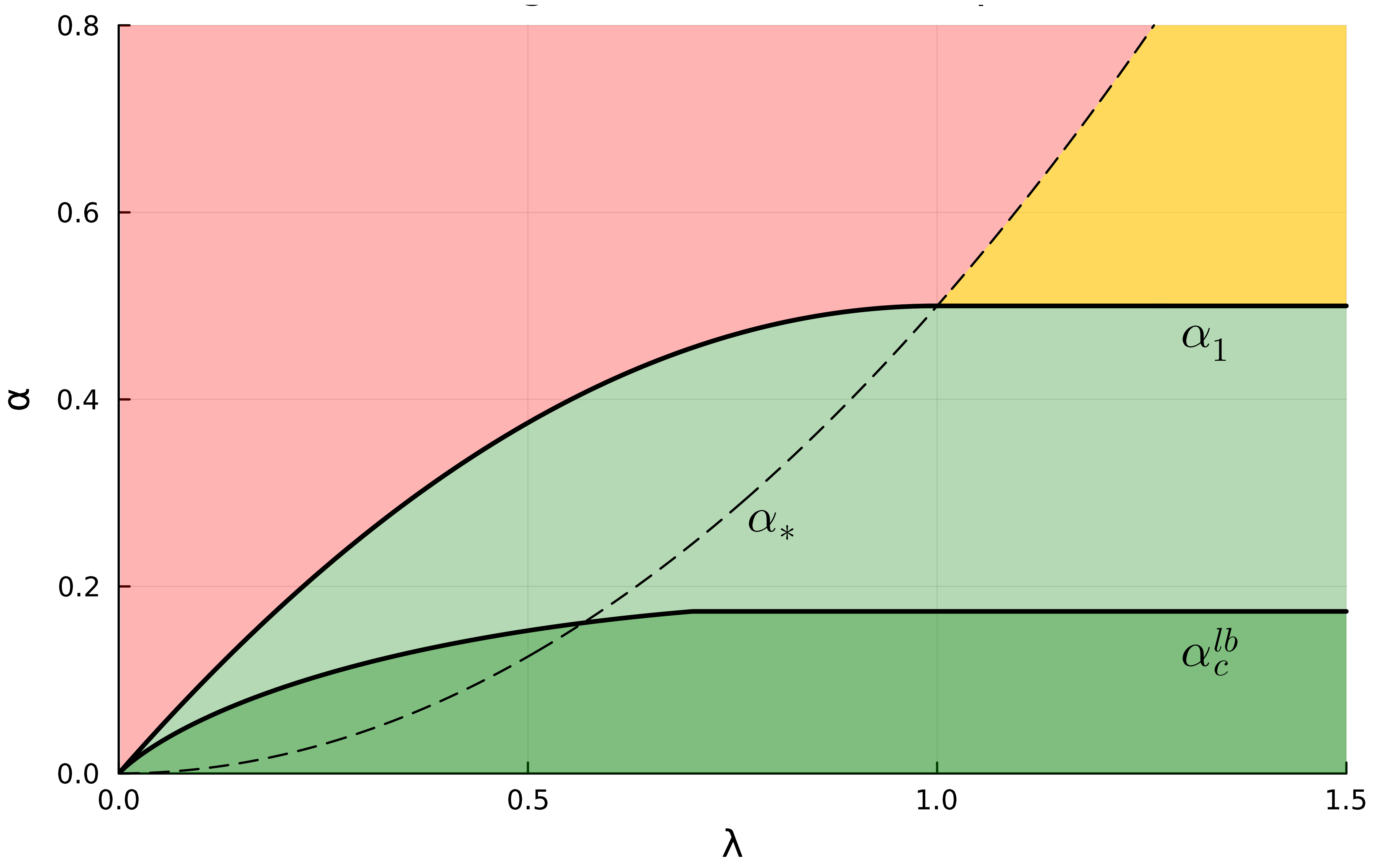
- **Single Pattern Retrieval.** Most memories correspond to minima of the energy.

- **All Patterns Retrieval.** All memories are minima of the energy.

- **Uncondensed phase.** No retrieval due to contributions from exponentially many other memories in the REM.

- **Condensed phase.** No retrieval due to sub-exponential number of other memories.

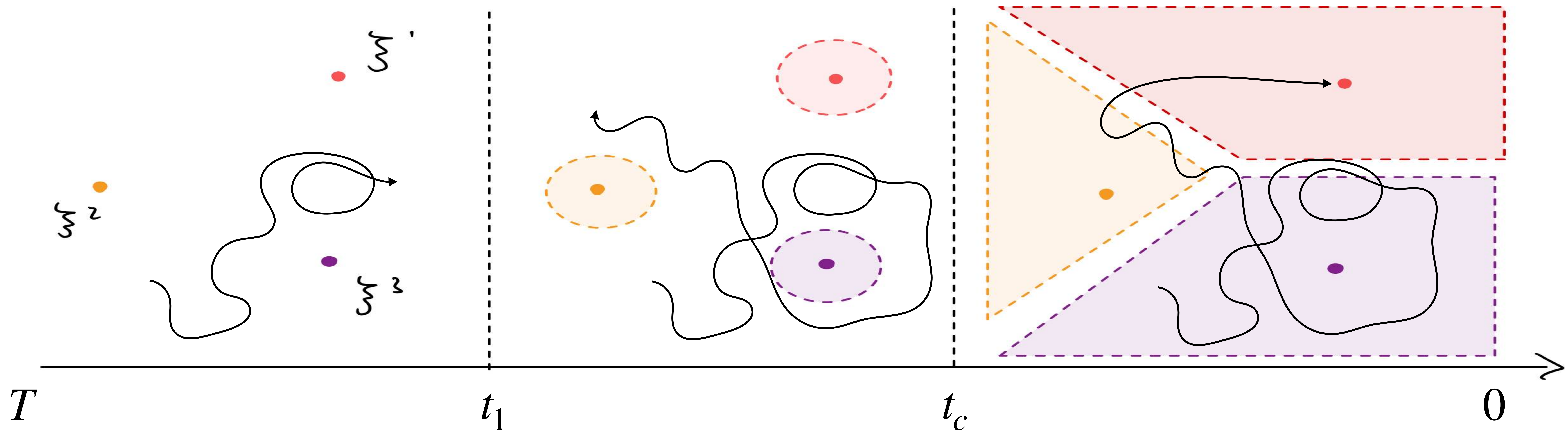
Gaussian Memories



[**Lucibello**, Mézard PRL'24]

Back To Diffusion with Empirical Score

Reverse Process Through Empirical Score



- Diffusion + drift to data manifold.
- $D_{KL}[p_t^{emp} | p_t^{true}] \approx 0$.
- The diffusive process is not aware of the finiteness of the dataset.

- Same as before
  +
- Traps appear in the dynamical landscape.
- Traps have no influence on typical trajectories.

- **Trajectories fall into memories.**
- $D_{KL}[p_t^{emp}, p_t^{true}] \gg 0$.
- Collapse time = REM Condensation time (due to BO)
- $t_c = O\left(e^{-\frac{\log P}{2D_{hidden}}}\right)$
- Need exponentially many datapoints for small t_c, mitigated by low data-manifold dimension.
- Minimum of $D_{KL}[p_t^{emp} | p_{data}]$ at $t = t_g < t_c$.

Analysis of diffusion with true score function

Stochastic Localization

- Target distribution on $\mathbb{R}^N$ we want to sample from:

$$p(w) = \frac{1}{Z}\psi(w); \qquad Z = \int dw\, \psi(w)$$

partition function, possibly disordered and hard to compute

- Consider the process (called Stochastic Localization [Eldan '13])

$$h_0 = 0$$
$$\mathrm{d}h_t = m_t(h_t)\mathrm{d}t + \mathrm{d}B_t$$

$$m_t(h) = \mathbb{E}_{p_{t,h}}[w]$$

$$p_{t,h}(w) \propto p(w)\, e^{h\cdot w - \frac{t}{2}\|w\|^2}$$

time-varying distribution

$\longrightarrow$ $p_{t,h} \xrightarrow[t\to\infty]{} \delta_{w^\star}$ with $w^\star \sim p$

- Bayesian structure [Montanari, El Alaoui '22][Montanari '23]:

$$h_t \sim t w^\star + \sqrt{t} g, \quad w^\star \sim p, \quad g \sim \mathcal{N}(0, I_N)$$

$$m_t(h_t) = \mathbb{E}[w^\star \mid h_t]$$ Bayesian denoiser

Algorithmic Stochastic Localization

- We use Approximate Message Passing (AMP) to estimate the posterior average, following [Montanari, El Alaoui '22] [Montanari, El Alaoui, Selke '23]
- AMP is an iterative algorithm that at the fixed point (provided it converges and converges to the correct FP) gives the marginals / magnetizations of the system.
- So our **ASL scheme** to generate a sample is:

★ Discretize in time the Stochastic Localization SDE for the field h_t.

★ At each discrete time, run AMP until convergence and obtain the drift $m_t(h_t)$.

★ Integrate the SDE up to some large time T and return a sample as $w = m_T(h_T)$.

- For the perceptron problems we will consider, the form of AMP is known as GAMP. It is conjecturally optimal among polynomial algorithms for this denoising task [Barbier et al' PNAS '19].

Asymptotic Analysis

Ricci-Tersenghi, Guilhem Semerjian, JSTAT ‘09
Ghio, Dandi, Krzakala, Zdeborová, PNAS ‘24
Straziota, Demyanenko, Baldassi, **CL,** arxiv ‘25

- The asymptotic (large N) performance of ASL can be characterized through a free-entropy:

$$\phi_t = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E}_{\psi,g} \int \frac{\psi\left(\mathrm{d}\boldsymbol{w}^\star\right)}{Z} \log \int \psi\left(\mathrm{d}\boldsymbol{w}\right) e^{\overbrace{(t\boldsymbol{w}^\star+\sqrt{t}\mathbf{g})}^{\boldsymbol{h}_t}\cdot\boldsymbol{w}-\frac{t}{2}\|\boldsymbol{w}\|^2}$$

$$= \lim_{N\to+\infty} \frac{1}{N} \lim_{s\to 0} \lim_{n\to 0} \partial_n \mathbb{E}_{\psi,g} \int \prod_{\alpha=1}^{s} \psi\left(\mathrm{d}\boldsymbol{w}_\alpha^\star\right) \prod_{a=1}^{n} \psi\left(\mathrm{d}\boldsymbol{w}_a\right) e^{(t\mathbf{w}_1^\star+\sqrt{t}\mathbf{g})\cdot\boldsymbol{w}_a-\frac{t}{2}\|\boldsymbol{w}_a\|^2}$$

$$\lim_{s\to 0} Z^{s-1} = \frac{1}{Z}$$

$$\lim_{n\to 0} \partial_n Z^n = \log Z$$

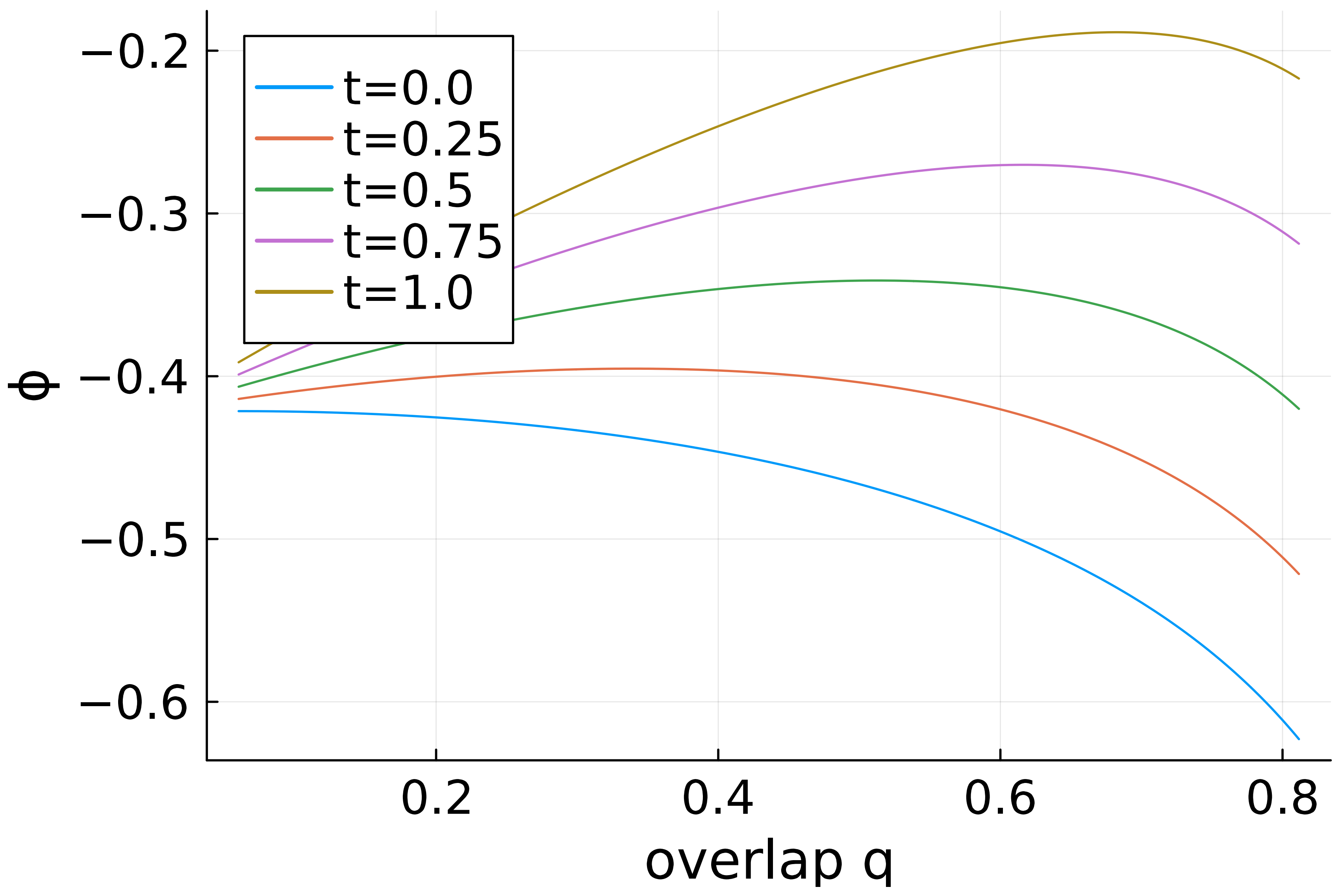
double application of replica trick
(à la [Franz-Parisi ’95])
[Straziota, Demyanenko, Baldassi, **CL** ’25]

- For dense graphical models, the computation reduces to finding a critical point of a function of few scalar parameters (overlaps). Problem simplified by Nishimori conditions.

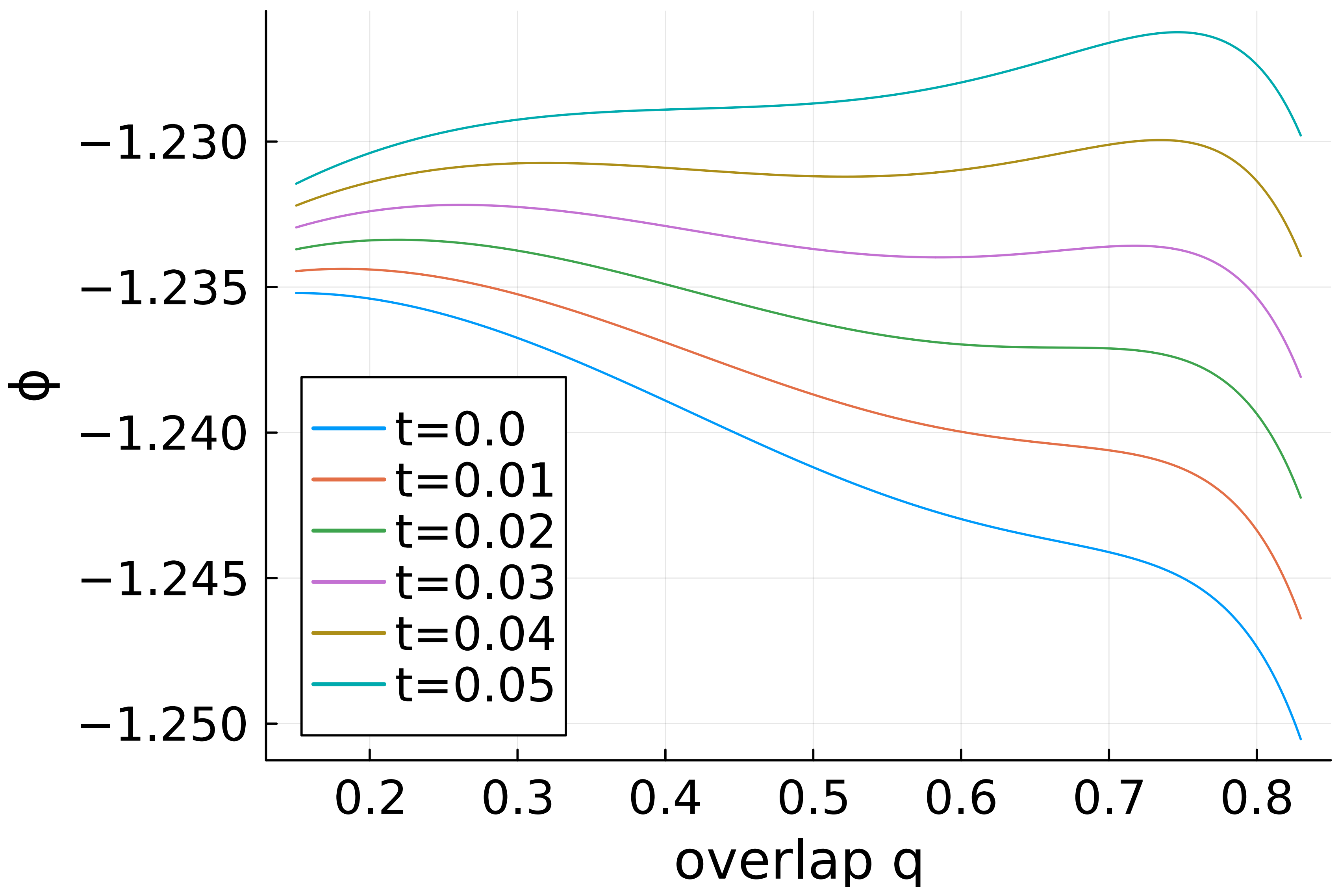
Success and Failure of ASL

Fixed points of AMP are in correspondence with free-entropy maxima.

Success



Failure



$$q = \frac{1}{N}\, w^\star \cdot w$$

Non-Convex Perceptron models

Take M patterns $x^\mu \sim \mathcal{N}(0, I_N)$ and a margin $\kappa \in \mathbb{R}$. The uniform distribution over the solutions of the constraint satisfaction problem is:

$$p(w) \propto P(w) \prod_{\mu=1}^{M} \mathbb{I}(s^\mu \geq k), \qquad s^\mu = \frac{w \cdot x^\mu}{\sqrt{N}}$$

stabilities

with priors:

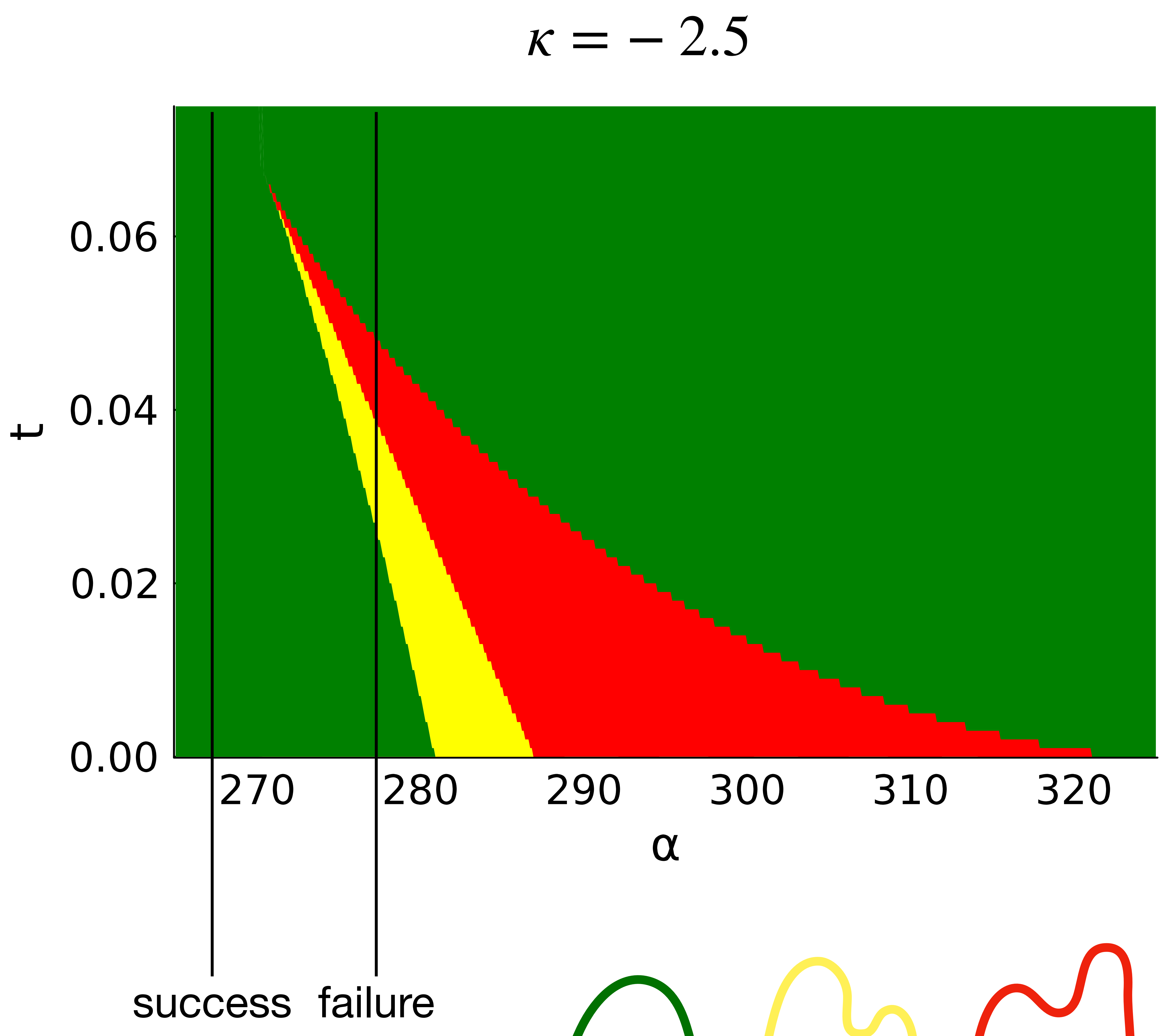
Spherical: $P(w) = \delta(\|w\|^2 - N)$. In this setting we also consider $\kappa < 0$ for non-convexity [Franz,Parisi ’16] [Montanari, Zhong, Zhou ’23].

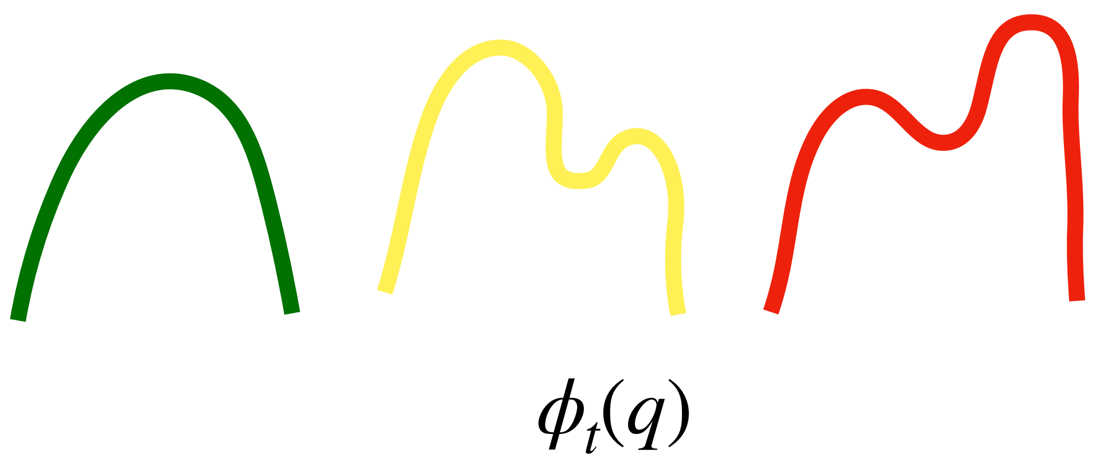
or

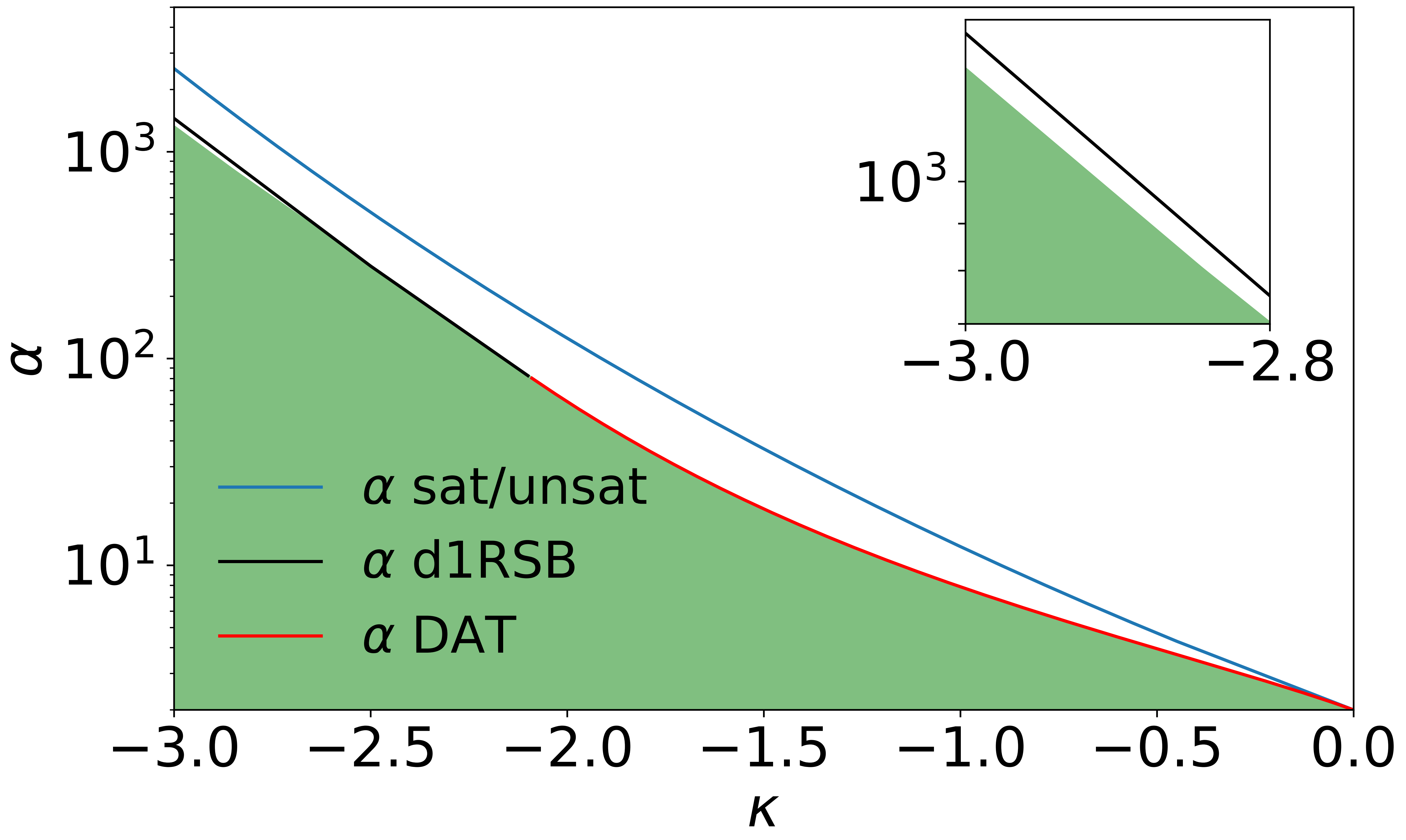
Binary: $P(w) = \prod_{i=1}^{N} \left(\delta(w_i - 1) + \delta(w_i - 1)\right)$. Here we take $\kappa = 0$ for simplicity.

We will take $N, M \to \infty$ at fixed density of constraints $\alpha = \dfrac{M}{N}$.

Spherical Perceptron with negative margin



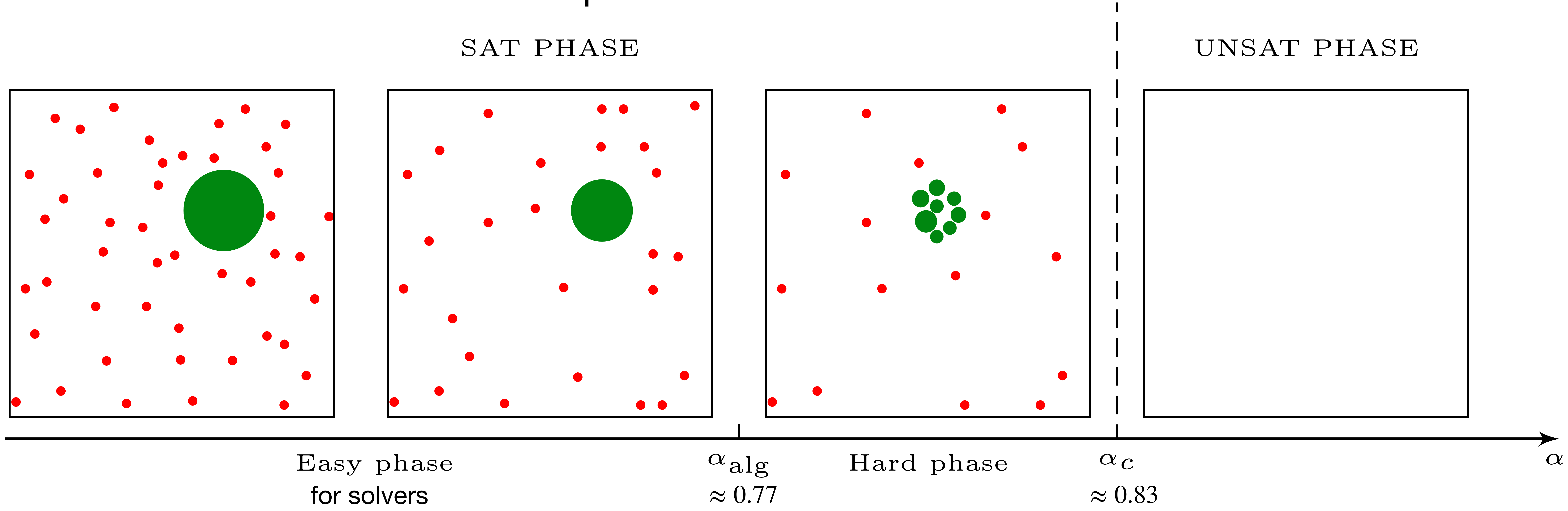




Straziota, Demyanenko, Baldassi, **CL,** arxiv ‘25

Solution Space for Binary Perceptron

- **Sampling from the uniform distribution fails at any $\alpha > 0$.** This is expected since:
  - Most configurations are isolated [Huang, Kabashima, PRE ’14].
  - Hardness due to Overlap Gap Property [Gamarnik, PNAS’21].
- **There exist though an algorithmically accessible dense cluster** [Baldassi et al. PRL ’15, PNAS ’16,…].
- Efficient solvers exist but their output is not well characterized.



Small epsilon analysis and tilted potential

- For the flat measure, there is always a second peak of the free-entropy at $q = 1$.

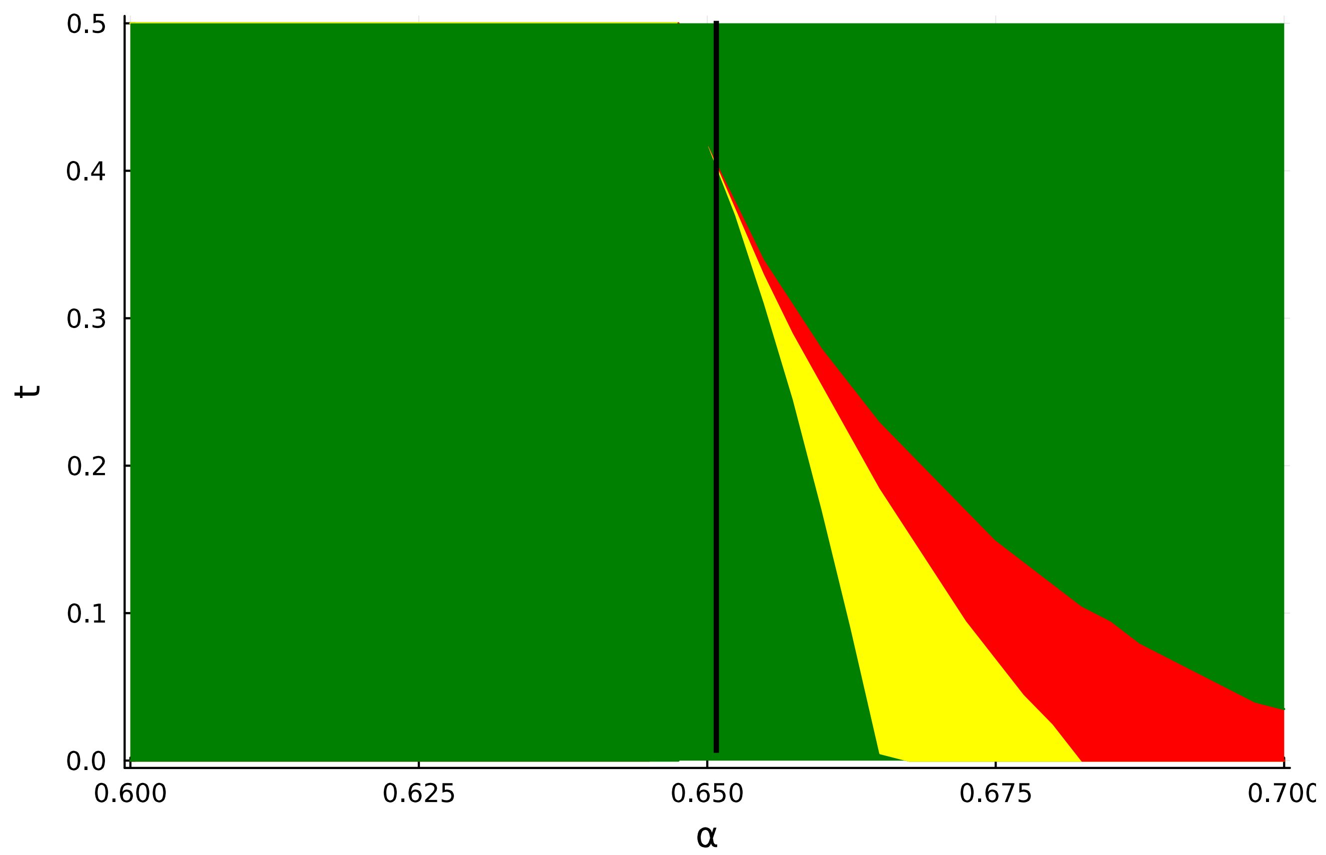
- Can we find an easy-to-sample distribution on the solution space?

- We add a potential:

$$p(w) \propto \prod_{\mu=1}^{M=\alpha N} \mathbb{I}(s^\mu \geq 0)\, e^{-\beta U(s^\mu)}, \qquad s^\mu = \frac{w \cdot x^\mu}{\sqrt{N}}.$$

- We perform an expansion of $\phi_t(q)$ around $q = 1$ and find a condition for removing the second peak:

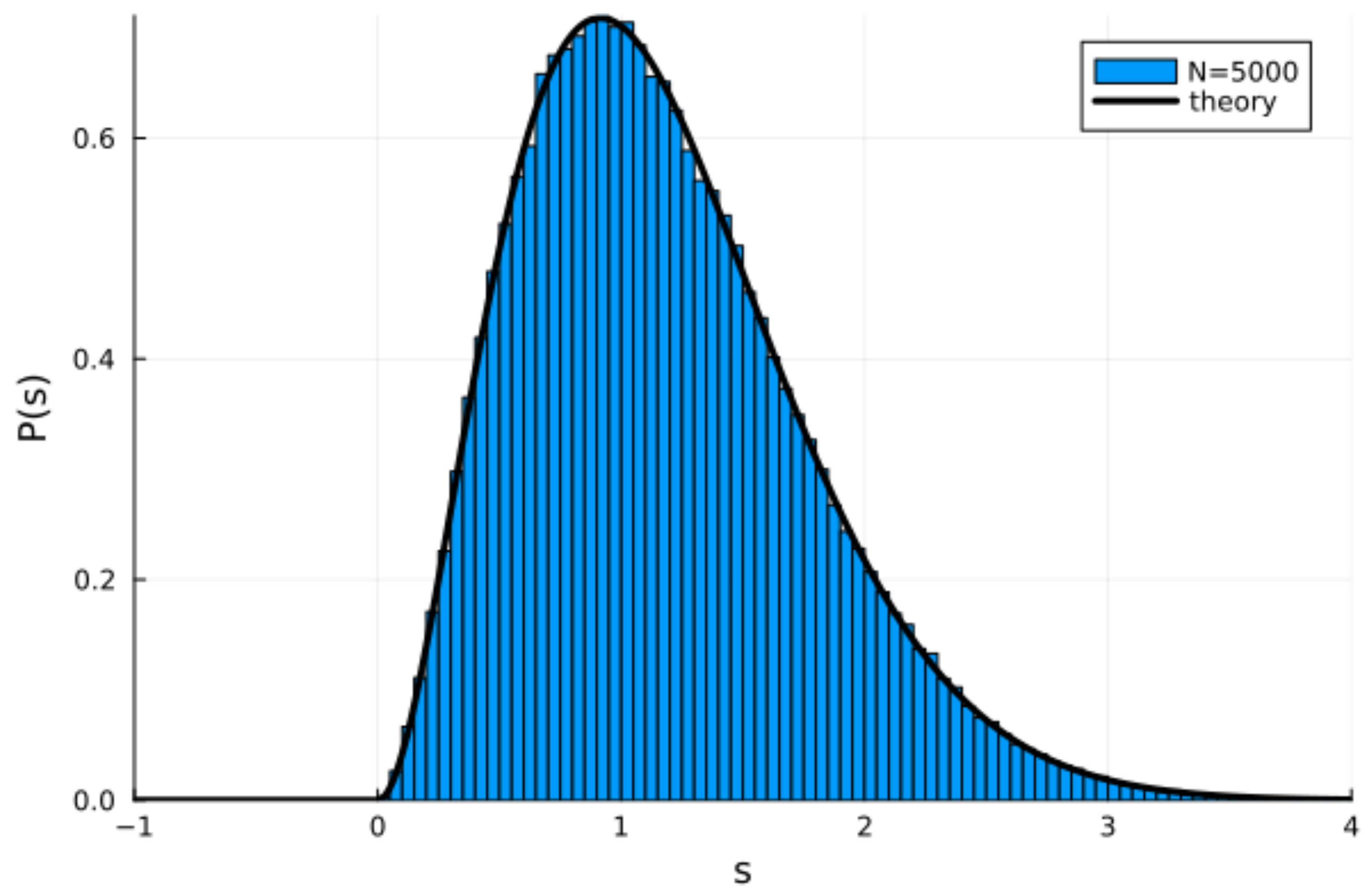
  Need potential at least as singular as $U(s) = -\log(s)$ near $s = 0$ and also $\beta > 1$.



$\beta = 2$

$\alpha_{ASL} \approx 0.65$

histogram of stabilities $\beta = 2,\ \alpha = 0.3$

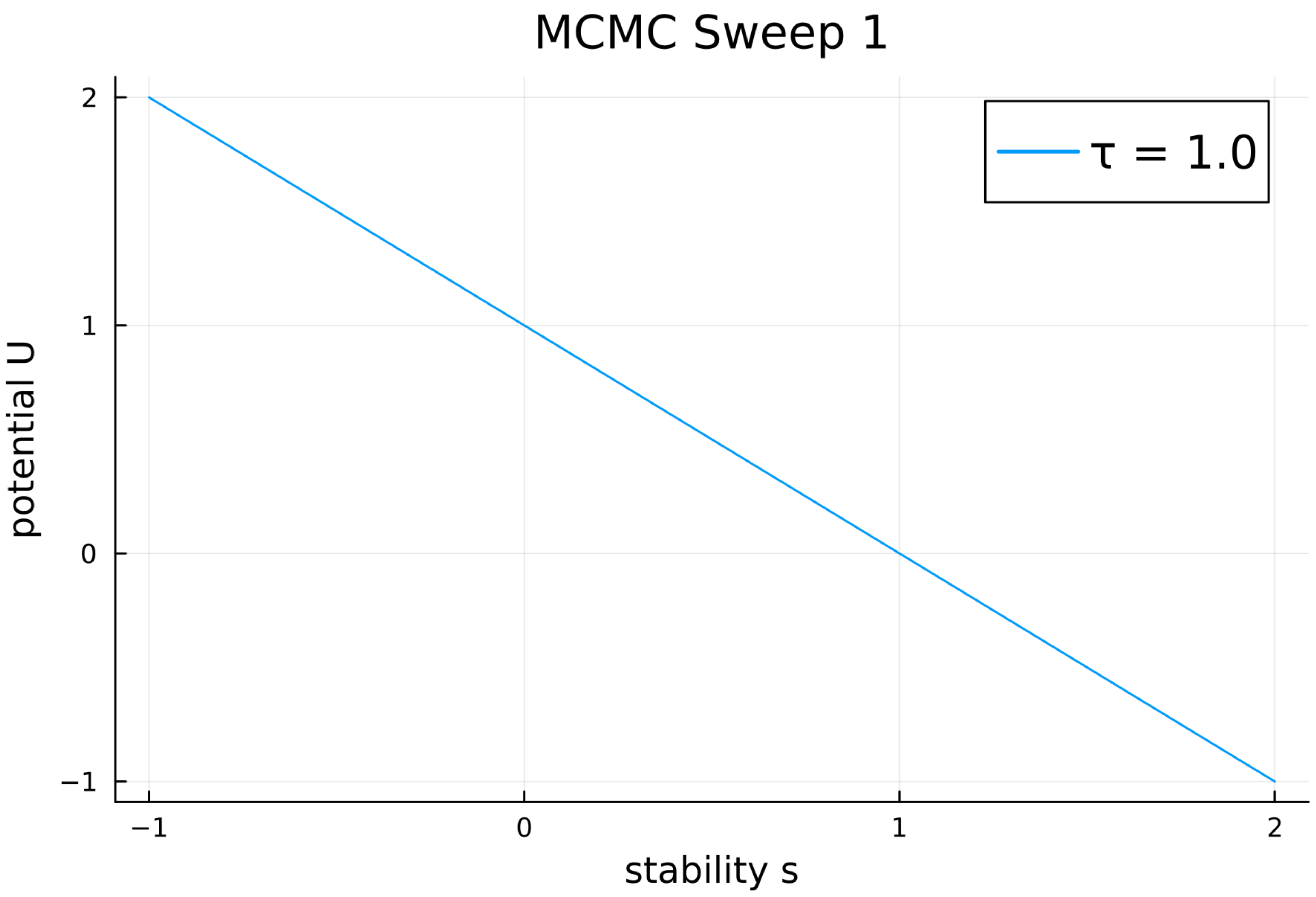


τ-annealing MCMC for binary perceptron

AMP is very frail (heavy statistical assumptions). Can we devise a MCMC scheme?

$$U_\tau(s) = \begin{cases} \frac{1}{\tau}(1 - s^\tau) & s > 0, \\ \frac{1}{\tau}(1 - s) & s \leq 0. \end{cases}$$

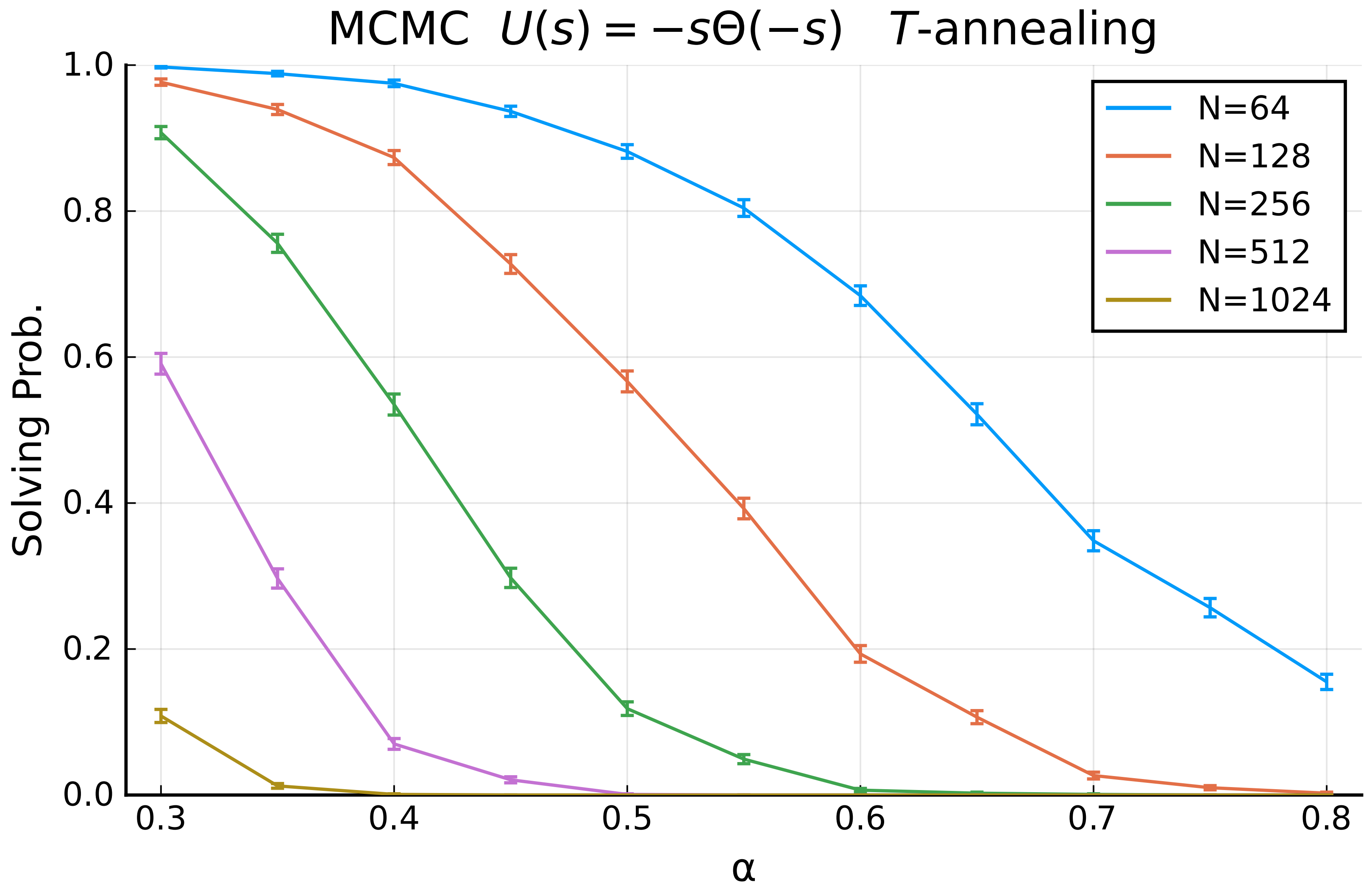
$$\lim_{\tau \to 0} U_\tau(s) = -\log(s)$$

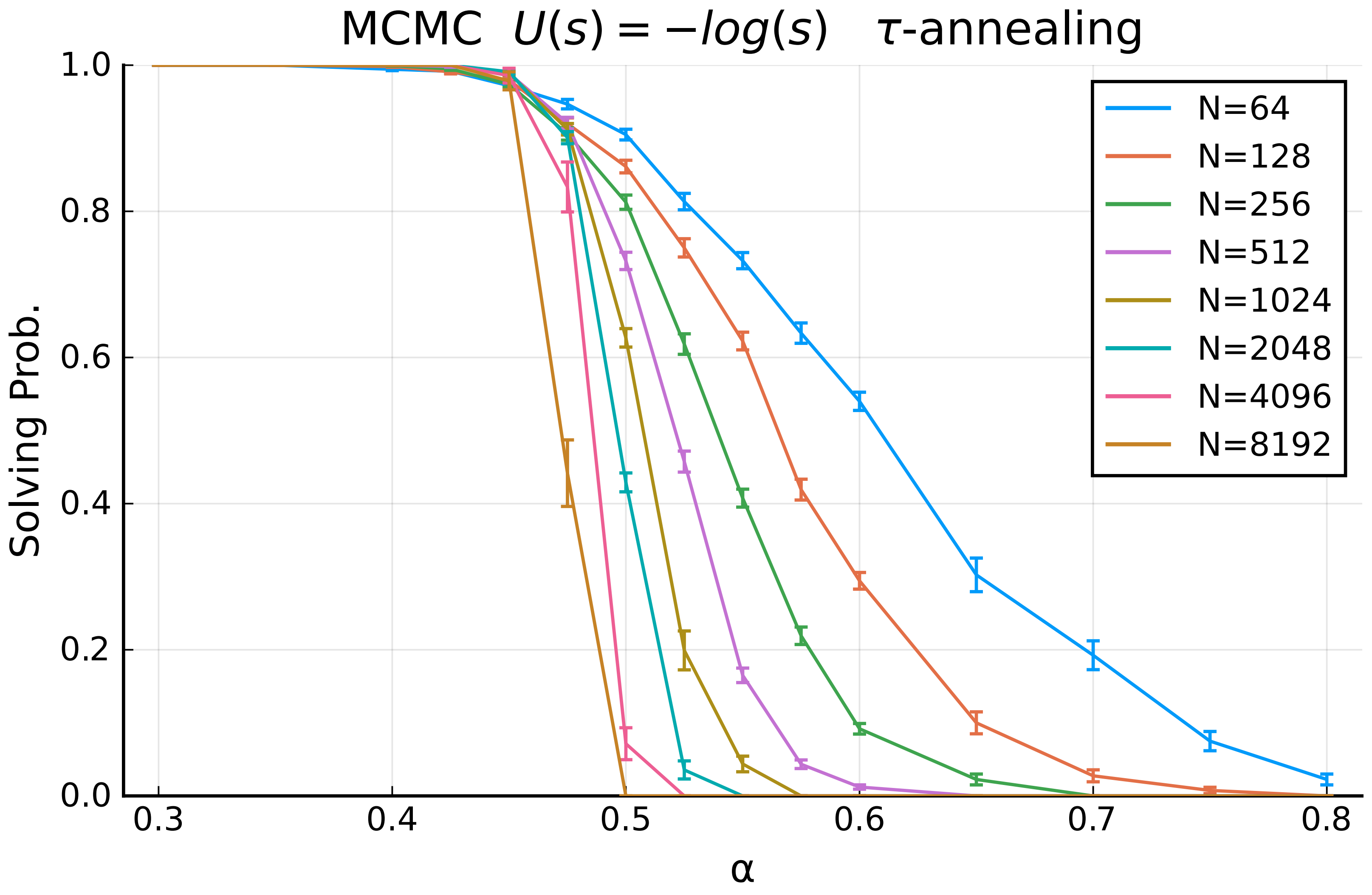


τ-annealing MCMC for binary perceptron

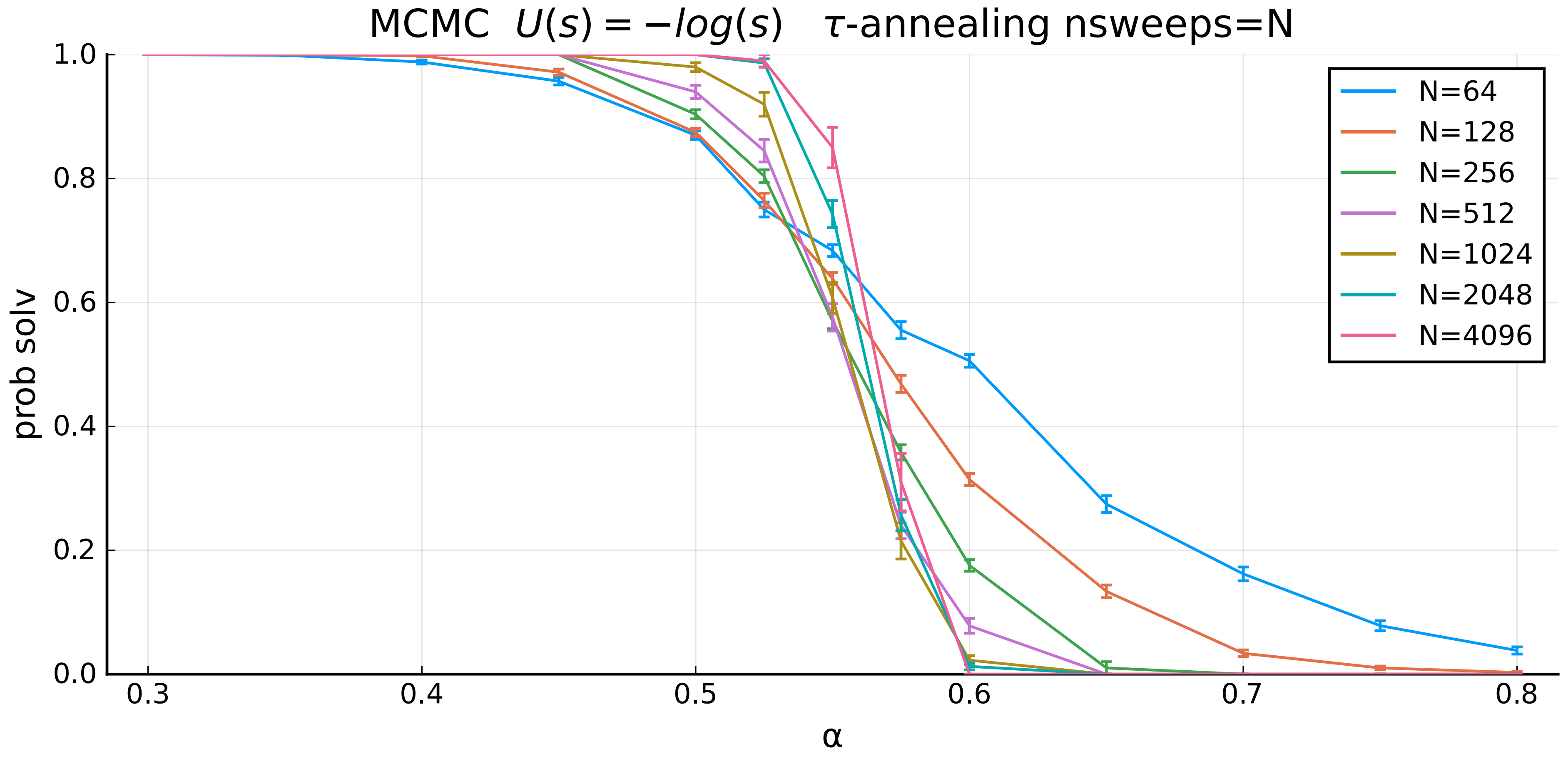
For the first time we have a simple and robust algorithm for producing diverse and under-control solutions to the binary perceptron problem.

num sweeps = 100





num sweeps = N



Straziota, Demyanenko, Baldassi, **CL,** arxiv ‘25

Thanks!

Carlo Baldassi

Elizaveta Demyanenko

Davide Straziota

Luca Ambrogioni

Beatrice Achilli

Marc Mézard

Enrico Ventura