Neural manifold packing by stochastic gradient descent

Guanming Zhang (gz2241@nyu.edu)
Center for Soft Matter Research, Department of Physics, New York University, New York 10003, USA
Department of Chemistry, New York University, New York 10003, USA

Stefano Martiniani (stefano.martiniani@nyu.edu)
Center for Soft Matter Research, Department of Physics, New York University, New York 10003, USA
Simons Center for Computational Physical Chemistry, Department of Chemistry, New York University, New York 10003, USA
Courant Institute of Mathematical Sciences, New York University, New York 10003, USA

Abstract

In self-supervised learning, different stimulus categories correspond to unique manifolds within an embedded neural state space. Accurate classification can be achieved by separating the manifolds from one another during learning, in a process that is analogous to a packing problem. To theoretically investigate the dynamics of 'neural manifold packing', we consider Stochastic Gradient Descent (SGD) for particle systems in physical dimension. In this framework, SGD aims to minimize a hinge-loss, L, proportional to the particles' overlaps, by performing gradient descent on a batch of randomly selected particles. The resulting stochastic dynamics exhibit a critical packing efficiency, ϕ_c, below which the system reaches an 'absorbing' state with zero classification error, $L = 0$, and above which the system settles in an 'active' steady-state with $L > 0$. We thus explore the connection between the dynamics of SGD and of a well-characterized absorbing state model known as Biased Random Organization (BRO) that evolves the particle positions with random kicks. We show that in the limit of small kick sizes and learning rates, BRO and SGD on a hinge-loss are equivalent, and exhibit the same critical packing efficiency $\phi_c \approx 0.64$. Further, we demonstrate that the behavior of SGD near the critical point is consistent with the Manna universality class. Thus, we propose that 'neural manifold packing' by SGD in high-dimensions is mean-field, given that Manna universality reduces to mean field critical behavior in $d > 4$. This work furthers our understanding of self-supervised learning dynamics and opens avenues for designing learning algorithms based on physical principles.

Keywords: self-supervised learning; neural manifolds; stochastic gradient descent.

Introduction

In self-supervised classification tasks, each class is represented by a distinct manifold within the embedded neural state space. Consequently, self-supervised learning can be understood as the process of separating neural manifolds (Chung, Lee, & Sompolinsky, 2016). This process is reminiscent of high-dimensional packing problems in mathematics, and of the packing of physical objects in three-dimensional space. To connect the physical packing problem with the packing of manifolds in neural state space, we consider the case where the embedding space is three-dimensional, $d = 3$, and the manifolds are approximated by spheres of the same size (n.b., we make this simplifying assumptions because the nature of the critical phenomena we study do not depend on the particle size distribution). Dynamical physics models that rely on the minimization of short-range, repulsive pairwise interactions are frequently employed to investigate such systems (O'Hern, Langer, Liu, & Nagel, 2002). An alternative viewpoint is provided by "absorbing state models" that evolve particle positions through stochastic dynamics rules without explicitly defining a potential, leading to a transition between an 'absorbing state" where halts as all geometric constraints are met, and an "active state" where it maintains dynamic equilibrium due to unresolved constraints. One such model is Biased Random Organization (BRO), in which overlapping particles are randomly displaced away from one another. (Wilken, Guerra, Levine, & Chaikin, 2021; Wilken, Guo, Levine, & Chaikin, 2023).

To address the neural manifold packing problem underlying self-supervised learning, we consider SGD on particle systems with pairwise, short-range interactions, corresponding to a hinge loss. We first demonstrate analytically that the dynamics of BRO closely resemble those of particles governed by a linear repulsive potential influenced by multiplicative self-quenching noise. Subsequently, we show that SGD of the hinge loss is equivalent to BRO in the limit of small learning rate (and small kick size for BRO). Notably, both BRO and SGD exhibit the same critical point in three dimensions, $\phi_c \approx 0.64$, corresponding to the maximum lossless packing efficiency. Further, we observe that SGD exhibits behavior consistent with the Manna universality class near the critical point.[1] This consistency persists across various batch sizes and even in the noiseless (gradient descent) scenarios. This equivalence thus allows us to conceptualize neural manifold packing by SGD in terms of a well-characterized self-organizing physical model, and to draw conclusions about neural manifold packing in high dimensions.

[1] Models belonging to the same universality class exhibit similar scale-invariant behavior near critical points, for instance in their temporal correlations and spatial structure. The Manna universality class represents a family of models that, through a series of cascading events, dynamically evolve towards a critical scale-invariant state.

The stochastic approximation of BRO and SGD

In the BRO model, spherical particles with radius R are randomly placed in a periodic box with volume fraction (viz., density) ϕ. Overlapping particles (i.e., whose centers are closer than 2R) are identified as being "active", and their dynamics evolved as follows. Active particle i is kicked by neighboring overlapping particles j, each contributing an equal unit vector along the direction connecting their centers, $\frac{\mathbf{x}^i-\mathbf{x}^j}{|\mathbf{x}^i-\mathbf{x}^j|}$. The summation over all neighboring unit vectors is scaled by a random number sampled from a uniform distribution over $[0,\varepsilon]$, where ε is the 'kick size'. By means of stochastic approximation (Li, Tai, & Weinan, 2017, 2019; Hu, Li, Li, & Liu, 2019), we find that BRO dynamics can be approximated by particles with pairwise, linear, repulsive interactions driven by a multiplicative noise process, which reads

$$\mathbf{x}^i_{k+1} = \mathbf{x}^i_k - \varepsilon\nabla^i L + \varepsilon\sqrt{\Sigma^i}\cdot\xi^i_k, \text{ where } \Sigma^i = \frac{1}{3}\nabla^i L \nabla^i L^T \quad (1)$$

where ξ^i_k is the standard Gaussian noise for particle i at step k, $L = \sum_i \sum_{j>i} U(\mathbf{x}^i_k - \mathbf{x}^j_k)$ is the total energy/loss obtained by summing the pairwise interactions,

$$U(\mathbf{r}) = \begin{cases} \frac{1}{2}(2R-|\mathbf{r}|), & \text{if } |\mathbf{r}| < 2R \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

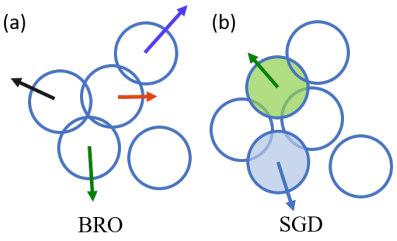

Figure 1: (a) In BRO each active particle is randomly kicked away from its overlapping neighbors. (b) In SGD two particles (shaded) are selected (batch size, $B = 2$) out of the $|Q_k| = 5$ active particles, and they are displaced by gradient descent.

We devise a particle-wise SGD method inspired by the commonly employed SGD approach in self-supervised learning. During each iteration k, a batch is formed by randomly selecting B (denoted as the batch size) particles from the set of all active particles. Each particle i in this batch undergoes a displacement $-\alpha\nabla^i L$ where α is the learning rate, and L is the total energy/loss. The unselected particles remain undisturbed at this iteration. This iterative process continues until the system's energy reaches $L = 0$, or a steady-state value $L > 0$. Following the same stochastic approximation approach as above, we approximate the particle-wise SGD dynamics as

$$\mathbf{x}^i_{k+1} = \mathbf{x}^i_k - \alpha b_f \nabla^i L + \alpha\sqrt{b_f - b_f^2}\sqrt{\nabla^i L \nabla^i L^T}\cdot\xi^i_k, \quad (3)$$

where $b_f = B/|Q_k|$. Q_k is the set of active particles at time step k, and $|Q_k|$ is the number of active particles (viz., the cardinality of Q_k). Interestingly, the noise term in Eq. 3 has the same functional form as in Eq. 1, revealing that the noise processes in SGD and BRO are of the same nature, despite the differences in their microscopic mechanism.

Critical behavior of SGD

In BRO, the absorbing phase transition between absorbing ($L = 0$) and active states ($L > 0$) approaches a limiting critical volume fraction (viz., packing efficiency), $\phi_c \approx 0.64$, for infinitesimal kick sizes ($\varepsilon \to 0$), in three-dimension. We show that SGD exhibits a transition at the same critical volume fraction, $\phi_c \to 0.64$, as the learning rate $\alpha = \varepsilon/b_c \to 0$. It has been shown that the behavior of BRO at criticality belongs to the Manna universality class (Wilken et al., 2021, 2023), and it is natural to ask whether SGD is also in the Manna class. We measure the fraction of active particles at the steady state, f^a, and the typical relaxation time, τ, required to reach the steady state. We test this hypothesis by performing finite-size scaling analysis at different batch sizes, $b_f = B/|Q_k|$, to check the critical behavior $f^a \sim (\phi - \phi_c)^\beta$ and $\tau \sim |\phi - \phi_c|^{-\nu_\parallel}$ in the proximity of ϕ_c using Manna exponents(Henkel, Hinrichsen, & Lubeck, 2009). Fig. 2 displays the scaled activity and relaxation time across various system sizes. These transition curves for varying system sizes collapse. Moreover, this collapse is observed for different batch sizes. It is worth noting that for $b_f = 1.0$, SGD reduces to gradient descent (GD), the zero noise limit of SGD. Therefore, the critical behavior of SGD and GD is consistent with the Manna universality class.

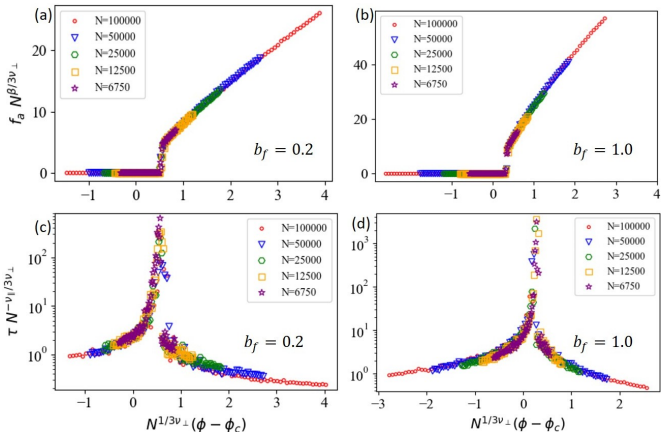

Figure 2: Finite size scaling analysis for SGD: Steady-state activity as a function of volume fraction for (a) $b_f = 0.2$ and (b) $b_f = 1.0$. Relaxation time as a function of volume fraction for (d) $b_f = 0.2$ and (e) $b_f = 1.0$. $\nu_\parallel = 1.08, \beta = 0.84$ and $\nu_\perp = 0.59$ are Manna exponents in $d = 3$.

Discussion

We show that BRO dynamics are equivalent to those of linearly repulsive particles driven by multiplicative noise. We then bridge SGD and BRO by showing that the batch selection noise in SGD shares the same form as the multiplicative noise in BRO. Furthermore, we demonstrate that the critical behavior of SGD is consistent with the Manna universality class. We

thus propose that when the dimensionality of the neural state space is $d > 4$ and the number of classes (equivalent to the number of particles in our model) becomes large, the critical behavior of neural manifold packing by SGD converges to the mean-field universality class given that the upper critical dimension for the Manna universality class is $d = 4$. In future work, we will generalize our analysis to ellipses to better capture the low-rank structure of neural manifolds. This work thus paves the way for understanding how neural manifolds could evolve through a mechanism equivalent to SGD or, even, only through multiplicative noise. Our results can be extended to higher dimensions, offering a way of designing learning algorithms based on physical principles.

Acknowledgments

G.Z. and S.M. thank Sam Wilken, Ashley Guo, Mathias Casiulis, and David Heeger for helpful discussions. G.Z. and S.M. acknowledge the support by NSF grant 2132995 and NIH grant EY035242. S.M. acknowledges the Simons Center for Computational Physical Chemistry for financial support. This work was supported in part through the NYU IT High-Performance Computing resources, services, and staff expertise.

References

Chung, S., Lee, D. D., & Sompolinsky, H. (2016). Linear readout of object manifolds. *Physical Review E*, *93*(6), 060301.

Henkel, M., Hinrichsen, H., & Lubeck, S. (2009). *Non-equilibrium phase transitions* (2009th ed.). New York, NY: Springer.

Hu, W., Li, C. J., Li, L., & Liu, J.-G. (2019). On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, *4*(1).

Li, Q., Tai, C., & Weinan, E. (2017). Stochastic modified equations and adaptive stochastic gradient algorithms. In *International conference on machine learning* (pp. 2101–2110).

Li, Q., Tai, C., & Weinan, E. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, *20*(1), 1474–1520.

O'Hern, C. S., Langer, S. A., Liu, A. J., & Nagel, S. R. (2002). Random packings of frictionless particles. *Physical Review Letters*, *88*(7), 075507.

Wilken, S., Guerra, R. E., Levine, D., & Chaikin, P. M. (2021). Random close packing as a dynamical phase transition. *Physical Review Letters*, *127*(3), 038002.

Wilken, S., Guo, A. Z., Levine, D., & Chaikin, P. M. (2023, Dec). Dynamical approach to the jamming problem. *Phys. Rev. Lett.*, *131*, 238202. doi: 10.1103/PhysRevLett.131.238202