Sampling through Algorithmic Diffusion in Non-Convex Perceptron Problems

Elizaveta Demyanenko, Davide Straziota, Carlo Lucibello, Carlo Baldassi
Department of Computing Sciences, Bocconi University
elizaveta.demyanenko@phd.unibocconi.it

Task introduction

We analyze the problem of sampling from the solution space of simple yet non-convex neural network models by employing a denoising diffusion process known as Algorithmic Stochastic Localization [2, 3], where the score function is provided by Approximate Message Passing. We introduce a formalism based on the replica method to characterize the process in the infinite-size limit in terms of a few order parameters, and, in particular, we provide criteria for the feasibility of sampling.

1. Stochastic Localization (SL)

- **Input:** μ probability measure on N-dimensional unit vectors
- **Output:** a sample $\boldsymbol{W} \sim \mu$

Stochastic Localization: maintain a *tilted measure* μ_t,

$$\mu_t(\mathrm{d}\boldsymbol{W}) \propto e^{\langle \mathbf{h}, \boldsymbol{W} \rangle} \mu(\mathrm{d}\boldsymbol{W}),$$

and a field $\mathbf{h}(t)$ that follows the stochastic differential equation

$$\mathrm{d}\mathbf{h}(t) = \mathbf{m}(t)\,\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \quad \mathbf{m}(t) = \text{mean of } \mu_t.$$

(μ_t) is *localizing* to μ: $\mu_t \underset{t\to\infty}{\longrightarrow} \delta_{\boldsymbol{W}^*}$ almost surely, with $\boldsymbol{W}^* \sim \mu$.

Algorithmic Stochastic Localization:

1. **Discretize** the differential equation.
2. Compute the mean of the tilted distribution with **Approximate Message Passing**.

2. The perceptron models

Input: $\boldsymbol{\xi}^\mu \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_N)$

Goal: find a *weight vector* $\mathbf{W} \in \mathbb{R}^N$ such that

$$\mathbf{W} \cdot \boldsymbol{\xi}^\mu > \kappa\sqrt{N} \quad \text{for all } \mu = 1, \dots, M.$$

κ - the *margin* of the problem.
$\alpha = M/N$ - constant in the thermodynamical limit ($N \to \infty$).

- **spherical** perceptron: $\|\boldsymbol{W}\|^2 = N, \kappa \geq 0$.
- **negative** perceptron: $\|\boldsymbol{W}\|^2 = N, \kappa < 0$.
- **binary** perceptron: $\boldsymbol{W} \in \{-1, 1\}^N$.

3. Negative Spherical Perceptron

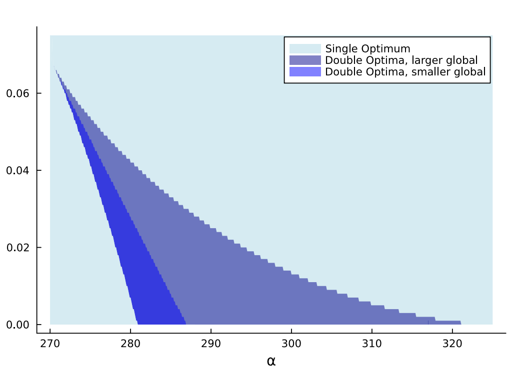
The measure to sample from is:

$$\mu(\mathbf{W}) \propto \prod_\mu \Theta\left(\sum_i \frac{W_i \xi_i^\mu}{\sqrt{N}} - \kappa\right) \delta\left(\|\boldsymbol{W}\|_2 = N\right) \tag{1}$$

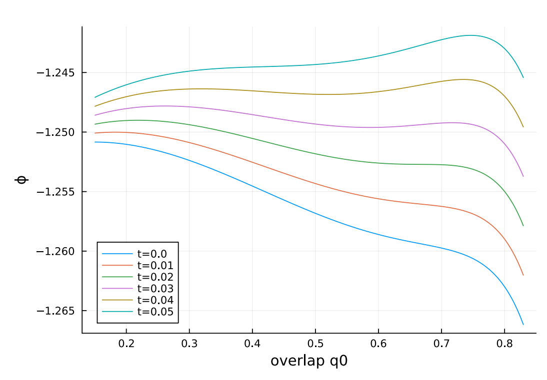
For the asymptotic analysis we define $\mathbf{W} \sim \mu_t$ and $\mathbf{W}^* \sim \mu$, respectively samples from the tilted and objective measures. The asymptotic analysis of the sampling procedure can be studied through replica analysis (under the RS Ansatz), i.e. we construct n copies of $\boldsymbol{W}_a$ and s copies of $\boldsymbol{W}^*_\alpha$. We can define the following overlaps and the optimality conditions

$$q_{ab} = \langle \boldsymbol{W}_a, \boldsymbol{W}_b \rangle \qquad r_{\alpha\beta} = \langle \boldsymbol{W}^*_\alpha, \boldsymbol{W}^*_\beta \rangle \qquad p_{\alpha a} = \langle \boldsymbol{W}_a, \boldsymbol{W}^*_\alpha \rangle \tag{2}$$

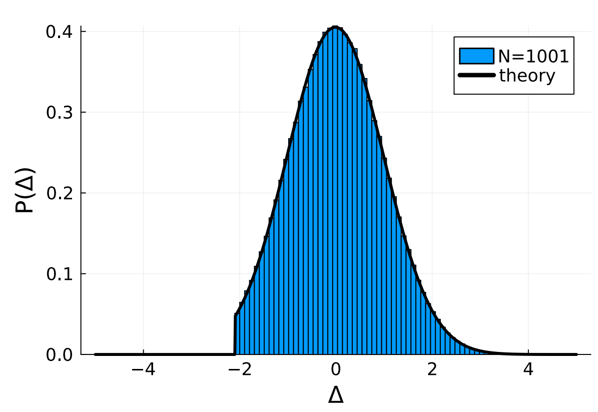
$$p_1 = q_0 \qquad p_0 = r_0 \tag{3}$$



Spherical perceptron phase diagram with $\kappa = -2.5$



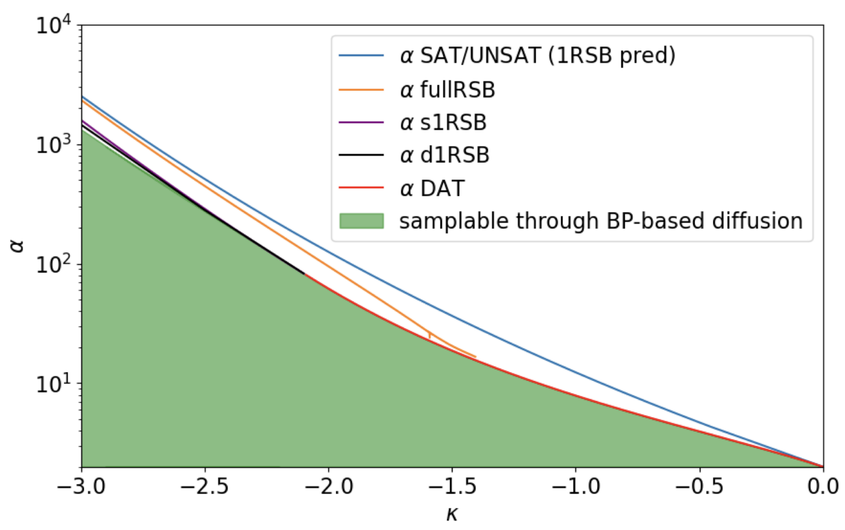
Free entropy ϕ as a function of q_0



Stabilities Distribution for $N = 1001$, $\kappa = -2.1$ and $\alpha = 80$

Cluster of solutions $\Longrightarrow$ Possible to sample from μ, depending on α

4. Spherical Perceptron Transition



Sampleable region and transitions in Negative spherical perceptron $\kappa = -2.5$

5. Binary Perceptron

The measure to sample from:

$$\mu(\mathbf{W}) \propto \prod_\mu \Theta\left(y^\mu \sum_i \frac{W_i \xi_i^\mu}{\sqrt{N}}\right)$$

Isolated Solutions $\Longrightarrow$ Impossible

6. Towards the binary perceptron

The frozen nature of the binary perceptron calls for alternative measures μ that target specific types of solutions [1].

$$\mu(\mathbf{W}) \propto \prod_\mu \exp\left(-\beta \frac{\log\left(1 + e^{-2\gamma\left(y^\mu \sum_i \frac{W_i \xi_i^\mu}{\sqrt{N}}\right)}\right)}{2\gamma}\right) \tag{4}$$

$$\mu(\mathbf{W}) \propto \prod_\mu \Theta\left(y^\mu \sum_i \frac{W_i \xi_i^\mu}{\sqrt{N}}\right) \left(y^\mu \sum_i \frac{W_i \xi_i^\mu}{\sqrt{N}}\right)^\beta \tag{5}$$

References

[1] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Physical Review Letters*, 115(12), September 2015.

[2] D. Ghio, Y. Dandi, F. Krzakala, and L. Zdeborová. Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective, 2023.

[3] A. Montanari. Sampling, diffusions, and stochastic localization, 2023.